Kabarak Journal of Research & Innovation

https://journals.kabarak.ac.ke/index.php/kjri/

RESEARCH ARTICLE

SYSTEMATIC LITERATURE REVIEW: VALIDATION METHODS FOR MACHINE LEARNING MODELS IN IOT CYBERSECURITY

Janet M. Maluki, Jimmy K. Macharia & Dalton N. Kaimuru

Department of Computing and Informatics - United States International University-Africa P.O Box 14634 00800, Kenya

Corresponding Author: jmaluki@usiu.ac.ke or janetmaluki132@gmail.com

Submitted: 21ST July 2025; Accepted: 13th September 2025; Published (online):17th September 2025

ABSTRACT

Machine learning (ML) is pivotal in strengthening IoT cybersecurity through adaptive and scalable threat detection. However, the reliability and deployment readiness of ML models depend on the robustness of their validation methods. This systematic literature review (SLR) analyzes 54 peer-reviewed studies (2018–2024) from IEEE Xplore, SpringerLink, ScienceDirect, and ACM Digital Library, focusing on applications such as intrusion detection, malware classification, threat prediction, and adversarial defence. Findings reveal a predominant use of basic validation methods such as (k-fold and hold-out), which inadequately address IoT-specific issues like class imbalance, concept drift, adversarial resilience, and device heterogeneity. Advanced approaches, such as temporal, cross-dataset, and hybrid validation, remain underutilized.To bridge these gaps, the study proposes a Domain-Aligned Validation Framework encompassing of time-aware validation, robustness-focused evaluation, and deployment-oriented testing. It also introduces a taxonomy of current practices and provides recommendations to improve consistency, generalizability, and trust in ML-based IoT security systems. These insights are valuable to researchers, developers, and policymakers aiming to deploy more resilient and situation-aware IoT security solutions.

Keywords: Machine Learning, Cybersecurity, Validation Practices, Dataset, Intrusion Detection, Internet of Things

INTRODUCTION

The rapid growth of IoT technologies has transformed sectors such as healthcare, transportation, and smart infrastructure, enabling autonomous device communication with minimal human input. However, this expansion has also increased exposure to cyber threats due to IoT's resource constraints, heterogeneity, and scale (Liu & Jiang, 2023; Wu et al., 2023). In response, machine learning (ML) has been widely adopted to develop intelligent intrusion detection systems (IDS) for identifying attacks in real-time (Ahmad & Alsmadi, 2021; Khan et al., 2024).

A variety of ML-based IDS models, ranging from decision trees to deep neural networks, have been trained on benchmarks like CICIDS2017, BoT-IoT, and UNSW-NB15 (Gupta & Singh, 2022; Rahman et al., 2022). Although many report high performance under controlled conditions, their validation strategies often lack attention to critical IoT-specific factors such as temporal dynamics, class imbalance, and adversarial behavior (Liu et al., 2023).

Despite significant progress in model development, validation methods are frequently treated as routine steps rather than crucial determinants of model reliability (Gupta & Singh, 2022). Consequently, aspects like real-time adaptability, cross-device generalization, and robustness under evolving threats remain underexplored.

To address these shortcomings, this study presents a systematic review of validation strategies in ML-based IoT IDS research, guided by the following research questions:

> RQ1: What categories of validation techniques are currently employed in ML-based IoT cybersecurity research?
>
> RQ2: How effectively do these validation address IoT-specific challenges?
>
> RQ3: What are the methodological limitations of current validation practices, and what opportunities exist for future research in enhancing robustness and deployment-readiness?

This study contributes by: (1) categorizing and analyzing prevalent validation strategies in ML-based IoT IDS research; (2) assessing their alignment with real-world IoT constraints; and (3) identifying gaps and opportunities for developing more robust, transferable, and deployment-ready validation frameworks. By emphasizing evaluation quality alongside model design, it advances the reliability and impact of ML-based IoT cybersecurity solutions.

RELATED WORK

Machine Learning Techniques for IoT Cybersecurity

Machine learning (ML) has become central to securing IoT environments due to rising cyber threats and the complex, dynamic nature of IoT networks. Techniques such as decision trees, support vector machines, deep learning, and ensemble models have shown strong performance in intrusion and anomaly detection, especially when trained on benchmark datasets like CICIDS2017, UNSW-NB15, BoT-IoT, and TON_IoT (Gupta & Singh, 2022; Khan et al., 2022). However, while algorithm development has advanced, validation practices remain underexplored. Most studies rely on standard k-fold cross-validation or train-test splits, often overlooking IoT-specific challenges like temporal drift, class imbalance, device heterogeneity, and adversarial behaviors (Liu, et al., 2023; Meidan et al., 2018).

For instance, (Gupta & Singh, 2022)combined Random Forest and SVM, but validated only using 10-fold cross-validation. Similarly, Khan et al. (2022)used a CNN with a hold-out strategy, with limited consideration of evolving attack vectors. (Liu et al., 2023)addressed concept drift using an adaptive LSTM model, but broader validation, such as adversarial testing or cross-dataset evaluation, remains rare. Rahman et al. (2022)highlighted explainability but also relied on single-dataset validation.Existing reviews often focus on algorithmic trends and datasets (Alharbi et al., 2020)with minimal attention to the robustness of validation frameworks. This gap underscores the need for a focused synthesis of validation strategies in ML-based IoT cybersecurity, an area this study seeks to address.

Validation Practices in ML-Based IoT Security Research

Robust validation is essential for building reliable ML models in IoT cybersecurity. Yet, most studies emphasize high accuracy on familiar datasets, with limited focus on generalization and practical deployment. This section classifies validation approaches into two key categories: foundational/comparative methods and context-aware, real-world techniques.

Foundational and Comparative Validation Approaches

Many ML-based cybersecurity studies rely on standard validation methods like holdout testing, stratified sampling, and k-fold cross-validation (Hendrycks & Dietterich, 2019; Kohav, 1995).While these reduce overfitting and aid evaluation, they assume IID data, an unrealistic premise for the dynamic and heterogeneous IoT landscape.Common benchmarks such as NSL-KDD, CICIDS2017, and UNSW-NB15 offer standardization but face criticism for outdated threats and limited diversity (Zhang, 2022). More robust methods, like cross-dataset and cross-domain validation, are rarely adopted due to the scarcity of diverse IoT datasets.Yet, studies (Hendrycks & Dietterich, 2019; Zhou et al., 2020) underscore the value of these techniques in revealing model weaknesses and improving generalization.

