ISSN 2277 – 8322

LEARNING-BASED APPROACHES FOR CYBER THREAT DETECTION: TECHNIQUES, APPLICATIONS, AND RESEARCH TRENDS

Chetan Swami
Assistant Professor
Computer Science and Engineering.
Jagannath University, Jaipur, Rajasthan, India
chetan.swami@jagannathuniversity.org

Tulsi Ram Sharma
Assistant Professor
Computer Science and Engineering.
Jagannath University, Jaipur, Rajasthan, India
tulsiram.sharma@jagannathuniversity.org

Abstract- Cyber threat detection has evolved from signature- and rule-based monitoring toward adaptive, data-driven systems that learn patterns from network traffic, endpoint telemetry, security logs, identity events, application traces, and threat intelligence. This review synthesizes learning-based approaches used for cyber threat detection, with emphasis on supervised machine learning, unsupervised and semi-supervised anomaly detection, deep learning, ensemble methods, graph neural networks, transformers, federated learning, explainable artificial intelligence, and emerging foundation-model and agentic approaches. The review discusses the end-to-end detection pipeline, representative algorithms, applications across network intrusion detection, malware detection, insider-threat detection, cloud and IoT security, security-log analysis, phishing and fraud detection, and threat-intelligence analytics. Public benchmark datasets, evaluation metrics, common experimental pitfalls, adversarial machine-learning risks, concept drift, class imbalance, data leakage, interpretability, privacy, and deployment constraints are examined. Recent literature indicates a shift from isolated event classification toward contextual, sequence-aware, behavior-centric and threat-informed detection. Transformer-based sequence models, graph learning, self-supervised learning, federated approaches, multimodal security analytics, and AI-assisted security operations are prominent research directions. However, high benchmark scores alone do not guarantee operational value; realistic temporal splits, cross-dataset validation, low false-positive operation, calibrated risk scoring, robustness testing, and human-centered evaluation are increasingly important.

Keywords: cyber threat detection; machine learning; deep learning; intrusion detection;

anomaly detection; transformers; graph neural networks; federated learning; explainable AI; cybersecurity.

I. INTRODUCTION

Modern digital environments generate large volumes of heterogeneous security telemetry. Traditional signature-based systems remain valuable for known threats, but they are less effective when adversaries modify infrastructure, exploit legitimate tools, use previously unseen behaviors, or operate slowly across multiple stages. Learning-based detection addresses this limitation by estimating patterns of normal and malicious behavior from data and updating the decision boundary as the environment changes. Surveys of machine learning for cybersecurity describe a broad transition from manually engineered rules toward supervised, unsupervised, reinforcement and deep-learning methods [1,2].

