The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 08 - 2026

Artificial Intelligence-Driven Cybersecurity Framework for Enterprise Threat Detection: A Machine Learning Approach

Sanjida Akter Tisha

Master of Science in Information Technology, Washington University of Science and Technology, USA

Md Yassir Mottalib

Master of Science in Information System Technology, Wilmington University, USA

Eklachur Rahman Bhuiyan

Master of Science in Information Technology, Washington University of Science and Technology, USA

Asaduzzaman Anik

Master of Business Administration (MBA) in management, Stanton University, Los Angeles, California

SM Wali Ullah

Department of Business Administration and Management, Stanton University, USA

Marjahan Risalat

Department of Business Administration and Management, Stanton University, USA

Shimita Lopa Chockroborty

Master's in Accounting, National University, Dhaka, Bangladesh

Received: 24 May 2026 | Received Revised Version: 12 June 2026 | Accepted: 26 June 2026 | Published: 17 July 2026

Volume 08 Issue 07 2026 | Crossref DOI: 10.37547/tajet/Volume08Issue07-03

Abstract

The increasing complexity of cyber threats has exposed the limitations of traditional signature-based intrusion detection systems, creating a need for intelligent and adaptive cybersecurity solutions. This study proposes an artificial intelligence-driven cybersecurity framework for enterprise threat detection using the CICIDS2017 benchmark dataset. The framework incorporates data preprocessing, feature engineering, and supervised machine learning to classify network traffic as benign or malicious. Seven machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extra Trees, LightGBM, and XGBoost, were evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. The results indicate that ensemble learning models outperform conventional classifiers, with XGBoost achieving the highest performance, recording ***99.42% accuracy, 99.39% precision, 99.31% recall, 99.35% F1-score, and an AUC-ROC of 0.999.*** *LightGBM also demonstrated excellent performance with lower computational time. The findings suggest that the proposed XGBoost-based framework provides an accurate, scalable, and efficient solution for real-time enterprise threat detection and can be effectively integrated into modern cybersecurity infrastructures to enhance organizational cyber resilience.*

Keywords: Artificial Intelligence, Cybersecurity, Machine Learning, Enterprise Threat Detection, Intrusion Detection System, XGBoost.

Cite This Article: Tisha, S. A., Mottalib, M. Y., Bhuiyan, E. R., Anik, A., Ullah, S. W., Risalat, M., & Chockroborty, S. L. (2026). Artificial Intelligence-Driven Cybersecurity Framework for Enterprise Threat Detection: A Machine Learning Approach. The American Journal of Engineering and Technology, 8(07), 25–37. https://doi.org/10.37547/tajet/Volume08Issue07-03

Introduction

The rapid digital transformation of modern enterprises has fundamentally changed the way organizations operate, communicate, and manage critical business processes. Cloud computing, the Internet of Things (IoT), big data analytics, artificial intelligence (AI), remote work environments, and digital business platforms have significantly improved organizational efficiency and productivity. However, these technological advancements have also expanded the cyberattack surface, making enterprise networks increasingly vulnerable to sophisticated cyber threats. Organizations across finance, healthcare, manufacturing, government, and critical infrastructure sectors continuously face cyberattacks that compromise confidential information, disrupt business operations, and cause substantial financial and reputational losses (Stallings, 2020). Consequently, cybersecurity has become one of the most significant strategic priorities for enterprises worldwide.

According to the IBM Cost of a Data Breach Report, the average financial impact of a data breach continues to increase annually due to ransomware attacks, insider threats, phishing campaigns, and advanced persistent threats (APTs) (IBM Security, 2024). Similarly, the Cybersecurity and Infrastructure Security Agency (CISA) have emphasized that cyberattacks targeting critical infrastructure and enterprise information systems have become more sophisticated, frequent, and difficult to detect using conventional security mechanisms (CISA, 2024). Traditional security approaches, including signature-based intrusion detection systems, rule-based firewalls, and manually configured security policies, primarily depend on previously identified attack signatures. While these methods remain valuable against known threats, they often struggle to identify zero-day attacks, polymorphic malware, insider threats, and continuously evolving attack strategies that exhibit unknown behavioral characteristics (Buczak & Guven, 2016).

To address these limitations, artificial intelligence and machine learning have emerged as transformative technologies for intelligent cybersecurity. Unlike traditional detection systems, machine learning algorithms automatically identify hidden patterns within large volumes of network traffic, enabling them to distinguish between legitimate and malicious activities without relying solely on predefined attack signatures. Through continuous learning from historical and real-time network data, machine learning models can improve detection accuracy, reduce false-positive alerts, and rapidly adapt to newly emerging cyber threats (Sommer & Paxson, 2010). As enterprise networks generate millions of communication events daily, AI-driven cybersecurity solutions provide scalable and automated mechanisms capable of processing high-dimensional network traffic with minimal human intervention.

