Developing machine learning-based intrusion detection systems for IoT environments

Zhe Deng[1]

[1]Tallinn University of Technology, Ehitajate tee 5, 12616, Tallinn, Estonia

Abstract

The rapid growth of Internet of Things (IoT) devices has expanded the attack surface of modern networks, underscoring the need for robust and adaptive security solutions. Machine learning-based Intrusion Detection Systems (IDS) offer promise but face challenges like data scarcity, imbalance, and concept drift. This paper outlines ongoing doctoral research focused on developing scalable and label-efficient IDS frameworks tailored to IoT environments. A systematic literature review evaluates the role of generative models, such as GANs, Autoencoders, and Transformers, in addressing these challenges. In addition, the paper presents a novel active learning-based detection pipeline, validated using mobile malware data. Combining uncertainty sampling, auto-labeling, and drift detection, the approach achieves over 97% accuracy with less than 3% labeled data. These results support the core hypothesis that adaptive, lightweight ML models may improve intrusion detection in IoT environments.

Keywords

machine learning (ML), intrusion detection system (IDS), Internet of Things (IoT), generative AI, mobile malware

1. Introduction

The proliferation of the Internet of Things (IoT) has fundamentally transformed the digital landscape, connecting billions of devices and industrial sensors into a massive, distributed computing environment [1, 2]. While this interconnectedness brings unprecedented convenience and innovation, it also introduces serious cybersecurity risks. IoT environments are increasingly targeted by adversaries exploiting their inherent limitations: heterogeneous architectures, minimal security configurations, limited computational resources, and inconsistent software updates. These factors render traditional Intrusion Detection Systems (IDS), which are often designed for stationary, high-power environments, ineffective when applied directly to IoT [3, 4].

Machine learning (ML)-based IDSs have emerged as promising tools to enhance the security of IoT environments due to their ability to learn complex attack patterns from data and adapt to previously unseen threats [5]. However, several challenges complicate the effective use of ML in this context. First, acquiring large, labeled datasets in IoT is difficult due to privacy concerns, labeling costs, and the domain-specific nature of threats. Second, IoT data is often highly imbalanced and non-stationary, with concept drift emerging over time as attackers evolve their techniques [6]. Third, the constrained resources of IoT devices demand lightweight, efficient models capable of real-time inference without degrading performance [7].

The central objective of my PhD research is to develop machine learning-based intrusion detection systems tailored specifically for IoT environments that are adaptive, label-efficient, and robust to evolving threat patterns and computational limitations. In support of this objective, my work has taken two significant directions thus far. First, a systematic literature review has examined the current state of generative artificial intelligence within IDS, with a particular focus on its relevance and applicability to the Internet of Things. This review offers an overview of recent advances, identifies common design challenges, and highlights unresolved problems that limit the deployment of generative models in

BIR-WS 2025: BIR 2025 Workshops and Doctoral Consortium, 24rd International Conference on Perspectives in Business Informatics Research (BIR 2025), September 17-19, 2025, Riga, Latvia

zhe.deng@taltech.ee (Z. Deng)
https://zhe2d.github.io (Z. Deng)
0000-0002-0990-6031 (Z. Deng)

resource-constrained environments [8]. Second, I have designed and evaluated an active learning-based detection framework that combines auto-labeling with data drift detection. This approach aims to reduce the reliance on expensive manual annotation while maintaining strong detection performance over time, even as the underlying data distribution shifts [9].

Together, these contributions lay the groundwork for developing intelligent IDS solutions capable of operating in complex and dynamic IoT ecosystems. They also help clarify the practical requirements and limitations that any future IDS must meet to be effective and sustainable in the long term. My ongoing research builds on these insights by investigating how hybrid architectures, including ensemble models and federated learning techniques, might further improve adaptability and efficiency in diverse deployment contexts [10, 11]. Through this research, I aim to contribute to the development of practical, scalable, and intelligent IDS frameworks that address the evolving security needs of the IoT landscape.

This paper is structured as follows: Section 2 introduces the research methodology and experimental framework employed. Section 3 reviews related literature on machine learning, specifically generative AI for IoT IDS, and outlines gaps. Section 4 presents the preliminary results based on the active learning framework developed in earlier work. Section 5 discusses the implications of these findings and proposes directions for future investigation.

1.1. Research questions

The following research questions (RQs) are derived from the limitations, challenges, and research gaps identified in the systematic literature review presented in Section 2. Specifically, gaps related to data scarcity, class imbalance, deployment constraints, and concept drift shaped the formulation of each question. These RQs guide the scope of this doctoral research and ensure alignment between literature-driven needs and methodological direction.

