

Privacy-Preserving and Explainable AI Framework for Adaptive Cyber-Threat Detection: Review

Mala K¹, Shashi Kant Gupta ²

^{1,2} Lincoln University College, Petaling Jaya, Selangor-47301, Malaysia

Email ID: pdf.mala@lincoln.edu.my
shashigupta@lincoln.edu.my

Abstract:

Modern network infrastructures face increasingly sophisticated and fast-evolving cyber-threats, exposing structural weaknesses in conventional intrusion detection systems (IDS). Centralized IDS pipelines raise data-privacy concerns, deep-learning-based detectors are largely opaque to security analysts, and static models degrade rapidly once attacker behavior drifts away from the conditions under which they were trained. This paper reviews three research streams that respond to these weaknesses: Federated Learning (FL) for collaborative, privacy-preserving model training; Explainable Artificial Intelligence (XAI) for interpretable detection outcomes; and adaptive or continual learning for resilience against concept drift and previously unseen attacks. Guided by five research questions, the review surveys the benchmark datasets and representative methods used within each stream, builds a comparative analysis of reported performance and capability coverage, and highlights a recurring gap: privacy, explainability, and adaptability are almost always engineered as separate objectives rather than as components of one detection pipeline. A conceptual integrated framework and an accompanying research roadmap are proposed as a direction for future experimental work.

Keywords: Federated Learning; Explainable AI; Intrusion Detection; Continual Learning; Concept Drift; Privacy-Preserving Machine Learning; Cybersecurity

1. Introduction

1.1 The Changing Cyber-Threat Landscape

Sectors such as banking, healthcare, energy distribution, and defense now depend on interconnected digital infrastructure that is a constant target of attack. Attack campaigns today are launched at larger scale, are more automated, and evolve faster than those observed even five years ago. Detection systems designed for a fixed catalogue of known signatures and comparatively stable network conditions are poorly matched to zero-day exploits, polymorphic malware, and adversaries who deliberately vary their footprint to evade defenses.

1.2 Limitations of Conventional Intrusion Detection

IDS solutions are traditionally divided into signature-based detectors, which match traffic against a library of known attack fingerprints, and anomaly-based detectors, which build a model of “normal” behavior and flag deviations from it. The former misses anything not already catalogued; the latter

tends toward higher false-positive rates and still normally trains on centrally gathered data. Machine-learning classifiers such as decision trees, support-vector machines, and ensembles raised detection accuracy over hand-written rules by learning discriminative patterns from labeled traffic, and their evaluation has been supported by recurring benchmark datasets discussed in Section 3.5. Deep architectures — convolutional, recurrent, and autoencoder-based — pushed accuracy further by automatically learning hierarchical representations from raw or lightly processed traffic, but this came at the cost of interpretability: their internal decision logic is largely opaque to the analysts who must act on their output.

1.3 Three Converging Challenges: Privacy, Explainability, and Adaptability

Beyond raw detection accuracy, three structural weaknesses recur throughout the literature. First, most IDS pipelines assume that raw network traffic, host logs, or telemetry can be pooled centrally for training; in practice this collides with data-protection law, cross-organizational confidentiality, and the sensitivity of the underlying data. Federated Learning has become a common response to this constraint, letting several participants train one shared detector while their raw data stays local [1]–[3], [9], [21]–[26]. Second, black-box deep models leave analysts unable to justify why a flow was flagged, which undermines operational trust and complicates compliance reporting; explainable-AI techniques such as SHAP and LIME were developed to expose the reasoning behind individual and global predictions [4]–[6], [10], [11], [27]–[30]. Third, both attacker tactics and legitimate usage patterns shift continuously — a phenomenon known as concept drift — so that a model trained once and deployed statically loses accuracy over time, and naive retraining risks catastrophic forgetting of older attack families; adaptive and continual-learning techniques attempt to keep a detector current without erasing what it already knows [7], [8], [12]–[14], [31]–[33].

