Online at: https://researchtrendsjournal.com
Indexed Journal

ISSN No: 2584-282X
Peer Reviewed Journal



INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 4; Issue 2; 2026; Page No. 135-140

Received: 18-01-2026
Accepted: 29-02-2026
Published: 15-03-2026

An Ai-Enabled Framework for Effective Cyber Threat Monitoring, Detection, And Risk Management

[1]Sunil Kumar and [2]Dr. Sunil Verma

[1]Research Scholar, Department of Computer Science, Monad University, Hapur, Uttar Pradesh, India
[2]Professor, Department of Computer Science, Monad University, Hapur, Uttar Pradesh, India

DOI: https://doi.org/10.5281/zenodo.23811446

Corresponding Author: Sunil Kumar

Abstract

Computer networks are exposed to an expanding range of cyber threats, including denial-of-service attacks, credential compromise, malware communication, reconnaissance, lateral movement, data exfiltration and previously unseen attack techniques. Conventional signature-based security controls remain important, but they may be insufficient when attack behaviour changes rapidly, indicators are unavailable or multiple weak signals occur across network components. Artificial intelligence (AI), including machine learning (ML), deep learning (DL), anomaly detection and natural language processing, offers opportunities to analyse large volumes of network telemetry and identify suspicious patterns. However, detection accuracy alone is not sufficient for operational security. A useful framework must also assess the likelihood and potential impact of threats, explain its alerts, control false positives, preserve privacy and support accountable response.

This paper proposes an integrated AI-based cyber threat detection and risk assessment framework for secure computer networks. The proposed architecture combines network data collection, preprocessing, feature engineering, supervised and unsupervised detection, risk scoring, explainable AI, threat-intelligence enrichment and human-approved response. It considers network flow records, logs, endpoint events, authentication activity and asset criticality as complementary sources of evidence. A mixed-method research methodology is proposed, involving a structured literature review, benchmark-dataset experiments and, where permission is available, controlled evaluation in a test network. Model performance will be assessed using precision, recall, F1-score, false-positive rate, detection latency, robustness and computational overhead. Risk assessment will consider threat likelihood, asset exposure, vulnerability severity and business impact. The expected contribution is a transparent, adaptive and risk-aware framework that can assist security teams without treating automated model output as a substitute for professional judgement.

Keywords: Artificial intelligence, cyber threat detection, network intrusion detection, machine learning, deep learning, anomaly detection, cybersecurity risk assessment, explainable AI, threat intelligence, secure computer networks, adversarial robustness

1. Introduction

Computer networks underpin communication, financial transactions, cloud services, industrial operations, education, healthcare and public administration. Their growing interconnection increases the scale and complexity of the attack surface. Enterprise networks now include on-premises servers, virtual machines, cloud workloads, mobile endpoints, Internet of Things (IoT) devices, remote access services and third-party connections. Each component produces security-relevant events, but these events may be distributed across different tools and formats. Attackers can exploit this fragmentation by moving between systems, using legitimate credentials, disguising malicious activity as normal traffic or exploiting newly disclosed vulnerabilities.

Traditional network security commonly combines firewalls, access controls, antivirus tools, intrusion detection systems, security information and event management platforms and signature-based rules. These controls remain essential because known indicators and policy violations can be detected efficiently. Their limitations become more visible

when attacks use novel payloads, encrypted channels, living-off-the-land techniques or subtle behavioural changes. Rule-based systems may also generate large volumes of alerts that require analysts to correlate and prioritise manually.

AI can support security monitoring by learning patterns from labelled attack examples, identifying deviations from normal behaviour and correlating events across multiple sources. Supervised models such as random forests, support vector machines and gradient-boosted trees can classify known traffic patterns. Deep learning architectures can learn complex representations from sequential or high-dimensional data. Unsupervised and semi-supervised methods can identify unusual behaviour when labels are incomplete. These methods are complementary rather than interchangeable: their suitability depends on data quality, attack prevalence, latency constraints, interpretability and deployment resources.

The proposed research treats cyber threat detection and risk assessment as connected but distinct tasks. Detection estimates whether observed activity is suspicious and what attack category may be involved. Risk assessment estimates the significance of that activity in a particular environment by considering the affected asset, exposure, vulnerability, threat likelihood and potential operational impact. A suspicious event on a low-value isolated test host may require a different priority from similar activity against a critical identity server or production database.

1.1 Aim and Research Objectives

The overall aim is to design and evaluate an AI-based intelligent framework that detects suspicious network activity and converts detection evidence into explainable, context-sensitive cyber risk assessments.

