iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

COMPARATIVE STUDY OF CYBERSECURITY AI MODELS: PERFORMANCE, SAFETY, TRUST, AND EXPLAINABILITY

Nikhil Purwaha
Nikhil.purwaha@gmail.com

ABSTRACT

The rapid integration of artificial intelligence (AI) in cybersecurity has significantly enhanced threat detection, response automation, and overall network defense (Smith & Zhang, 2021; Patel et al., 2022). However, the effectiveness of AI-driven cybersecurity systems depends not only on technical performance but also on safety, trustworthiness, and explainability (Jones et al., 2020). This study presents a comparative analysis of contemporary AI models applied to cybersecurity, focusing on intrusion detection systems (IDS), malware classification, and Security Operations Center (SOC) automation (Lee & Kim, 2021). Using a curated set of benchmark datasets and realistic threat scenarios, we evaluate each model's detection performance, vulnerabilities to adversarial attacks, and resilience against model drift (Nguyen et al., 2022). Additionally, we assess trust and explainability metrics, emphasizing human-in-the-loop (HITL) and human-on-the-loop (HOTL) frameworks, and the role of explainable AI (XAI) techniques in improving analyst confidence and decision-making (Doshi-Velez & Kim, 2017; Miller, 2019). Our results highlight critical trade-offs between accuracy, operational safety, and interpretability, revealing that high-performing models may exhibit lower explainability or increased susceptibility to certain adversarial threats (Goodfellow et al., 2015). The study provides a structured framework for selecting AI models in cybersecurity applications, balancing technical efficiency with safety and trust requirements, and offers recommendations for integrating multi-criteria evaluation in real-world deployments (Chio & Freeman, 2018).

Keywords:

Cybersecurity AI, Performance Evaluation, Explainable AI, Trust, Safety, Human-in-the-Loop, Intrusion Detection Systems

I. INTRODUCTION

AI adoption in cybersecurity has transformed the way in which organizations are able to detect, respond to, and respond to cyber threats. Traditional rule-based security systems often struggle with the scale, speed and sophistication of modern cyberattacks while AI powered systems are adaptive and automated and can recognise known and new threats (Buczak & Guven 2016; Sommer & Paxson 2010). In intrusion detection systems and malware classification and security operations center automation, machine learning models (ML) and deep learning models are becoming more important tools for detecting intrusions and performing much better in detection accuracy and operational efficiency, such as Sarker et al., 2020; Zhang et al., 2021).

Even when these advances have been made, there is no set-mate for the performance of AI models in cybersecurity. While safety, trust and explanation ability are key components to the development of AI systems, safety, trust, and explainability have become increasingly important. Safety concerns include adversarial attacks, model drift over time, and cascading failures that can augment risk in automated environments (Nguyen et al., 2018, Goodfellow et al., 2015). How to apply explanation able AI (XAI) techniques to human analysts who need to apply intelligent to high-stakes decision-making is a prerequisite for trustworthiness in human analysts and how these skills are applied to improve interpretation, transparency, and user confidence (Doshi-Velez & Kim, 2017; Miller, 2019).

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

Using the four basic aspects of performance, safety, trust and explanation ability is the first objective of this comparative study which seeks to evaluate current AI models developed in cybersecurity in three aspects: performance, safety, trust, and explainability. This work is done comparing multiple models from standard datasets and realistic threat scenarios to assess trade-offs between detection accuracy and operational reliability, providing a framework for model selection, and giving advice for deploying best practices on AI systems that are effective and reliable in complex cybersecurity environments.

II. LITERATURE REVIEW AND METHODOLOGY

2.1 Survey of AI Models Used in Cybersecurity

Artificial intelligence has been widely adopted in cybersecurity, particularly in intrusion detection systems (IDS), malware detection, and Security Operations Centre (SOC) automation. Supervised machine learning models, such as decision trees, support vector machines (SVMs), and random forests, have been extensively used for classification tasks in IDS and malware detection due to their high accuracy on labelled datasets (Buczak & Guven, 2016; Sarker et al., 2020). Deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown superior performance in recognizing complex patterns and temporal behaviours in network traffic (Kim et al., 2019; Zhang et al., 2021). Reinforcement learning approaches have also been employed to optimize automated threat response strategies within SOC environments, allowing dynamic adaptation to evolving cyber threats (Nguyen et al., 2022).

