IJSES
International Journal of Scientific Engineering and Science
Volume 9, Issue 4, pp. 102-106, 2025. ISSN (Online): 2456-7361

Methods of Proactive Detection of Cyber Threats Using Machine Learning

Rajesh Kumar C G
CEO, INOVITSI
Sydney, Australia
Email address: rajeshkumar.cg@inovitsi.com

Abstract— *This study examines modern methods for proactively detecting cyber threats in critical information systems (hereinafter referred to as CIP) based on machine learning algorithms, including Random Forest, Support Vector Machine, Multi-Layer Perceptron, AdaBoost, and hybrid approaches. The advantages and limitations of these methods are analyzed in the context of early anomaly detection, reduction of false positives, and adaptation to dynamic attack types. Particular attention is given to the specifics of infrastructure characterized by high failure risks and the inadmissibility of even short-term downtimes. The study presents the results of a comparative performance analysis of various ML models, including key metrics such as Accuracy, Recall, Precision, F1-score, and ROC-AUC, along with their practical applicability in real-world scenarios. Technical aspects and economic efficiency are considered, emphasizing the importance of algorithm parameter tuning and continuous model retraining. A recommended implementation plan is provided, covering data preparation, automation of ML module deployment, false positive control, and periodic security audits. The findings confirm the high value of proactive cyber threat detection, ensuring timely and accurate threat identification in complex infrastructure environments. This study will be of interest to researchers focused on the theoretical justification and enhancement of incident prediction models, as well as to professionals seeking to integrate these technologies into corporate and government monitoring and security systems.*

Keywords— *Proactive threat detection, machine learning, cybersecurity, infrastructure, Random Forest, AdaBoost, hybrid models, SIEM, SOAR.*

I. INTRODUCTION

Modern cyber threats are characterized by an increasing number of multi-stage, targeted (APT) and high-precision attacks capable of disrupting both commercial organizations and critical infrastructure, including energy, transportation, and healthcare sectors. Traditional security measures such as antivirus software, signature-based intrusion detection systems, and rule-based filters primarily focus on known patterns of malicious activity. However, these methods often fail to respond in time to the emergence of zero-day vulnerabilities, creating favorable conditions for cyberattacks that are unpredictable in vector, complex in execution, and rapidly spreading. This is particularly critical in environments requiring continuous operation, such as SCADA systems and industrial IoT devices.

The transition from reactive to proactive attack detection methods enables the identification of anomalous activity before clear indicators of compromise appear. Machine learning (ML), combined with big data techniques, facilitates the automated analysis of extensive logs, network traffic, and telemetry in near real-time, revealing hidden patterns and scenarios that are difficult to detect manually.

Shan A. and Myeong S. [1] propose a hybrid algorithm combining traditional statistical methods with modern machine learning models for threat hunting in critical infrastructure. In parallel, Jeffrey N., Tan Q., and Villar J. R. [2] review anomaly detection strategies in cyber-physical systems, emphasizing the necessity of integrating traditional detection methods with machine learning algorithms for timely threat identification. Goenka R., Chawla M., and Tiwari N. [3] provide a comprehensive review of phishing attacks, introducing a new taxonomy that allows for accurate threat classification and the development of protective mechanisms. Similarly, Khraisat A. et al. [4] analyze intrusion detection systems, highlighting the issue of dataset and method diversity, which reflects a scientific gap in standardizing approaches.

Nasereddin M. et al. [5], in a systematic review of SQL injection detection and prevention techniques, focus on improving system reliability by applying machine learning for proactive attack prediction. In the field of corporate security, a significant contribution is made by the review conducted by Nour B., Pourzandi M., and Debbabi M. [6], which explores modern cybersecurity methods in corporate networks. The authors emphasize that integrating real-time operational data with predictive models helps minimize the impact of cyber incidents. Tahmasebi M. [7] presents a concept that extends beyond traditional defense mechanisms. An innovative approach is also demonstrated by Kumar P. et al. [8], who describe the integration of blockchain technology with artificial intelligence, addressing the black-box problem in traditional models and setting new standards in risk assessment. An interdisciplinary perspective is presented in the work of Zhang J. et al. [9], where machine learning methods and metaheuristic algorithms are applied to optimize parameters in an engineering task. Despite the thematic distance, this approach illustrates the potential for adapting similar methods to enhance cybersecurity model parameterization.

