

ARTICLE

A Scan-Based Analysis of Internet-Exposed IoT Devices Using Shodan Data

Richelle Williams^{*} , Fernando Koch 

Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA;
kochf@fau.edu

ABSTRACT

Determining whether scan-observable network configurations capture meaningful characteristics of security exposure across large-scale Internet of Things (IoT) populations remains an open problem in measurement research. This study analyzes Internet-exposed IoT endpoints associated with the TR-069 protocol across a global population exceeding 37 million hosts, using a structured sample of 7,634 hosts derived from approximately 10,000 hosts identified via Shodan Search and Shodan InternetDB. Data were collected during a fixed observation window in April 2026 across ten countries with the highest observed prevalence of transmission control protocol (TCP) port 7547 exposure. Scan-derived metadata, including open ports, service indicators, customer premises equipment (CPE) identifiers, vulnerability-associated fields, and network attribution metadata, were used to construct a composite exposure proxy representing externally observable service-surface characteristics. The analysis combines descriptive statistics, the Kruskal–Wallis H test, and exploratory feature separability analysis to assess geographic variation and whether service-level characteristics differentiate exposure groups. Results show statistically significant cross-country variation in exposure structure. Logistic regression and random forest analyses produced limited feature separability, with the random forest approach achieving a balanced accuracy of 0.58. These findings suggest that Internet-wide scan data can support systematic population-level analysis of IoT exposure while highlighting the limitations of port-based features as indicators of broader exposure characteristics. The study provides a reproducible framework for large-scale exposure measurement using passive scan data without requiring device interaction

*CORRESPONDING AUTHOR:

Richelle Williams, Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA;
Email: williamsr2023@fau.edu

ARTICLE INFO

Received: 23 April 2026 | Revised: 6 August 2026 | Accepted: 13 August 2026 | Published Online: 15 September 2026
DOI: <https://doi.org/10.30564/jeis.v8i2.13448>

CITATION

Williams, R., Koch, F., 2026. A Scan-Based Analysis of Internet-Exposed IoT Devices Using Shodan Data. *Journal of Electronic & Information Systems*. 8(2): 81–93. DOI: <https://doi.org/10.30564/jeis.v8i2.13448>

COPYRIGHT

Copyright © 2026 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

or exploit-based validation.

Keywords: Internet of Things; Security; Risk Assessment; Data Analysis; Cybersecurity; Artificial Intelligence for IT Operations (AIOps)

1. Introduction

The continued expansion of Internet-connected embedded systems has made IoT device security a persistent and measurable infrastructure problem. Devices deployed for remote management, home automation, and industrial monitoring are increasingly reachable from public network address space, exposing management interfaces and auxiliary services to unsolicited probing and automated scanning. Unlike traditional computing infrastructure, IoT devices are frequently deployed with default configurations, receive infrequent firmware updates, and remain in operation beyond the

effective security lifecycle of their software components^[1]. These conditions collectively produce a large and evolving attack surface that is directly observable from the public Internet without requiring privileged access^[2].

The external visibility of this attack surface creates a measurement opportunity. Internet-wide scanning enables systematic observation of exposed services across heterogeneous networks and administrative domains, producing scan-derived data that reflects real-world deployment configurations at scale^[3]. **Figure 1** illustrates this variation, showing consistent differences in mean exposure across the ten sampled countries under uniform observation conditions.