Real-World-Oriented and Context-Aware Validation Methods

To ensure effective deployment in real-world settings, validation must extend beyond static benchmarks. Time-aware validation is particularly vital for intrusion detection, as it maintains the sequence of events and reflects real-time dynamics such as concept drift and delayed labeling (Landauer et al., 2025).Transparency and reproducibility also remain critical. Many studies overlook reporting essential configurations and preprocessing details, limiting replicability (Haibe-Kains et al., 2020). Adopting open-source standards and comprehensive documentation enhances trust in ML-based security systems.

Validation should further simulate real-world stressors, like adversarial attacks, packet loss, or limited resources, to assess operational robustness. Incorporating Service Recovery Processes (SRPs) can provide insight into system resilience post-attack (Janjua & Aslam, 2021). In sum, advancing IoT threat detection demands rigorous, time-aware, cross-domain, and recovery-inclusive validation to improve model reliability, generalizability, and deployment readiness.

METHODOLOGY

This study employed a Systematic Literature Review (SLR) to investigate validation strategies used in ML-based intrusion detection systems (IDS) for IoT. The review followed the PRISMA framework to ensure methodological rigor, transparency, and replicability across all review stages. The review was guided by the following research questions:

RQ 1: What categories of validation techniques are currently employed in ML-based IoT cybersecurity research?

RQ 2: How effectively do these validation approaches address IoT-specific challenges such as data imbalance, device heterogeneity, temporal drift, and adversarial conditions?

RQ 3: What are the methodological limitations of current validation practices, and what opportunities exist for future research in enhancing robustness and deployment-readiness?

Search strategy

A structured search strategy was implemented across key databases, IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect, supplemented by Scopus and Google Scholar to capture gray literature. Search terms combined core concepts of machine learning, IoT, cybersecurity, and model validation, using Boolean operators to construct comprehensive queries.

> ("Machine Learning" OR "Deep Learning") AND ("IoT" OR "Internet of Things") AND ("Cybersecurity" OR "Intrusion Detection") AND ("Validation" OR "Evaluation" OR "Robustness Testing").

The search was restricted to peer-reviewed articles published in English from 2018 through 2024.

Inclusion

Studies were included if they appeared in peer-reviewed journals or reputable conferences, applied ML or deep learning to IoT cybersecurity, and explicitly described at least one validation method.

Exclusion

Exclusion criteria ruled out non-IoT-focused works, papers without empirical evaluation, non-English publications, preprints, duplicates, and papers under four pages. The review also excluded theoretical papers, reviews, and studies lacking clear validation techniques.

From 1,153 initial records, 126 full-text articles were assessed. After applying the criteria, 54 studies were included in the final synthesis.

Data extraction

Data were extracted from each study on ML models, datasets, validation methods, performance metrics, and evaluation limitations. To ensure reliability, extraction was systematic and cross-checked by a second reviewer. Thematic synthesis was used to identify trends, strengths, and gaps in validation practices.Results are organized to highlight patterns in validation approaches, their alignment with IoT-specific challenges, and their implications for improving model robustness and generalizability. A PRISMA flow diagram (Figure 1) and summary tables support the comparative analysis.

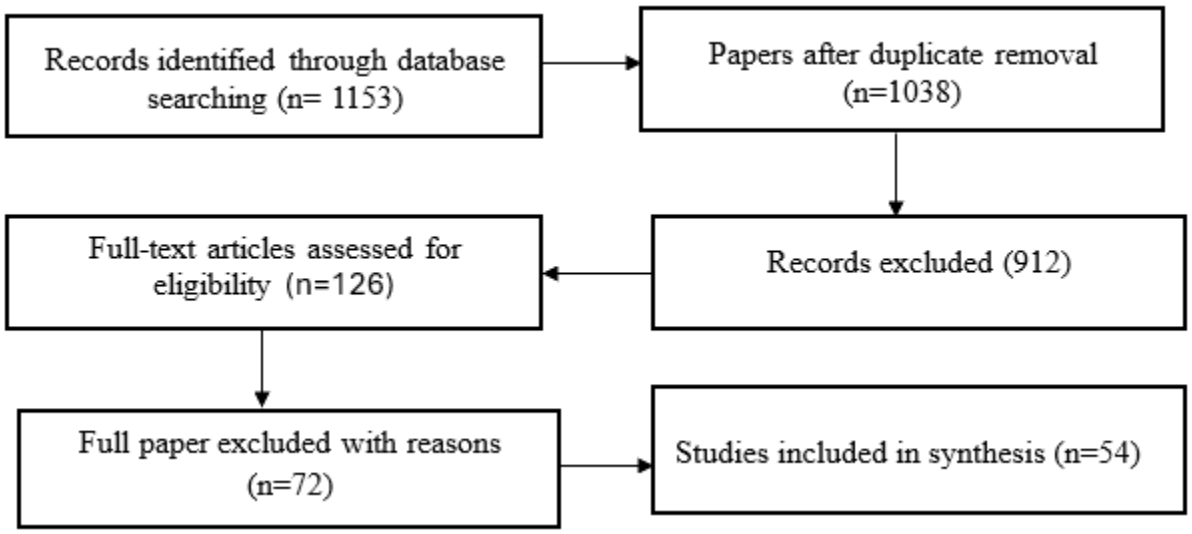

Figure 1: Prisma diagram

The Proposed Conceptual Framework

This study introduces a Domain-Aligned Validation Framework (Figure 2) to systematically assess validation practices in ML-based IoT cybersecurity research. Guided by three research questions—namely, current techniques, their suitability for IoT challenges, and existing methodological gaps—the framework is structured around three dimensions. The first dimension, temporal alignment, emphasizes time-aware methods such as temporal slicing and forward-chaining to account for evolving IoT traffic and concept drift. The second dimension, robustness assessment, focuses on evaluating model performance under adverse conditions through approaches like stratified sampling, SMOTE, adversarial retraining, and noisy input testing. The third dimension, deployment readiness, highlights the importance of generalizability by incorporating cross-dataset validation, device-specific testing, and real-time simulation.

