North Asian International Research Journal of Sciences, Engineering & I.T.

IRJIF. I.F. : 3.821

ISSN: 2454-7514 Vol. 10, Issue-11, November-2024

Index Copernicus Value: 52.88 Indian Citation Index Thomson Reuters ID: S-8304-2016

NAIRJC UGC-Aligned | HEI-Approved Peer Reviewed Refereed Journal (Under New UGC Guidelines)

DOI: doiglobal.org/10.2024/NAIRJCSEIT.04

ARTIFICIAL INTELLIGENCE IN CYBERSECURITY: APPLICATIONS, CHALLENGES AND A TRUSTWORTHY DEFENCE FRAMEWORK

Nagaraja. K.V.

Associate professor, Dept of Computer science, Government First Grade Women's College, Tumkur.

ABSTRACT

Artificial intelligence (AI) is increasingly used to detect malware, identify anomalous network behaviour, prioritise security alerts, analyse threat intelligence and automate incident response. These capabilities are valuable because modern organisations generate volumes of security telemetry that exceed the capacity of manual analysis. At the same time, AI-enabled security systems introduce new risks, including adversarial examples, poisoned training data, model extraction, privacy leakage, automation bias and poor explainability. This paper critically examines the dual role of AI as both a defensive capability and an attack surface in cybersecurity. It adopts a structured narrative review and design-science approach using authoritative standards, official cybersecurity knowledge bases, benchmark dataset documentation and peer-reviewed research. The study compares major AI techniques, intrusion-detection datasets, evaluation metrics and operational use cases. Rather than presenting fabricated laboratory findings, the results are expressed as a transparent evidence synthesis and a proposed trustworthy AI-cybersecurity architecture. The analysis finds that tree-based ensembles are often strong baselines for tabular security data, deep learning is effective for high-dimensional and sequential data, and unsupervised methods are useful for previously unseen anomalies. However, accuracy alone is an insufficient measure because class imbalance, dataset age, concept drift and adversarial manipulation can produce misleading results. A practical deployment should combine data governance, threat-informed modelling, human oversight, explainability, continuous monitoring and adversarial testing. The paper proposes a five-layer architecture linking telemetry collection, preprocessing, AI analytics, explainable decision support and orchestrated response, governed through the NIST AI Risk Management Framework functions of govern, map, measure and manage. The study concludes that AI should augment rather than replace cybersecurity professionals and that trustworthy deployment requires security-by-design across the entire model life cycle.

Keywords: ***artificial intelligence; cybersecurity; machine learning; intrusion detection; adversarial machine learning; threat intelligence; explainable AI; incident response.***

1. INTRODUCTION

Cybersecurity has become a data-intensive discipline. Enterprise networks, cloud platforms, endpoint agents, identity systems and applications continuously generate events that must be distinguished as benign, suspicious or

malicious. Traditional signature-based controls remain essential, but they are limited when attackers change infrastructure, obfuscate payloads or exploit previously unseen vulnerabilities. AI extends defensive capability by learning patterns from historical data, identifying deviations and correlating weak signals across multiple sources.
AI is not a single technique. In cybersecurity it includes supervised classification, unsupervised anomaly detection, deep neural networks, natural-language processing, reinforcement learning and generative models. These techniques can reduce analyst workload and improve detection speed, yet their effectiveness depends on representative data, appropriate metrics and secure operational integration. NIST identifies secure and resilient behaviour as a core characteristic of trustworthy AI, while its AI Risk Management Framework organises risk work around governance, mapping, measurement and management.

2. Problem Statement and Research Gap

Many published intrusion-detection studies report very high accuracy, but the reported values are frequently difficult to compare. Studies may use different datasets, preprocessing procedures, train-test splits, attack classes and definitions of false positives. Older datasets may not represent current encrypted, cloud-native or zero-trust environments. Highly imbalanced data can also allow a model to achieve high accuracy while failing to detect rare attacks. A second gap is that model performance is often studied separately from model security. A classifier that performs well under normal testing may fail when an attacker modifies features, poisons training data or probes the decision boundary.