RQ1: *What are the key limitations and emerging trends in applying machine learning, including generative AI, to intrusion detection in IoT environments?*

RQ2: *How can ML models be adapted to detect cyber threats in IoT environments with limited labeled data, data imbalance, and concept drift?*

RQ3: *What features and data representations are most effective for training ML-based IDS in heterogeneous and resource-constrained IoT environments?*

RQ4: *To what extent can hybrid ML techniques improve detection performance, robustness, and adaptability in evolving IoT threat landscapes?*

These questions are addressed through a combination of systematic literature analysis and empirical evaluation of proposed ML-based IDS frameworks. The answers to these questions will contribute to the design of next-generation intrusion detection solutions that are both effective and practical in real-world IoT environments.

2. Research methodology

This research adopts the Applied Research method [12], targeting real-world challenges in designing and deploying machine learning-based intrusion detection systems for IoT environments. It is a six-phase cyclic model:

(1) Problem analysis: This work is motivated by limitations in existing IDSs, particularly their inability to handle non-stationary data, reliance on labeled datasets, and inefficiency in constrained environments. IoT environments introduce further complexity, including heterogeneous devices, multimodal data, and evolving threats—necessitating IDSs that are adaptive and resource-aware.

(2) Literature review: A structured, ongoing literature review underpins all research stages, covering foundational theories and emerging approaches such as generative AI, active/semi-supervised learning, drift detection, and explainable AI. Particular attention is given to IoT-specific challenges like device diversity and limited computational capacity. The initial research cycle involved a systematic review identifying limitations in current ML-based IDSs, guiding both hypothesis formation and experimental design.

(3) Hypothesis formulation: Each cycle defines testable hypotheses addressing design goals such as label efficiency, robustness, and adaptability. For example, we hypothesize that active learning reduces annotation costs and that hybrid models enhance performance across heterogeneous domains. These hypotheses align with the overarching research questions.
(4) Dataset collection and feature engineering: The evaluation leverages a dual-domain setup: mobile malware detection (KronoDroid) and IoT intrusion datasets (UNSW-NB15 [13], Edge-IIoT [14]). Mobile malware offers a high-fidelity proxy for IoT dynamics, supporting experiments on drift and hybrid features. IoT network datasets complement this with protocol-specific traffic and deployment realism. Known dataset limitations are mitigated through preprocessing and feature engineering.

Features across all datasets were unified into hybrid representations combining static (e.g., permissions) and dynamic (e.g., system calls) indicators. Temporal segmentation simulates real-world evolution, enabling longitudinal evaluations of adaptability and generalizability.
(5) Development and experimentation: Machine learning-based IDS prototypes were developed with modular designs incorporating active learning, drift detection, auto-labeling, and (future) generative augmentation. Experiments span various datasets, feature sets, and supervision levels (e.g., full, uncertainty-based). Android malware currently serves as the testbed, with plans to extend to federated IoT setups.
(6) Evaluation and validation: Evaluation includes standard metrics (accuracy, F1, etc.) alongside deployment-oriented criteria like label cost, adaptability, and computational overhead. Explainability is also considered to ensure interpretability in critical contexts. Longitudinal tests across time-segmented data support analysis under concept drift.

Each cycle's findings refine subsequent methodology, supporting an iterative and practice-driven research process. Implementation is conducted in Python using Scikit-learn, TensorFlow, and PyTorch.

3. Related research

3.1. Generative AI in IDS for IoT: a systematic literature review

The growing interest in applying machine learning (ML) to intrusion detection for Internet of Things (IoT) systems has resulted in a fragmented and evolving body of literature. Given the complexity of the ML landscape and the increasing prominence of generative artificial intelligence (AI) within it, this study conducted a systematic literature review [8] to establish a clear foundation for this doctoral research.

The review protocol was developed to ensure methodological rigor and relevance to the research questions at hand, following Kitchenham's guidelines [15]. We conducted a comprehensive search of major scientific databases, including IEEE Xplore, Scopus, ACM Digital Library, SpringerLink, and ScienceDirect, focusing on the period between 2018 and 2023, beginning with the first recognized scientific evaluation of GAN applicability in IoT IDS [16]. The query string "*(generative AI OR GAN OR Transformers) AND (Internet of Things OR IoT) AND (intrusion detection OR IDS)*" was used to identify studies at the intersection of these domains. After applying inclusion, exclusion, and quality assessment criteria, 100 primary studies were selected for detailed analysis.