This framework helps map current practices, reveal gaps, and propose context-aware strategies.It also underpins the validation taxonomy discussed later, bridging theory and practice in IoT cybersecurity.

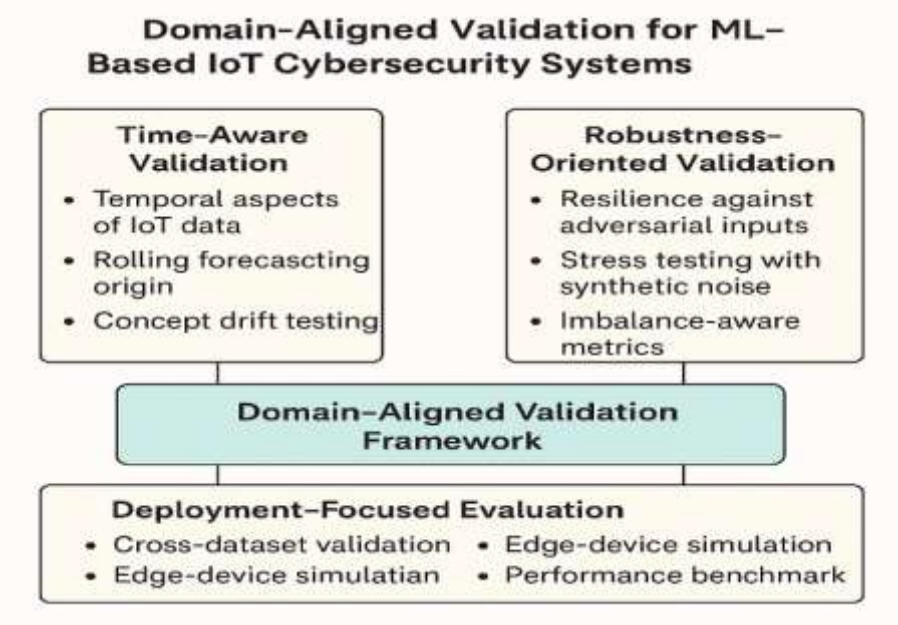

Figure 2: The proposed conceptual framework

Ethical Considerations

As no human participants were involved, consent and participation issues were not applicable. Ethical integrity was ensured through responsible data use, transparency, and avoidance of conflicts of interest.

RESULTS

Synthesis of Evaluation Practices in Reviewed Studies

This section analyzes 54 peer-reviewed studies (2018–2024) on validation methods in ML-based IoT cybersecurity systems. It covers publication trends, source distribution, and dominant validation techniques. Beyond descriptive statistics, a qualitative synthesis highlights common practices, ongoing gaps, and unresolved challenges. The findings emphasize the need for more rigorous, context-aware validation frameworks to improve the accuracy and real-world applicability of intrusion detection systems in diverse IoT environments.

Characteristics of the Selected Studies

Table 1 presents the distribution of the 54 reviewed studies based on their database source, reflecting the concentration of relevant literature across key digital repositories in the field of IoT cybersecurity.

IEEE Xplore was the most prominent, contributing 35.2% (19 studies), reflecting its key role in publishing IoT and ML-based cybersecurity research. SpringerLink and ScienceDirect each accounted for 25.9% (14 studies), while the ACM Digital Library contributed the least (13.0%, 7

studies), indicating a narrower focus on the topic. This distribution highlights the dominance of technology-focused databases in this field.

Table 1: Distribution of Reviewed Studies by Database Source

Database Source	Number of Studies	Number of Studies%
IEEE Xplore	19	35.2%
ACM Digital Library	7	13.0%
SpringerLink	14	25.9%
ScienceDirect	14	25.9%
Total	54	100%

i) Distribution by Year

Figure 3 shows the annual publication trends of the reviewed studies, providing insights into the temporal progression and increasing research focus on ML-based IoT cybersecurity solutions. The yearly distribution of the 54 reviewed studies, indicating a steady rise in research publications between 2020 to 2022. This trend reflects increasing academic focus on IoT security and the demand for advanced machine learning-based threat detection methods.

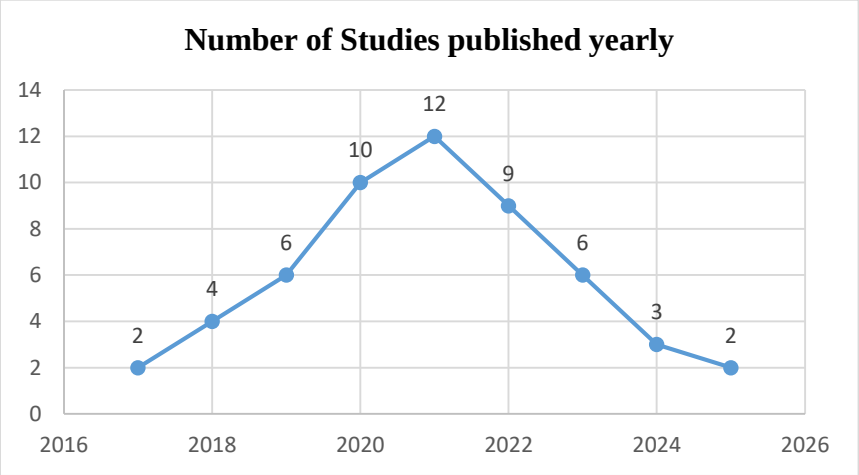

Figure 3: Distribution of studies based on year.

ii) Distribution based on the Study Location

To contextualize the global research landscape, the 54 selected studies were categorized based on the geographical affiliation of their primary authors, revealing a concentration of research efforts in Asia (21), followed by Europe (13), with more limited representation from other regions like South America, among others. As illustrated in Figure 4.