To examine machine-learning, deep-learning and anomaly-detection techniques for identifying malicious and abnormal network behaviour.

To develop an integrated risk-scoring approach that combines model outputs with asset criticality, vulnerability context and threat intelligence.

To evaluate the proposed framework using recognised datasets and controlled experiments against detection quality, false alarms, latency, explain ability and robustness.

1.2 Scope of the Study

The study focuses on network-oriented threat detection for enterprise and hybrid computer networks. It considers flow-level traffic, network and security logs, authentication events and asset context. It does not seek to replace endpoint protection, vulnerability management, identity governance or incident response teams. Instead, it proposes a decision-support layer that can integrate evidence from these controls. Experiments will use lawful, publicly available datasets and isolated test environments; no unauthorised scanning, exploitation or collection of private network traffic is required.

2. Review of Previous Work

Research on machine learning for intrusion detection has developed from early statistical and rule-based methods toward classification, clustering, ensemble learning and deep neural networks. Buczak and Guven (2016) [1] provide a broad survey of data-mining and machine-learning approaches for cyber intrusion detection, discussing the importance of feature selection, labelled data and evaluation. Their review demonstrates that model choice cannot be separated from the characteristics of the data and the operational problem.

DL methods have been investigated to reduce reliance on manual feature engineering. Shone *et al*. (2018) [4] proposed a non-symmetric deep autoencoder approach for unsupervised feature learning in network intrusion detection. Such approaches are relevant where attack labels are limited, although reconstruction-based anomaly detection may flag benign but unusual activity and requires careful threshold calibration.

Salem *et al*. (2024) [5] review AI-driven cyber detection techniques, including ML, DL and metaheuristic approaches, and compare their application to multiple categories of cyberattacks. Their review reinforces the potential of AI while identifying the continuing need to improve robustness and practical deployment. Muneer *et al*. (2024) [6] examine AI-based intrusion detection, including federated learning and explain ability, and note the trade-offs between computational demands, labelled data and generalisation to unseen attacks.

Explain ability has become a major research theme because security analysts need to understand why an alert was raised and which evidence contributed to a decision. The IEEE survey on explainable AI for Internet traffic classification and intrusion detection (2025 issue) reviews techniques and requirements for interpreting network-traffic models. E-XAI (2024) [8] evaluates black-box explain ability approaches for network intrusion detection and highlights the need to assess explanations rather than assume that a feature-importance plot is automatically useful.

Recent systematic reviews also point to dataset and deployment limitations. A 2025 Journal of Big Data literature study reports challenges including class imbalance, outdated attack patterns, false positives, overfitting, zero-day generalisation and computational constraints. These issues indicate that high scores on a single benchmark do not establish readiness for production networks. Dataset provenance, train-test separation, temporal drift and cross-dataset validation are important for meaningful evaluation.

2.1 Common AI Approaches

2.1.1 Supervised learning: Random Forest, support vector machines, decision trees, gradient boosting and neural networks classify traffic or events using labelled examples. Performance depends on label quality and coverage of attack types.

2.1.2 Unsupervised and semi-supervised learning: Clustering, isolation-based methods and autoencoders identify deviations from learned normal patterns. These approaches can help when labels are scarce, but unusual legitimate activity may generate false alarms.

2.1.3 Deep learning: Recurrent, convolutional and attention-based architectures can model sequential or complex feature relationships. They may require greater training data, processing capacity and monitoring for drift.

2.1.4 Ensemble and hybrid models: Combine classifiers or detection paradigms to improve stability and coverage, although they can increase complexity and reduce interpretability.

2.1.5 Federated learning: Enables multiple participants to train shared models without centralising raw network data. It introduces challenges involving communication cost, non-identically distributed data and the security of model updates.

2.2 Risk Assessment and Explainable Security

Cybersecurity risk assessment requires more than a model confidence score. NIST SP 800-30 Rev. 1 describes a structured approach to assessing threats, vulnerabilities, likelihood and impact. NIST Cybersecurity Framework 2.0 provides an outcome-oriented structure for governing, identifying, protecting, detecting, responding to and recovering from cybersecurity risks. These sources support a contextual risk layer in which AI-generated evidence is interpreted alongside asset importance, exposure, vulnerability information and organisational risk tolerance.

Explainable AI methods such as SHAP and LIME can help describe which input features influenced a model output. Explanations should be evaluated for consistency, stability and usefulness to analysts. A model that labels traffic as malicious without a comprehensible reason may be difficult to operationalise, particularly when an alert could trigger containment or affect service availability.