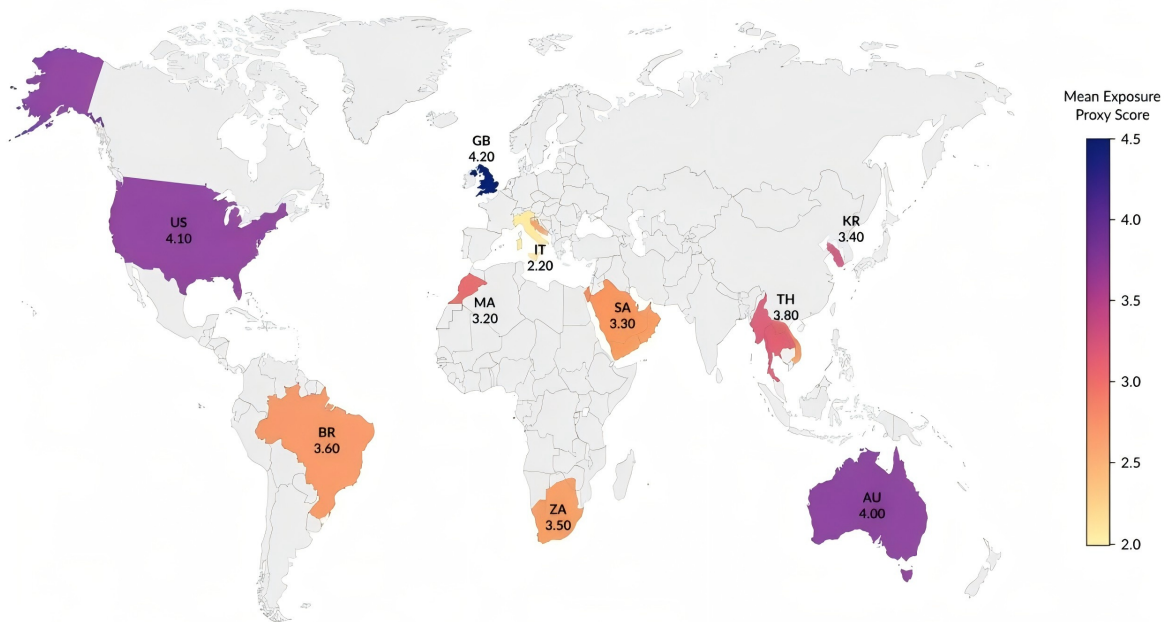


Figure 1. Geographic distribution of mean exposure proxy scores across the ten sampled countries (TR-069/TCP 7547 Sample).

Note: Countries are shaded according to the average exposure proxy value per host, with darker colors indicating higher exposure levels. The map illustrates systematic regional variation in exposure structure under uniform observation conditions.

Whether these observable configurations encode meaningful differences in security exposure across device populations and geographic regions remains an open problem in IoT measurement research. Addressing this question has practical significance for network operators, security researchers, and policymakers seeking scalable methods for assessing IoT

risk without relying on device access, vendor cooperation, or exploit execution^[4].

Prior work in this domain has focused primarily on vulnerability discovery, device fingerprinting, and exploit-driven compromise^[5]. While these approaches have produced important insights, they are often constrained by re-

liance on authenticated access, active exploitation, or complete firmware metadata. Scan-based measurement provides a complementary approach grounded in passive observation of externally visible service configurations. However, systematic cross-country comparison of exposure structure using controlled sampling remains limited, and geographic variation is often treated as incidental rather than as a primary analytical dimension^[6].

This study addresses this gap by examining whether scan-derived service configurations encode measurable and geographically differentiated exposure signals. Using a dataset of 7,634 Internet-exposed hosts identified via TCP port 7547 and sampled proportionally across ten countries from a global population exceeding 37 million TR-069-exposed devices, the analysis applies feature relevance assessment, nonparametric hypothesis testing, and exploratory feature separability analysis. Data were collected through the Shodan Search Application Programming Interface (API) and Shodan InternetDB during April 2026, relying exclusively on passive observation of publicly available scan data^[7].

The remainder of this paper is organized as follows. Section 2 reviews relevant background on IoT security measurement, Internet-wide scanning, and analytical approaches for evaluating observable exposure characteristics. Section 3 describes the data collection procedure, sampling strategy, feature construction, and statistical analysis methods. Section 4 presents the results of the exposure analysis and feature separability assessment. Section 5 discusses the findings, limitations, and directions for future work.

2. Background

The rapid growth of Internet-connected embedded systems has significantly expanded the population of IoT devices reachable from public network address space. Constraints in hardware capability, irregular patch cycles, and extended deployment lifespans contribute to persistent security risk across both consumer and enterprise environments^[8]. These structural characteristics make large-scale mitigation difficult and sustain long-term exposure across heterogeneous device ecosystems^[9,10].

Internet-wide scanning has emerged as a standard approach for measuring this exposure. Tools capable of high-speed probing enable discovery of publicly accessible ser-

vices across the global address space, while indexing platforms aggregate host-level attributes including service banners, open ports, protocol fingerprints, and vulnerability-associated metadata. These capabilities support large-scale comparative analysis of network-exposed systems without requiring direct device interaction or authenticated access^[11]. Observational studies based on such data consistently demonstrate that insecure configurations and exposed services are widespread across diverse IoT deployments, indicating that misconfiguration is a systemic rather than isolated issue.