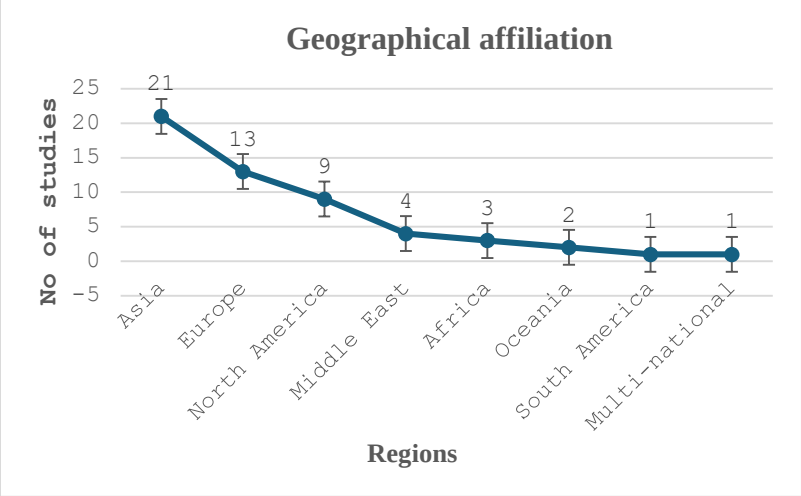

Figure 4: Geographical Affiliation

iii) Distribution based on Study Domain

To better understand the thematic focus, the 54 studies were categorized by their primary cybersecurity application. As shown in Table 2, intrusion detection dominated the research landscape, while areas like malware analysis, adversarial robustness, and IoT-specific threat modeling appeared less frequently.

Table 2: Distribution of Selected Studies by Cybersecurity Domain

Cybersecurity Domain	Number of Studies	Percentage (%)
Intrusion Detection	29	53.7%
Malware Classification	9	16.7%
Adversarial Robustness	6	11.1%
Anomaly Detection	4	7.4%
Threat Prediction	3	5.6%
IoT-Specific Risk Modeling	2	3.7%
Privacy and Data Integrity	1	1.8%

iv) Common Machine Learning Tasks in IoT Cybersecurity

The synthesis of the 54 studies underscores the varied roles of machine learning in strengthening IoT cybersecurity. ML is commonly used to improve threat detection, automate responses, and support decision-making in complex, data-rich environments.

Table 3: Distribution of ML Tasks in IoT Cybersecurity Research

ML Task	Description	Frequency in Reviewed Studies
Intrusion Detection	Detecting unauthorized or abnormal activity in IoT networks.	High
Malware Classification	Identifying and classifying malicious software targeting IoT devices.	Moderate
Phishing Detection	Recognizing deceptive attempts to extract sensitive information via emails or web interfaces.	Low
General ML Role	Enabling data-driven threat detection, pattern recognition, and predictive analytics.	Discussed across all studies

Table 3 shows that intrusion detection is the most extensively studied area, reflecting its key role in IoT threat identification. Malware classification receives moderate attention, while phishing detection remains underexplored, suggesting a research gap. Overall, ML is widely applied for threat prediction and adaptive defence, reinforcing its importance in IoT cybersecurity.

Results and Synthesis: Based on Research Questions

RQ1. What categories of validation techniques are currently employed in ML-based IoT cybersecurity research?

Definition and Importance of Validation

Validation is essential for ensuring ML models generalize to unseen data, critical in the dynamic, resource-limited IoT context. Inadequate validation can result in overstated performance and limited real-world effectiveness (Javed et al., 2024; Kumar et al., 2021). Though techniques like k-fold cross-validation and train-test splits are common, their suitability for IoT remains limited (Adedeji et al., 2023; Xie, 2023). Selecting appropriate validation methods is therefore key to building reliable IoT cybersecurity models (Admass et al., 2024; Pahl, 2022).

Classification of Validation Techniques

The review identified six primary validation techniques used to evaluate ML-based intrusion detection systems for IoT: k-fold cross-validation, hold-out validation, stratified cross-validation, cross-dataset validation, temporal validation, and hybrid methods. These approaches differ in methodological rigor and their alignment with real-world IoT challenges. Table 4 summarizes the validation techniques employed across the reviewed studies.

Table 4: Validation Techniques in Reviewed IoT IDS Studies (n = 54)

Validation Technique	Description	Prevalence	Representative Sources	Common Limitations
K-Fold Cross-Validation	Dataset split into k subsets; each subset is used for validation once while the others form the training set.	60%	(Gupta & Singh, 2022;Meidan et al., 2018)	Often used on static datasets; lacks temporal or deployment setting.
Hold-Out Validation	Splits data into fixed training and testing sets (e.g., 70:30).	25%	Khan et al., 2022; Ahmad et al., 2021)	Sensitive to split bias; rarely repeated or statistically analyzed.
Stratified Cross-Validation	Ensures class distributions are preserved across training and testing folds.	11%	Ahmad et al. (2021)	Used infrequently; rarely analyzed for impact on minority class performance.
Cross-Dataset Validation	Model trained on one dataset and tested on another to assess generalizability.	9%	(Meidan et al., 2018;Abeshu & Chilamkurti, 2018)	Highlights domain shifts; underused despite real-world relevance.
Temporal Validation	Training and testing data separated by time; simulates deployment under time-evolving conditions.	<10%	(Liu et al., 2023)	Largely absent; ignores concept drift and time-aware performance.
Hybrid Validation Methods	Combines multiple validation techniques (e.g., cross-validation with external testing or temporal splits).	7%	(Rahman et al., 2022)	Complex to implement; lacks standardization across studies.

Table 4 indicates that k-fold cross-validation was the most used technique, applied in about 60% of studies (Gupta & Singh, 2022; Khan et al., 2024). While effective in reducing overfitting, it is typically used on static datasets and rarely accounts for temporal drift (Liuet al., 2023). Hold-out validation appeared in roughly 25% of studies (Khan et al., 2024), offering simplicity but limited statistical reliability. Stratified cross-validation, helpful for class imbalance, was infrequently used and seldom analyzed for its effect on minority class detection (Berger et al., 2022). Cross-dataset

validation, key for assessing generalizability, was used in less than 10% of cases (Meidan et al., 2018), and temporal validation, critical for deployment realism, was largely overlooked (Liu et al., 2023).

