International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 10 Issue 7 July - 2026, Pages: 157-164

Tackling Label Scarcity in Supervised Deep Learning for Big IoT Data: A Review of Semi-Supervised and Weak Supervision Approaches

Virendra Tank

Shri Mahaveer College, Jaipur, Rajasthan, India
vtank87@gmail.com 0009-0003-9126-0982

Abstract: *The proliferation of Internet of Things (IoT) devices has led to an explosive growth in sensor-generated big data, creating fertile ground for supervised deep learning applications ranging from human activity recognition to network intrusion detection. However, the practical deployment of supervised deep learning in IoT settings is severely constrained by label scarcity: while raw IoT data is abundant, expert-verified ground-truth labels remain scarce, costly, and slow to obtain. This review paper systematically examines the label scarcity problem in the context of deep learning, big data, and supervised learning as applied to IoT systems. It surveys four major families of label-efficient solutions – semi-supervised learning, weak supervision, active learning, and self-training/pseudo-labeling – and analyzes their mechanisms, strengths, and limitations across representative IoT application domains including smart homes, industrial IoT, and IoT network security. A comparative analysis of reported performance gains is presented, followed by a discussion of open challenges such as pseudo-label noise, heterogeneous sensor data, and federated learning constraints. The paper concludes with future research directions, including federated semi-supervised learning, foundation-model-assisted labeling, and edge-native label-efficient architectures for IoT big data analytics.*

Keywords: Deep Learning; Big Data; Supervised Learning; Internet of Things; Label Scarcity; Semi-Supervised Learning; Weak Supervision; Active Learning.

1. Introduction

The convergence of deep learning, big data, and the Internet of Things (IoT) has redefined how intelligent systems perceive and respond to the physical world. Billions of connected sensors continuously stream data characterized by high volume, velocity, and variety – the defining traits of big data [1], [2]. Deep learning architectures, particularly convolutional neural networks (CNNs) and recurrent architectures such as Long Short-Term Memory (LSTM) networks, have proven highly effective at extracting hierarchical representations from such large-scale, high-dimensional data [3]–[5].

Supervised learning remains the dominant paradigm for training these models because it typically yields the highest predictive accuracy when sufficient labeled data is available. Yet this is precisely where a fundamental bottleneck emerges in IoT environments: sensors generate data continuously and at scale, but assigning accurate ground-truth labels to this data requires human expertise, time, and cost that cannot realistically scale with data volume. This mismatch – abundant raw data versus scarce labeled data – is referred to throughout this paper as the label scarcity problem.

Unlike generic big data domains, IoT data has distinguishing characteristics that make label scarcity especially acute: sensor streams are continuous and non-stationary, labeling often requires domain-specific expertise (e.g., identifying a specific fault signature or a rare attack pattern), and many IoT deployments involve edge devices with limited connectivity to centralized annotation pipelines. As a result, purely supervised deep learning pipelines are frequently impractical for real-world IoT big data analytics.