The findings highlight a growing use of generative models, especially GANs—for addressing core challenges in IoT IDS, including limited labeled data, class imbalance, and evolving threat behavior (RQ2, RQ3). GANs are primarily employed for data augmentation and anomaly detection, while Transformer-based models are emerging for time-series and sequential traffic analysis, reflecting increased attention to temporal patterns and real-world applicability (RQ1).

Applications of generative AI span synthetic attack generation, unsupervised feature learning, and latent representation modeling. These methods support the development of IDS that can adapt to heterogeneity and resource constraints common in IoT environments (RQ2, RQ3). However, the review also exposes methodological gaps: many studies lack evaluation across datasets, fail to assess robustness or explainability, and rely on static benchmarks, limiting insight into model generalizability and operational readiness (RQ4).

This review thus establishes a foundation for the present research, reinforcing the need for adaptable, label-efficient, and rigorously evaluated ML-based IDS tailored for dynamic IoT environments. The following subsections provide a detailed synthesis of three dimensions of the literature review: generative model architectures and techniques, their applications in IoT IDS, and the approaches used to evaluate their performance and relevance.

3.1.1. Generative AI model architectures and techniques

Four primary generative model families dominate the research landscape: GANs, AEs (including VAEs and CVAEs), Transformer-based models, and hybrid or alternative architectures. GAN-based approaches constitute the majority, with over 58% of the reviewed studies employing variants such as Conditional GANs (CGAN), Auxiliary Classifier GANs (ACGAN), and Wasserstein GANs (WGAN-GP) [17]. These models typically pair a generator with a discriminator to synthesize realistic intrusion traffic, often to augment minority-class samples in imbalanced datasets.

Autoencoder-based approaches, including standard AEs and variational variants, are primarily used for feature compression and reconstruction. While they are less prevalent as standalone generative models, AEs often appear in hybrid systems, where their capacity to model latent representations complements the data synthesis capabilities of GANs. Transformer-based architectures, though relatively recent, have shown potential in modeling sequential patterns in IoT traffic. Their parallel computation capability and attention mechanisms offer a pathway to capturing long-range patterns in complex IoT traffic data.

Hybrid architectures further combine the strengths of multiple model types. Notable examples include GAN+AE models, which leverage adversarial learning for sample generation alongside autoencoding for anomaly detection. Similarly, Transformer-GAN hybrids incorporate temporal modeling with data synthesis to create robust, explainable, and adaptable detection pipelines.

3.1.2. Applications in IoT intrusion detection

The reviewed literature identifies three core application domains for generative models within IoT IDS: data augmentation and class balancing, anomaly detection and reconstruction, and adversarial attack simulation.

The most prevalent application is data augmentation, where synthetic samples are generated to expand limited training sets. GANs and VAEs are commonly employed to address extreme class imbalance, a common issue in IDS datasets where benign traffic vastly outnumbers malicious instances. Techniques such as Conditional Tabular GAN (CTGAN) [18] are employed to selectively synthesize minority class data while preserving distributional fidelity. Several studies report gains in recall and generalization, particularly in class-imbalanced contexts.

Reconstruction-based applications primarily involve AEs and VAEs. These models are trained to learn the distribution of benign traffic and identify anomalies based on reconstruction error. This is particularly useful for zero-day attack detection, where no labeled attack data exists. Additionally, AEs contribute to dimensionality reduction and feature extraction, both of which are critical for efficient deployment in resource-constrained IoT settings.

A smaller but growing segment of research explores adversarial attack generation. Here, GANs are trained to generate adversarial samples that evade detection, serving both offensive research and defensive countermeasure development. These works underscore the vulnerability of existing ML-based IDS and highlight the importance of robust, adversarially trained models for deployment in real-world scenarios.

3.1.3. Evaluation approaches and performance metrics

While the reviewed studies report promising detection accuracies, often exceeding 95% on the benchmark datasets, they frequently lack robustness evaluations across diverse scenarios and real-world conditions. The quality and comprehensiveness of evaluation methodologies vary widely. The most frequently

used datasets include NSL-KDD [19], BoT-IoT [20], CICIDS2017 [21], and UNSW-NB15 [13], many of which suffer from known limitations such as outdated attack scenarios, synthetic traffic, or limited diversity.