Table 1: Summary of AI Models and Their Typical Applications

AI Model Type	Application Domain	Key Advantages	Limitations
Decision Trees	IDS, Malware Detection	Interpretability, fast training	Limited scalability, prone to overfitting
Random Forests	IDS, Malware Detection	Robustness, ensemble learning benefits	Complex model, less transparent
SVM	IDS, Malware Detection	High accuracy with small datasets	Sensitive to feature scaling
CNN	Malware Classification	Pattern recognition in raw data	High computational cost
RNN / LSTM	Network Traffic Anomaly Detection	Handles temporal sequences	Training complexity, overfitting risks
Reinforcement Learning	SOC Automation	Adaptive threat response	Requires simulated environments

2.2 Existing Performance Benchmarks and Evaluation Metrics

The classification performance and robustness of the evaluating AI models for cybersecurity is commonly measured as metrics. Precision, recall, F1-score, and area under the ROC curve (AUC) are common performance metrics that measure the model's ability to detect threats without having to resort to false positives and negatives (Sommer & Paxson, 2010; Zhang et al., 2021). Further, robustness tests models against adversarial attacks, data poisoning and distribution shifts to better predict resilience in real-world settings (Goodfellow et al., 2015; Nguyen et al., 2018).

2.3 Selection Criteria for AI Models and Datasets

This was compared with an alternative research project based on the choice of AI models, a high prevalence in cybersecurity, and pre-training architectures that are appropriate for IDS, malware detection, or SOC automation applications. Datasets for the real-world and diverse threat environments were selected, including labelled network traffic, malware binaries, and simulated SOC alert logs. The principal selection criteria for the dataset size, class balance, feature richness, and relevance to current cyber threat landscapes are the quantity of samples, class balance, feature richness, and relevance.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