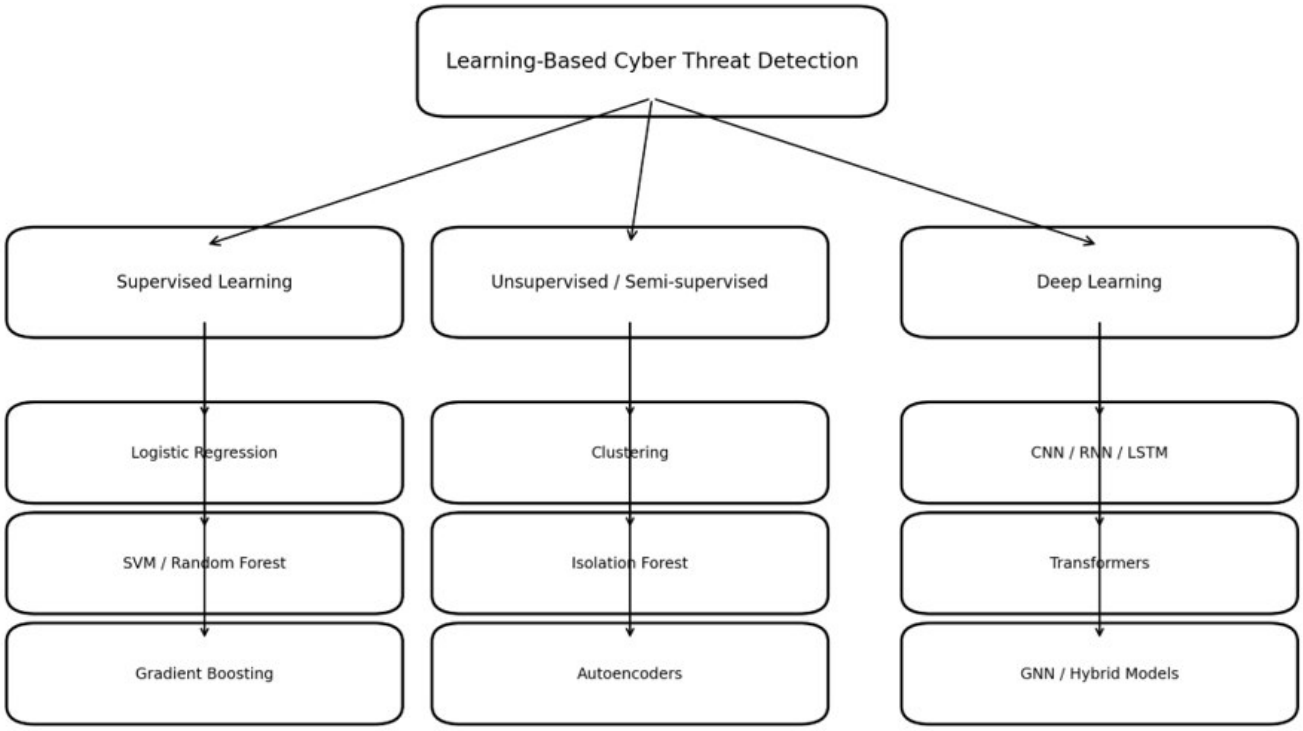

Figure 1. Taxonomy of learning-based approaches for cyber threat detection.

The practical objective is not simply to maximize classification accuracy. A useful detector must identify relevant threats with acceptable false-positive rates, provide evidence that analysts can investigate, operate with manageable latency and compute cost, and remain reliable when traffic distributions drift. MITRE ATT&CK provides a widely used

behavioral vocabulary for describing adversary tactics and techniques, and current detection engineering increasingly connects telemetry and analytics to adversary behavior rather than treating every alert as an isolated event [3–5].

This review therefore treats cyber threat detection as a complete socio-technical pipeline: data acquisition, pre-processing, representation learning, detection, contextualization, risk scoring, explanation, analyst validation and response. The goal is to summarize established methods while highlighting research trends that are likely to influence next-generation security operations.

II. REVIEW SCOPE AND ANALYTICAL FRAMEWORK

This article is a structured narrative review rather than a statistical meta-analysis. Literature and authoritative technical resources were considered across machine learning, deep learning, anomaly detection, network intrusion detection, security-log analysis, adversarial machine learning, threat-informed defense, and trustworthy AI. Recent work from 2024–2026 was emphasized for research-trend discussion, while established methods and benchmark datasets were included to provide historical context.

The analytical framework used in this review compares approaches according to learning paradigm, telemetry type, supervision requirement, ability to detect known versus novel threats, interpretability, computational requirements, and deployment maturity. Recent reviews of deep-learning NIDS identify reconstruction/generative models, transformer-based sequence models, convolutional/deep neural classifiers, and hybrid deployment-oriented approaches as important categories [6].

III. LEARNING PARADIGMS FOR CYBER THREAT DETECTION

3.1 Supervised Learning

Supervised learning maps labeled observations to threat classes. Common models include logistic regression, decision trees, random forests, support-vector machines, k-nearest neighbors, gradient boosting and neural networks. These methods are attractive when labeled security events are available and when organizations need multi-class classification such as benign, denial-of-service, reconnaissance, exploitation or malware categories. Random forests and gradient-boosting models often provide strong tabular-data baselines with relatively modest training cost and useful feature-importance estimates.

3.2 Unsupervised and Semi-supervised Learning

Because malicious examples are rare and new attack types may not be labeled, unsupervised and semi-supervised methods are central to cyber anomaly detection. Clustering, density estimation, isolation forests, one-class classifiers and autoencoders learn a representation of normal or dominant behavior. An observation that deviates strongly from the learned distribution is assigned a high anomaly score. These approaches can identify previously unseen activity, but benign operational changes can also appear anomalous, producing alert fatigue.

3.3 Deep Learning

Deep neural networks reduce the dependence on manual feature engineering by learning representations from raw or minimally processed data. CNNs are useful for local patterns in traffic or transformed byte/packet representations; RNNs and LSTMs model sequential behavior; autoencoders learn reconstruction-based normality; and hybrid models combine temporal, spatial and contextual features. Deep learning has also become prominent in security-log analysis, where sequence models learn patterns of events from very large collections of logs [7].