The identified research gap indicates that despite advancements in applying machine learning for network intrusion detection, there is still a lack of studies focused specifically on proactive cyber threat detection in critical infrastructure using hybrid ML algorithms. Existing research either concentrates on reactive scenarios, where an attack is already in progress, or is limited to comparisons of basic ML models. Additionally, the challenge of reducing false positives

while maintaining high sensitivity to novel anomalous patterns remains unresolved.

The objective of this study is to analyze the effectiveness of various machine learning methods for proactive threat detection in critical infrastructure systems and provide recommendations for selecting the most effective approach, considering the need to reduce false positives while improving anomaly detection accuracy.

The scientific contribution of this study lies in conducting an extensive review of recent research publications and subsequently offering recommendations on the application of machine learning methods for proactive cyber threat detection.

The working hypothesis is that hybrid ML algorithms combining the strengths of ensemble methods (such as RF and AdaBoost) with optimization and neural network techniques (such as MLP-based or genetic algorithms) can achieve higher accuracy and lower false positive rates compared to using each of these approaches individually.

As a methodological basis, this study includes a comparative analysis of open-access scientific publications.

II. Review of existing approaches to proactive cyber threat detection

Modern cybersecurity practices demonstrate an evolution from traditional reactive attack detection methods to fundamentally new proactive approaches aimed at identifying potentially malicious activity before a security incident occurs. Figure 1 presents the main groups of approaches used for predictive threat detection and their place within the framework of critical system protection.