3. Identification of Research Gap

The literature demonstrates substantial progress in AI-based intrusion detection, yet important gaps remain between experimental classification and dependable operational cyber defence. First, many studies prioritise accuracy on a selected dataset, while operational systems must also control false-positive rates, detection latency, resource consumption and alert workload. Second, benchmark datasets may not represent current network architectures, encrypted traffic, cloud workloads, IoT devices or changing attacker behaviour. Third, models trained and tested on randomly split records may benefit from data leakage or fail to generalise across time, networks and attack families.

Fourth, detection and risk assessment are frequently treated separately. A classifier may identify an attack category but fail to account for the importance of the affected asset, the exploitability of a vulnerability or the likely business impact. Fifth, black-box models can limit analyst trust and make it difficult to justify response decisions. Sixth, adversarial manipulation, poisoning, evasion and concept drift can reduce model reliability after deployment. Finally, privacy and data-sharing constraints may prevent centralised collection of network telemetry from multiple organisations.

The research gap addressed here is the need for a unified, explainable and risk-aware framework that combines multi-source network evidence, complementary AI detection methods, asset-contextual risk scoring and controlled response recommendations. The study will explicitly evaluate generalisation and operational trade-offs rather than relying on a single accuracy measure. The proposed contribution is a framework and evaluation methodology; its effectiveness must be established through experiments and should not be assumed in advance.

4. Expected Impact on Academics

The study is expected to contribute to cybersecurity and information-systems research by connecting AI-based intrusion detection with contextual risk assessment. It will distinguish detection confidence from organisational risk and propose a structured method for combining both. This distinction may help future researchers develop evaluation models that better reflect how security teams prioritise incidents.

The research may also contribute an integrated conceptual architecture combining supervised classification, anomaly detection, threat-intelligence enrichment, explain ability and human-reviewed response. It will provide a comparative evaluation approach that considers precision, recall, F1-score, false-positive rate, detection latency, resource consumption, robustness and explanation quality. Such multidimensional evaluation can help avoid overemphasis on benchmark accuracy.

For teaching and academic practice, the framework can support coursework and research projects in network security, machine learning, digital forensics, risk management and responsible AI. It can be used to demonstrate data preprocessing, model validation, threat modelling and explain ability in a controlled laboratory setting. The work may also provide a foundation for future studies on federated intrusion detection, privacy-preserving analytics, drift adaptation and AI assurance.

5. Major Inputs

The proposed framework will be organised into interconnected functional layers. Each layer will have defined inputs, outputs and safeguards so that the system can be evaluated as a complete decision-support process.

5.1 Proposed Framework Components

5.1.1 Data acquisition layer: Collects authorised network-flow records, firewall and IDS alerts, DNS and authentication logs, endpoint security events and asset inventory information. Data collection will follow privacy, retention and access-control requirements.

5.1.2 Preprocessing and feature engineering layer: Removes duplicates, handles missing values, normalises numerical fields, encodes categorical variables, aligns timestamps and derives flow statistics such as duration, packet counts, byte counts and connection frequency.

5.1.3 AI detection layer: Uses a supervised classifier for known attack categories and an anomaly detector for unusual behaviour. An ensemble or rule-assisted combination may be tested to compare complementary strengths.

5.1.4 Threat-intelligence and context layer: Enriches alerts with approved indicators, vulnerability information, asset criticality, network zone, exposure and known threat context. External intelligence will be treated as evidence requiring validation, not as unquestionable truth.

5.1.5 Risk assessment layer: Estimates likelihood and impact using detection evidence, asset value, vulnerability severity, exposure and confidence. Risk categories will be calibrated and documented with explicit assumptions.

5.1.6 Explain ability and analyst interface: Provides alert reasons, influential features, event timelines, confidence information and recommended investigation steps in a form suitable for security analysts.

5.1.7 Response and feedback layer: Recommends containment or investigation actions, but high-impact actions such as disabling accounts or isolating production systems require authorised human approval. Confirmed incidents and analyst feedback can support controlled model updates.

5.2 Risk Scoring Concept

A proposed risk score may combine threat likelihood and impact on a common, documented scale. For example: Risk Score = Likelihood × Impact, where likelihood is informed by model confidence, threat evidence, exposure and vulnerability context, while impact is informed by asset criticality, data sensitivity, service dependency and recovery consequences. The score should be calibrated using organisational risk criteria and should not be interpreted as a universal probability. Confidence and data quality should be displayed separately so that uncertain evidence does not create a false sense of precision.