3.4 Transformers and Attention-Based Models

Transformers model long-range relationships through attention and are increasingly used for security event sequences and multivariate time series. The Anomaly Transformer introduced an association-discrepancy mechanism for unsupervised time-series anomaly detection [8]. Recent NIDS reviews identify transformer-based sequence models as a major contemporary category because they can represent temporal dependencies without the strict recurrent structure of traditional RNNs [6].

3.5 Graph Neural Networks

Cyber environments are naturally relational: users access devices, processes create network connections, hosts communicate with domains, and identities interact with cloud resources. Graph neural networks (GNNs) can therefore represent entities as nodes and relationships as edges, enabling detection of suspicious subgraphs, lateral movement and coordinated behavior. Their main challenge is maintaining accurate, scalable graphs from rapidly changing telemetry.

3.6 Federated and Privacy-Preserving Learning

Federated learning allows multiple organizations or endpoints to contribute to

model training without directly pooling all raw data. It is attractive for cross-organizational threat intelligence and privacy-sensitive environments. However, non-IID data, communication overhead, poisoning attacks, client drift and model-update leakage must be managed.

Table 1. Comparison of major learning paradigms

Approach	Typical algorithms	Strengths	Limitations	Best-fit use cases
Supervised ML	RF, SVM, XGBoost, LR	Strong tabular classification; fast inference	Needs labeled data; weak on novel attacks	Known attack classification, malware, phishing
Unsupervised	Clustering, Isolation Forest, One-Class SVM	Works with scarce labels; novelty detection	Sensitive to drift and benign anomalies	Anomaly detection, insider threat
Autoencoders	AE, VAE, denoising AE	Learns compact normal-behavior representations	Threshold selection; reconstruction may hide anomalies	Network and log anomalies
Sequence DL	LSTM, GRU, CNN-LSTM	Captures temporal patterns	Training cost; sequence dependence	Logs, traffic sequences, user behavior
Transformers	Self-attention, encoder/decoder models	Long-range contextual modeling	Compute and data requirements	Complex event sequences, multivariate telemetry
GNNs	GCN, GAT, temporal GNN	Models entity relationships and attack paths	Graph construction and scalability	Lateral movement, identity/cloud graphs
Federated learning	FedAvg and variants	Data locality and privacy	Non-IID data, poisoning, communication	Multi-site security analytics

ISSN 2277 – 8322

IV. END-TO-END LEARNING-BASED DETECTION PIPELINE

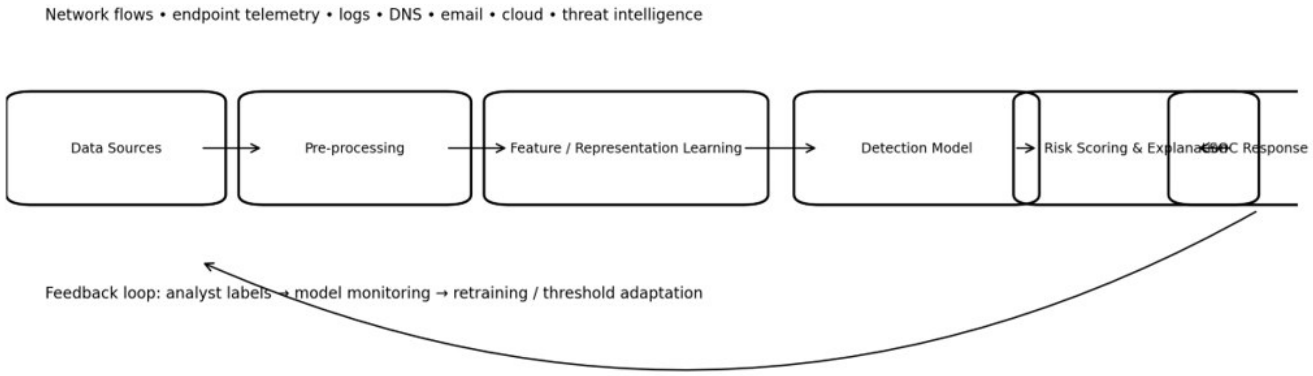

Figure 2. Conceptual end-to-end pipeline for learning-based cyber threat detection.