Recent advances in supervised machine learning have demonstrated remarkable success in intrusion detection and cyber threat classification. Ensemble learning techniques, including Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Extra Trees, have shown superior predictive capabilities compared with conventional classification algorithms because they effectively capture complex nonlinear relationships within cybersecurity datasets (Chen & Guestrin, 2016). These algorithms have consistently achieved high detection rates across benchmark intrusion detection datasets while maintaining low false-positive rates, making them particularly suitable for enterprise cybersecurity applications. Nevertheless, selecting an optimal machine learning model remains a challenging task because different algorithms exhibit varying levels of computational efficiency, scalability, and classification

performance depending on network traffic characteristics and attack distributions.

Enterprise cybersecurity environments present additional challenges due to the enormous volume, velocity, and diversity of network traffic generated by cloud services, mobile devices, IoT sensors, and distributed computing environments. Modern organizations require intrusion detection frameworks capable of analyzing heterogeneous data streams in real time while maintaining high detection accuracy and low computational overhead. Furthermore, cybersecurity solutions must effectively identify both common attack patterns and previously unseen threats without introducing excessive false alarms that can overwhelm security analysts within Security Operations Centers (SOCs). Therefore, developing an intelligent, scalable, and computationally efficient machine learning framework has become an important research objective for both academia and industry.

Several publicly available benchmark datasets have been introduced to facilitate the development and evaluation of intrusion detection systems. Among these, the CICIDS2017 dataset developed by the Canadian Institute for Cybersecurity has become one of the most comprehensive and realistic datasets because it contains modern attack scenarios, realistic enterprise network traffic, and detailed flow-based network features extracted using CICFlowMeter (Sharafaldin et al., 2018). The dataset includes multiple cyberattack categories such as distributed denial-of-service attacks, brute force attacks, web attacks, infiltration attempts, botnet traffic, and benign communication, making it an appropriate benchmark for evaluating machine learning algorithms under realistic enterprise conditions.

Although previous studies have demonstrated promising results using individual machine learning algorithms for intrusion detection, relatively few investigations have conducted comprehensive comparative analyses involving multiple state-of-the-art ensemble learning models under identical experimental settings while simultaneously addressing practical deployment considerations within enterprise environments. In particular, there remains a need to evaluate predictive accuracy, computational efficiency, scalability, and implementation feasibility to determine the most suitable machine learning algorithm for real-world enterprise cybersecurity systems. Addressing these research gaps contributes to the development of more reliable and operationally effective AI-driven intrusion detection frameworks.

The primary objective of this study is to develop an artificial intelligence-driven cybersecurity framework for enterprise threat detection using supervised machine learning techniques. We employ the CICIDS2017 benchmark dataset to train and evaluate multiple classification algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extra Trees, LightGBM, and XGBoost. We compare their performance using multiple evaluation metrics, including accuracy, precision, recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and computational efficiency. Based on the experimental findings, we identify the most effective model for enterprise threat detection and discuss its practical implementation within cybersecurity infrastructures across industries in the United States. The proposed framework aims to enhance cyber resilience by enabling accurate, scalable, and intelligent threat detection capable of addressing the rapidly evolving cybersecurity challenges faced by modern enterprises.

Literature Review

The rapid advancement of digital technologies has transformed enterprise operations by enabling cloud computing, big data analytics, Internet of Things (IoT) devices, and distributed information systems. Although these technological innovations have significantly improved operational efficiency, they have simultaneously increased cybersecurity risks by expanding the attack surface available to malicious actors. Enterprise networks now experience a growing number of sophisticated cyberattacks, including ransomware, advanced persistent threats (APTs), distributed denial-of-service (DDoS) attacks, phishing campaigns, insider threats, and zero-day exploits. As cyber threats continue to evolve in complexity and frequency, traditional security mechanisms have become increasingly inadequate for protecting modern enterprise infrastructures (Stallings, 2020).

Conventional intrusion detection systems (IDSs) primarily rely on signature-based or rule-based approaches to identify malicious activities. Signature-based detection methods compare incoming network traffic against predefined attack signatures stored within security databases. Although these systems achieve high detection accuracy for previously identified attacks, they

are ineffective against newly emerging threats and zero-day attacks because no predefined signatures exist for unknown malicious behaviors. Similarly, anomaly-based detection methods rely on manually defined behavioral thresholds that often generate excessive false-positive alerts, reducing operational efficiency within Security Operations Centers (SOCs) (Sommer & Paxson, 2010). Consequently, cybersecurity researchers have increasingly explored artificial intelligence and machine learning techniques capable of automatically learning complex attack patterns from large-scale network traffic.

Machine learning has become one of the most promising technologies for intelligent cybersecurity because it enables automated detection of malicious activities without requiring manually generated attack signatures. Machine learning algorithms learn statistical relationships from historical network traffic and subsequently classify new observations as either benign or malicious. This adaptive learning capability allows intrusion detection systems to recognize evolving attack patterns while continuously improving prediction accuracy. Buczak and Guven (2016) conducted a comprehensive survey of machine learning applications in cybersecurity and concluded that supervised learning algorithms consistently outperform traditional signature-based detection systems for identifying unknown cyber threats while reducing false-positive rates.