Standard evaluation metrics, accuracy, precision, recall, and F1 score, are commonly reported, but are often insufficient to fully characterize model performance in dynamic IoT environments. Few studies assess robustness to concept drift, generalizability [3] across domains, or computational efficiency. Moreover, explainability [22], a crucial criterion for operational trust in IDS, is rarely addressed. Transformer-based models and some hybrid architectures offer potential for interpretable outputs (e.g., through attention weights), yet few studies capitalize on this capacity.

Another overlooked aspect is deployment feasibility. While generative AI models show strong performance in centralized training, few consider lightweight [11] or distributed deployment. This is a significant omission, as many IoT devices operate under severe resource constraints, necessitating models with low memory and computational footprints. Federated and edge-aware generative frameworks, while emerging, remain largely underexplored.

In summary, the systematic literature review reveals that while generative models offer substantial advantages for IoT IDS, particularly in enhancing data availability and improving detection of evolving threats, significant gaps remain in evaluation rigor, deployment realism, and operational explainability. These findings directly inform the methodological direction of this doctoral research.

Despite promising results, most studies remain limited in terms of practical deployment. Evaluation practices are frequently narrow in scope, relying on isolated metrics such as accuracy or F1-score without examining robustness, generalizability, or explainability across different IoT scenarios (RQ4). Furthermore, only a few studies report performance under adversarial conditions or across multiple datasets, and almost none consider model interpretability, a critical factor for real-world applications. These limitations suggest that while generative AI holds potential for advancing IDS design, existing approaches often fall short in ensuring operational feasibility and broad applicability.

This review establishes a foundation for this doctoral study by identifying where generative techniques can meaningfully contribute to IoT security and where methodological gaps persist. Specifically, it highlights the need for IDS frameworks that are not only accurate but also adaptive, explainable, and efficient enough to be deployed at scale across diverse IoT environments.

3.2. Active learning and concept drift in ML-based IDS

Intrusion detection in IoT environments is challenged by dynamic data distributions, severe class imbalance, scarce labeled data, and resource constraints at the edge. These limitations make conventional, static ML-based IDS architectures inadequate for real-world deployment. Concept drift caused by evolving attacker behaviors, software updates, or shifting device patterns can degrade model performance over time, necessitating adaptive learning strategies (RQ1, RQ2).

Active learning offers a label-efficient alternative by querying only the most informative samples from incoming data, thereby reducing annotation effort while preserving accuracy. This approach has been explored in both batch-based and online learning settings, the latter processing data one instance at a time to accommodate environments with limited storage or computational resources [23]. When applied to non-stationary data streams, active learning frameworks can incorporate drift-aware retraining and ensemble methods to remain effective [7, 24]. Nonetheless, labeling the entire stream remains costly, reinforcing the utility of selective querying.

While active learning holds promise, it is not immune to adversarial risks. Attackers may inject malicious samples into the data pool, corrupt the labeling oracle, or craft adversarial examples that evade detection [25, 26]. These threats highlight the need for robust, trustworthy integration of active learning in IDS pipelines.

In terms of practical applications, active learning has been successfully employed in network intrusion detection to improve detection quality with minimal supervision [27]. This study leverages mobile malware detection as a proxy domain to explore active learning under non-stationary conditions—a challenge shared with many real-world IoT deployments Recent research has shown that dynamic

behavioral features (e.g., system calls)[28, 29] and hybrid representations combining static and dynamic data[30] enhance model performance in such settings. Studies that leverage human-in-the-loop active learning have also demonstrated improved detection of IoT-specific threats, including botnets, with significantly reduced data requirements [31].

Explainability is another important consideration. Active learning inherently supports human oversight by surfacing the most uncertain or ambiguous samples for labeling. This selective transparency not only improves model robustness but also facilitates trust and auditability in high-stakes applications, such as healthcare, industrial automation, and autonomous systems (RQ4).

In summary, active learning provides a promising avenue for addressing the limitations of traditional IDS in IoT environments. However, existing methods often lack comprehensive integration of drift detection, adversarial resilience, and explainability. This doctoral research aims to develop adaptive, label-efficient, and resource-aware intrusion detection frameworks for IoT environments, with a focus on robustness to evolving threats, concept drift, and deployment constraints. The solutions will be empirically validated and designed to be practically deployable in heterogeneous real-world contexts.