A practical detector begins with heterogeneous telemetry. Network data may include flow statistics, DNS queries, HTTP metadata and connection sequences; endpoints contribute process, file, registry, authentication and system-call events; cloud environments add identity, API and resource telemetry; and security operations centers contribute alerts and analyst labels. Feature engineering may include normalization, categorical encoding, temporal aggregation, frequency statistics, embeddings, sequence windows and graph construction.

Model outputs should ideally be treated as evidence rather than absolute truth. A high anomaly score can trigger additional context collection, correlation with threat intelligence, ATT&CK mapping and analyst review. This is especially important for ambiguous behaviors in which the same legitimate administrative action can also be used by an attacker. MITRE's recent work emphasizes context and telemetry requirements for distinguishing benign from malicious use of ambiguous techniques [5].

ISSN 2277 – 8322

V. APPLICATIONS

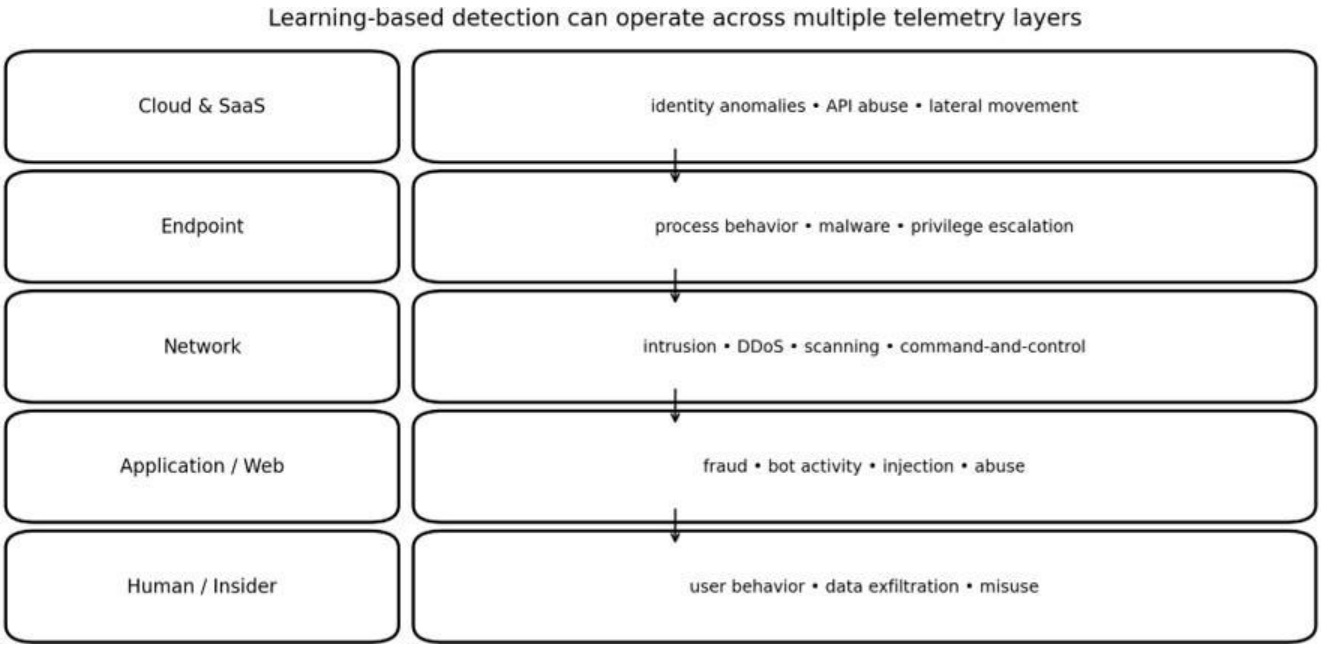

Figure 4. Representative application layers for learning-based cyber threat detection.

Table 2. Applications of learning-based cyber threat detection

Application	Typical learning-based role
Network Intrusion Detection	Classifies or scores flows and sessions for scanning, DoS/DDoS, exploitation, botnet and command-and-control activity. CNNs, tree ensembles, autoencoders, LSTM models and transformers are common choices.
Malware and Endpoint Detection	Learns executable, process, API-call, memory, file-system and behavioral features. Static, dynamic and hybrid analysis can be combined; learning can assist malware family classification and behavioral anomaly detection.
Security-Log Analysis	Learns event sequences and temporal patterns from system, application and authentication logs. Deep learning is useful when logs are too large or complex for manual inspection [7].
Insider-Threat Detection	Models user and entity behavior using authentication, file access, communication and resource-use patterns. Unsupervised and semi-supervised approaches are valuable because malicious insider examples are scarce.
Cloud and IoT Security	Detects abnormal identity use, API activity, device behavior, lateral movement and resource access. Distributed and federated learning can be useful where data remains at multiple sites.
Phishing, Fraud and	NLP and multimodal models analyze email text, URLs, sender behavior,

Abuse	domain features and transaction patterns. Detection often combines content with behavioral context.
Threat Intelligence and Attack-Path Analytics	Knowledge graphs and GNNs can connect indicators, entities, techniques and events. Mapping detections to MITRE ATT&CK improves behavioral context and threat-informed analysis [3,4].