The practical research problem is therefore not simply how to maximise classification accuracy. It is how to design AI-enabled cybersecurity that remains useful, explainable, privacy-conscious and resilient under changing and adversarial conditions.

3. Objectives

1. To examine major defensive applications of AI in cybersecurity.
2. To compare machine-learning and deep-learning approaches used in intrusion and anomaly detection.
3. To evaluate commonly used cybersecurity datasets and performance metrics.
4. To identify technical, operational, ethical and adversarial risks.
5. To propose a trustworthy, threat-informed architecture for AI-enabled cyber defence.
6. To formulate implementation recommendations and future research directions.

4. Research Questions

1. Which AI techniques are most appropriate for different cybersecurity tasks?
2. Why are benchmark accuracy values insufficient for production decisions?
3. Which weaknesses in data and models reduce the reliability of AI-based defence?
4. How can human oversight, explainability and threat intelligence be integrated?
5. What architecture can support continuous and trustworthy deployment?

5. Contributions of the Study

Contribution	Description
Evidence synthesis	Connects standards, benchmark datasets and current research without claiming invented experiments.
Comparative framework	Compares algorithms, datasets, use cases, metrics and deployment risks.
Architecture	Proposes a five-layer AI-cybersecurity workflow with governance and feedback.
Evaluation guidance	Emphasises precision, recall, F1, false-positive rate, latency, robustness and drift.
Research agenda	Identifies challenges in adversarial robustness, privacy, explainability and real-world validation.

6. Review of Literature

Machine learning has long been applied to spam filtering, malware classification and network intrusion detection. Recent reviews describe a shift from manually engineered signatures toward data-driven systems capable of learning complex relationships and generalising to partially unseen events. Supervised models such as decision trees, random forests, support-vector machines and gradient boosting are widely used when labelled examples are available. Unsupervised methods such as clustering, isolation forests and autoencoders are valuable where labels are incomplete and the objective is to detect deviations from normal behaviour.

Deep learning extends this capability to high-dimensional and sequential information. Convolutional networks can learn local patterns in byte sequences or transformed traffic representations; recurrent and transformer models can analyse temporal sequences, logs and threat reports. Natural-language processing supports phishing detection, vulnerability extraction and mapping of threat reports to tactics and techniques. MITRE ATT&CK provides a knowledge base of adversary behaviour based on real-world observations, while MITRE D3FEND catalogues defensive techniques. MITRE ATLAS extends threat-informed knowledge to attacks against AI-enabled systems.
The literature also warns that AI systems can be attacked. NIST's adversarial machine-learning taxonomy distinguishes attacks by life-cycle stage, attacker objective, knowledge and capability. Important categories include training-data poisoning, evasion at inference time, privacy attacks, model extraction and supply-chain compromise. Research on network intrusion detection shows that small, carefully designed feature changes can cause misclassification even when ordinary test performance is high.

Table 1. Major AI Techniques in Cybersecurity

Technique	Typical use	Strength	Important limitation
Decision trees / Random forests	Malware, phishing and flow classification	Fast, interpretable baseline for tabular data	May overfit; limited temporal modelling
Gradient boosting	Risk scoring and intrusion classification	High predictive power on structured features	Parameter sensitivity and reduced transparency
SVM	Binary or multiclass classification	Effective in high-dimensional spaces	Scaling cost on very large datasets
Autoencoders / clustering	Anomaly and zero-day discovery	Can operate with limited labels	May generate many false positives
CNN / RNN / Transformers	Sequences, logs, payloads and CTI text	Learns complex spatial or temporal patterns	Compute cost, explainability and adversarial risk
Reinforcement learning	Adaptive response and resource allocation	Optimises sequential decisions	Unsafe exploration and reward-design risk
Generative AI / LLMs	Alert summarisation, query support, CTI extraction	Natural-language interface and synthesis	Hallucination, prompt injection and data leakage

7. Research Methodology

The study adopts a structured narrative review combined with design-science reasoning. The unit of analysis is the AI-enabled cybersecurity system rather than an individual algorithm. Sources were selected from official standards and knowledge bases, benchmark dataset owners and peer-reviewed or academically archived studies. The review focused on four themes: defensive applications, data and evaluation, adversarial risks, and operational governance.