4. Results (preliminary)

Active learning-based intrusion detection for mobile malware

As the beginning of this doctoral research, we developed and evaluated an active learning-based malware detection framework aimed at improving adaptability and reducing labeling costs in machine learning-based intrusion detection systems (IDS) for dynamic environments. The framework was experimentally validated in the context of mobile malware detection, using real-world Android telemetry data as a testbed to simulate IoT-relevant operational conditions such as data imbalance, concept drift, and constrained labeling resources. The results presented in this section are drawn from our publication [9].

The study utilizes the KronoDroid dataset [32], which is labeled with timestamps and rich hybrid features (permissions and system calls). These timestamps enabled the simulation of a continuous, evolving data stream, divided into 44 chronological periods, thus allowing for realistic modeling of non-stationary malware behavior and drift in data distributions.

Our approach integrates a pool-based active learning mechanism with auto-labeling and drift-aware thresholding strategies. During each learning iteration, the model selectively queries the most uncertain samples for human annotation while automatically labeling high-confidence instances. A key innovation lies in dynamically adjusting the auto-labeling threshold across periods to balance labeling cost with detection performance. We also incorporated basic oversampling and undersampling strategies to address class imbalance during initial training.

Three training strategies were evaluated: (1) traditional batch learning with full supervision (upper-bound baseline), (2) active learning with uncertainty-based querying, and (3) active learning with random sampling (lower-bound baseline). We carried out the benchmark training across feature sets and balancing methods. In particular, our best-performing configuration, using hybrid features and undersampling, achieved a 97.4% F1 score and 96.8% accuracy while querying only 6.9% of the labeled data, indicating reduced annotation costs while maintaining comparable performance levels.

To further reduce labeling overhead, we experimented with static, time-dynamic, and iteration-dynamic auto-labeling thresholds. Among these, a polynomial function-based dynamic threshold offered the best trade-off between false label rate and detection accuracy. This approach achieved up to 97.9% F1 using only 2.37% of queried labels, a notable reduction in manual annotation compared to baseline active learning pipelines.

Table 1 displays results using static thresholds for model confidence.

While lower thresholds increase the number of auto-labeled samples, they also risk performance drops due to mislabeling.

Dynamic strategies adjusting the threshold across time periods are summarized in Table 2.

Descending thresholds improved performance while reducing mislabeled data, indicating their advantage in later-stage learning.

Table 1
Training results for different static threshold values (Hybrid, Undersampling)

Static Threshold	Label		F1(%)	Accuracy(%)	Auto-label	Miss-label
	Numbers	Proportion(%)			Numbers	Numbers
0.90	946	2.98	89.9	91.2	27550	972
0.92	1079	3.40	91.0	92.5	26392	698
0.95	1673	5.27	92.4	94.1	21825	419
0.97	1655	5.21	94.3	95.7	13790	55

Table 2
Training results for thresholds increasing or decreasing through time (Hybrid)

Balancing Method	Dynamic Threshold	Label		F1(%)	Accuracy(%)	Auto-label	Miss-label
		Numbers	Proportion(%)			Numbers	Numbers
Oversampling	Ascending	1439	4.53	92.0	93.3	27466	592
	Descending	1145	3.60	**94.3**	**95.1**	10306	**181**
Undersampling	Ascending	1405	4.60	92.6	94.0	24792	387
	Descending	927	3.00	90.9	94.0	18751	323

We also implemented the dynamic thresholds changing through iterations. The most effective threshold shaping was found via polynomial and linear iteration-based functions.

Table 3
Training results for thresholds changing through iterations by optimized shape (Hybrid)

Shapes	F1 (%)	Accuracy (%)	Label	
			Numbers	Proportion(%)
no auto-labeling	98.30	97.40	1334	4.27
fourth	97.20	96.00	502	1.61
fourth maxed	97.70	96.70	718	2.30
squared	97.30	96.10	531	1.70
lin 1	97.50	96.30	711	2.28
lin 2	97.70	96.60	716	2.29
lin 3	97.70	96.60	576	1.85
level	**97.97**	96.90	662	2.12
T_desc	97.40	96.30	897	2.87

According to Table 3, the *level* function achieves nearly identical F1 to the baseline, but with around 50% fewer queried labels.

To enhance robustness, we integrated a drift-aware auto-labeling strategy, which explored the integration of concept drift detection into the auto-labeling process. Using statistical monitoring to estimate drift magnitude, the system temporarily suspended auto-labeling when the drift exceeded a predefined threshold. Results are shown in Table 4.