VI. DATASETS AND EXPERIMENTAL RESOURCES

Benchmark datasets have accelerated research but also create an important evaluation problem: a model can achieve excellent results on a dataset while failing in a new environment. CICIDS2017 contains benign traffic and several common attack categories, with labeled flows and more than 80 extracted network-flow features [9]. UNSW-NB15 contains modern normal activity mixed with nine attack categories and provides 49 features and predefined training/testing partitions [10]. These resources are useful for reproducibility, but researchers should report their generation process, class distributions, temporal assumptions and feature leakage risks.

Table 3. Representative datasets and resources

Dataset/resource	Data type	Labels/content	Representative features	Main use
CICIDS2017	Network flows / PCAP	Benign + common attacks	80+ flow features	Intrusion classification and anomaly detection
UNSW-NB15	PCAP + flow features	9 attack families + normal	49 features	Network IDS benchmarking
CERT Insider Threat	User/system activity	Insider scenarios	User behavior features	Insider-threat detection
Log datasets (e.g., HDFS/BGL)	System logs	Normal/anomalous events	Parsed event sequences	Log anomaly detection
MITRE ATT&CK	Knowledge base / TTPs	Adversary tactics and techniques	Behavioral ontology	Threat-informed detection and coverage

Dataset realism is a major research issue. Public datasets can be highly imbalanced, contain duplicated or correlated records, represent laboratory traffic rather than production behavior, or allow inadvertent leakage between training and testing. A recent critical analysis of security-ML evaluation demonstrates why strong baselines and realistic experimental design are necessary before claiming superiority of complex deep models [11].

VII. EVALUATION METRICS

Accuracy is often misleading in cybersecurity because benign traffic can dominate the dataset. A detector should be evaluated using precision, recall (detection rate), F1-score, specificity, false-positive rate, ROC-AUC and preferably precision-recall AUC for imbalanced data. Operational measures such as detection latency, throughput, memory use, analyst workload and alert reduction are equally important.

For binary classification:

Precision = TP / (TP + FP)

Recall (TPR) = TP / (TP + FN)

Specificity = TN / (TN + FP)

F1-score = 2 × Precision × Recall / (Precision + Recall)

FPR = FP / (FP + TN)

Evaluation should also include cross-dataset testing, time-based holdout testing, attack-family holdouts and robustness testing. For operational deployment, the false-positive rate at a fixed low-alert threshold may be more informative than a global ROC-AUC.

VIII. MAJOR CHALLENGES AND LIMITATIONS

Table 4. Key research and deployment challenges

Challenge	Why it matters	Promising response
Class imbalance	Malicious events are rare relative to benign activity.	Cost-sensitive learning, focal loss, resampling, anomaly detection, PR-AUC.
Concept drift	Traffic, users, applications and attacker behavior change over time.	Continuous monitoring, adaptive thresholds, online learning, scheduled retraining.
Adversarial ML	Attackers may evade, poison or probe learning systems.	Adversarial training, robust features, monitoring, ensemble defenses.
Interpretability	Analysts need evidence explaining why an alert was generated.	SHAP/LIME, feature attribution, counterfactuals, attention analysis, evidence retrieval.
Data leakage	Duplicate/correlated records can inflate test performance.	Entity/time-based splits, deduplication, strict preprocessing pipelines.
Cross-environment generalization	A detector trained in one network may fail elsewhere.	Domain adaptation, transfer learning, federated learning, diverse benchmarks.
Latency and scale	SOC systems process high-volume streaming data.	Edge inference, model compression, streaming features, approximate algorithms.
Privacy	Security telemetry may contain	Federated learning, minimization, access

	sensitive employee/customer information.	controls, privacy-preserving analytics.