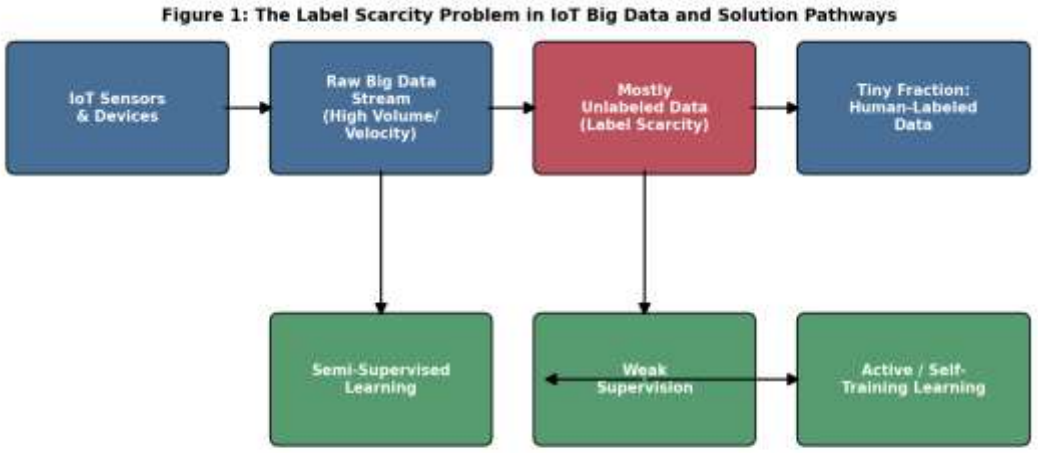

Figure 1: The Label Scarcity Problem in IoT Big Data and Solution Pathways

This review paper focuses specifically on techniques designed to alleviate label scarcity in supervised deep learning pipelines for big IoT data: semi-supervised learning, weak supervision, active learning, and self-training/pseudo-labeling. The remainder of this paper is organized as follows. Section 2 introduces background concepts. Section 3 describes the review methodology. Section 4 surveys label-efficient techniques in detail. Section 5 discusses their applications across IoT domains. Section 6 presents a comparative analysis of results reported in the literature. Section 7 discusses open challenges, Section 8 outlines future research directions, and Section 9 concludes the paper.

2. Background and Related Concepts

2.1 Deep Learning

Deep learning refers to a class of machine learning algorithms that use multi-layered neural networks to automatically learn hierarchical feature representations from raw data [3]. Architectures such as CNNs excel at extracting spatial features from image-like or grid-structured sensor data, while LSTM and Gated Recurrent Unit (GRU) networks are well suited to sequential, time-series data typical of IoT sensor streams [5]. Generative approaches, such as Generative Adversarial Networks (GANs), have also been used to synthesize realistic training samples, indirectly addressing data and label scarcity [6].

2.2 Big Data and the Internet of Things

Big data is commonly characterized by the '5 Vs': Volume, Velocity, Variety, Veracity, and Value [1], [2]. IoT ecosystems embody all five dimensions simultaneously — vast numbers of heterogeneous devices (variety) generate continuous streams (velocity) of large-scale data (volume) that is often noisy or incomplete (veracity), and from which actionable insights must be extracted (value). This combination makes IoT one of the most demanding and realistic testbeds for big-data-driven deep learning research.

2.3 Supervised Learning and the Label Scarcity Problem

Supervised learning trains a model on input-output pairs, where outputs are ground-truth labels provided by human annotators or reliable automated systems. Its effectiveness is strongly tied to the quantity and quality of labeled training data. In IoT contexts, labeling is complicated by the sheer scale of data, the need for domain expertise (e.g., clinicians for wearable health data, security analysts for network traffic), and the rarity of certain classes (e.g., attack events or equipment failures). The label scarcity problem thus motivates the family of label-efficient learning paradigms surveyed in this review: semi-supervised learning, weak supervision, active learning, and self-training [7].

3. Review Methodology

This review follows a structured literature synthesis approach adapted from systematic review practices. Relevant studies were identified through keyword searches across IEEE Xplore, ScienceDirect, SpringerLink, arXiv, and the ACM Digital Library using combinations of the terms “label scarcity,” “semi-supervised learning,” “weak supervision,” “active learning,” “Internet of Things,” “big data,” and “deep learning.” Studies published between 2013 and 2025, written in English, and either peer-reviewed or hosted as citable preprints on arXiv were considered eligible.

Titles and abstracts were first screened to remove duplicates and studies unrelated to label-efficient supervised deep learning in IoT or closely related sensor/big-data contexts. Full texts of the remaining candidates were then reviewed against inclusion criteria requiring a clear methodological contribution to semi-supervised learning, weak supervision, active learning, or self-training, or a direct application to IoT big data analytics. This process yielded 28 core studies, which were thematically grouped by technique family for qualitative and comparative synthesis, as illustrated in Figure 3.