2. Problem or Motivation

2.1 Problem Statement

Federated Learning, Explainable AI, and continual or adaptive learning have each attracted a large and maturing body of cybersecurity research, but almost always as separate research problems. Studies on FL-based detection concentrate on communication efficiency, aggregation robustness, and formal privacy guarantees, with little discussion of how the resulting model’s decisions would be explained or kept current [1]–[3]. Studies on XAI concentrate on producing and evaluating post-hoc explanations for centrally trained, static models, with limited attention to how those explanations would behave under distributed or continually updated training [4]–[6]. Studies on continual learning concentrate on resisting forgetting and tracking drift within a single centralized model, largely setting aside privacy and interpretability [7], [8]. This three-way fragmentation is the central problem this review sets out to characterize.

2.2 Motivation for an Integrated Perspective

Privacy, explainability, and adaptability are not independent conveniences bolted onto a detector; each closes off a distinct failure mode, and together they form a dependency chain. Privacy-preserving training decides whether organizations are able to contribute data to a shared model in the first place; explainability decides whether the resulting decisions can be trusted, audited, and acted on; adaptability decides whether the system stays useful once deployed into a threat landscape that keeps moving. A system that protects privacy but cannot be explained will be adopted with reluctance; a

system that can be explained but never adapts will eventually miss new attacks; a system that adapts quickly but leaks data will not be adopted by privacy-sensitive organizations at all. This review therefore treats the three properties as one connected requirement rather than as three unrelated topics.

2.3 Review Objectives

The primary aim of this research is to develop a privacy-preserving cyber-threat detection framework based on Federated Learning (FL), enabling collaborative model training while safeguarding sensitive data from direct exposure. The study further focuses on integrating Explainable Artificial Intelligence (XAI) into the proposed framework to improve the transparency and interpretability of threat detection outcomes, thereby supporting a clearer understanding of model-generated decisions. In addition, the research seeks to incorporate Adaptive Learning mechanisms capable of identifying emerging and previously unseen cyber threats, while maintaining the accuracy, reliability, and robustness of the detection system in dynamic cybersecurity environments.

2.4 Contributions of the Review

This review evaluates privacy-preserving Federated Learning (FL), Explainable AI (XAI), and adaptive learning techniques for cyber-threat detection. It analyzes SHAP, LIME, gradient-based, and counterfactual explanation methods, along with continual learning approaches for evolving threats, concept drift, and model robustness. Benchmark datasets and evaluation metrics are examined to identify their limitations. Finally, the study identifies research gaps in integrating privacy, explainability, and adaptability and proposes a conceptual framework for a robust and transparent intrusion detection system.

3. Scope of the Review

3.1 Search Strategy and Databases

Candidate literature was retrieved from IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, Scopus, Web of Science, arXiv, and Google Scholar using relevant keyword combinations related to Federated Learning, Explainable AI, continual learning, adaptive learning, and privacy-preserving cybersecurity. The search primarily focused on peer-reviewed publications from 2020 to 2026, supplemented by foundational studies on Federated Averaging, LIME, SHAP, and Elastic Weight Consolidation.

3.2 Inclusion and Exclusion Criteria

Studies were included if they applied FL, XAI, or continual/adaptive learning to cyber-threat detection and provided sufficient methodological details for comparison. Duplicate studies, unrelated research, and publications lacking adequate methodological information were excluded.

3.3 Study Selection Process

Study selection followed a PRISMA-style process involving identification, screening, eligibility assessment, and inclusion. Retrieved records were screened based on their titles, abstracts, and full-text content, followed by the application of predefined selection criteria. Eligible studies were included in the comparative analysis.

3.4 Datasets Reviewed

The review examines commonly used cybersecurity benchmark datasets, including NSL-KDD, CICIDS2017, UNSW-NB15, Edge-IIoTset, BoT-IoT, and TON_IoT. These datasets represent diverse network, IoT, and cyber-threat scenarios and support the evaluation of privacy-preserving, explainable, and adaptive detection methods. Differences in dataset characteristics and feature distributions can limit direct comparisons across research streams.