Among the most security-relevant components of the Internet-exposed IoT surface is the TR-069 CPE WAN Management Protocol (TR-069/CWMP). This protocol is widely deployed by Internet service providers for remote configuration of customer-premises equipment and operates over TCP port 7547. Despite its operational importance, it has been repeatedly targeted in large-scale scanning and exploitation campaigns. Weaknesses in deployment practices, including exposure of management interfaces and insufficient input validation, have enabled compromise of devices at scale. The persistence of newly disclosed vulnerabilities across multiple implementations indicates that this attack surface remains active and operationally significant^[12,13].

Direct mapping of scan-derived metadata to known vulnerabilities presents inherent limitations. Banner information is often incomplete or obfuscated due to device configuration, intermediary network components, or protocol-specific behavior, making reliable version-based vulnerability attribution difficult. As a result, feature-based approaches to exposure characterization have gained traction. Analytical methods that leverage network-observable attributes—such as open ports, service metadata, and protocol behavior—provide an alternative means of distinguishing exposure profiles across heterogeneous device populations^[14].

Despite these advances, systematic cross-country comparison of IoT exposure structure using controlled sampling remains limited. Existing work largely focuses on vulnerability discovery or device-level compromise, rather than population-level analysis of externally observable service configurations. This study addresses that gap by applying scan-derived feature analysis to a proportionally sampled dataset of TR-069-exposed hosts across ten countries, enabling comparative evaluation of exposure patterns across diverse network environments^[15].

3. Method

3.1. Data Source and Collection Window

This study adopts a scan-based observational design to examine Internet-exposed IoT endpoints associated with the TR-069/CWMP remote management protocol. Host discovery was performed through the Shodan Search API using the query “port:7547,” targeting devices that expose TCP port 7547 on the public Internet^[16].

Data collection took place during a fixed observation window in April 2026. Constraining collection to a defined window reduces temporal inconsistencies arising from host churn and dynamic address reassignment, ensuring that the dataset reflects a consistent cross-sectional snapshot of the exposed host population (**Table 1**).

Table 1. Shodan Search API query result for port:7547, April 2026.

Query	Total Hosts	Collection Date
port:7547	37,807,141	April 2026

Scan-derived metadata were retrieved for each host directly from Shodan, including IP address, country code, organization, Internet service provider, autonomous system number, hostname, and associated timestamps. These fields capture externally observable network configuration characteristics without requiring authenticated access, device interaction, or any form of active probing.

All retrieved records were stored in raw tabular format prior to preprocessing and feature construction. The study is grounded exclusively in passive observation of publicly available scan data and does not involve active exploitation or intrusion of any kind. Preprocessing steps and the country-stratified sampling procedure applied to construct the analytical dataset are described in Section 3.2.

3.2. Study Population and Filtering

Given the scale of the global host population, exhaustive collection was not feasible via the Shodan API. Instead, an initial dataset of approximately 10,000 hosts was retrieved through paginated queries using the Shodan Search API. This dataset represents a systematically collected subset of Internet-exposed hosts observed during the defined collection window.

To construct a focused analytical dataset, hosts were

filtered based on geographic prevalence. The ten countries with the highest observed frequency of TCP port 7547 exposure were selected based on the initial count query described in Section 3.1. The resulting analytical dataset, summarized in **Table 2**, includes South Korea (36.5%), the United States (25.4%), Great Britain (13.5%), Brazil (7.0%), Australia (4.8%), Saudi Arabia (3.8%), Thailand (2.6%), Italy (2.3%), South Africa (2.0%), and Morocco (2.0%). Together, these countries represent the high-prevalence regions examined in this study.

Table 2. Final analytical dataset distribution across ten countries following preprocessing and filtering, April 2026.