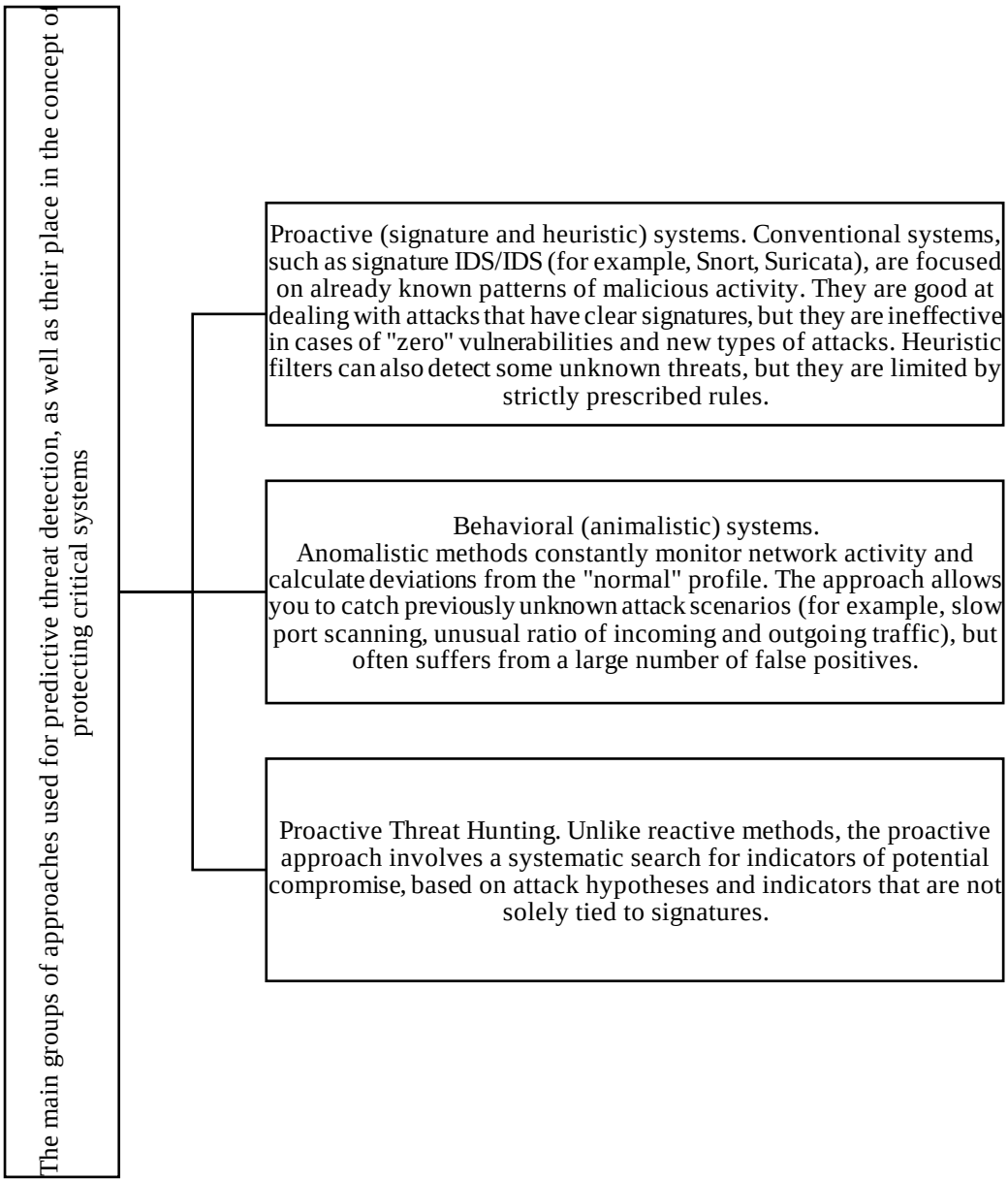

Fig. 1. The main groups of approaches used for predictive threat detection, as well as their place in the concept of protecting critical systems [1].

International Journal of Scientific Engineering and Science
Volume 9, Issue 4, pp. 102-106, 2025. ISSN (Online): 2456-7361

At the same time, unsupervised algorithms such as clustering and autoencoders allow for anomaly detection without explicit labeling by identifying statistical deviations [5]. These algorithms are more effective for early identification of emerging threats but tend to generate more false positives.

Hybrid ML models combine the advantages of multiple approaches, such as ensemble methods (Random Forest, Gradient Boosting) and optimization algorithms (metaheuristics, genetic algorithms). It has been demonstrated that this combination reduces overfitting and increases overall detection accuracy. In the context of proactive threat detection, hybrid solutions provide additional flexibility by adapting to various attack scenarios and identifying new threat patterns [7].

Beyond classical ML algorithms, Deep Learning methods (such as convolutional and recurrent neural networks) have gained significant traction in recent years, particularly in analyzing large volumes of logs and network traffic in real time [6]. However, these methods require substantial computational resources and present challenges in interpreting their internal structures. In resource-constrained environments, such as IoT devices, lightweight models or online learning mechanisms are often used to continuously adjust classifiers as new data becomes available [8].

Proactive threat detection rarely operates in isolation and is typically integrated into a broader ecosystem of Security Information and Event Management (SIEM) or Security Orchestration, Automation, and Response (SOAR) solutions [1]. SIEM systems aggregate security events from multiple sources, including IDS/IPS, firewalls, and application logs, while proactive machine learning modules analyze anomalies within the combined data stream [4]. If signs of an attack are detected, automation systems enable an immediate response, such as blocking suspicious traffic or issuing alerts about a potential intrusion.

This integration is particularly critical in infrastructures where even short-term downtime is unacceptable and where attacks can have catastrophic consequences. Proactive methods integrated with SIEM and SOAR solutions significantly reduce threat detection and response times, thereby mitigating potential damage [6].

To provide a detailed understanding of proactive threat detection principles in cybersecurity, Table 1 summarizes popular methods used for predictive security analysis.