Table 4
Auto-labeling driven by drift detection (Hybrid, *level*)

Drift Threshold	F1 (%)	Accuracy (%)	Label	
			Numbers	Proportion(%)
0.20	**97.8**	96.7	699	**2.24**
0.25	**97.9**	**96.9**	741	**2.37**
0.30	**97.7**	96.6	734	**2.35**
0.50	**97.8**	96.7	691	**2.21**

While this strategy did not significantly improve predictive accuracy beyond our already optimized thresholding mechanisms, it provided improved interpretability and stability across simulated drift scenarios.

These results provide empirical support for the working hypothesis that combining active learning with dynamic auto-labeling and drift awareness can enable cost-effective and adaptive intrusion detection in dynamic IoT environments. Overall, they demonstrate that a combined strategy of active learning,

automated labeling, and lightweight drift detection provides a scalable and resource-efficient foundation for ML-based IDS in dynamic, data-constrained environments. Although the current study is situated within the domain of Android malware detection, the underlying methodology is designed to generalize to broader IoT contexts, where similar constraints apply and adaptability is critical. These findings provide preliminary empirical support for the core hypothesis of this PhD research: that adaptive, label-efficient, and drift-aware ML models can significantly enhance the feasibility and effectiveness of intrusion detection in real-world IoT deployments.

5. Discussion and future work

This doctoral research has made initial strides toward developing machine learning-based intrusion detection systems (IDS) that are better suited to the evolving, heterogeneous, and resource-constrained nature of Internet of Things (IoT) environments. The work to date integrates both conceptual and empirical investigations that lay the groundwork for building adaptive and efficient IDS tailored for real-world IoT deployments.

The first major contribution comes from a systematic literature review focused on the use of generative artificial intelligence (AI) in intrusion detection for IoT environments. This review analyzed recent advances in applying models such as Generative Adversarial Networks (GANs) and Transformers for tasks like synthetic data generation, anomaly detection, and feature augmentation. It highlighted the growing reliance on generative methods to overcome limitations such as class imbalance and data scarcity, common challenges in IoT security. The review also underscored significant gaps in current research, particularly in areas such as model generalizability, real-world validation, and explainability. These findings have provided a comprehensive understanding of the state-of-the-art, directly informing the design choices and evaluation priorities in the subsequent empirical work.

Building on this foundation, the second contribution is an experimental framework for malware detection using active learning integrated with auto-labeling and concept drift detection. This approach addresses practical limitations identified in the review, especially the high cost of labeled data and the non-stationarity of real-world threat environments. Using a time-structured Android malware dataset to simulate evolving data streams, the system achieved competitive detection performance while drastically reducing the labeling effort. Dynamic thresholding, hybrid feature modeling, and uncertainty-based sampling strategies collectively contributed to improved adaptability and operational efficiency, key requirements for deployable IoT IDS. This study offers initial evidence that machine learning models, when combined with uncertainty sampling, adaptive thresholding, and drift awareness, can maintain effective detection performance over time with limited human supervision. These results are consistent with the central hypothesis of the research, although further validation remains necessary in real-world IoT scenarios.

Together, these two studies reflect the complementary nature of theoretical landscape mapping and practical method development. The literature review clarified where generative AI can be applied and where it remains underexplored, while the empirical study offered one such implementation of data-efficient, adaptive learning within a real-world security context.

5.1. Future work and milestones

Looking ahead, several future directions are envisioned to extend the contributions of the current research. These directions are categorized by priority and feasibility within the remaining PhD timeframe (until Q1 2027), and each is linked to specific research questions (RQs). A milestone-based plan and risk assessment are also provided to ensure a realistic and focused trajectory.
Core directions which has high priority and are feasible:

Federated learning in IoT contexts (RQ2, RQ4). The experimental framework will be expanded to distributed IoT domains such as smart home and industrial systems using federated learning [10]. These environments introduce challenges in terms of limited resources, communication constraints,

and device heterogeneity. This work will be integrated into the applied research cycle through iterative experimentation and evaluation.

Generative AI for data augmentation (RQ2, RQ3). While the current work focused on active learning, planned efforts include applying GANs or diffusion models to synthesize attack traffic for underrepresented classes. Pretrained Transformer architectures will also be used for feature representation in heterogeneous IoT data. These generative methods will be evaluated as part of the extended development phase, targeting data-scarce or adversarial scenarios.

Online and continual learning (RQ1, RQ3). Batch-based models will be extended to stream-processing settings using online learning [33]. This includes integrating active learning, auto-labeling, and drift detection into a unified real-time learning loop. Preliminary prototypes will be developed in simulated environments using partitioned data streams.