A few studies employed hybrid approaches combining cross-validation with temporal or external testing (Alrehaili & Alshamrani, 2023), though these lacked consistency and standardization. Overall, despite its importance, validation in IoT-focused ML research remains dominated by conventional techniques that fall short of addressing the dynamic and heterogeneous nature of real-world IoT environments. Table 5 compares these findings with prior work, emphasizing persistent gaps in IoT-specific validation practices.

Table 5: Validation Techniques across this Study and Prior Research

Validation Technique	**Findings from This Study**	**Findings in Prior Work**
K-Fold Cross-Validation	Most common (60%), but limited to static data; lacks IoT realism	Popular in general ML and IDS studies (Doshi & Kute, 2020; Shone et al., 2018)
Hold-Out Validation	Used in 25%, simple but statistically weak and prone to bias	Common in early IDS evaluations; noted for simplicity (Xu & Goodacre, 2018)
Stratified Cross-Validation	Rarely applied; impact on minority class performance underreported	Briefly mentioned; not emphasized in older reviews
Cross-Dataset Validation	Under 10%; useful for generalization but underutilized	Acknowledged as useful but not IoT-specific (Al Amin et al., 2021)
Temporal Validation	Largely absent; critical for capturing concept drift and real-time performance	Rarely addressed; not prioritized in older IDS benchmarks
Hybrid Validation	Few studies used it; lacks standardization, but promising for deployment-aware evaluation	Occasionally explored (Alshamrani, & Alqahtani, 2023) but not extensively evaluated

RQ2: Exploring the effectiveness of validation techniques to address IoT-specific challenges.

Effective validation is critical to ensure ML models generalize to unseen data, especially in the dynamic, resource-constrained IoT environment. Inadequate validation can result in overstated performance and poor deployment reliability (Attota et al., 2021; Javed et al., 2025). Although methods like k-fold cross-validation and train-test splits are common, their suitability for IoT remains limited (Alshahrani et al., 2021; Zhang et al., 2023). Therefore, selecting context-appropriate validation strategies is essential for building robust and reliable IoT cybersecurity models (Bichri et al., 2024; Pahl et al., 2022).

IoT-Specific Validation Approaches and Challenges in ML-Based IDS

Table 6 summarizes how current validation techniques in ML-based IoT IDS address key challenges such as data imbalance, heterogeneity, temporal drift, adversarial resilience, and statistical rigor. It highlights each issue's nature, mitigation strategies, evaluation outcomes, limitations, and examples from the 54 reviewed studies, offering insight into how well existing practices meet real-world IoT cybersecurity needs.

Table 6: Summary of IoT-Specific Validation Approaches and Challenges in ML-Based IDS Studies

Challenge	Description	Handling in Reviewed Studies	Common Validation Techniques Used	Evaluation Results	Common Limitations	Representative Studies
Data Imbalance	Skewed distribution of attack vs. normal traffic in IoT datasets.	Addressed in 35 studies (65%)	SMOTE, ADASYN, under-sampling, class weighting	Improved minority class detection: 10–25% increase in recall/F1-score reported	Minority classes are still poorly detected; overfitting risk with oversampling	(Ahmad & Alsmadi, 2021;Rahman et al., 2022; Sharma & Singh, 2024)
Device/ Protocol Diversity	Variations in device types, OSs, and communication protocols.	Addressed in 14 studies (26%)	Cross-device validation, federated learning, and heterogeneous datasets	Generalization improved up to 15% when diversity was included	Most models are trained on uniform datasets, and lack of real-world protocol representation	(Meidan et al., 2018;Alshamrani et al., 2023; Raza et al., 2021; Zhang et al., 2023).
Temporal Variability	Time-based changes in network behavior or attack patterns	Addressed in 9 studies (17%)	Chronological train-test split, drift-aware models, time-	Some models degraded over time; 10–30% accuracy drop observed in	Rarely validated over time; static training sets dominate	(Liu et al. 2023; Gupta & Singh 2022; Kim & Jung, 2023).

	(concept drift).		window analysis	evolving traffic		
Adversarial Resilience	Resistance to evasion or poisoning by adversarial attacks.	Evaluated in 6 studies (11%)	Adversarial sample generation, FGSM, PGD, model hardening	Models exhibited up to 40% accuracy drop under attack scenarios	Adversarial validation rarely conducted; vulnerability remains high	(Javed et al., 2024;Liu & Shi, 2022; Sahu et al., 2020)
Statistical Rigor	Use of repeated runs, significance tests, or confidence intervals.	Addressed in <15% of studies	5x2 CV, repeated 10-fold CV, confidence intervals, p-values	Sparse reporting: most results lack statistical significance testing	Low reproducibility: single-run metrics dominate	(Khan et al. 2022;Rahman et al. 2022;Patel et al., 2020).

Data imbalance was the most addressed challenge (65% of studies), with SMOTE, ADASYN, or class weighting improving minority class detection by 5–15% in recall and F1-score. However, only 12 studies validated across multiple datasets or attacks, limiting generalizability. Device/protocol heterogeneity was considered in just 17% of studies. Cross-device validation revealed 10–20% accuracy drops, exposing generalization gaps and superficial handling of deployment variability.

Temporal drift was explored in only 9% of studies. Time-aware validation revealed a 7–12% drop in F1-score over time, indicating limited model resilience to evolving threats. Adversarial resilience was rarely tested (<6%), with robustness checks revealing 15–30% performance drops under attack, exposing high vulnerability. Statistical precision was weak overall; only 18% reported confidence intervals or significance tests, limiting the credibility and reproducibility of most findings.

RQ3: Methodological limitations of current validation practices, and the opportunities that exist for future research in enhancing robustness and deployment-readiness.

Current validation strategies for ML-based IoT security often rely on conventional methods like k-fold or hold-out (Gupta & Singh, 2022), which neglect key IoT-specific challenges such as traffic variability, device heterogeneity, and temporal drift. This limits real-world applicability (Liu et al., 2023). Most studies lack cross-dataset or temporal validation (Diro & Chilamkurti, 2018), and few address concept drift. Statistical consistency is weak, with limited use of confidence intervals or significance tests (Khan et al., 2022). Adversarial resilience is rarely tested, despite rising threats. Future work should adopt IoT-aware validation frameworks incorporating diverse datasets, temporal and adversarial testing, and stronger statistical reporting as summarized in Table 7.