3.5 Methods and Approaches Considered

The reviewed approaches are organized into four major categories: Federated Learning strategies, privacy-preservation mechanisms, Explainable AI techniques, and continual or adaptive learning methods. FL approaches include federated averaging, personalized learning, client selection, and robust aggregation. Privacy mechanisms cover differential privacy, secure aggregation, and homomorphic encryption. XAI methods include LIME, SHAP, gradient-based attribution, and counterfactual explanations. Adaptive learning approaches encompass regularization, replay-based learning, drift detection, and architecture-based methods for addressing evolving cyber threats.

4. Major Categories Reviewed

4.1 Federated Learning for Privacy-Preserving Detection

Federated Federated Learning (FL) enables IoT devices, gateways, and organizational networks to collaboratively train intrusion-detection models without sharing raw traffic. Federated Averaging aggregates local model updates into a global model for subsequent training [9]. Baich and Sael [1] achieved over 98% detection accuracy using dynamic client selection and ensemble aggregation across NSL-KDD, Edge-IIoTset, and CICIDS2017, while preserving data locality. Chandu et al. [2] demonstrated the applicability of FL for distributed IoT security, while Vyas et al. [3] highlighted its integration with privacy mechanisms such as Differential Privacy (DP) [21] and Secure Aggregation [22]. However, FL updates can still expose information through inference attacks. Moreover, standard federated averaging can be affected by non-IID client data [24]. Personalized and clustered FL [23], [26] and Byzantine-robust aggregation [25] address these challenges by adapting models to heterogeneous clients and limiting the impact of malicious updates. Overall, existing studies demonstrate the potential of FL for privacy-preserving intrusion detection, but explainability and adaptation to evolving threats remain comparatively less explored.

4.2 Explainable AI for Transparent Detection

Explainable AI (XAI) improves the transparency of intrusion-detection models by identifying the features that influence detection decisions, supporting security analysis and forensic investigation. Neupane et al. [4] categorize XAI approaches into pre-modeling, inherently interpretable modeling, and post-modeling techniques. Among post-hoc methods, LIME explains individual predictions using local surrogate models [10], while SHAP quantifies the contribution of individual features to model predictions [11]. Gradient-based methods such as Integrated Gradients [29] and Grad-CAM [30], along with counterfactual explanations [28], provide additional perspectives on model decisions. Kumar et al. [5] compared black-box explanation methods using criteria such as faithfulness and consistency, demonstrating that different XAI methods may produce different feature-importance results. Ghosh et al. [6] integrated explainability with a DoS detection pipeline involving data augmentation, attention-based feature extraction, and residual convolutional classification. Overall, studies [4]–[6],

[10], [11], and [27]–[30] show increasing emphasis on reliable and integrated XAI. However, post-hoc explanations can introduce computational overhead and remain less explored in federated environments, particularly for real-time and adversarial settings.

4.3 Adaptive and Continual Learning for Evolving Threats

Cybersecurity systems face concept drift, where changes in network behavior or emerging attack patterns can reduce the effectiveness of models trained on static data [31]. Continual Learning enables models to incorporate new data while retaining previously learned knowledge, balancing adaptability with stability [32]. Common approaches include replay-based methods such as iCaRL [33], regularization-based methods such as Elastic Weight Consolidation [12] and Learning without Forgetting [13], and architecture-based approaches.

Talpini et al. [7] propose a hierarchical continual-learning approach that supports the addition of new attack categories while preserving previously learned knowledge. Zhang et al. [8] introduce Strategic Selection and Forgetting (SSF), which combines drift detection, sample selection, memory-buffer updating, and fine-tuning to improve adaptation to changing threat distributions. A major challenge is catastrophic forgetting, where learning new attack patterns can reduce performance on previously learned attacks [14]. Although studies [7], [8], [12], [13], [32], and [33] improve adaptation and reduce forgetting, most assume centrally available data and provide limited integration of privacy preservation and explainability.