TABLE 1. A brief overview of popular methods of proactive threat detection (compiled by the author based on [1, 6, 8])

Method	Description	Advantages	Disadvantages
Anomaly-based IDS	Monitors deviations from statistical "normal" behavior in network traffic and logs	Detects previously unknown attack types; does not require constant signature updates	High rate of false positives; requires a long training phase to define "normal" patterns
ML Classifiers (RF, SVM, AdaBoost, etc.)	Supervised learning trained on labeled datasets (normal/malicious)	High accuracy with well-prepared training data; extensive tools for hyperparameter tuning	Limited ability to detect novel threats outside the training dataset; potential overfitting
Hybrid (Machine Learning + Optimization)	Combines multiple ML algorithms with optimization techniques (e.g., genetic algorithms)	Improved accuracy and reduced overfitting; flexible adaptation to different threat types	Increased computational costs; more complex development and integration
Deep Neural Networks (DL)	Multi-layer neural networks (CNN, RNN) often used for analyzing large-scale logs and network traffic	Effective for big data analysis; capable of detecting complex anomalous patterns	High computational requirements; difficult to interpret decision-making logic
Threat Intelligence (TI) Approach	Utilizes external intelligence on emerging threats (IoC lists, APT group activity) and integrates this data into monitoring systems	Enhances ML models with real-world indicators; accelerates response to known vulnerabilities	Dependent on the quality and timeliness of external threat data; potential challenges in correlating global and local indicators
Proactive Threat Hunting	Continuous system monitoring for potential indicators of compromise and hypothesis-based attack analysis	Detects anomalies before an actual incident occurs; reduces investigation time in case of a real attack	Requires expert involvement ("threat hunters"); necessitates SIEM/SOAR integration for efficient data aggregation

Collectively, these methods form the foundation of modern active cyber defense systems. The combination of anomaly-based approaches, ML models, external threat intelligence (Threat Intelligence), and manual analysis by threat hunters ensures optimal effectiveness when properly integrated and continuously trained. Machine learning, in conjunction with external sources and anomaly-based methods, enables the development of comprehensive early incident detection systems. Hybrid approaches demonstrate high potential but require increased computational resources and careful configuration.

III. Machine Learning Methods for Proactive Threat Hunting: A Comparative Analysis

In the context of proactive threat detection in infrastructure environments, machine learning (ML) algorithms are gaining increasing popularity. When properly configured, ML models can identify both known and previously unseen attack patterns, reducing response time and improving detection accuracy.

Random Forest (RF) is an ensemble method based on constructing multiple independent decision trees. Each tree-classifier is trained on a bootstrap sample of the original dataset, and classification decisions are made through majority voting.

Support Vector Machine (SVM) constructs a hyperplane or a set of hyperplanes that separate feature spaces into "attack" and "normal" classes [8].

MLP is a classical fully connected feedforward neural network with at least one hidden layer, trained using the backpropagation algorithm.

AdaBoost (Adaptive Boosting) is a boosting algorithm that iteratively builds a strong ensemble from weak classifiers, increasing the weights of misclassified instances.

Hybrid models are constructed by combining multiple ML algorithms (e.g., Random Forest + SVM) or integrating

machine learning with optimization techniques such as genetic algorithms or particle swarm optimization [6].

To illustrate the strengths and weaknesses of the listed ML algorithms in proactive cyber threat detection for infrastructure, Table 2 provides a comparative overview.

TABLE 2. Comparison of ML algorithms in the context of proactive threat search (compiled by the author based on [1, 4])

Algorithm	Accuracy (Accuracy / ROC-AUC)	Configuration Complexity	Computational Cost	Noise Sensitivity	Real-Time Applicability
Random Forest	High	Moderate (number of trees, depth, etc.)	Medium (parallelizable)	Low (robust to outliers)	Good (especially with parallelization)
SVM	Medium/High	High (kernel selection, C, etc.)	Medium	Medium (sensitive to kernel parameter selection)	Satisfactory (may be problematic for large datasets)
MLP	Medium (strong dependence on hyperparameters)	High (number of layers, neurons, learning rate)	Can be high for large networks	May overfit (requires regularization)	Satisfactory (with proper optimization)
AdaBoost	High	Moderate (base classifier type, number of iterations)	Low–Medium	Medium (sensitive to noise in data)	Good (for small datasets)
Hybrid Models	Very High	Very High (combination of multiple methods, optimization tuning)	High (requires resources for ensemble)	Low–Medium (depends on specific implementation)	Good (but requires optimization and resources)

As shown in Table 2, hybrid models often achieve the highest detection accuracy and are best suited for adapting to evolving attack profiles. However, their implementation can be challenging due to computational overhead and the complexity of configuration. Random Forest and AdaBoost serve as balanced solutions, offering relatively high accuracy while maintaining moderate resource requirements, as confirmed by multiple experiments in proactive threat detection within industrial networks.