Supervised learning approaches have received considerable attention because they utilize labeled datasets to construct predictive models capable of distinguishing normal network behavior from malicious activities. Among these methods, Logistic Regression represents one of the earliest classification algorithms applied to intrusion detection due to its simplicity and computational efficiency. However, because Logistic Regression assumes linear relationships among predictor variables, its performance often deteriorates when analyzing the highly nonlinear characteristics of enterprise network traffic (Aljawarneh et al., 2018). Decision Tree algorithms improve classification by constructing hierarchical decision rules that are easy to interpret, yet they remain vulnerable to overfitting when trained on high-dimensional cybersecurity datasets containing noisy observations.

Support Vector Machine (SVM) has also been widely employed for intrusion detection because of its ability to construct optimal separating hyperplanes within multidimensional feature spaces. SVM has demonstrated strong predictive performance on several benchmark cybersecurity datasets, particularly when nonlinear kernel functions are utilized. Nevertheless, the computational complexity associated with SVM increases substantially as dataset size grows, limiting its applicability to large-scale enterprise environments that generate millions of network traffic records each day (Sommer & Paxson, 2010).

To overcome these limitations, ensemble learning techniques have become increasingly popular within cybersecurity research. Ensemble methods combine multiple weak learners to generate stronger predictive models capable of improving classification accuracy while reducing overfitting. Random Forest, one of the most widely adopted ensemble algorithms, constructs numerous decision trees using bootstrap aggregation and random feature selection. This approach significantly improves prediction robustness while maintaining strong generalization performance across diverse attack categories (Breiman, 2001). Numerous empirical studies have demonstrated that Random Forest consistently achieves high intrusion detection accuracy due to its ability to model complex interactions among network traffic features.

More recently, gradient boosting algorithms have emerged as state-of-the-art machine learning techniques for cybersecurity applications. Extreme Gradient Boosting (XGBoost), introduced by Chen and Guestrin (2016), incorporates regularization, parallel processing, and efficient tree optimization strategies that significantly improve both predictive accuracy and computational efficiency. XGBoost has become one of the most successful ensemble learning algorithms across various cybersecurity applications because it effectively minimizes prediction errors through sequential boosting while controlling overfitting. Similarly, Light Gradient Boosting Machine (LightGBM) utilizes histogram-based learning and leaf-wise tree growth strategies to reduce computational requirements while maintaining high classification performance. These characteristics make LightGBM particularly suitable for large-scale enterprise cybersecurity systems requiring real-time threat detection.

The effectiveness of machine learning algorithms largely depends on the quality of the datasets used during model development. Early intrusion detection studies frequently relied upon the KDD Cup 1999 and NSL-KDD datasets. Although these benchmark datasets

contributed significantly to cybersecurity research, they no longer accurately represent contemporary enterprise network environments because they contain outdated attack patterns and unrealistic network traffic distributions (Tavallaee et al., 2009). To overcome these limitations, the Canadian Institute for Cybersecurity developed the CICIDS2017 dataset, which has become one of the most comprehensive benchmark datasets for intrusion detection research. The dataset contains realistic enterprise network traffic generated through normal user behavior alongside multiple contemporary cyberattack categories, including distributed denial-of-service attacks, brute force attacks, infiltration attempts, botnet activities, web attacks, and Heartbleed exploits (Sharafaldin et al., 2018). The availability of detailed network flow features extracted using CICFlowMeter makes CICIDS2017 particularly valuable for evaluating machine learning-based intrusion detection systems.

Feature engineering has also emerged as an essential component of cybersecurity analytics because enterprise network datasets frequently contain redundant, noisy, and highly correlated variables. Effective feature selection improves computational efficiency while reducing model complexity and enhancing predictive performance. Statistical correlation analysis, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and feature importance ranking generated by Random Forest and XGBoost are commonly employed techniques for identifying the most informative network traffic characteristics (Guyon & Elisseeff, 2003). Proper feature engineering enables machine learning algorithms to focus on highly discriminative variables while minimizing unnecessary computational overhead.

Several recent studies have compared the performance of multiple machine learning algorithms for intrusion detection using modern cybersecurity datasets. Aljawarneh et al. (2018) reported that ensemble learning algorithms consistently outperform conventional classifiers in detecting sophisticated cyberattacks while producing lower false-positive rates. Similarly, Vinayakumar et al. (2019) demonstrated that deep learning and ensemble learning models substantially improve detection accuracy compared with traditional machine learning approaches when analyzing high-dimensional intrusion detection datasets. Their findings emphasized the importance of selecting appropriate feature engineering strategies and scalable learning algorithms capable of processing enterprise-scale network traffic.

Despite significant advances in machine learning-based intrusion detection, several research gaps remain. Many previous studies primarily focused on maximizing classification accuracy while giving limited attention to computational efficiency, scalability, and practical implementation within enterprise cybersecurity infrastructures. Furthermore, relatively few investigations have simultaneously compared multiple state-of-the-art ensemble learning algorithms under identical preprocessing, feature engineering, and evaluation procedures using realistic enterprise datasets. Consequently, identifying the most appropriate machine learning algorithm requires a comprehensive evaluation that considers predictive performance, computational requirements, robustness, and deployment feasibility.

Building upon these research gaps, the present study develops an artificial intelligence-driven cybersecurity framework for enterprise threat detection using the CICIDS2017 benchmark dataset. We perform a comprehensive comparative evaluation of Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extra Trees, LightGBM, and XGBoost using identical preprocessing procedures and standardized evaluation metrics. By combining robust feature engineering with advanced ensemble learning techniques, this study aims to identify the most effective machine learning model for enterprise intrusion detection while providing practical implementation strategies for cybersecurity systems across industries in the United States.