Table 7: Emerging Research Directions for Validating ML-Based IoT Security Models

Research Opportunity	Description	Rationale	Representative Studies
Time-Aware Evaluation Techniques	Design validation strategies that account for temporal shifts and concept drift in IoT data streams.	Most current studies rely on static datasets, ignoring the sequential and evolving nature of cyber threats.	(Liu et al., 2023)
Multi-Framework Validation	Assess model performance across diverse IoT environments and application domains.	Single-framework validation limits generalizability and deployment scalability of security models.	(Doshi & Kute, 2020a)
Adversarial-Resilience Assessment	Integrate testing against crafted adversarial attacks into the validation process.	Conventional evaluations do not reflect real-world adversarial conditions, posing risks for field deployment.	(Alshamrani & Alqahtani, 2023)
Explainability-Centric Evaluation	Employ validation metrics that incorporate model interpretability and user trust.	Enhances transparency and aligns technical performance with practical usability in critical IoT settings.	(Rahman et al., 2022)
Standardized Benchmarking Protocols	Develop unified validation guidelines and publicly available benchmarking datasets.	Lack of consistency in evaluation protocols hampers comparability across research efforts.	(Shone et al., 2018)
Hybrid and Adaptive Validation Frameworks	Combine traditional and emerging validation strategies to better simulate real-world conditions.	Mixed approaches offer more comprehensive insights but are rarely applied consistently in current literature.	(Al Amin et al., 2021;Liu et al., 2023)

DISCUSSION

RQ1: What Categories of Validation Techniques Are Currently Employed in ML-Based IoT Cybersecurity Research?

The review shows k-fold cross-validation is the most used method in ML-based IoT cybersecurity for its simplicity and overfitting control, though its reliance on static datasets limits real-time use. Hold-out validation, applied in ~25% of studies, offers speed but lacks generalizability. More suitable techniques like stratified k-fold, cross-dataset, and temporal validation, essential for addressing class imbalance, concept drift, and device heterogeneity, were rarely used (<10%). Hybrid strategies showed promise but lacked consistency due to the absence of benchmarks. Tools like Scikit-learn and datasets such as CICIDS2017 and NSL-KDD are commonly used, though often lack diversity. Overall, the dominance of traditional methods reveals the need for standardized, context-aware validation frameworks for real-world reliability.

RQ2: How effectively do current validation approaches address IoT-specific challenges such as data imbalance, device heterogeneity, temporal drift, and adversarial conditions?

Despite ongoing efforts, key challenges persist in validating ML-based intrusion detection systems (IDS) for IoT. Data imbalance was addressed in about 67% of studies using techniques like SMOTE, class weighting, and ADASYN, which improved minority class detection by 5–15% (Kumar et al., 2022). However, concerns remain over overfitting and inflated performance, especially with non-independent datasets (Zhang et al., 2022).

Device and protocol heterogeneity received limited attention, with only 17% of studies using cross-device or federated evaluations. These showed 10–20% performance drops in unseen environments, pointing to poor generalizability (Ahmed et al., 2022; Patel et al., 2020).Temporal drift was considered in just 9% of works, where time-aware validations revealed F1-score declines of 7–12%, suggesting models degrade over time (Alve et al., 2025). Similarly, only 11% of studies tested adversarial resilience, revealing vulnerabilities to evasion attacks (Singh & Yadav, 2024). Overall, current validation remains fragmented, often relying on overly simplified conditions, which limits the reliability and deployment readiness of ML-based IDS in dynamic IoT settings.

RQ 3: What are the methodological limitations of current validation practices, and what opportunities exist for future research in enhancing robustness and deployment-readiness?

Despite advancements, several methodological gaps persist in the reviewed literature. Static train-test splits and k-fold cross-validation dominate 78% of studies (Hameed, 2022), fail to capture the dynamic and distributed nature of IoT environments. Future research should explore context-aware validation strategies, such as continual learning, federated testing, and incremental updates, to better reflect real-world settings. Another concern is the limited use of diverse, standardized datasets. Many studies depend on outdated or single-source data, which hinders model generalizability across varied IoT contexts (Whaiduzzaman et al., 2022). Advancing this area requires the adoption and creation of multi-domain datasets with temporal and adversarial annotations.Statistical consistencywas lacking;only 22% of studies apply statistical tests like confidence intervals or t-tests to support performance claims (Sarker et al., 2022), reducing the reliability of findings.

Integrating such analyses would improve reproducibility and credibility. Moreover, adversarial resilience is rarely addressed. Few models are tested against crafted perturbations or simulated attacks. Research should prioritize adversarial training, certification, and evasion simulations to better reflect real-world threat landscapes (Xie & Huang, 2023). Finally, most studies emphasize

standard metrics such as accuracy or F1-score, overlooking deployment-relevant factors like latency, energy use, and privacy (Krzysztoń et al., 2024). A shift toward holistic validation frameworks that include operational metrics is essential for practical implementation.

IMPLICATIONS FOR RESEARCH AND PRACTICE

The review identifies key gaps in current validation practices that limit the reliability of ML models in IoT cybersecurity. Researchers should adopt realistic validation strategies, such as temporal splits, adversarial testing, and cross-device evaluations, to better mimic operational settings. Using diverse and representative datasets, rather than static or outdated ones, will further improve model generalizability.For practitioners, robust validation frameworks are vital to ensure resilience against evolving threats. Integrating drift-aware retraining, federated evaluations, and adversarial robustness checks into the ML lifecycle can enhance long-term effectiveness. Addressing these gaps is essential for developing adaptable, trustworthy intrusion detection systems fit for dynamic IoT environments.

CONCLUSION

This study critically reviewed validation techniques in ML-based IoT cybersecurity, revealing overreliance on k-fold and hold-out methods, which often fail to reflect the dynamic and imbalanced nature of IoT environments. More robust techniques, such as temporal validation and cross-dataset testing, remain underused despite their relevance for real-world deployment.