Adversarial machine learning deserves special attention because the detector itself becomes an attack surface. NIST's 2025 adversarial-machine-learning taxonomy organizes attacks and mitigations according to lifecycle stage, attacker goals, capabilities and knowledge [12]. Therefore, a production security model should be protected, monitored and tested like any other security-critical component.

Trustworthiness is also broader than predictive performance. NIST's AI Risk Management Framework identifies validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness with harmful bias managed as important characteristics of trustworthy AI [13,14]. These principles are directly relevant to automated cyber detection because a detector can cause operational harm through systematic false positives, opaque decisions or unsafe automated response.

IX. RESEARCH TRENDS

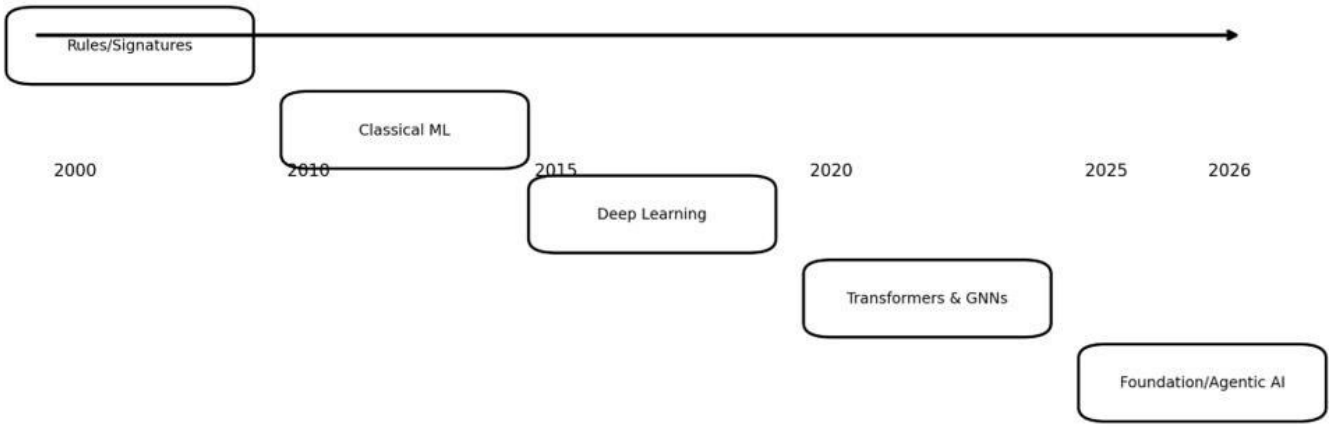

Figure 3. Conceptual evolution of cyber threat detection toward contextual and adaptive AI.

Table 5. Emerging research directions

Trend	Research significance
Self-supervised and few-shot learning	Reduce dependence on costly attack labels by learning general representations from large amounts of unlabeled security telemetry, then adapting with few labeled examples.
Transformer-based security analytics	Model long-range event dependencies, multivariate time series and contextual sequences; current NIDS surveys identify

	transformers as a major emerging category [6,8].
Graph and temporal-graph learning	Represent entities and interactions to detect lateral movement, coordinated attacks and attack paths.
Multimodal detection	Combine network flows, endpoint telemetry, logs, email/text, threat intelligence and identity signals rather than relying on a single modality.
Explainable and human-centered AI	Generate evidence, ranked contributing factors and investigation context instead of only an anomaly score.
Federated and collaborative learning	Train across organizations or sites while reducing the need to centralize sensitive raw telemetry.
Foundation models and security copilots	Use large pretrained models for security-log understanding, threat-intelligence summarization, query generation and analyst assistance. Detection claims still require domain-specific validation.
Threat-informed detection	Align learned signals with adversary behavior and ATT&CK techniques. MITRE's current work increasingly emphasizes behavioral context, telemetry quality and measurable detection robustness [4,5].
Continuous evaluation and adaptive defense	Move from one-time benchmark testing to continuous monitoring of drift, calibration, robustness, coverage and analyst outcomes.

X. PROPOSED RESEARCH FRAMEWORK FOR FUTURE STUDIES

A rigorous future study should separate the research question into four layers: (1) detection capability, (2) generalization, (3) robustness, and (4) operational usefulness. The following workflow is recommended:

1. Define the threat model and operational objective, including the cost of false positives and false negatives.
2. Select diverse telemetry and establish a reproducible preprocessing pipeline without test-set leakage.
3. Build strong classical baselines before evaluating deep or foundation models.
4. Use temporal and entity-aware train/test splits and, where possible, cross-dataset or cross-environment validation.
5. Report precision, recall, F1, PR-AUC, low-FPR performance, latency, throughput and resource consumption.
6. Evaluate robustness against drift, adversarial perturbations, missing telemetry and changes in attack infrastructure.
7. Add explainability and evidence retrieval so that analysts can understand and validate detections.