IV. Practical Implementation and Recommendations for Deployment

Implementing a proactive cyber threat detection strategy using machine learning presents significant challenges due to a range of integration, computational, and organizational factors. For continuous monitoring and analysis of large data volumes, including logs, network traffic, and IoT telemetry, cluster-based or cloud-based solutions with GPU/TPU support are preferable. In high-load environments such as SCADA systems or distributed industrial networks, parallelized computing is required to enable machine learning algorithms, such as Random Forest or MLP, to process data streams in near real-time [6]. In cases where resources are limited, such as secure IoT networks in remote locations, lightweight models or edge analytics mechanisms can be utilized to filter part of the data before transmitting it to a central repository [1].

A crucial element in data collection and processing architecture is a system capable of aggregating security events from multiple sources, including network sensors (IDS/IPS), firewalls, SIEM logs, and IoT telemetry. The most commonly used architecture involves stream processing combined with a data lake for historical analysis. Before ingestion, data undergoes preprocessing, including cleaning, normalization, and handling of missing values [4,8]. Particular attention is given to securing the transmission channel through encryption and agent authentication to prevent log compromise or label tampering, as outlined in NIST SP 800-53 Rev. 5 (2020).

To ensure timely updates and retraining of ML models, it is advisable to use continuous integration (CI) and continuous delivery (CD) principles. For example, each model update—whether it involves incorporating new data or adjusting hyperparameters—should be automatically validated before being deployed into an operational environment. Performance metrics such as Precision, Recall, and F1-score, along with false positive rates, should be monitored in a dedicated validation environment. If accuracy falls below a predefined threshold, the system should revert to the last stable model version [3].

Regarding integration with SIEM and SOAR systems, SIEM solutions such as IBM QRadar, Splunk, and ArcSight allow for the aggregation of security events across different infrastructure layers. For proactive cyber threat detection, an additional custom ML module should be configured as an extension, enabling real-time anomaly analysis of incoming events [5]. Integration with external intelligence platforms, including IoC lists and APT group reports, further enhances detection by flagging potentially malicious IP addresses, domains, or files [1,5].

When suspicious activity is detected, a SOAR system, such as Palo Alto Networks XSOAR or Splunk Phantom, can automatically enforce predefined security policies, such as blocking an offender's IP address or quarantining a compromised node [1]. The ML model transmits anomaly detection alerts to the SOAR platform via an API or event queue. Threat analysts then assess the severity of the incident and determine the appropriate course of action, accelerating response time in critical scenarios [8].

The system should provide interpretable indicators of anomalous behavior to human operators [6]. While AdaBoost and Random Forest offer relatively transparent decision-making (feature importance, tree paths), MLP and deep learning models often require additional Explainable AI (XAI) techniques for interpretability [1,7]. Proactive threat detection prevents escalation, reducing financial losses caused by system downtime or data breaches. Automating routine actions, such as blocking an IP address or generating an incident management ticket, alleviates the workload on SOC operators and enhances operational efficiency [2,5].

For total cost of ownership (TCO) assessment, factors such as hardware expenses, expert time for model deployment and training, and potential financial losses due to undetected attacks

should be considered [1,4]. The NIST risk assessment framework can be adapted to meet the specific requirements of critical infrastructure protection (NIST SP 800-53 Rev. 5, 2020). Excessive false positives impose additional strain on security teams and may nullify the cost savings achieved through automation if they exceed acceptable tolerance thresholds [4]. Regular hyperparameter tuning and the integration of expert feedback loops improve the balance between sensitivity (detecting true threats) and specificity [1,8].