There are some secondary or optional directions, which can be done if time permits:

Explainability. Model-agnostic interpretability tools (e.g., SHAP, LIME) will be evaluated to generate human-readable justifications for model decisions. This is a longer-term direction and will be pursued if core experiments are completed ahead of schedule.

Milestones are set with risk analysis and mitigation strategies:

Q2–Q3 2025: Finalize Android-based baseline and submit paper on active learning + drift (RQ1, RQ3).
Q4 2025: Conduct cross-domain generalization experiments on IoT datasets (RQ2).
Q1 2026: Integrate generative models for augmentation; evaluate on class imbalance tasks (RQ2, RQ3).
Q2 2026: Develop federated IDS prototype and some simulations (RQ2, RQ4).
Q3–Q4 2026: Deployment and explore online learning loop (RQ1, RQ3, RQ4).
Q1 2027: Final evaluation, thesis writing.

If Generative AI models such as GANs or Transformers fail to generalize, traditional augmentation (e.g., SMOTE + ensemble learning) will be used. Optional directions (e.g., explainability) will only be pursued if core milestones are achieved early.

The future work plan is tightly integrated with the applied research methodology and structured to ensure a realistic path to completion. By prioritizing core objectives and preparing for foreseeable challenges, this research aims to deliver adaptive, scalable, and resource-aware IDS frameworks deployable across real-world IoT environments.

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Dr. Ants Torim, Prof. Dr. Sadok Ben Yahia, and Prof. Dr. Hayretdin Bahsi, for their invaluable guidance, support, and constructive feedback throughout this research. Their expertise and encouragement have been vital to the development of my PhD work. I also thank the Department of Software Science at Tallinn University of Technology for supporting this research through the doctoral study program, and special thanks to Dr. Gunnar Piho.

Declaration on Generative AI

During the preparation of this work, the author used *Grammarly* to check spelling and grammar only. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Computer Networks 54 (2010) 2787–2805. doi:10.1016/j.comnet.2010.05.010.

[2] Statista Research Department, Number of connected iot devices worldwide 2019–2030, https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/, 2023. Accessed: 2024-12-10.

[3] F. Hussain, G. Sarwar, S. U. Khan, A review of performance metrics for machine learning-based ids in iot networks, IEEE Access 8 (2020) 71564–71581. doi:10.1109/ACCESS.2020.2988028.

[4] J. Singh, G. S. Aujla, N. Kumar, A survey on machine learning-based intrusion detection systems for iot networks, Computer Communications 164 (2020) 114–135. doi:10.1016/j.comcom.2020.10.002.

[5] R. Sommer, V. Paxson, Outside the closed world: On using machine learning for network intrusion detection, IEEE Symposium on Security and Privacy (2010) 305–316. doi:10.1109/SP.2010.25.

[6] R. Mohammad, S. K. Datta, Handling concept drift in iot security with adaptive learning models, Journal of Network and Computer Applications 186 (2021) 103082. doi:10.1016/j.jnca.2021.103082.

[7] W. Liu, H. Zhang, Z. Ding, Q. Liu, C. Zhu, A comprehensive active learning method for multiclass imbalanced data streams with concept drift, Knowledge-Based Systems 215 (2021) 106778. doi:10.1016/j.knosys.2021.106778.

[8] Z. Deng, A. Torim, S. Ben Yahia, H. Bahsi, Generative ai in intrusion detection systems for internet of things: A systematic literature review, IEEE Open Journal of the Communications Society 6 (2025) 4689–4717. doi:10.1109/OJCOMS.2025.3573194.

[9] Z. Deng, A. Hubert, S. Ben Yahia, H. Bahsi, Active learning-based mobile malware detection utilizing auto-labeling and data drift detection, in: 2024 IEEE International Conference on Cyber Security and Resilience (CSR), 2024, pp. 146–151. doi:10.1109/CSR61664.2024.10679343.

[10] S. Sharma, D. Puthal, Federated learning for intrusion detection in iot: Concepts, applications, and challenges, IEEE Transactions on Industrial Informatics 18 (2022) 4711–4721. doi:10.1109/TII.2021.3123014.

[11] Y. Li, J. Wang, H. Zhu, Lightweight transformer-based model for real-time intrusion detection in iot networks, IEEE Internet of Things Journal 10 (2023) 6893–6905. doi:10.1109/JIOT.2023.3255912.

[12] A. Everitt, P. Hardiker, J. Littlewood, A. Mullender, Applied Research for Better Practice, Bloomsbury Publishing, 1992.