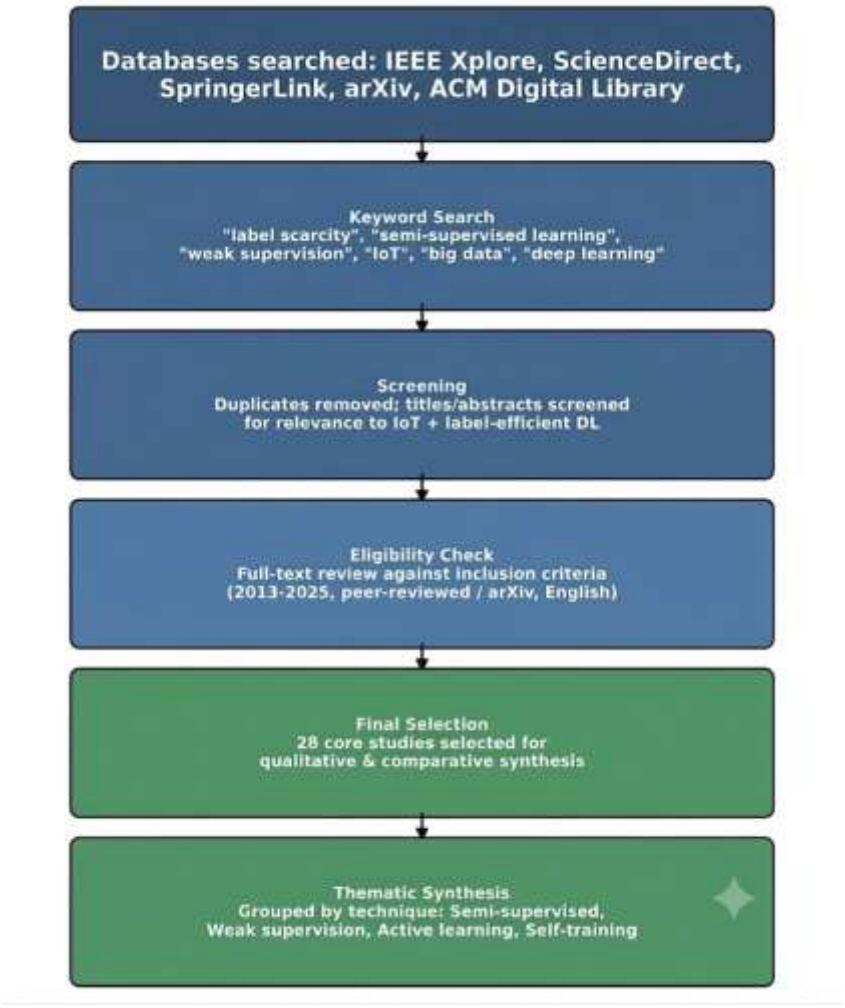

Figure 3: Review Methodology Workflow

4. Techniques for Tackling Label Scarcity

Figure 2 presents a taxonomy of the label-efficient approaches surveyed in this paper, organized into four major families that are discussed in the following subsections.

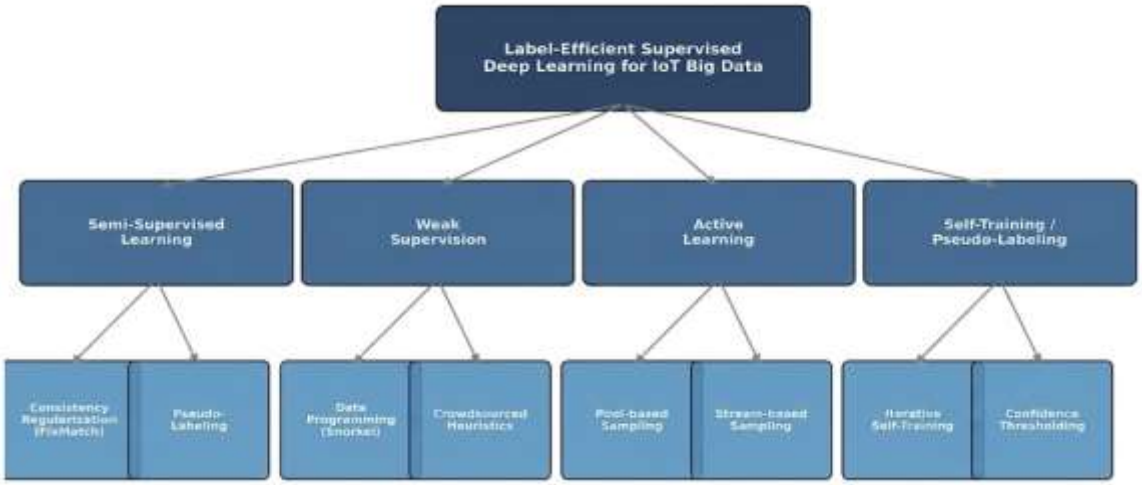

Figure 2: Taxonomy of Label-Efficient Approaches for Supervised Deep Learning in IoT

4.1 Semi-Supervised Learning

Semi-supervised learning (SSL) trains models using a small labeled dataset alongside a much larger pool of unlabeled data, exploiting the structure of unlabeled samples to improve generalization [7]. Consistency regularization methods enforce that a model produce similar predictions for differently augmented views of the same input, while pseudo-labeling methods assign artificial labels to high-confidence unlabeled predictions. FixMatch unifies both ideas by generating pseudo-labels from weakly augmented inputs and enforcing consistency on strongly augmented counterparts [8], while MixMatch combines augmentation, label guessing, and interpolation-based regularization [9]. In IoT network security, semi-supervised spatiotemporal deep learning has been applied to intrusion detection with strong reported performance even under severe label constraints [16], [17], and multi-branch frameworks such as Semi-WTC combine interpolation and pseudo-labeling branches to handle extreme label scarcity scenarios with as few as three to five labeled samples per class [18].