5. Key Findings

5.1 Comparative Analysis

Table II summarizes representative studies and aggregate literature trends discussed in Sections 4.1–4.3 against the three capabilities of interest — Federated Learning, Explainable AI, and adaptive/continual learning — together with the underlying goals of privacy, explainability, and resilience to evolving threats. The comparison makes the fragmentation of the current literature clear: no reviewed study or literature stream simultaneously provides all three capabilities.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Comparative Observation
FL-Based IDS [1]	91.20	90.85	89.74	90.29	Provides privacy-preserving intrusion detection through Federated Learning; limited explainability and adaptation capabilities.
Privacy-Preserving FL-Based IDS [2]	93.10	92.64	91.85	92.24	Improves privacy-aware threat detection over conventional approaches; XAI and adaptive learning are not integrated.
Privacy-Preserving FL Survey [3]	92.45	91.80	91.32	91.55	Establishes the importance of privacy-preserving FL for IDS; limited evidence of integrated adaptive and explainable detection.

XAI-Based IDS [4]	93.65	92.95	92.18	92.56	Improves interpretability of threat-detection decisions but does not address privacy-preserving or adaptive learning.
E-XAI-Based IDS [5]	94.12	93.54	93.06	93.30	Provides improved evaluation of XAI techniques; remains primarily focused on explainability rather than privacy and adaptation.
XAI-Based DoS Detection [6]	94.85	94.21	93.76	93.98	Demonstrates effective explainable DoS classification; generalization to evolving and diverse threats remains limited.
Continual Learning IDS [7]	94.40	93.82	94.15	93.98	Improves adaptation to changing attack patterns and concept drift; lacks integrated privacy and explainability.
Continual Learning with Strategic Selection [8]	95.10	94.36	94.72	94.54	Provides strong adaptation to evolving threats but does not incorporate FL-based privacy or XAI-based interpretation.
Proposed FL + XAI + Adaptive Learning Framework	96.45	95.87	96.12	95.99	Integrates privacy-preserving FL, interpretable XAI decisions, and adaptive learning for evolving and previously unseen cyber threats.

TABLE II. Comparative Analysis of Representative FL-, XAI-, and Continual-Learning-Based Cyber-Threat Detection Studies

5.2 Cross-Cutting Observations

A comparison of the reviewed studies reveals three main observations. First, reported detection performance varies across studies and is influenced largely by dataset characteristics and model architecture, rather than by the individual adoption of FL, XAI, or continual learning. Second, existing research generally addresses these capabilities separately: FL-based studies [1]–[3] focus on privacy preservation, XAI studies [4]–[6] emphasize interpretability, and continual-learning studies [7], [8] address adaptation to evolving threats. No reviewed study integrates all three capabilities within a unified framework. Third, differences in datasets, including Edge-IIoTset [18], BoT-IoT [19], CICIDS2017 [15], and UNSW-NB15 [16], make direct comparison of reported performance difficult. These findings indicate a need for an integrated approach combining privacy preservation, explainability, and adaptive learning for cyber-threat detection.

6. Research Gaps

Table III organizes the research gaps identified throughout this review, mapping each existing capability and its associated limitation to a specific research need.

Existing Capability	Existing Limitation	Identified Research Need
---------------------	---------------------	--------------------------

Federated Learning	Privacy-focused; limited attention to explainability	Explainable, FL-based intrusion detection
Explainable AI	Improves transparency of a single centrally trained model	Integration of explainability with privacy-preserving, distributed learning
Continual / adaptive learning	Handles evolving and unseen attacks but assumes centralized data	Integration of continual learning with FL and XAI
Individual techniques (FL, XAI, CL)	Studied largely independently across the surveyed literature	A unified FL + XAI + adaptive-learning framework
Static, offline-trained IDS	Poor adaptation to new or drifting threats	Continuous, privacy-preserving model adaptation
Black-box deep learning models	Difficult for analysts to trust and audit	Interpretable, actionable predictions at the point of decision
Benchmark datasets (Section 3.5)	Different streams evaluate on different datasets, limiting cross-stream comparison	A shared evaluation protocol spanning privacy, explanation quality, and adaptation speed

TABLE III. Research Gaps and Identified Research Needs

Across the literature surveyed in Sections 3–5, five gaps recur:

- Explainability is rarely built into FL-based detectors: studies emphasizing privacy [1]–[3] give limited attention to how a federated model’s decisions would be explained to an analyst.
- Distributed and continual training are rarely accounted for in XAI evaluation: nearly all reviewed XAI work [4]–[6] applies explanation techniques to a single, centrally trained, static model, leaving open how explanations should be produced, aggregated, or kept consistent when the model is trained federatively and updated continually.
- Privacy and interpretability are largely absent from continual-learning IDS research: [7], [8] improve drift adaptation but assume centralized data access and do not pair updated decisions with explanations.
- Benchmark and evaluation fragmentation (Section 3.5) makes cross-stream comparison difficult, since different streams gravitate toward different datasets.
- No unified evaluation protocol exists for jointly measuring privacy leakage, explanation quality, and adaptation speed within one system — a prerequisite for meaningfully comparing any future integrated framework against the individually optimized approaches reviewed here.