[13] N. Moustafa, J. Slay, Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set), Military Communications and Information Systems Conference (MilCIS) (2015) 1–6. doi:10.1109/MilCIS.2015.7348942.

[14] M. Alsheikh, V. Sivaraman, H. Haddadi, D. K. Kanhere, Edge-iiotset: A new industrial iot dataset for edge intrusion detection, Computers & Security 129 (2023) 102894. doi:10.1016/j.cose.2023.102894.

[15] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele Univ. 33 (2004).

[16] V. Belenko, V. Chernenko, M. Kalinin, V. Krundyshev, Evaluation of GAN applicability for intrusion detection in self-organizing networks of cyber physical systems, in: 2018 International Russian Automation Conference, RusAutoCon 2018, 2018. doi:10.1109/RUSAUTOCON.2018.8501783.

[17] S. Kim, S. Yoon, S. Yoon, Wasserstein gan-based synthetic sample generation to improve intrusion detection performance, IEEE Access 8 (2020) 21966–21975. doi:10.1109/ACCESS.2020.2969295.

[18] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, in: NeurIPS, 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/254edc18f30c5aceee5bff6c5d430b67-Paper.pdf.

[19] G. Mohi-ud din, Nsl-kdd, 2018. URL: https://dx.doi.org/10.21227/425a-3e55. doi:10.21227/425a-3e55.

[20] N. Koroniotis, N. Moustafa, E. Sitnikova, B. Turnbull, Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset, 2018. arXiv:1811.00701.

[21] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in: International Conference on Information Systems Security and Privacy, 2018. URL: https://api.semanticscholar.org/CorpusID:4707749.

[22] M. Piskozub, J. Kremp, W. Mazurczyk, Explainable intrusion detection with transformers: A case study in iot, Journal of Cybersecurity and Privacy 3 (2023) 152–172. doi:10.3390/jcp3010010.

[23] S. C. H. Hoi, D. Sahoo, J. Lu, P. Zhao, Online learning: A comprehensive survey, 2018. arXiv:1802.02871.

[24] A. Guerra-Manzanares, H. Bahsi, On the application of active learning to handle data evolution in android malware detection, in: International Conference on Digital Forensics and Cyber Crime, Springer, 2022, pp. 256–273.

[25] B. Miller, A. Kantchelian, S. Afroz, R. Bachwani, E. Dauber, L. Huang, M. Tschantz, A. Joseph, J.D.Tygar, Adversarial active learning, volume 2014, 2014. doi:10.1145/2666652.2666656.

[26] N. Wang, Y. Chen, Y. Hu, W. Lou, Y. T. Hou, Manda: On adversarial example detection for network intrusion detection system, in: IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, IEEE Press, 2021, p. 1–10. URL: https://doi.org/10.1109/INFOCOM42981.2021.9488874. doi:10.1109/INFOCOM42981.2021.9488874.

[27] A. Ziai, Active learning for network intrusion detection, 2019. arXiv:1904.01555.

[28] X. Xiao, S. Zhang, F. Mercaldo, G. Hu, A. Kumar, Android malware detection based on system call sequences and lstm, Multimedia Tools and Applications 78 (2019) 1–21. doi:10.1007/s11042-017-5104-0.

[29] V. P., A. Zemmari, M. Conti, A machine learning based approach to detect malicious android apps using discriminant system calls, Future Generation Computer Systems 94 (2019) 333–350. URL: https://www.sciencedirect.com/science/article/pii/S0167739X18306216. doi:https://doi.org/10.1016/j.future.2018.11.021.

[30] Y.-C. Shyong, T.-H. Jeng, Y.-M. Chen, Combining static permissions and dynamic packet analysis to improve android malware detection, 2020, pp. 75–81. doi:10.1109/ICCCI49374.2020.9145994.

[31] A. Guerra-Manzanares, H. Bahsi, On the application of active learning to handle data evolution in android malware detection, in: International Conference on Digital Forensics and Cyber Crime, Springer, 2022, pp. 256–273.

[32] A. Guerra-Manzanares, H. Bahsi, S. Nõmm, Kronodroid: Time-based hybrid-featured dataset for effective android malware detection and characterization, Computers & Security 110 (2021) 102399.

[33] K. Wang, Y. Jia, Z. Tian, Streamal: Online active learning for evolving network intrusion detection, Expert Systems with Applications 188 (2022) 116012. doi:10.1016/j.eswa.2021.116012.