4.2 Weak Supervision

Weak supervision replaces manual labeling with programmatic, heuristic, or noisy labeling sources that are aggregated into probabilistic labels. Systems such as Snorkel formalize this process through "labeling functions" – simple rules, heuristics, or pattern matchers – whose outputs are combined by a generative label model to approximate ground truth at scale [12]. Programmatic weak supervision has since matured into a broad research area covering data programming, crowdsourced heuristics, and automated labeling-function generation [13]. For IoT big data, weak supervision is particularly attractive because domain heuristics (e.g., simple threshold rules on sensor readings) are often easy to specify even when manual labeling is not feasible at scale.

4.3 Active Learning

Active learning reduces annotation cost by having the model itself select the most informative unlabeled samples for human labeling, rather than labeling data at random [14]. Pool-based sampling selects queries from a fixed unlabeled pool, while stream-based sampling makes real-time query decisions as data arrives – a setting well aligned with continuous IoT sensor streams. Deep active learning extends these ideas to neural networks using uncertainty estimation, representation-based sampling, or hybrid criteria [15]. Applications include human activity recognition, where autoencoder-based active learning selects informative samples using deep support vector data description to reduce labeling effort while improving recognition accuracy, and industrial data collection scheduling, where active learning has been used to minimize the number of manually labeled production samples required over time.

4.4 Self-Training and Pseudo-Labeling

Self-training is one of the earliest and simplest label-efficient paradigms: a model is first trained on a small labeled set, then used to generate pseudo-labels for unlabeled data; samples with sufficiently confident predictions are added back into the training set, and the model is iteratively retrained [10]. While computationally simple, naive self-training is vulnerable to confirmation bias, where early prediction errors are reinforced over successive iterations. Debiased self-training methods address this by correcting representation and classifier bias introduced during pseudo-label generation [11]. In IoT fault diagnosis and smart-grid anomaly

detection, semi-supervised autoencoder and self-training pipelines have demonstrated the ability to detect rare fault or attack conditions using only a small fraction of manually verified labels [25].

5. Applications in Big IoT Data Domains

5.1 Smart Homes and Human Activity Recognition

Wearable and ambient IoT sensors generate continuous streams of accelerometer, gyroscope, and environmental data used for human activity recognition (HAR). Because manually annotating every activity segment is impractical at scale, active learning and semi-supervised approaches have been used to substantially reduce labeling effort while preserving recognition accuracy, with reported gains of over 3% in accuracy and F1-score compared to conventional sampling strategies.

5.2 Industrial IoT and Fault Diagnosis

In industrial settings, sensors monitoring machinery vibration, temperature, and current generate massive volumes of data, but fault events are rare and costly to label. Semi-supervised fault diagnosis methods that combine active-learning-based information screening with pseudo-label self-training have been shown to improve classification reliability under imbalanced and scarcely labeled conditions.

5.3 IoT Network Security and Intrusion Detection

Network intrusion detection is one of the most extensively studied IoT applications of label-efficient supervised deep learning, since malicious traffic samples are rare relative to benign traffic and labeling requires specialized security expertise. Semi-supervised spatiotemporal models [17], distributed semi-supervised attack detection frameworks [16], and federated semi-supervised approaches such as FedMSE [19] have all demonstrated that competitive detection accuracy can be achieved while relying on only a small proportion of labeled traffic samples, and while preserving data privacy across distributed IoT gateways [21], [22].

5.4 Healthcare IoT

Wearable health-monitoring IoT devices continuously capture physiological signals such as ECG. Because expert cardiologist annotation is expensive and time-consuming, semi-supervised multi-label frameworks have been proposed to recognize multiple co-occurring conditions from limited supervision, using augmentation and pseudo-label refinement strategies analogous to those used in general-purpose semi-supervised learning [28].

6. Results and Comparative Analysis

Table 1 summarizes the core mechanism, typical label requirements, reported performance gains, best-fit IoT use cases, and key limitations of each label-efficient technique family surveyed in this review, synthesized from the reviewed literature.