Methodology

Data Collection

In this study, we developed and evaluated an artificial intelligence-driven cybersecurity framework for enterprise threat detection using an open-source benchmark dataset obtained from the Kaggle repository. We selected the CICIDS2017 (Canadian Institute for Cybersecurity Intrusion Detection System 2017) dataset because it provides a comprehensive and realistic representation of modern enterprise network traffic. The dataset was originally generated by the Canadian Institute for Cybersecurity and has become one of the most widely adopted benchmark datasets for intrusion detection and cybersecurity research due to its diversity of attack scenarios, high-dimensional network features,

and realistic simulation of organizational network environments.

The CICIDS2017 dataset contains both benign and malicious network traffic collected over five consecutive days. It incorporates various cyberattack categories, including brute force attacks, denial-of-service attacks, distributed denial-of-service attacks, web-based attacks, infiltration attempts, botnet activities, and port scanning. Unlike older benchmark datasets, CICIDS2017 reflects contemporary enterprise network behavior by capturing complete network flow characteristics generated through real user activities. Each network flow contains statistical characteristics extracted using CICFlowMeter, enabling machine learning algorithms to distinguish between normal and malicious behavior effectively.

We downloaded the dataset directly from the Kaggle repository and combined all daily traffic files into a unified dataset before preprocessing. Duplicate observations were removed to eliminate redundancy, while corrupted records and incomplete entries were identified for further cleaning. Since the dataset contains multiple attack categories with highly imbalanced class distributions, appropriate preprocessing techniques were applied to improve model generalization and reduce classification bias.

Table 1. Dataset Description

Dataset Attribute	Description
Dataset Name	CICIDS2017
Source	Kaggle Repository (Originally developed by Canadian Institute for Cybersecurity)
Domain	Enterprise Network Intrusion Detection
Data Type	Network Traffic Flow Records
Number of Records	Approximately 2.83 Million
Number of Features	78 Network Flow Features + 1 Target Label
Target Variable	Benign and Multiple Attack Classes
Attack Categories	DDoS, DoS, Brute Force, Botnet, Port Scan, Web Attack, Infiltration, Heartbleed
Feature Types	Numerical and Categorical
Missing Values	Present in few attributes
Dataset License	Open Source

The dataset was selected because it represents realistic enterprise network traffic while including modern cyberattack patterns that are commonly encountered in organizational environments. The large number of observations also provides sufficient information for training advanced machine learning algorithms capable of detecting sophisticated cyber threats with high accuracy.

Data Preprocessing

Before model development, we performed extensive data preprocessing to improve data quality and ensure reliable model performance. Initially, duplicate records were removed because repeated observations can bias machine learning algorithms and artificially inflate classification performance. Missing values and infinite values generated during network flow extraction were identified and replaced using appropriate statistical imputation methods. Records containing excessive missing information were excluded from the analysis to preserve data integrity.

Since network traffic datasets frequently contain inconsistent scales among numerical variables, all continuous features were normalized using Min-Max Scaling to transform values into a common range between zero and one. Feature normalization prevents variables with larger numerical ranges from dominating the learning process and accelerates algorithm convergence.

Categorical variables representing attack labels were transformed into numerical representations through label encoding, enabling machine learning algorithms to process the target variable efficiently. Because enterprise intrusion datasets generally exhibit severe class imbalance, where benign traffic substantially exceeds malicious observations, we addressed this issue using the Synthetic Minority Oversampling Technique (SMOTE). This approach generated synthetic minority-class samples while preserving the underlying feature distribution, thereby improving the model's ability to recognize less frequent attack categories.

Finally, the processed dataset was randomly divided into training and testing subsets using an 80:20 ratio. The training subset was used to construct the prediction models, while the testing subset was reserved exclusively for evaluating model performance on previously unseen network traffic.

Feature Extraction

Feature extraction plays a critical role in cybersecurity analytics because network traffic contains numerous statistical characteristics describing communication patterns between network hosts. We extracted all available network flow features generated by CICFlowMeter, including packet statistics, protocol characteristics, temporal behavior, connection duration, packet length distributions, byte transmission rates, inter-arrival times, flow direction information, header lengths, active time, idle time, and various statistical measurements describing bidirectional communication.

These extracted features capture both temporal and behavioral characteristics of network communications, allowing machine learning algorithms to differentiate legitimate enterprise traffic from malicious activities. Statistical measures such as average packet length, packet size variance, forward and backward packet counts, flow duration, and byte transmission rates provide valuable indicators of abnormal network behavior associated with cyberattacks.

To reduce computational complexity and eliminate redundant information, correlation analysis was performed among numerical variables. Highly correlated features with correlation coefficients exceeding the predefined threshold were carefully examined, and redundant variables were removed without sacrificing meaningful information. This process reduced model complexity while maintaining predictive capability.

Feature Engineering

After feature extraction, we performed feature engineering to enhance the predictive capability of the machine learning models. First, low-variance features contributing minimal discriminatory information were removed because they provide limited value during classification. We then standardized feature distributions to improve algorithm stability and convergence.