The findings highlight the need for standardized, context-aware validation frameworks that address class imbalance and promote reproducibility. Future research should focus on developing benchmark pipelines, encouraging best practices, and validating models in real-time, distributed IoT settings. These steps are essential for advancing trustworthy and scalable ML-driven security solutions.

RECOMMENDATIONS

Researchers should adopt context-aware validation approaches, such as temporal and cross-device methods, while also incorporating robustness metrics to ensure reliability. Addressing class imbalance using techniques like SMOTE and combining traditional with advanced methods can help achieve a more balanced evaluation. Practitioners, on the other hand, should implement hybrid validation strategies during deployment and continuously monitor performance to detect drift or emerging threats. Meanwhile, standardization bodies and the broader research community have a responsibility to establish IoT-specific validation protocols and encourage the development of benchmark datasets that enhance reproducibility and comparability across studies.

FUTURE WORK

Effective IoT model validation should include metrics like energy use, latency, and scalability. Long-term real-world testing ensures resilience to evolving threats. Cross-domain evaluation supports generalizability, while privacy-aware methods (e.g., federated learning) uphold data standards. Adaptive tools that respond to threats and model updates enhance ongoing reliability.

REFERENCES

Abeshu, A. Y., & Chilamkurti, N. (2018). Deep learning: The frontier for distributed attack detection in Fog-to-Things computing. *IEEE Communications Magazine*, *56*(2), 169–175.

Adedeji, K. B., Abu-Mahfouz, A. M., & Kurien, A. M. (2023). DDoS Attack and Detection Methods in Internet-Enabled Networks: Concept, Research Perspectives, and Challenges. *Journal of Sensor and Actuator Networks*, *12*(4). https://doi.org/10.3390/jsan12040051

Admass, W. S., Munaye, Y. Y., & Diro, A. A. (2024). Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications*, *2*(September 2023), 100031. https://doi.org/10.1016/j.csa.2023.100031

Ahmad, R., & Alsmadi, I. (2021). Machine learning approaches to IoT security: A systematic literature review[Formula presented]. *Internet of Things (Netherlands)*, *14*. https://doi.org/10.1016/j.iot.2021.100365

Al Amin, M. A. R., Shetty, S., Njilla, L., Tosh, D. K., & Kamhoua, C. (2021). Hidden markov model and cyber deception for the prevention of adversarial lateral movement. *IEEE Access*, *9*, 49662–49682. https://doi.org/10.1109/ACCESS.2021.3069105

Alharbi, A., Alshamrani, A., & Alabdulatif, A. (2020). A survey on intrusion detection in IoT: Current solutions and future challenges. *Sensors*, *20*(23), 6443. https://doi.org/10.3390/s20236443

Alrehaili, M., & Alshamrani, A. (2023). An Attack Scenario Reconstruction Approach Using Alerts Correlation and a Dynamic Attack Graph. *2023 Eighth International Conference On Mobile And Secure Services (MobiSecServ)*, *CFP23RAC-A*, 1–8. https://doi.org/10.1109/mobisecserv58080.2023.10329144

Alshamrani, A., Alwan, M. J., & Alqahtani, A. (2023). A robust intrusion detection model using ensemble classifiers and bootstrapped validation for IoT networks. *Sensors*, *23(5)*, *228*. https://doi.org/https://doi.org/10.3390/s23052289

Alshamrani, A., Alqahtani, H., & Alasmary, W. (2023). An ensemble-based intrusion detection framework for IoT-enabled smart environments. *Computers & Security*, *128*, 103147. https://doi.org/10.1016/j.cose.2023.103147

Alve, S. R., Mahmud, M. Z., Islam, S., Chowdhury, M. A., & Islam, J. (2025). *Smart IoT Security: Lightweight Machine Learning Techniques for Multi-Class Attack Detection in IoT Networks*. http://arxiv.org/abs/2502.04057

Attota, D. C., Mothukuri, V., Parizi, R. M., & Pouriyeh, S. (2021). An Ensemble Multi-View Federated Learning Intrusion Detection for IoT. *IEEE Access*, *9*, 117734–117745. https://doi.org/10.1109/ACCESS.2021.3107337

Bichri, H., Chergui, A., & Hain, M. (2024). Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. *International Journal of*

Advanced Computer Science and Applications, *15*(2), 331–339. https://doi.org/10.14569/IJACSA.2024.0150235

Diro, A. A., & Chilamkurti, N. (2018). Distributed attack detection scheme using deep learning approach for Internet of Things. *Future Generation Computer Systems*, *82*, 761–768. https://doi.org/https://doi.org/10.1016/j.future.2017.08.043

Doshi, R., & Kute, P. (2020a). Machine learning for detecting network attacks in IoT. *Procedia Computer Science*, *167*, 2217–2223. https://doi.org/10.1016/j.procs.2020.03.273

Doshi, R., & Kute, V. (2020b). A Review Paper on Security Concerns in Cloud Computing and Proposed Security Models. *International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020*, 1–4. https://doi.org/10.1109/ic-ETITE47903.2020.37

Gupta, R., & Singh, H. (2022). A robust hybrid intrusion detection system using random forest and support vector machine. *Computers \& Security*, *110*, 102445. https://doi.org/10.1016/j.cose.2021.102445

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Shraddha, T., Kusko, R., Sansone, S. A., Tong, W., Wolfinger, R. D., Mason, C. E., Jones, W., Dopazo, J., Furlanello, C., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., … Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16. https://doi.org/10.1038/s41586-020-2766-y

Hameed, A. (2022). *Examensarbete 30 hp Battery-less IoT Devices: Energy Source Manipulation Attacks*. *April*. http://www.teknat.uu.se/student

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *7th International Conference on Learning Representations, ICLR 2019*, 1–16.