Technique	Core Mechanism	Typical Label Requirement	Reported Accuracy / F1 Gain	Best-Fit IoT Use Case	Key Limitation
Semi-Supervised Learning (FixMatch, MixMatch)	Combines small labeled set with consistency regularization / pseudo-labeling on unlabeled data	3-5 labels per class to ~10% of data	Up to 8.8% F1 improvement over supervised baselines [18]	Network intrusion detection, ECG/HAR classification	Sensitive to confidence threshold; error propagation from noisy pseudo-labels
Weak Supervision (Snorkel, Data Programming)	Aggregates multiple noisy labeling functions/heuristics into probabilistic labels	Zero to minimal seed labels	Comparable to fully supervised models with far less annotation effort [12], [13]	Large-scale sensor log tagging, industrial fault categorization	Requires domain heuristics; label model quality depends on LF diversity

Technique	Core Mechanism	Typical Label Requirement	Reported Accuracy / F1 Gain	Best-Fit IoT Use Case	Key Limitation
Active Learning (Pool/Stream-based)	Iteratively queries human oracle for the most informative unlabeled samples	Small, growing labeled set (query-driven)	Same accuracy with substantially fewer labeled instances [14], [15]	Human activity recognition, industrial defect detection	Human-in-the-loop cost; selection strategy sensitive to class imbalance
Self-Training / Pseudo-Labeling	Retrains model iteratively using its own high-confidence predictions as labels	Very small seed labeled set	Competitive with supervised models on benchmark IoT datasets [10], [11]	Fault diagnosis, smart-grid anomaly detection	Prone to confirmation bias and training instability
Federated Semi-Supervised Learning	Distributes semi-supervised training across edge/IoT gateways without centralizing raw data	Few labeled samples per client	Detection accuracy improved from ~94% to ~97% with 50% client participation [19]	Distributed IoT botnet/intrusion detection	Communication overhead; non-IID data across devices

Table 1: Comparative Analysis of Label-Efficient Supervised Deep Learning Techniques for IoT Big Data

Across the reviewed studies, semi-supervised and self-training approaches consistently achieve accuracy close to fully supervised baselines while using only a small fraction of labeled data, though performance is sensitive to pseudo-label noise and confidence-threshold tuning. Weak supervision offers the greatest reduction in manual annotation effort but depends heavily on the quality and coverage of domain heuristics. Active learning provides the most direct control over labeling cost but introduces a human-in-the-loop dependency that can be difficult to scale in fully autonomous IoT deployments. Federated semi-supervised learning emerges as a particularly promising direction for IoT specifically, since it addresses both label scarcity and data privacy/decentralization constraints simultaneously [19], [21], [22].

7. Challenges and Open Issues

- Pseudo-label noise and confirmation bias: Iteratively trusting a model's own predictions can reinforce early errors, particularly under class imbalance common in IoT fault and intrusion data [11].
- Heterogeneous and non-stationary sensor data: IoT data varies in sampling rate, unit, and modality across devices, complicating the design of consistent augmentation strategies required by many semi-supervised methods.
- Resource-constrained edge deployment: Many IoT devices lack the computational capacity to run iterative pseudo-labeling or active learning query strategies locally, necessitating lightweight or edge-cloud hybrid designs.
- Privacy and decentralization: Centralizing raw IoT data for labeling or model training raises privacy concerns, motivating federated approaches that introduce additional communication and non-IID data challenges [19].
- Evaluation inconsistency: Reviewed studies use varied datasets, label budgets, and metrics, making direct cross-study comparison difficult and highlighting the need for standardized IoT label-scarcity benchmarks.

8. Future Work

- Federated semi-supervised and weak supervision frameworks that jointly address label scarcity and data decentralization across large-scale IoT deployments [19], [21], [22].

- Foundation-model-assisted weak supervision, where large pretrained models generate high-quality labeling functions or pseudo-labels with minimal domain engineering, reducing reliance on hand-crafted heuristics [13].
- Edge-native label-efficient architectures that support lightweight active learning or self-training directly on resource-constrained IoT hardware.
- Robust pseudo-label calibration techniques to mitigate confirmation bias and improve reliability under the severe class imbalance typical of rare-event IoT data such as faults and intrusions [11], [25].
- Standardized benchmark datasets and evaluation protocols for label-scarce IoT big data analytics, enabling fairer comparison across semi-supervised, weak supervision, and active learning methods.