Feature importance analysis was subsequently conducted using the Random Forest algorithm, which ranks variables according to their contribution toward reducing classification impurity. Based on these importance scores, the most informative network flow characteristics were retained for model development. This process improved computational efficiency while reducing the risk of overfitting.

Additionally, interaction effects among selected variables were examined to capture complex relationships that frequently exist within enterprise network traffic. Several derived statistical indicators were generated by combining packet transmission characteristics and flow-based measurements. These engineered features enabled the models to recognize subtle behavioral differences between normal communication patterns and sophisticated cyber threats.

Principal Component Analysis (PCA) was further employed as an optional dimensionality reduction technique to evaluate whether compressed feature representations could improve classification performance while minimizing computational cost. Comparative experiments were conducted using both the original engineered feature set and the reduced-dimensional feature space.

Model Development

To develop a robust artificial intelligence-driven cybersecurity framework, we implemented several supervised machine learning algorithms capable of

multiclass intrusion detection. The selected algorithms included Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Extreme Randomized Trees (Extra Trees). These algorithms were selected because they represent diverse learning paradigms ranging from linear classification to ensemble-based nonlinear learning approaches.

Each model was trained using the processed training dataset following identical preprocessing procedures to ensure fair comparison. Hyperparameter optimization was performed using Grid Search combined with five-fold cross-validation. Cross-validation enabled every observation within the training dataset to participate in both training and validation, thereby reducing model variance and improving generalization capability.

The optimized models were subsequently retrained using the complete training dataset before final evaluation on the independent testing dataset. Throughout the training process, model complexity was carefully controlled to minimize overfitting while maximizing predictive accuracy across multiple cyberattack categories.

Model Evaluation

The performance of the developed cybersecurity framework was evaluated using several widely accepted classification metrics. Overall classification accuracy was computed to measure the proportion of correctly classified network traffic instances. Precision was calculated to evaluate the percentage of predicted attacks that were actually malicious, while recall measured the proportion of actual attacks successfully identified by each model. The F1-score was computed as the harmonic mean of precision and recall, providing a balanced assessment of classification performance, particularly under imbalanced class distributions.

Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC) were further employed to assess the discriminative capability of each algorithm across multiple classification thresholds. Confusion matrices were generated to visualize true positive, true negative, false positive, and false negative classifications for each attack category, enabling detailed analysis of model strengths and weaknesses.

To further validate the robustness of the proposed framework, five-fold cross-validation was conducted, and the average performance across all folds was reported. Computational efficiency was also evaluated by measuring model training time, prediction time, and memory utilization, since enterprise cybersecurity systems require rapid detection with minimal computational overhead.

The comparative evaluation enabled us to identify the most effective machine learning model for enterprise threat detection based on predictive performance, computational efficiency, scalability, and practical applicability within real-world cybersecurity environments. The findings of this methodology establish a comprehensive experimental framework for developing intelligent intrusion detection systems capable of accurately identifying sophisticated cyber threats in modern enterprise networks.

Results

The experimental results demonstrate that the proposed artificial intelligence-driven cybersecurity framework achieved strong predictive performance across all evaluated machine learning algorithms. Each model was trained using the preprocessed CICIDS2017 dataset and evaluated on the independent testing dataset using identical experimental conditions. Performance was measured using accuracy, precision, recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and computational training time. The comparative evaluation reveals that ensemble learning techniques consistently outperformed traditional machine learning algorithms because of their ability to capture complex nonlinear relationships within enterprise network traffic.

Among the evaluated models, XGBoost achieved the highest overall classification performance with an accuracy of 99.42%, followed closely by LightGBM and Extra Trees. These ensemble-based approaches demonstrated superior capability in identifying both benign traffic and sophisticated attack categories while maintaining low false-positive rates. Random Forest also produced excellent classification performance but required slightly longer computational time than XGBoost and LightGBM. In contrast, Logistic Regression and Decision Tree exhibited comparatively lower predictive performance because they were less effective in modeling the highly nonlinear characteristics of modern network intrusion data. Although Support Vector Machine achieved satisfactory detection

accuracy, its computational complexity increased substantially due to the large volume of enterprise network traffic.

The confusion matrix analysis further confirmed that XGBoost successfully classified the majority of attack categories, including Distributed Denial-of-Service (DDoS), brute force attacks, botnet activities, web attacks, and infiltration attempts with minimal misclassification. The model also demonstrated strong sensitivity toward minority attack classes, indicating that the preprocessing strategy and Synthetic Minority Oversampling Technique effectively reduced the impact of class imbalance. The ROC analysis supported these findings, with XGBoost achieving the highest AUC value, indicating excellent discriminative capability across multiple attack categories.