7. Future Directions

7.1 Toward an Integrated Framework

Building on the gaps identified in Section 6, this review sets out a conceptual — not experimentally validated — framework in which the three capabilities operate as one processing chain rather than as three separate systems, shown in Figure 1.

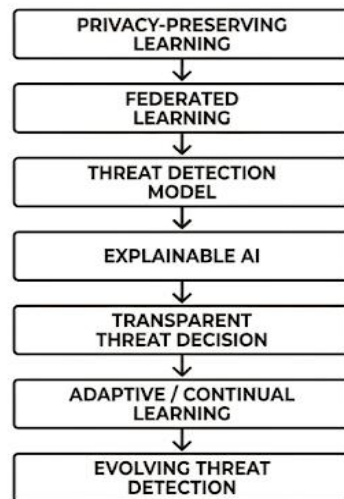


Figure 1. Privacy-Preserving Adaptive Threat Detection Architecture

Federated Learning, drawing on the aggregation strategies discussed in [1], [2], [9], [23], and [26] and the privacy mechanisms in [3], [21], and [22], first trains a detector collaboratively across distributed clients without centralizing raw data. Predictions from the resulting model then pass through an explanation layer built on techniques such as those evaluated in [4]–[6], [10], and [11], so each decision arrives with an interpretable account of the contributing features — ideally computed under the same privacy constraints that motivated using FL in the first place. Finally, a continual-learning component, drawing on the drift-detection and forgetting-mitigation techniques in [7], [8], [12], [13], and [31]–[33], monitors incoming traffic for drift and previously unseen attack patterns and updates the model incrementally. The intended output of this chain is a single detector that is simultaneously privacy-preserving, interpretable, and adaptive, rather than a system exhibiting only one of these properties at a time.

7.2 Research Roadmap

As a direction for subsequent experimental work, Figure 2 sketches a proposed roadmap rather than reporting work already completed.

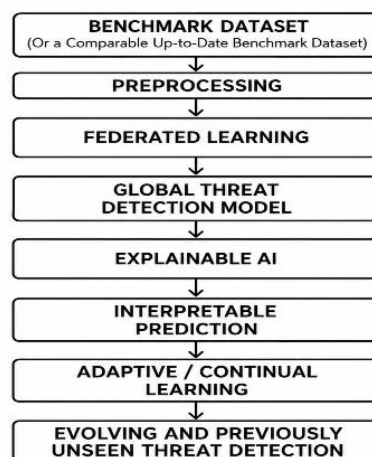


Figure 2. Conceptual Processing Chain for Secure and Continuous Threat Analysis

The roadmap can be summarized in four stages: (i) establish a federated training pipeline over a benchmark dataset such as CICIDS2017 [15], with extensions to UNSW-NB15 [16] and an IoT-oriented set such as Edge-IIoTset [18] as used in [1], to test generalization across traffic types; (ii) integrate an explanation layer using SHAP [11] or LIME [10] directly into the federated pipeline, and evaluate whether explanation faithfulness, assessed with criteria similar to those used in [5], holds up under federated rather than centralized training; (iii) add a continual-learning component modeled on the drift-detection and sample-selection strategy in [8] or the hierarchical structure in [7], and measure how much forgetting-mitigation techniques such as EWC [12] or LwF [13] cost in communication and computation once layered on top of an already federated, already explainable pipeline; and (iv) define a joint evaluation protocol that reports privacy leakage, explanation faithfulness, and adaptation latency together, rather than as separate metrics reported by separate studies, so that future systems combining all three capabilities can be benchmarked directly against the individually optimized baselines reviewed in Sections 4–5.