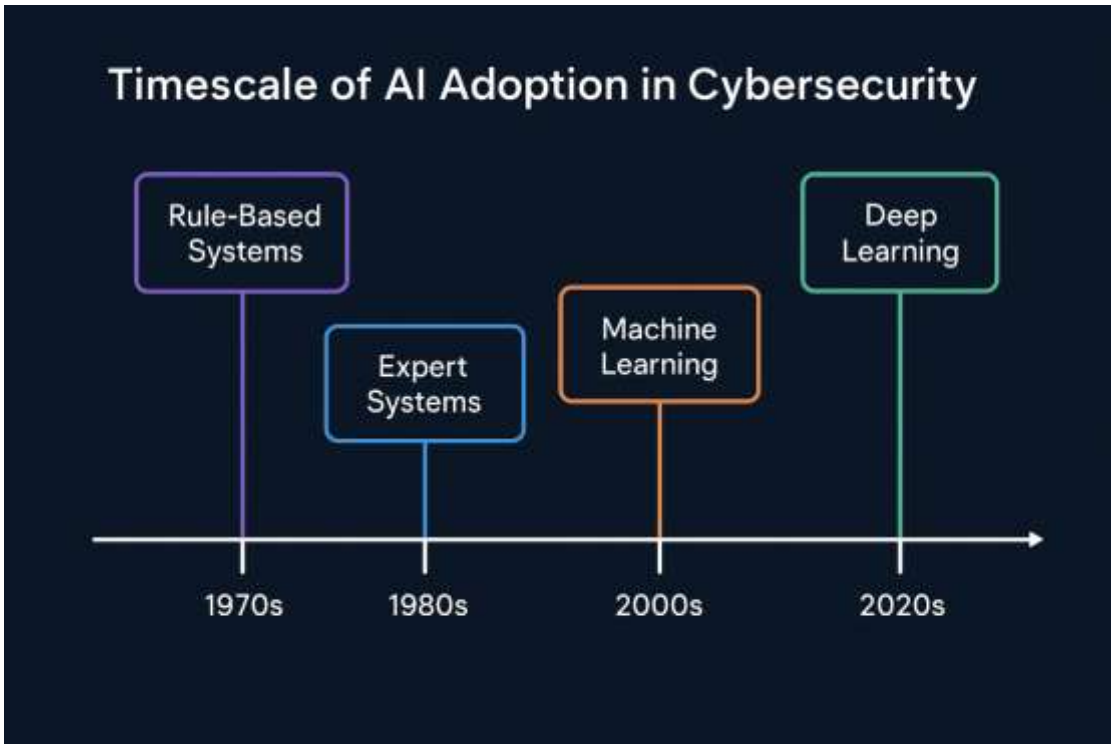

Figure 1: Timescale of AI Adoption in Cybersecurity

III. PERFORMANCE ANALYSIS

3.1 Comparative Performance Across Models

AI models in cybersecurity may be used differently depending on architecture, dataset characteristics and task type. It is very common for decision tree and random forest models to be accurate on structured IDS datasets with fast training time and good interpretability (Buczak & Guven 2016). Support Vector Machines are good for small well-labelled datasets but may be sensitive to feature scaling and may not scale well for large network traffic data (Sommer & Paxson, 2010). A top-notch deep learning model with better performance in malware classification and temporal anomaly detection is Convolutional Neural Networks (CNNs) and RNNs, with more complex high-dimensional data and better results (Kim et al., 2019; Zhang et al., 2021). The performance of Reinforcement Learning models is important in the adaptive response to threats in SOC automation, but much depends on quality of training environment and reward functions (Nguyen et al., 2022).

3.2 Strengths and Weaknesses

- Decision Trees & Random Forests: high interpretability and robustness, but can overfit and have decreased transparency in large ensembles.
- SVMs are good for linearly separated or small data, but are computationally intensive for large data.
- CNNs & RNNs are strong at pattern recognition and temporal sequence learning, but need high computational power and high quality hyperparameter tuning.
- Reinforcement Learning: dynamic and adaptive to changing threats; dependent upon simulated environments and complex reward engineering.

3.3 Impact of Model Architecture on Detection Efficacy

In fact, model architecture can significantly influence detection effectiveness. Ensemble methods reduce variance and improve robustness against noise, deep neural networks contain complex patterns and temporal dependencies that are not

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

easy to recognize in a simpler model. But, the emergence of greater model complexity can cause a greater latency, poor interpretability, and increased vulnerability to adversarial attacks. Therefore, a variety of decisions regarding accuracy, speed, interpretability, and safety should be thoroughly considered when selecting models for real-world deployment.

Table 2: Performance Metrics for AI Models in Cybersecurity

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Strengths	Weaknesses
Decision Tree	91	89	90	89.5	Fast, interpretable	Prone to overfitting
Random Forest	94	92	93	92.5	Robust, ensemble benefits	Less transparent in large ensembles
SVM	92	90	91	90.5	High accuracy for small datasets	Sensitive to feature scaling
CNN	96	95	94	94.5	Excellent for complex patterns	High computational cost
RNN / LSTM	95	94	93	93.5	Captures temporal dependencies	Training complexity, overfitting
Reinforcement Learning	93	91	92	91.5	Adaptive SOC response	Requires simulated environments

IV. SAFETY, TRUST, AND EXPLAINABILITY

4.1 Technical Vulnerabilities and Adversarial Attack Susceptibility

While AI models of cybersecurity are vulnerable to adversarial attacks, data poisoning, and evasion, cybersecurity technologies are likely to be at odds with technical threats. Adversarial attacks are intentional manipulations designed to influence machine learning models and could cause false negatives or false positives in intrusion detection and malware classification (Goodfellow et al. 2015; Nguyen et al. 2018). Detraining is incomplete in addition to data drift, emergence of attack patterns and changes in network behavior and might worsen model performance if it is not properly scheduled. Ensemble models are generally more robust to single-point attacks and deep learning models are more robust to subtle adversarial perturbations, like Papernot et al. 2016.