7.1 Inclusion and Exclusion Criteria

Criterion	Included	Excluded
Source quality	Standards bodies, official dataset documentation, peer-reviewed studies and established research archives	Unverifiable performance claims and marketing-only material
Topical relevance	AI for cyber defence, security of AI, IDS datasets, threat intelligence and incident response	General AI studies without cybersecurity relevance
Evidence	Clearly stated method, dataset, framework or operational implication	Claims without enough methodological context
Time	Foundational work plus recent work needed for current risk analysis	Outdated claims presented as current practice without qualification

7.2 Analytical Procedure

First, AI applications were mapped to cybersecurity functions: prevention, detection, analysis, response and recovery. Second, benchmark datasets were compared by traffic type, attack coverage and known limitations. Third, algorithms were evaluated qualitatively against accuracy, scalability, interpretability and adversarial robustness. Fourth, risks were mapped to model-life-cycle stages. Finally, the findings were transformed into a proposed architecture and implementation checklist.

7.3 Evaluation Metrics

For binary detection, true positives (TP) are attacks correctly detected, false positives (FP) are benign events incorrectly flagged, true negatives (TN) are benign events correctly ignored, and false negatives (FN) are attacks missed. Accuracy is useful only when class distributions are considered. Precision measures the reliability of alerts; recall measures detection coverage; F1 balances precision and recall. False-positive rate and detection latency are operationally critical because excessive alerts create analyst fatigue.

Metric	Expression	Operational meaning
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Overall correctness; can mislead on imbalanced data
Precision	TP / (TP + FP)	How many alerts are truly malicious
Recall	TP / (TP + FN)	How many attacks are detected
F1-score	2 × Precision × Recall / (Precision + Recall)	Balanced alert quality and coverage
False-positive rate	FP / (FP + TN)	Burden imposed on analysts
Latency	Time from event to detection/response	Suitability for real-time defence

8. Benchmark Datasets and Data Quality

Public datasets support reproducibility, but they should not be treated as perfect representations of production networks. NSL-KDD was designed to reduce redundant records in KDD'99 and remains useful for methodological comparison, although its traffic is old. UNSW-NB15 contains modern normal activity and nine attack families generated in a cyber range, but researchers have identified class imbalance and class overlap. CICIDS2017 contains labelled flows and several contemporary attack scenarios, yet later analyses have documented data-quality and collection-pipeline concerns. These limitations demonstrate that a model should be tested on multiple datasets and, where possible, on local temporal holdouts.

Table 2. Comparison of Widely Used IDS Datasets

Dataset	Main characteristics	Research value	Key limitation
NSL-KDD	Derived from KDD'99; reduced redundancy; labelled normal and attack records	Simple benchmark and reproducible comparison	Old traffic and limited representation of modern networks
UNSW-NB15	Hybrid realistic and synthetic traffic; nine attack categories; large raw capture	More contemporary feature set and multiclass evaluation	Class imbalance, overlap and lab-generated behaviour
CICIDS2017	Benign traffic plus brute force, botnet, DoS/DDoS, web attacks, infiltration and others	Flow-level labels and diverse attack scenarios	Potential artefacts, imbalance and documented pipeline issues
Organisation-specific telemetry	Current logs, identity, endpoint, cloud and network data	Highest contextual relevance	Privacy, labelling cost, limited public reproducibility

8.1 Data Preparation Requirements

Stage	Required control	Reason
Collection	Document sensors, time range, environment and consent	Supports traceability and legal compliance
Cleaning	Remove duplicates, invalid values and leakage features	Prevents inflated performance
Encoding	Apply reproducible categorical and numeric transformations	Ensures consistent training and inference
Balancing	Use class weights, sampling or suitable loss functions	Improves minority-attack detection
Splitting	Prefer time-based and environment-based holdouts	Reduces leakage and tests generalisation
Monitoring	Detect drift in features, labels and alert outcomes	Maintains performance after deployment

8.2 Threat-Informed Labelling

Labels should not be limited to 'benign' and 'malicious'. Mapping events to ATT&CK tactics and techniques improves interpretability and helps analysts connect model outputs to adversary behaviour. Defensive actions may also be mapped to D3FEND techniques. For AI-system attacks, ATLAS provides a specialised knowledge base covering tactics, techniques, mitigations and case studies.