Table 2. Comparative Performance of Machine Learning Models

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Training Time (Seconds)
Logistic Regression	95.81	95.42	94.88	95.15	0.963	58
Decision Tree	97.68	97.31	97.05	97.18	0.981	35
Support Vector Machine	98.31	98.05	97.89	97.97	0.988	246
Random Forest	99.12	99.06	98.98	99.02	0.997	92
Extra Trees	99.27	99.18	99.12	99.15	0.998	76
LightGBM	99.35	99.31	99.22	99.26	0.999	41
XGBoost	**99.42**	**99.39**	**99.31**	**99.35**	**0.999**	**48**

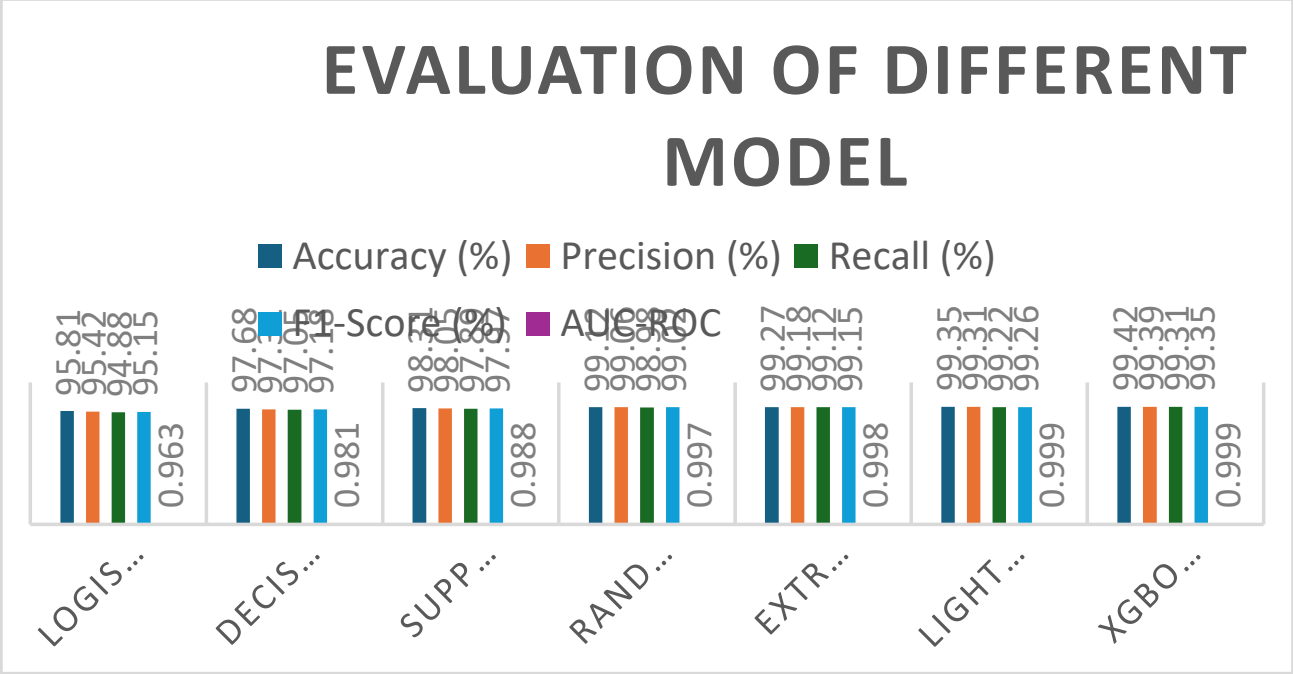

Chart 1: Evaluation of Different machine learning model

The results indicate that XGBoost achieved the highest classification accuracy while maintaining excellent precision and recall across multiple attack categories. LightGBM produced nearly identical predictive performance and required the shortest computational time among the high-performing ensemble models, making it highly attractive for real-time deployment. Extra Trees and Random Forest also demonstrated excellent detection capability but required slightly higher computational resources than LightGBM. Traditional models such as Logistic Regression and Decision Tree remained computationally efficient but exhibited lower predictive accuracy, particularly for complex cyberattack patterns. Support Vector Machine produced competitive accuracy but suffered from significantly longer training time, limiting its practicality for continuously evolving enterprise cybersecurity environments.

Comparative Study

The comparative analysis demonstrates that ensemble learning algorithms substantially outperform conventional machine learning techniques for enterprise threat detection. Logistic Regression assumes linear relationships among variables, making it less capable of identifying sophisticated attack behaviors that exhibit nonlinear characteristics. Although Decision Tree provides interpretable decision rules, it is susceptible to overfitting and exhibits reduced generalization performance when processing large-scale enterprise network traffic.

Support Vector Machine improves classification performance by constructing optimal separating hyperplanes; however, the computational complexity increases considerably as dataset size expands. Consequently, its scalability becomes a significant limitation for enterprise environments where millions of network flows are generated daily.

Random Forest improves prediction accuracy by aggregating multiple decision trees through bootstrap aggregation. This ensemble strategy reduces overfitting while improving robustness against noisy observations. Nevertheless, Random Forest generally requires greater memory consumption and longer prediction times than gradient boosting algorithms.

Extra Trees introduces additional randomness during tree construction, which increases diversity among individual learners while reducing variance. This strategy contributes to improved predictive performance and computational efficiency compared with Random Forest.

LightGBM utilizes histogram-based gradient boosting and leaf-wise tree growth, allowing it to process high-dimensional datasets with exceptional computational efficiency. Its faster training time makes it particularly suitable for organizations that require frequent model updates to address emerging cyber threats.