5.3 Evaluation Measures

Detection performance: precision, recall, F1-score, ROC-AUC and precision-recall AUC where appropriate.
Operational performance: false-positive rate, false-negative rate, alert volume, inference latency and throughput.
Robustness and generalisation: temporal holdout, cross-dataset testing, sensitivity to class imbalance and controlled perturbation tests.
Resource efficiency: CPU, memory and storage requirements, model size and retraining time.
Explainability: consistency of feature attributions, analyst usefulness and clarity of alert rationale.

6. Methodology of Research Work

The study will use a quantitative experimental design supported by a structured literature review. The literature review will identify established AI approaches, network intrusion datasets, risk-assessment standards and explain ability methods. The experimental stage will compare selected algorithms under a reproducible evaluation protocol. If access and ethics approval permit, a small controlled laboratory network may be used to test data collection and alert workflows without exposing operational systems.

6.1 Research Design and Data Sources

Candidate public datasets may include CICIDS2017, CSE-CIC-IDS2018, UNSW-NB15 and other datasets suitable for the selected network scenarios. Dataset selection will be based on documentation, feature availability, attack coverage, licensing and relevance to the research questions. The study will document dataset limitations, including synthetic or testbed origins, class imbalance, outdated attack patterns and possible overlap between records.
Data will be divided into training, validation and test sets using a strategy that reduces leakage. Where timestamps are available, temporal separation will be preferred for at least one evaluation so that the model is tested on later observations. Feature preprocessing will be fitted on training data only and then applied to validation and test data. Class imbalance may be addressed using class weights or carefully applied resampling within the training data; the test set will retain a realistic distribution.

6.2 Model Development

Baseline models may include logistic regression, decision tree, random forest and gradient boosting. A support vector machine may be included where computational feasibility permits. An anomaly-detection model such as Isolation Forest or an autoencoder will be evaluated for identifying previously unseen or poorly labelled patterns. A hybrid model will combine outputs through a documented decision rule or calibrated ensemble. Model complexity will be justified by measured improvement rather than assumed superiority.
Hyperparameters will be tuned using training and validation data only. Reproducibility will be supported through documented software versions, fixed random seeds where relevant, preprocessing pipelines and model configuration records. The study will assess the effects of feature selection and compare models using the same data splits. If a model is substantially more computationally expensive, that cost will be reported alongside detection performance.

6.3 Risk Assessment and Explain ability

For each detected event, the risk layer will combine the model output with contextual variables such as asset criticality, network exposure, vulnerability severity and business impact. A transparent rule-based scoring matrix will initially be defined, reviewed and documented. Sensitivity analysis will examine how risk rankings change when likelihood or impact assumptions are varied. This approach will make the risk calculation auditable and reduce the danger of presenting an arbitrary score as objective truth.
Explain ability methods may include SHAP or LIME for selected model families, together with rule-based alert explanations. The evaluation will examine whether explanations are stable under small input changes and whether they identify features that security practitioners can interpret. Any automated response will remain advisory in the research prototype unless a controlled test environment and explicit authorisation support an automated action.

6.4 Data Analysis and Ethical Considerations

Results will be summarised using confusion matrices, class-wise precision and recall, F1-score, ROC and precision-recall curves, false-positive rates, latency measurements and resource utilisation. Confidence intervals or repeated runs may be used where appropriate. Cross-dataset and temporal tests will be used to assess generalisation. The results will be interpreted in light of dataset limitations and will not be generalised automatically to every production network.
The research will use public or explicitly authorised data.

Personal identifiers and sensitive payload content will be excluded or anonymised where possible. Any laboratory traffic generation will take place in an isolated environment. Access to logs and experimental artefacts will be restricted, and data will be retained only as required by institutional policy. The research will follow applicable institutional ethics requirements and relevant cybersecurity laws and policies.

7. Major Inputs Required

7.1 Technical Resources

Computing environment: workstation or cloud virtual machine with sufficient CPU, memory and storage for dataset processing and model experiments.

Software: Python, data-science libraries, machine-learning frameworks, packet/flow analysis tools, notebook environment, version control and a reproducible experiment-tracking method.

Datasets and documentation: public intrusion-detection datasets, dataset licences, feature descriptions, vulnerability references and relevant cybersecurity standards.

Test environment: isolated virtual network, sample hosts and logging infrastructure for validating the data pipeline and alert workflow, subject to available resources.

7.2 Human and Institutional Resources

The project will require the researcher, academic supervisor and, where available, a cybersecurity subject expert to review the threat model and risk-scoring criteria. Security analysts or network administrators may be consulted to assess the usefulness of alert explanations and response recommendations. Institutional access to academic databases and software resources will support the literature review and experiments.