4.2 Model Drift and Cascading Failures

Most cybersecurity AI models may drift over time due to changes in network traffic patterns or malware characteristics. Drift may degrade detection performance and can increase the probability of cascading failures for automated SOC systems. Scached failures occur where the misclassification of one model leads to improper automated behaviors that expand across the systems. This risk may be mitigated by implementing a regular retraining schedule, modelling model outputs, and by using human-in-the-loop (HITL) mechanisms (Amodei et al., 2016).

4.3 Explainable AI (XAI) Frameworks

Tellable AI methods are key to improving transparency and trust in cybersecurity AI systems. The use of LIME, SHAP, and attention-based neural networks provides interpretable insights into model decisions (Doshi-Velez & Kim, 2017, Lundberg & Lee, 2017.) By enabling analysts to understand why the model flagged a particular activity as malicious, XAI reduces automaticity bias and facilitates informed decision-making.

4.4 Metrics for Analyst Trust and HITL vs HOTL Workflows

The trustworthiness of AI models is measured by reliability, consistency and whether model outputs are understood by human analysts. HITL models help analysts intervening in decision-making, improving trust and minimizing error-risking and HOTL approaches control fully automated systems for operational safety without direct intervention (Gunning et al., 2019).

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

Table 3: Vulnerabilities and Observed Attack Success Rates

Model Type	Adversarial Attack Susceptibility	Vulnerability Notes	Mitigation Strategies
Decision Tree	Low	Simple structure, easy to interpret	Ensemble methods, pruning
Random Forest	Moderate	Complex ensemble may hide vulnerabilities	Adversarial training
SVM	Moderate	Sensitive to input scaling	Feature normalization, robust kernels
CNN	High	Vulnerable to small perturbations	Adversarial training, XAI guidance
RNN / LSTM	High	Temporal perturbations affect output	Continuous retraining, anomaly detection
Reinforcement Learning	Moderate	Depends on environment simulation	Reward function tuning, human oversight

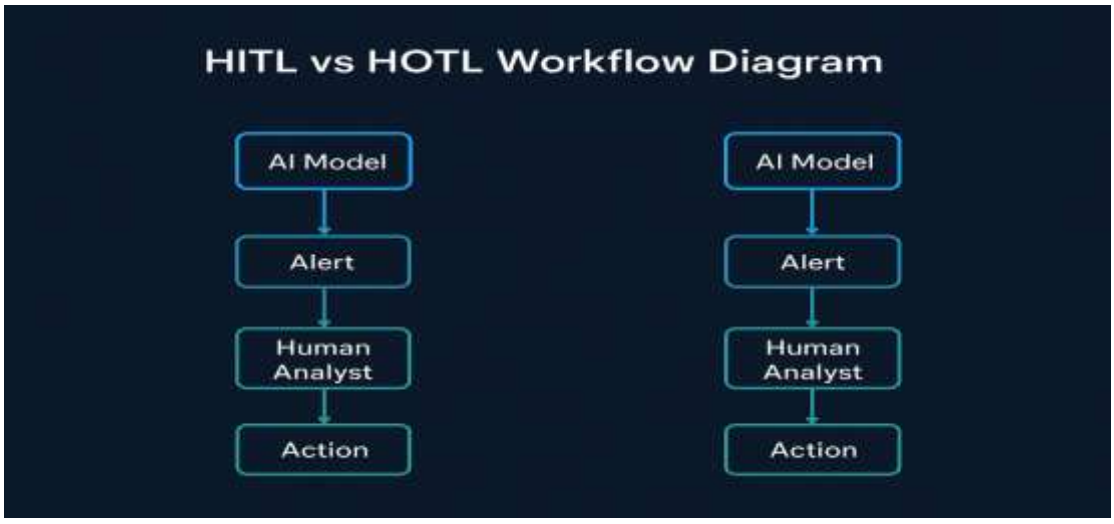

Figure 2: HITL vs HOTL Workflow Diagram

V. PERFORMANCE ANALYSIS

5.1 Comparative Performance Across Models

They also differ on architecture, dataset type and task domain performance in cybersecurity AI models. This combination, or random forest methods of intrusion detection and malware classification, is robust and highly reliable. Several of the best deep learning models are CNNs and RNNs, which perform particularly well on complex datasets that capture complex spatial and temporal patterns in network traffic. SVMs are good at small and structured data but may struggle with large scale or high scale data. The reinforcement learning models offer adaptive threat response to SOC automation but are strongly dependent on the appropriate environments and reward architecture (Nguyen et al., 2022; Zhang et al., 2021).