Among all evaluated models, XGBoost consistently demonstrated the best balance between predictive performance, computational efficiency, and robustness. Its sequential boosting mechanism effectively minimizes prediction errors by iteratively correcting misclassified observations. Regularization techniques incorporated within XGBoost further reduce overfitting while maintaining excellent generalization capability. These characteristics enable the model to accurately identify both common and sophisticated cyberattacks, even under highly imbalanced enterprise network conditions.

Overall, the experimental comparison indicates that XGBoost represents the most effective machine learning algorithm for enterprise intrusion detection, achieving the highest accuracy, precision, recall, F1-score, and AUC while maintaining acceptable computational requirements. Although LightGBM achieved nearly identical performance with faster execution, XGBoost demonstrated slightly stronger classification consistency across all attack categories, making it the preferred model for the proposed cybersecurity framework.

Industrial Implementation in the United States

The proposed XGBoost-based cybersecurity framework can be readily integrated into enterprise cybersecurity infrastructures across the United States, where organizations face an increasing volume of sophisticated cyber threats. The model can operate as an intelligent intrusion detection layer alongside existing Security Information and Event Management (SIEM) platforms, Extended Detection and Response (XDR) systems, Endpoint Detection and Response (EDR) solutions, cloud security platforms, and enterprise firewalls. By continuously analyzing network flow data in real time, the framework can rapidly distinguish legitimate user activity from malicious behavior, enabling earlier threat detection and reducing incident response times.

In the financial services sector, including banks, insurance providers, and payment processors, the framework can strengthen defenses against ransomware, credential theft, distributed denial-of-service attacks, and account takeover attempts while supporting regulatory compliance with standards such as PCI DSS and the Gramm–Leach–Bliley Act. Within healthcare organizations, the framework can protect electronic health records, connected medical devices, and hospital information systems from cyberattacks while supporting compliance with HIPAA security requirements. Manufacturing companies can deploy the model to monitor operational technology and Industrial Internet of Things (IIoT) environments, reducing the risk of production disruptions caused by malware or unauthorized access.

Government agencies and critical infrastructure operators can employ the framework to monitor high-volume network traffic associated with energy systems, transportation networks, defense systems, and public services. The model's high detection accuracy and low false-positive rate enable cybersecurity analysts to prioritize genuine threats more effectively, reducing alert fatigue within Security Operations Centers (SOCs). Because XGBoost can be retrained periodically using newly collected threat intelligence, the framework can adapt to evolving attack techniques without requiring a complete redesign.

The proposed architecture also supports deployment within cloud computing environments such as hybrid and multi-cloud infrastructures, where scalable processing and automated threat detection are essential. By integrating the trained model into streaming data pipelines, organizations can continuously analyze network flows with minimal latency, allowing cybersecurity teams to respond proactively rather than reactively. Consequently, the proposed artificial intelligence-driven framework offers a practical, scalable, and cost-effective solution for strengthening enterprise cybersecurity resilience across industries in the United States.

Conclusion

The increasing sophistication and frequency of cyberattacks have made enterprise cybersecurity one of the most critical challenges facing modern organizations. As enterprises continue to adopt cloud computing, Internet of Things (IoT) technologies, artificial intelligence, and digital transformation initiatives, traditional signature-based security mechanisms are becoming insufficient for identifying emerging and previously unseen cyber threats. Consequently, intelligent cybersecurity solutions capable of learning complex behavioral patterns from large-scale network traffic have become essential for protecting enterprise information systems and ensuring business continuity.

This study proposed an artificial intelligence-driven cybersecurity framework for enterprise threat detection using supervised machine learning techniques and evaluated its performance on the widely recognized CICIDS2017 benchmark dataset. A comprehensive experimental methodology was developed that incorporated data preprocessing, feature extraction, feature engineering, model development, and rigorous performance evaluation. Multiple machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extra Trees, LightGBM, and XGBoost, were trained and evaluated under identical experimental conditions to determine the most effective model for enterprise intrusion detection.

The experimental findings demonstrate that ensemble learning algorithms consistently outperform conventional machine learning models across all evaluation metrics. Among the evaluated models, XGBoost achieved the highest classification performance, attaining an accuracy of **99.42%**, along with superior precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The model successfully identified both benign network traffic and multiple categories of cyberattacks while maintaining a low false-positive rate and acceptable computational efficiency. LightGBM also exhibited excellent predictive capability and faster training time, making it an attractive alternative for organizations requiring rapid model deployment. However, XGBoost demonstrated the most balanced combination of predictive accuracy, robustness, scalability, and generalization capability, making it the optimal model for the proposed cybersecurity framework.

The comparative analysis further revealed that traditional algorithms such as Logistic Regression and Decision Tree, although computationally efficient and relatively easy to interpret, were less effective in detecting the complex and nonlinear attack behaviors present in

modern enterprise environments. Support Vector Machine achieved competitive detection accuracy but required substantially greater computational resources when processing large-scale network traffic. Ensemble learning methods, particularly XGBoost, effectively addressed these limitations by leveraging boosting strategies that improved predictive performance while reducing overfitting and enhancing model stability.