Below are recommendations for implementing ML modules for proactive threat detection in infrastructure environments.

A fundamental aspect of the proposed recommendations is the integration of big data processing methods with machine learning algorithms. Clustering, classification, and regression techniques play a crucial role in identifying anomalous patterns in network traffic behavior. The application of deep neural networks and reinforcement learning algorithms enhances detection accuracy and ensures system adaptability to new attack types.

Particular attention is given to the quality of data used for model training. It is recommended to create reliable, representative datasets and perform preprocessing steps such as normalization, outlier removal, and class balancing. These measures help reduce false positive rates and improve model stability in real-time conditions, which is critical for effective cybersecurity monitoring and attack prevention.

Another key aspect is the development of a scalable system architecture that can be integrated into existing enterprise infrastructure. The use of distributed computing platforms and cloud technologies is recommended for real-time analysis of large datasets. This architecture not only enables rapid incident response but also facilitates dynamic model adjustments, allowing them to adapt to changes in attacker behavior.

Finally, collaboration among cybersecurity experts, statisticians, and software developers ensures a comprehensive approach to the problem, contributing to the creation of effective, adaptive, and resilient security systems. Continuous algorithm improvement, experience sharing, and the integration of advanced technological solutions are essential for successfully countering cyber threats in an increasingly complex digital environment.

V. Conclusion

This study examined a range of machine learning algorithms used for proactive cyber threat detection in critical systems. A literature review demonstrated that combining multiple methods in hybrid models improves detection efficiency and reduces the risk of rare attack scenarios. However, such solutions require increased computational resources, complex configuration, and a well-structured data collection and processing architecture.

A comparative assessment of Random Forest, Support Vector Machine, Multi-Layer Perceptron, AdaBoost, and hybrid approaches confirmed that ensemble models, such as Random Forest and AdaBoost, often provide the best balance between accuracy, processing speed, and false positive rates. While MLP and SVM offer high sensitivity, they are prone to overfitting and are highly dependent on hyperparameter selection, necessitating regular adjustments.

The examined methods integrate closely with SIEM and SOAR systems, simplifying event correlation from various sources and accelerating incident response. Infrastructure organization plays a critical role, including parallel data processing, encrypted communication channels, periodic audits, and regular model updates.

References

1. Shan A., Myeong S. Proactive threat hunting in critical infrastructure protection through hybrid machine learning algorithm application //Sensors. – 2024. – Vol. 24 (15). – pp. 1-24.
2. Jeffrey N., Tan Q., Villar J. R. A review of anomaly detection strategies to detect threats to cyber-physical systems //Electronics. – 2023. – Vol. 12 (15). – pp. 1-10.
3. Goenka R., Chawla M., Tiwari N. A comprehensive survey of phishing: Mediums, intended targets, attack and defence techniques and a novel taxonomy //International Journal of Information Security. – 2024. – Vol. 23 (2). – pp. 819-848.
4. Khraisat A. et al. Survey of intrusion detection systems: techniques, datasets and challenges //Cybersecurity. – 2019. – Vol. 2 (1). – pp. 1-22.
5. Nasereddin M. et al. A systematic review of detection and prevention techniques of SQL injection attacks //Information Security Journal: A Global Perspective. – 2023. – Vol. 32 (4). – pp. 252-265.
6. Nour B., Pourzandi M., Debbabi M. A survey on threat hunting in enterprise networks //IEEE communications surveys & tutorials. – 2023. – Vol. 25 (4). – pp. 2299-2324.
7. Tahmasebi M. Beyond defense: Proactive approaches to disaster recovery and threat intelligence in modern enterprises //Journal of Information Security. – 2024. – Vol. 15 (2). – pp. 106-133.
8. Kumar P. et al. Blockchain and explainable AI for enhanced decision making in cyber threat detection //Software: Practice and Experience. – 2024. – Vol. 54 (8). – pp. 1337-1360.
9. Zhang J. et al. Multi-objective optimization of concrete mixture proportions using machine learning and metaheuristic algorithms //Construction and Building Materials. – 2020. – Vol. 253. – pp. 1-8.