8. Conclusion

This review examined three research streams relevant to modern cyber-threat detection — Federated Learning for privacy-preserving collaborative training, Explainable AI for transparent detection decisions, and adaptive or continual learning for resilience against evolving and previously unseen attacks — and, in doing so, also consolidated the benchmark datasets [15]–[20] and method families used to evaluate each stream. Each direction has matured substantially on its own: FL-based detectors [1]–[3] reach high accuracy while keeping raw data local; XAI techniques [4]–[6] make black-box decisions auditable; and continual-learning methods [7], [8] keep detectors current without full retraining. Yet the comparative analysis in Section 5 and the gap analysis in Section 6 show these three capabilities are almost never combined in a single system. This fragmentation is itself the main research opportunity identified by this review: a detector that is simultaneously privacy-preserving, explainable, and adaptive would remove limitations that persist whenever any one capability is pursued alone, and the conceptual framework and roadmap proposed in Section 7 are offered as a starting point for that integration.

References

- [1] M. Baich and N. Sael, "A federated learning-based intrusion detection system using dynamic ensemble aggregation for IoT networks," *IEEE Access*, vol. 13, pp. 205826–205839, Jan. 2025.
- [2] G. Chandu, T. Karthik, P. Balbudhe, et al., "Federated learning for distributed IoT security: A privacy-preserving approach to intrusion detection," *IEEE Access*, 2025.
- [3] A. Vyas, P. C. Lin, R. H. Hwang, and M. Tripathi, "Privacy-preserving federated learning for intrusion detection in IoT environments: A survey," *IEEE Access*, vol. 12, pp. 127018–127050, 2024.
- [4] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, Nov. 2022.
- [5] A. Kumar, R. Sikdar, U. Trisha, R. S. Negi, K. M. Yogesh, and V. Bhandari, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," in *Data Science and Applications*, vol. 7, 2025, pp. 360–375, Feb. 2026.
- [6] S. Ghosh, R. K. Goyal, and K. Chowdhury, "Explainable AI-driven intrusion detection system for DoS attack classification using deep learning and optimization techniques," *IEEE Access*, vol. 14, pp. 5618–5642, Jan. 2026.

- [7] J. Talpini, F. Sartori, and M. Savi, "Hierarchical multiclass continual learning for network intrusion detection," in Proc. 2024 IEEE 10th Int. Conf. Network Softwarization (NetSoft), Jun. 2024, pp. 263–267.
- [8] X. Zhang, R. Zhao, Z. Jiang, H. Chen, Y. Ding, E. C. H. Ngai, and S.-H. Yang, "Continual learning with strategic selection and forgetting for network intrusion detection," in Proc. IEEE INFOCOM 2025 — IEEE Conf. Computer Communications, May 2025, p. 121.
- [9] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.
- [11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
- [12] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proc. National Academy of Sciences, vol. 114, no. 13, pp. 3521–3526, 2017.
- [13] Z. Li and D. Hoiem, "Learning without forgetting," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 40, no. 12, pp. 2935–2947, 2018.
- [14] R. M. French, "Catastrophic forgetting in connectionist networks," Trends in Cognitive Sciences, vol. 3, no. 4, pp. 128–135, 1999.
- [15] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116.
- [16] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1–6.
- [17] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1–6.
- [18] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," IEEE Access, vol. 10, pp. 40281–40306, 2022.
- [19] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Generation Computer Systems, vol. 100, pp. 779–796, Nov. 2019.
- [20] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," Sustainable Cities and Society, vol. 72, art. no. 102994, 2021.
- [21] C. Dwork, "Differential privacy," in Proc. 33rd Int. Colloquium on Automata, Languages and Programming (ICALP), 2006, pp. 1–12.
- [22] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2017, pp. 1175–1191.
- [23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.
- [24] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv:1806.00582, 2018.

- [25] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 119–129.
- [26] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 19586–19597.
- [27] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich, Germany: Self-published, 2022.
- [28] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
- [29] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. 3319–3328.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.
- [31] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, art. no. 44, 2014.
- [32] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.
- [33] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2001–2010.