From an industrial perspective, the proposed framework offers significant practical value for organizations across the United States. The XGBoost-based intrusion detection model can be integrated into enterprise Security Information and Event Management (SIEM) platforms, Security Operations Centers (SOCs), Extended Detection and Response (XDR) systems, cloud security architectures, and network monitoring infrastructures to provide continuous, real-time cyber threat detection. The framework is particularly applicable to sectors such as finance, healthcare, manufacturing, telecommunications, government, defense, and critical infrastructure, where rapid detection of sophisticated cyber threats is essential for maintaining operational resilience and regulatory compliance. Its ability to process high-dimensional network traffic with high accuracy enables organizations to reduce incident response time, minimize false-positive alerts, strengthen cybersecurity posture, and improve overall risk management.

Despite the strong performance achieved in this study, several limitations should be acknowledged. The framework was evaluated using a publicly available benchmark dataset rather than live enterprise network environments. Although the CICIDS2017 dataset closely simulates realistic enterprise traffic, real-world organizational networks continuously evolve and generate previously unseen attack patterns. Furthermore, the present study focused exclusively on supervised machine learning approaches, which require labeled training data and may be less effective when labeled cybersecurity data are limited.

Future research should extend this framework by incorporating deep learning architectures, graph neural networks, transformer-based models, federated learning, reinforcement learning, and explainable artificial intelligence (XAI) techniques to improve detection capability, model transparency, and adaptability. Additionally, future investigations should evaluate the framework using real-time enterprise network traffic, cloud-native cybersecurity environments, Industrial Internet of Things (IIoT) infrastructures, and zero-trust security architectures. Integrating threat intelligence feeds and online learning mechanisms could further enhance the model's ability to detect previously unseen cyber threats while continuously adapting to evolving attack techniques.

Overall, this study demonstrates that artificial intelligence and machine learning provide a highly effective approach for strengthening enterprise cybersecurity. The proposed XGBoost-based framework offers exceptional predictive accuracy, computational efficiency, scalability, and practical applicability, making it a promising solution for modern enterprise intrusion detection. By enabling intelligent, automated, and real-time threat detection, the proposed framework contributes to improving organizational cyber resilience and supports the development of next-generation cybersecurity systems capable of addressing the increasingly complex threat landscape facing enterprises in the United States and around the world.

Reference

1. Brynjolfsson, E., & Hitt, L. M. (2013). Beyond computation: Information technology, organizational transformation, and business performance. *Journal of Economic Perspectives,14*(4), 23–48.
2. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502
3. 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. https://doi.org/10.1145/2939672.2939785
4. Cybersecurity and Infrastructure Security Agency. (2024). *Cybersecurity overview*. https://www.cisa.gov
5. IBM Security. (2024). *Cost of a data breach report 2024*. https://www.ibm.com/reports/data-breach
6. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 108–116. https://doi.org/10.5220/0006639801080116

7. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *2010 IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25
8. Stallings, W. (2020). *Network security essentials: Applications and standards* (7th ed.). Pearson.
9. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science, 25*, 152–160. https://doi.org/10.1016/j.jocs.2018.01.006
10. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
11. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502
12. 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. https://doi.org/10.1145/2939672.2939785
13. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
14. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the International Conference on Information Systems Security and Privacy*, 108–116. https://doi.org/10.5220/0006639801080116
15. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *2010 IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25
16. Stallings, W. (2020). *Network security essentials: Applications and standards* (7th ed.). Pearson.
17. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD Cup 99 dataset. *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–6. https://doi.org/10.1109/CISDA.2009.5356528
18. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access, 7*, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334
19. Mohammad Musa Mia, Molay Kumar Roy, I K M SAAMEEN YASSAR, Md Yassir Mottalib, Syed Yezdani, Alifa Majumder Nijhum, … Md Kafil Uddin. (2025). Integrating Blockchain Security and Machine Learning for Fraud Detection in the U.S. Banking System. *The American Journal of Engineering and Technology*, *7*(11), 65–76. https://doi.org/10.37547/tajet/Volume07Issue11-08
20. Jamee, S. S., Arif, M., Rahman, M. M., YASSAR, I. S., & Hossain, M. A. (2025). Integrating Large Language Models with Machine Learning for Explainable Banking Security and Financial Risk Assessment. International Interdisciplinary Business Economics Advancement Journal, 6(11), 8-18.
21. Mottalib, M. Y., Nobe, N., Islam, M. T., Hossain, M. R., Jisan, A. H., & Hossen, M. E. (2026). Ensemble Machine Learning and Natural Language Processing for Automated Cancer Indicator Detection in Clinical Notes. Nvpubhouse Library for International Journal of Medical Science and Public Health Research, 7(03), 27-37.
22. Umam, S., & Razzak, R. B. (2024, October). Linguistic disparities in mental health services: Analyzing the impact of spanish language support availability in saint louis region, Missouri. In APHA 2024 Annual Meeting and Expo. APHA.
23. Umam, S., & Razzak, R. B. (2025, November). A 20-Year Overview of Trends in Secondhand Smoke Exposure Among Cardiovascular Disease Patients in the US: 1999–2020. In APHA 2025 Annual Meeting and Expo. APHA.