9. Proposed Trustworthy AI-Cybersecurity Architecture

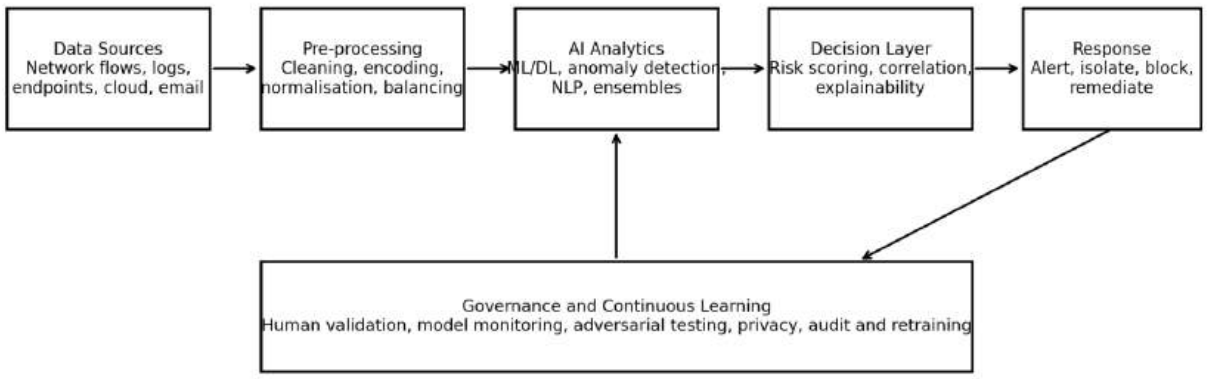

Figure 1. Proposed five-layer architecture with governance and continuous-learning feedback.

9.1 Architecture Components

Layer	Functions	Security controls
1. Data sources	Collect network flows, logs, endpoint, cloud, identity, email and CTI data	Asset inventory, access control, encryption and provenance
2. Pre-processing	Clean, encode, normalise, deduplicate, enrich and balance	Schema validation, leakage checks and privacy minimisation
3. AI analytics	Run supervised, unsupervised, deep and language models	Versioning, adversarial testing and reproducible pipelines
4. Decision layer	Correlate signals, assign risk, explain predictions and prioritise cases	Confidence thresholds, human review and audit logs
5. Response	Alert, isolate endpoints, block indicators and initiate playbooks	Least privilege, approval gates, rollback and post-incident review
Governance loop	Monitor drift, errors, bias, attacks and operational outcomes	NIST AI RMF: govern, map, measure and manage

9.2 Design Principles

The architecture follows six principles. First, AI outputs are advisory unless a response action has been safely

validated. Second, every model must have an owner, documented purpose and defined failure conditions. Third, high-impact automated actions require approval gates or reversible controls. Fourth, explanations should be tailored to analysts rather than presented only as mathematical feature weights. Fifth, model monitoring must include both performance drift and active attack indicators. Sixth, retraining should be controlled like a software release, with testing, rollback and signed artefacts.

10. Results: Comparative Evidence Synthesis

The 'results' in this review are derived from systematic comparison rather than a new laboratory experiment. The synthesis indicates that no single algorithm dominates across all environments. Selection should be based on data structure, attack prevalence, response-time requirements, interpretability and available computing resources.