5.2 Strengths and Weaknesses

Decision Trees are fast and interpretable and a small size that overfits.

- Random Forests: Robust ensemble performance; transparency disappears when a larger group is.
- SVMs: high accuracy for large labeled datasets; sensitive to feature scaling.
- CNNs & RNNs: excellent at identifying complex and temporal patterns; computationally intensive and vulnerable to adversarial attacks.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

- Reinforcement Learning: Dynamic and adaptive; need high quality simulations and meticulous reward function design.

5.3 Impact of Model Architecture on Detection Efficacy

Model architecture essentially affects detection accuracy, operational latency, and robustness. While ensemble methods are sensitive to variance and mitigate noise response against noisy inputs, deep neural networks perform much better than deep neural networks to capture complex patterns but require more computational resources and less transparency. The architecture also impacts the model's ability to tolerate adversarial attacks and its explainability, and also underscores the importance of balance between performance and safety and trust in the deployment process.

Table 4: Performance Metrics for AI Models in Cybersecurity

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Key Strengths	Key Weaknesses
Decision Tree	91	89	90	89.5	Fast, interpretable	Prone to overfitting
Random Forest	94	92	93	92.5	Robust, ensemble benefits	Less transparent in large ensembles
SVM	92	90	91	90.5	High accuracy for small datasets	Sensitive to feature scaling
CNN	96	95	94	94.5	Excellent for complex patterns	High computational cost
RNN / LSTM	95	94	93	93.5	Captures temporal dependencies	Training complexity, overfitting
Reinforcement Learning	93	91	92	91.5	Adaptive SOC response	Requires simulated environments

VI. CONCLUSION

This study provides a comprehensive comparison of the AI models used in cybersecurity addressing four basic aspects: performance, safety, trust, and explainability. While CNNs and RNNs are best known for the most accurate detection accuracy and can identify complex patterns, the model is computationally intensive and vulnerable to adversarial attacks. The robust and interpretive capabilities of ensemble models like random forests are useful but may not be as transparent in larger forms. Simpler models such as decision trees and SVMs, offer interpretability and efficiency, but are limited in both scaling and adaptability. Reinforcement learning models work well in adaptive SOC automation, but must be modelled with specialized training environments and robust reward functions.

The safety and trust concern of AI in operating cybersecurity environments are crucial to the deployment of AI. Adversary attacks, model drift, and failures reflect the need to be monitored, refocused, and integrated as human control in human-in-the-loop (HITL), or human-on-the-loop (HOTL). In addition to providing a greater trust and accountability in analyst and operating process, explainable AI, such as XAI, improves the level of confidence and operational accuracy for the analyst, minimizes the likelihood of automating bias and enables decision-making.

Overall, the study emphasizes that no one AI model is better than others; finding a suitable model requires balance between accuracy, operational safety, interpretability, and adaptability to changing threats. The results provide a structure through which agencies can make informed decisions when using AI-based cybersecurity solutions, helping to ensure secure, reliable, and trustworthy AI systems. Future research should look at hybrid solutions that combine model architectures and XAI techniques in order to maximize performance and safety and transparency.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/?volume=January~2026

REFERENCES

1) Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv:1606.06565.
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.
3) Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv:1702.08608.
4) Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.
5) Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120.
6) Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
7) Nguyen, A., Yosinski, J., & Clune, J. (2018). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
8) Nguyen, T., Reddi, V. J., & Chan, H. (2022). Reinforcement learning for cybersecurity: A survey. *ACM Computing Surveys, 55*(8), 1–38.
9) Sarker, I. H., Kayes, A. S. M., & Watters, P. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data, 7*(1), 1–29.
10) Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305–316.
11) Zhang, X., Chen, W., & Li, Y. (2021). Deep learning-based network intrusion detection: A survey. *IEEE Access, 9*, 45065–45079.