7.3 Estimated Resource and Risk Controls

The principal cost categories are computing, storage, internet access, software where open-source alternatives are insufficient, and possible expert consultation. The project can be designed to minimise cost by prioritising public datasets and open-source tools. Key project risks include poor dataset quality, insufficient compute capacity, class imbalance, overfitting, difficulty obtaining expert feedback and changes in software dependencies. Mitigation will include early dataset inspection, baseline models, a staged experiment plan and clear documentation.

8. Progress Monitoring

Progress will be monitored through monthly milestones and a research log. The log will record dataset versions, preprocessing decisions, model configurations, experiment outputs, errors, changes to the risk-scoring method and supervisor comments. Any change in the dataset, objective or evaluation protocol will be documented to preserve transparency. If an intended dataset is unsuitable or unavailable, a replacement will be selected using predefined criteria rather than changing datasets solely to improve reported performance.

9. Conclusion

The research is expected to deliver a conceptual and experimental framework that links AI-based network threat detection with contextual cyber risk assessment. The framework should demonstrate how supervised classification and anomaly detection can be used together, how model outputs can be enriched with asset and vulnerability context, and how explanations can help analysts understand the evidence behind alerts.

A second expected outcome is a comparative evaluation of selected AI methods across multiple performance dimensions. The study may identify trade-offs between detection quality, false alarms, latency, computational cost and generalisation. It may also show that a simpler model performs more reliably or efficiently than a complex model under particular conditions. Such findings would be valuable because practical security systems must operate within resource and response constraints.

A third expected outcome is a transparent risk-scoring approach that prioritises alerts according to both threat evidence and potential impact. This may help security teams distinguish high-confidence low-impact events from uncertain but potentially severe events. The scoring method will include documented assumptions and should be adapted to the risk tolerance and asset classification of the organisation using it.

The study is also expected to produce recommendations for safe deployment: use temporal and cross-dataset validation; monitor model drift; review false positives and false negatives; restrict access to telemetry; protect model updates; retain analyst approval for high-impact actions; and periodically reassess risk criteria. The research will identify limitations and future directions, including privacy-preserving federated learning, adversarially robust training, lightweight edge detection and integration with security orchestration platforms.

No numerical performance improvement is claimed in advance. The actual effectiveness of the proposed framework will be established only after experiments are completed and results are reported with appropriate limitations.

10. References

1. Buczak AL, Guven E. A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials. 2016;18(2):1153-1176. doi:10.1109/COMST.2015.2494502.
2. National Institute of Standards and Technology. Guide for Conducting Risk Assessments. NIST Special Publication 800-30 Rev. 1. Gaithersburg (MD): National Institute of Standards and Technology; c2012. doi:10.6028/NIST.SP.800-30r1.
3. National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0. NIST Cybersecurity White Paper 29. Gaithersburg (MD): National Institute of Standards and Technology; c2024. doi:10.6028/NIST.CSWP.29.
4. Shone N, Ngoc TN, Phai VD, Shi Q. A deep learning approach to network intrusion detection. IEEE Transactions on Emerging Topics in Computational Intelligence. 2018;2(1):41-50. doi:10.1109/TETCI.2017.2772792.
5. Salem AH, Azzam SM, Emam OE, Abohany AA. Advancing cybersecurity: a comprehensive review of

AI-driven detection techniques. Journal of Big Data. 2024;11:105. doi:10.1186/s40537-024-00957-y.
6. Muneer A, *et al*. A critical review of artificial intelligence-based approaches in intrusion detection: a comprehensive analysis. Journal of Engineering. 2024;2024:3909173. doi:10.1155/2024/3909173.
7. IEEE. A survey on explainable artificial intelligence for Internet traffic classification and prediction, and intrusion detection. IEEE Communications Surveys & Tutorials. 2025;27(5):3165-3198. doi:10.1109/COMST.2024.3504955.
8. E-XAI: evaluating black-box explainable AI frameworks for network intrusion detection. IEEE Access. 2024;12:23954-23988. doi:10.1109/ACCESS.2024.3365140.
9. A systematic literature study of machine learning techniques-based intrusion detection: datasets, models, challenges, and future directions. Journal of Big Data. 2025. doi:10.1186/s40537-025-01323-2.
10. Stallings W. Effective Cybersecurity: A Guide to Using Best Practices and Standards. Addison-Wesley Professional; c2018.

Creative Commons (CC) License
This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.