Table 3. Comparative Suitability of AI Model Families

Model family	Tabular flows	Sequential logs	Unknown attacks	Explainability	Production note
Tree ensembles	High	Low-Medium	Medium	Medium-High	Strong baseline; efficient inference
SVM	Medium-High	Low	Low-Medium	Medium	Useful with careful scaling and feature design
Autoencoder / Isolation methods	Medium	Medium	High	Low-Medium	Useful for anomaly discovery; tune thresholds
CNN / RNN	Medium	High	Medium-High	Low	Powerful but resource-intensive
Transformers / LLMs	Low-Medium	High for text/logs	Medium	Medium with tooling	Best for CTI extraction and analyst assistance
Hybrid / ensemble	High	High	High	Variable	Often most robust but operationally complex

Table 4. AI Application Areas and Expected Benefits

Application	AI contribution	Primary benefit	Main risk
Network intrusion detection	Classify flows and detect anomalies	Earlier discovery of suspicious behaviour	False positives and adversarial evasion
Malware analysis	Classify binaries, behaviour and code patterns	Faster triage at scale	Packing, obfuscation and dataset shift
Phishing and email security	Analyse text, links, sender and visual features	Improved detection of social engineering	Generative phishing and language drift
Threat intelligence	Extract entities and map reports to ATT&CK	Faster knowledge integration	Hallucinated or mis-mapped techniques
SOC alert triage	Prioritise, cluster and summarise cases	Reduced analyst workload	Automation bias and missed context
Incident response	Recommend or execute response playbooks	Reduced containment time	Unsafe automation and cascading disruption
Vulnerability management	Predict exploitability and prioritise remediation	Focus on highest-risk weaknesses	Biased historical labels and incomplete context

Table 5. Key Result of the Comparative Analysis

Finding	Evidence-based interpretation
High accuracy is not sufficient	Imbalance and leakage can hide poor minority-class recall.
Data quality determines reliability	Old, synthetic or artefact-rich data limit generalisation.
Hybrid models offer broader coverage	Supervised detection and anomaly detection address different threat types.
Human oversight remains essential	Context, business impact and uncertain cases require analyst judgement.
AI systems must themselves be secured	Poisoning, evasion, extraction and prompt injection expand the attack surface.

11. Challenges, Discussion and Recommendations

11.1 Adversarial and Technical Challenges

Risk	Life-cycle stage	Possible consequence	Recommended mitigation
Data poisoning	Collection / training	Backdoors or reduced detection	Provenance, validation, robust training and dataset review
Evasion examples	Inference	Malicious traffic classified as benign	Adversarial testing, ensembles and feature constraints
Model extraction	Deployment	Theft of decision boundary or intellectual property	Rate limiting, monitoring and output restriction
Privacy leakage	Training / inference	Exposure of sensitive records	Minimisation, access control and privacy-preserving learning
Concept drift	Operation	Declining accuracy as behaviour changes	Drift detection, temporal validation and controlled retraining
Prompt injection / hallucination	Generative AI use	Unsafe advice or data disclosure	Tool isolation, grounding, filters and human approval
Supply-chain compromise	Development	Malicious libraries, models or artefacts	SBOM/model inventory, signatures and dependency controls

11.2 Discussion

The evidence supports a socio-technical view of AI cybersecurity. A mathematically strong model can fail if sensors are misconfigured, labels are wrong, analysts distrust the alerts or response automation lacks safeguards. Conversely, a modest model embedded in a disciplined process may provide substantial value. The most appropriate objective is therefore not autonomous replacement of the security operations centre, but augmentation of human reasoning and acceleration of repetitive analysis.

Threat-informed modelling improves this relationship. An alert that identifies likely ATT&CK techniques, supporting telemetry, confidence and recommended D3FEND countermeasures is more actionable than an unexplained probability. Explainability should answer operational questions: what evidence triggered the result,

what alternative explanation exists, how urgent is the case, and what safe action is recommended?