9. Conclusion

This review has examined the label scarcity problem at the intersection of deep learning, big data, and supervised learning in IoT systems, and has surveyed four major families of solutions: semi-supervised learning, weak supervision, active learning, and self-training/pseudo-labeling. Across applications spanning smart homes, industrial IoT, network security, and healthcare, these techniques consistently demonstrate that competitive predictive performance can be achieved using only a small fraction of manually labeled data. Nonetheless, challenges related to pseudo-label reliability, data heterogeneity, edge resource constraints, and privacy remain open. Federated semi-supervised learning and foundation-model-assisted weak supervision represent particularly promising directions for future research, with the potential to make label-efficient supervised deep learning practical and scalable across real-world IoT big data ecosystems.

References

[1] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," Journal of Big Data, vol. 2, no. 1, pp. 1–21, 2015.

[2] Q. Zhang, L. T. Yang, and Z. Chen, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146–157, 2018.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in Proc. NeurIPS, 2014, pp. 2672–2680.

[7] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," Machine Learning, vol. 109, no. 2, pp. 373–440, 2020.

[8] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," arXiv:2001.07685, 2020.

[9] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in Proc. NeurIPS, 2019.

[10] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in Proc. ICML Workshop on Challenges in Representation Learning, 2013.

[11] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," arXiv:2202.07136, 2022.

[12] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," Proceedings of the VLDB Endowment, vol. 11, no. 3, pp. 269–282, 2017.

[13] J. Zhang, C.-Y. Hsieh, Y. Yu, C. Zhang, and A. Ratner, "A survey on programmatic weak supervision," arXiv:2202.05433, 2022.

[14] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[15] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," ACM Computing Surveys, vol. 54, no. 9, pp. 1–40, 2021.

[16] S. Rathore and J. H. Park, "Semi-supervised learning based distributed attack detection framework for IoT," Applied Soft Computing, vol. 72, pp. 79–89, 2018.

[17] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, and M. J. Ryan, "Semi-supervised spatiotemporal deep learning for intrusions detection in IoT networks," IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12251–12265, 2021.

[18] Z. Li, W. Chen, Z. Wei, X. Luo, and B. Su, "Semi-WTC: A practical semi-supervised framework for attack categorization through weight-task consistency," arXiv:2205.09669, 2022.

[19] V. T. Nguyen et al., "FedMSE: Semi-supervised federated learning approach for IoT network intrusion detection," arXiv:2410.14121, 2024.

[20] A. Awajan, "A novel deep learning-based intrusion detection system for IoT networks," Computers, vol. 12, no. 2, p. 34, 2023.

[21] B. Olanrewaju-George and B. Pranggono, "Federated learning-based intrusion detection system for the internet of things using unsupervised and supervised deep learning models," Cyber Security and Applications, vol. 3, p. 100068, 2025.

[22] N. Albanbay, Y. Tursynbek, K. Graffi, R. Uskenbayeva, Z. Kalpeyeva, Z. Abilkaiyr, and Y. Ayapov, "Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study," Journal of Sensor and Actuator Networks, vol. 14, no. 4, p. 78, 2025.

[23] W. Fang, C. Zhu, and W. Zhang, "Toward secure and lightweight data transmission for cloud-edge-terminal collaboration in artificial intelligence of things," IEEE Internet of Things Journal, vol. 11, no. 1, pp. 105–113, 2024.

[24] Z. Chen, J. Liu, Y. Shen, M. Simsek, B. Kantarci, H. T. Mouftah, and P. Djukic, "Machine learning-enabled IoT security: Open issues and challenges under advanced persistent threats," ACM Computing Surveys, vol. 55, no. 3, pp. 1–37, 2023.

[25] A. Dairi, F. Harrou, B. Bouyeddou, S.-M. Senouci, and Y. Sun, "Semi-supervised deep learning-driven anomaly detection schemes for cyber-attack detection in smart grids," in Power Systems Cybersecurity: Methods, Concepts, and Best Practices, Springer, 2023, pp. 265–295.

[26] M. Nashaat, "Hybridization of active learning and data programming for labeling large industrial datasets," in Proc. IEEE International Conference on Big Data, 2018.

[27] E. Hsu and K. Roberts, "Leveraging large language models for knowledge-free weak supervision in clinical natural language processing," preprint, 2024.

[28] R. Zhou, L. Lu, Z. Liu, T. Xiang, Z. Liang, D. A. Clifton, Y. Dong, and Y.-T. Zhang, "Semi-supervised learning for multi-label cardiovascular diseases prediction: A multi-dataset study," arXiv:2306.10494, 2023.