8. Map meaningful detections to a threat-informed framework such as MITRE ATT&CK and assess coverage rather than only aggregate accuracy.
9. Perform human-centered evaluation: alert reduction, investigation time, analyst agreement and response quality.
10. Monitor deployed models continuously and define retraining, rollback and governance procedures.

XI. DISCUSSION

The literature shows that no single learning paradigm dominates across all cyber detection tasks. Supervised models remain highly competitive on well-labeled tabular datasets, while anomaly detection is better suited to settings where attack labels are scarce. Deep sequence models are valuable when temporal context matters, and graph models become natural when relationships among identities, processes, devices and services are central to the threat. Transformers offer flexible contextual modeling but introduce computational and evaluation challenges.

A key lesson is that algorithmic complexity should not be confused with detection quality. Security evaluation can be misleading when datasets are easy to separate, attack instances are temporally correlated, or baselines are weak. Recent research on multi-step attack detection also emphasizes high-dimensionality, sequential attack behavior, false positives, cross-dataset evaluation and the need to understand attack chains rather than isolated events [15].

Operationally, the most promising direction is a layered system in which learning-based models generate candidate signals, contextual engines correlate events, threat intelligence supplies adversary knowledge, and human analysts validate high-impact actions. This architecture allows AI to improve scale and consistency without assuming that automated classification is equivalent to ground truth.

XII. CONCLUSION

Learning-based cyber threat detection has progressed from classical statistical and machine-learning classifiers to deep, sequence-aware, graph-based and transformer-driven systems. These methods can improve the discovery of known and previously unseen threats across network, endpoint, cloud, IoT, identity, application and insider-threat domains. Nevertheless, real-world performance depends on data quality, realistic evaluation, robustness, low false-positive operation, interpretability, privacy and continuous adaptation. The research frontier is moving toward self-

supervised and few-shot learning, multimodal and graph reasoning, federated privacy-preserving analytics, foundation models, and threat-informed detection integrated with security operations. Future work should therefore prioritize reproducibility, cross-environment generalization and measurable operational outcomes over benchmark accuracy alone.

References

[1] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[2] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity, 2*, Article 20. https://doi.org/10.1186/s42400-019-0038-7

[3] MITRE. (2020). *MITRE ATT&CK: Design and philosophy*. MITRE Corporation.

[4] MITRE. (2025). *MITRE ATT&CK*. MITRE Corporation. https://attack.mitre.org/

[5] MITRE Center for Threat-Informed Defense. (2026). *Context to confidence: The next phase of ambiguous techniques research*. MITRE Corporation.

[6] *A survey on the applications of deep learning in network intrusion detection systems to enhance network security*. (2025). *IEEE Access, 13*, 185357–185373. https://doi.org/10.1109/ACCESS.2025.3624952

[7] Landauer, M., Onder, S., Skopik, F., & Wurzenberger, M. (2022). *Deep learning for anomaly detection in log data: A survey* [Preprint]. arXiv. https://arxiv.org/abs/2207.03820

[8] Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=LzQQ89U1qm_

[9] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)* (pp. 108–116). SCITEPRESS.

[10] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In

2015 Military Communications and Information Systems Conference (MilCIS) (pp. 1–6). IEEE. https://doi.org/10.1109/MilCIS.2015.7348942

[11] Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., & Rieck, K. (2022). Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 3971–3988). USENIX Association.

[12] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (NIST AI 100-2 E2025). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025

[13] Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[14] National Institute of Standards and Technology. (2026). *Artificial intelligence risk management framework (AI RMF)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

[15] *A review on multi-step attack detection*. (2025). *IEEE Access, 13*, 161779–161805. https://doi.org/10.1109/ACCESS.2025.3607497

[16] Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017). *Deep learning for unsupervised insider threat detection in structured cybersecurity data streams* [Preprint]. arXiv. https://arxiv.org/abs/1710.00811

[17] Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=LzQQ89U1qm_

[18] National Institute of Standards and Technology. (2024). *Cybersecurity framework (CSF) 2.0*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29

[19]