11.3 Recommendations

1. Start with a clearly defined security decision, not with an algorithm.
2. Establish strong baselines and compare AI against rules and existing controls.
3. Use multiple datasets and temporal or cross-environment validation.
4. Report precision, recall, F1, false-positive rate, latency and robustness, not accuracy alone.
5. Maintain model cards, data documentation, ownership and audit trails.
6. Map detections to ATT&CK and defensive actions to D3FEND where appropriate.
7. Perform adversarial testing before release and after major retraining.
8. Place approval gates around disruptive automated responses.
9. Monitor drift, alert usefulness, analyst overrides and missed incidents.
10. Apply the NIST AI RMF throughout design, deployment and operation.

12. Conclusion

Artificial intelligence can materially strengthen cybersecurity by analysing large-scale telemetry, identifying anomalous behaviour, extracting threat intelligence and supporting faster incident response. Its value is greatest when the chosen technique matches the data and operational decision. Tree ensembles provide strong and interpretable baselines for structured data; deep models support complex sequences and text; anomaly-detection methods help discover previously unseen behaviour; and generative systems can assist analysts with summarisation and query-based investigation.

The review also demonstrates that AI creates new dependencies and vulnerabilities. Benchmark accuracy can be misleading when datasets are imbalanced, outdated or affected by leakage. Models may degrade through concept drift and may be manipulated through poisoning, evasion, extraction, prompt injection or supply-chain compromise. Consequently, trustworthy AI-enabled defence requires governance, human oversight, secure data engineering, threat-informed explanations, adversarial testing, continuous monitoring and reversible response mechanisms.

The proposed architecture integrates these requirements across data collection, preprocessing, AI analytics, decision support and response. It positions governance and continuous learning as cross-cutting controls rather than final compliance steps. The central conclusion is that AI should augment cybersecurity professionals, not replace accountable judgement. Future progress depends less on isolated accuracy improvements and more on robust evaluation under realistic, changing and hostile conditions.

13. Future Research

Future studies should evaluate cross-dataset generalisation, encrypted-traffic analysis, federated and privacy-preserving learning, AI security in cloud and IoT environments, trustworthy large-language-model assistants, real-time adversarial defence, causal explanation and the measurement of analyst-AI team performance. Greater attention is also required for reproducible production studies using temporal holdouts and transparent failure

analysis.

References

1. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
2. Vassilev, A., et al. (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2e2025). NIST.
3. MITRE. (2026). MITRE ATT&CK: A knowledge base of adversary tactics and techniques.
4. MITRE. (2026). MITRE D3FEND: A knowledge graph of cybersecurity countermeasures.
5. MITRE. (2026). MITRE ATLAS: Adversary tactics and techniques against AI-enabled systems.
6. ENISA. (2023). Artificial Intelligence and Cybersecurity Research.
7. Canadian Institute for Cybersecurity. (2017). CICIDS2017 intrusion detection evaluation dataset.
8. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. MilCIS.
9. Tavallaee, M., et al. (2009). A detailed analysis of the KDD CUP 99 data set. IEEE CISDA.
10. Engelen, G., Rimmer, V., & Joosen, W. (2021). Troubleshooting an intrusion detection dataset: The CICIDS2017 case. IEEE.
11. Zoghi, Z., & Serpen, G. (2021). UNSW-NB15 computer security dataset: Analysis through visualization.
12. Ennaji, S., De Gaspari, F., Hitaj, D., Kbidi, A., & Mancini, L. V. (2024). Adversarial challenges in network intrusion detection systems.
13. DiMaggio, P., et al. (2004). Digital inequality: From unequal access to differentiated use. [Methodological analogy for socio-technical systems].
14. Wang, Z. (2018). Deep learning-based intrusion detection with adversaries. NIST.
15. Olatunji, A. P., et al. (2024). A systematic review of the role of artificial intelligence in cybersecurity. IEEE Access.
16. Dhanushkodi, K., et al. (2024). AI-enabled threat detection: Leveraging artificial intelligence in cybersecurity. IEEE Access.
17. More, S., et al. (2024). Enhanced intrusion detection systems performance with machine learning on UNSW-NB15. Algorithms, 17(2), 64.
18. Castells, M. (2010). The rise of the network society. Wiley-Blackwell. [Context for networked security].