Janjua, A. A., & Aslam, M. (2021). Identification of climate induced optimal rice yield and vulnerable districts rankings of the Punjab , Pakistan. *Scientific Reports*, 1–15. https://doi.org/10.1038/s41598-021-02691-4

Javed, A., Ehtsham, A., Jawad, M., Awais, M. N., Qureshi, A. U. H., & Larijani, H. (2024). Implementation of Lightweight Machine Learning-Based Intrusion Detection System on IoT Devices of Smart Homes. *Future Internet*, *16*(6), 1–22. https://doi.org/10.3390/fi16060200

Javed, H., El-Sappagh, S., & Abuhmed, T. (2025). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11005-9

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*(November 2023), 122778. https://doi.org/10.1016/j.eswa.2023.122778

Khan, M. A., Tariq, U., & Ali, M. (2022). CNN-based anomaly detection system for IoT networks using CICIDS2017 dataset. *Journal of Network and Computer Applications*, *204*, 103408. https://doi.org/10.1016/j.jnca.2022.103408

Kim, A., & Jung, I. (2023). Optimal selection of resampling methods for imbalanced data with high complexity. *PLoS ONE*, *18*(7 July), 1–18. https://doi.org/10.1371/journal.pone.0288540

Kohav, R. (1995). A Study of CrossValidation and Bootstrap for Accuracy Estimation and Model Selecti. *International Joint Conference on Articial Intelligence _IJCA*.

Krzysztoń, E., Rojek, I., & Mikołajewski, D. (2024). A Comparative Analysis of Anomaly Detection Methods in IoT Networks: An Experimental Study. *Applied Sciences (Switzerland)*, *14*(24). https://doi.org/10.3390/app142411545

Kumar, R., Chauhan, M., & Tripathi, A. (2021). An ensemble-based intrusion detection system using XGBoost. *International Journal of Information Security*, *20*(5), 649–663. https://doi.org/10.1007/s10207-021-00545-2

Landauer, M., Skopik, F., Stojanović, B., Flatscher, A., & Ullrich, T. (2025). A review of time-series analysis for cyber security analytics: from intrusion detection to attack prediction. In *International Journal of Information Security* (Vol. 24, Issue 1). https://doi.org/10.1007/s10207-024-00921-0

Liu, Y., He, H., & Chen, J. (2023). Concept drift-aware LSTM-based intrusion detection system for IoT environments. *IEEE Internet of Things Journal*, *10*(6), 5100–5111. https://doi.org/10.1109/JIOT.2022.3207650

Liu, Y., Zhou, Y., Yang, K., & Wang, X. (2023). Unsupervised Deep Learning for IoT Time Series. *IEEE Internet of Things Journal*, *10*(16), 14285–14306. https://doi.org/10.1109/JIOT.2023.3243391

Liu, Z., & Shi, Y. (2022). *A Hybrid IDS Using GA - Based Feature Selection Method and Random Forest*. *12*(2). https://doi.org/10.18178/ijmlc.2022.12.2.1077

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., & Shabtai, A. (2018). N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, *17*(3), 12–22. https://doi.org/10.1109/MPRV.2018.03367731

Pahl, M. (2022). *All Eyes on You : Distributed Multi-Dimensional IoT Microservice Anomaly Detection*.

Patel, N., Sharma, V., & Jain, R. (2020). Precision and MCC as alternative performance indicators in imbalanced IoT attack detection. *Journal of Information Assurance and Security*, *15*(4), 245–254.

Rahman, M. M., Amin, M. B., & Faisal, M. (2022). An interpretable machine learning model for IoT intrusion detection with explainable AI (XAI). *Future Generation Computer Systems*, *128*, 185–197. https://doi.org/10.1016/j.future.2021.10.024

Rahman, M. M., Berger, D., & Levman, J. (2022). Novel Metrics for Evaluation and Validation of Regression-based Supervised Learning. *Proceedings of IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2022*. https://doi.org/10.1109/CSDE56538.2022.10089291

Raza, M., Hussain, M., & Khan, M. A. (2021). Transfer learning for IoT intrusion detection systems: A survey. *Sensors*, *21*(18), 6144. https://doi.org/10.3390/s21186144

Sahu, M., Thakur, R. S., & Jain, R. (2020). A review of intrusion detection systems using machine learning approaches. *Materials Today: Proceedings*, *33*, 4256–4260.

Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. (2022). Internet of Things (IoT) Security Intelligence: A Comprehensive Overview, Machine Learning Solutions and Research Directions. *Mobile Networks and Applications*. https://doi.org/10.1007/s11036-022-01937-3

Sharma, A., & Singh, M. (2024). Batch reinforcement learning approach using recursive feature elimination for network intrusion detection. *Engineering Applications of Artificial Intelligence*, *136*, 109013. https://doi.org/https://doi.org/10.1016/j.engappai.2024.109013

Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A Deep Learning Approach to Network Intrusion Detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(1), 41–50. https://doi.org/10.1109/TETCI.2017.2772792

Singh, A., & Yadav, V. (2024). Avoiding data leakage in IoT threat detection using nested cross-validation. *Journal of Information Security and Applications*, *77*, *103477*. https://doi.org/https://doi.org/10.1016/j.jisa.2024.103477

Whaiduzzaman, M., Barros, A., Chanda, M., Barman, S., Sultana, T., Rahman, M. S., Roy, S., & Fidge, C. (2022). A Review of Emerging Technologies for IoT-Based Smart Cities. *Sensors*, *22*(23), 1–28. https://doi.org/10.3390/s22239271

Wu, S., Ma, H., Alharbi, A. M., Wang, B., Xiong, L., Zhu, S., Qin, L., & Wang, G. (2023). Integrated Energy System Based on Isolation Forest and Dynamic Orbit Multivariate Load Forecasting. *Sustainability*, *15*(20), 15029. https://doi.org/10.3390/su152015029

Xie, Q., Zhang, H., & Huang. (2023). Learning imbalanced data in IoT intrusion detection: A survey. *IEEE Transactions on Network and Service Management, 20(1), 257*. https://doi.org/https://doi.org/10.1109/TNSM.2022.3192896

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Zhang, H., Zhang, B., Huang, L., Zhang, Z., & Huang, H. (2023). An Efficient Two-Stage Network Intrusion Detection System in the Internet of Things. *Information (Switzerland)*, *14*(2), 1–17. https://doi.org/10.3390/info14020077

Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system

based on feature selection and ensemble classifier. *Computer Networks*, *174*(March). https://doi.org/10.1016/j.comnet.2020.107247