Country	Code	Host Count	Sample (%)
South Korea	KR	2,789	36.5
United States	US	1,937	25.4
Great Britain	GB	1,031	13.5
Brazil	BR	534	7.0
Australia	AU	370	4.8
Saudi Arabia	SA	290	3.8
Thailand	TH	197	2.6
Italy	IT	179	2.3
South Africa	ZA	154	2.0
Morocco	MA	153	2.0
Total		7,634	100

Following country-based filtering, preprocessing steps were applied to ensure dataset integrity and consistency. Records with missing, invalid, or undefined country codes were removed. Hosts lacking essential metadata fields required for downstream feature construction, specifically autonomous system number and organization information, were excluded. These steps reduced the dataset from approximately 10,000 hosts to a final analytical set of 7,634 hosts, distributed across the ten selected countries as summarized in **Table 2**.

This distribution preserves the natural imbalance present in the observed data, reflecting real-world deployment patterns rather than enforcing artificial uniformity across regions. While the dataset does not represent the full global population, it captures a substantial subset of high-prevalence regions, enabling meaningful comparative analysis of exposure characteristics across heterogeneous network environments.

3.3. Feature Construction and Exposure Proxy Definition

To enable quantitative analysis of externally observable exposure characteristics, a set of features was constructed

from scan-derived metadata and observable service information. These features capture both attribution-related properties and direct indicators of network service exposure.

For each host i , four binary metadata indicators were defined:

- $x_{1,i}$: Presence of a resolvable hostname.
- $x_{2,i}$: Presence of organization (org) metadata.
- $x_{3,i}$: Presence of internet service provider (ISP) information.
- $x_{4,i}$: Presence of an autonomous system number (ASN).

Each binary variable was encoded as:

$x_{k,i} \in \{0, 1\}$, where $x_{k,i} = 1$ indicates the presence of the corresponding metadata field, and $x_{k,i} = 0$ otherwise.

These indicators capture externally observable attribution characteristics derived from Shodan data and do not require direct interaction with the underlying device.

These metadata indicators should not be interpreted as direct measures of security exposure or vulnerability. Rather, they reflect the availability of attribution-related information within scan-derived records. As a result, they capture aspects of host identifiability and metadata completeness rather than confirmed security posture.

In addition to metadata indicators, service-level ex-

posure features were defined based on observable network services:

- $x_{5,i}$: Normalized number of open ports observed for host i .
- $x_{6,i}$: Normalized number of risky ports observed for host i .

The open port count reflects the externally visible service surface of a host. The risky port count captures exposure to services commonly associated with remote access, device management, or historically exploited protocols in Internet-facing systems.

In this study, risky ports were defined as listed in **Table 3**. These services are widely documented in prior Internet measurement and IoT security research as frequent sources of misconfiguration, unauthorized access, or large-scale exploitation^[17]. Remote access services such as Telnet and SSH are common targets of credential-based attacks, while web interfaces often expose administrative functionality^[18]. TR-069 is specifically associated with remote device management vulnerabilities, and UPnP is known to introduce exposure through misconfiguration and amplification-based attack vectors.

Table 3. Ports classified as risky in this study, with associated protocols and documented risk rationale.

Port	Protocol	Risk Rationale
22	SSH	Frequent target of credential-based brute-force attacks on Internet-facing devices
23	Telnet	Transmits credentials in plaintext; exploited at scale by the Mirai botnet and subsequent campaigns
80	HTTP	Exposes web-based administrative interfaces without encryption
443	HTTPS	Encrypted web interface exposure associated with misconfigured or self-signed certificate deployments
7547	TR-069/CWMP	Remote device management protocol associated with large-scale scanning and exploitation of customer-premises equipment
1900	UPnP	Introduces exposure through misconfiguration and is associated with amplification-based attack vectors

To ensure comparability between binary metadata indicators and port-based features, the port-count features were normalized using min–max scaling across the dataset:

$$x'_{k,i} = \frac{x_{k,i} - \min(x_k)}{\max(x_k) - \min(x_k)} \quad (1)$$

where $x'_{k,i}$ is the normalized value of feature k for host i , $x_{k,i}$ is the original value of feature k , and $\min(x_k)$ and $\max(x_k)$ denote the minimum and maximum values of feature k across the analytical dataset.

To aggregate these features into a unified representation, an exposure proxy score was defined as:

$$E_i = \sum_{k=1}^4 x_{k,i} + x'_{5,i} + x'_{6,i} \quad (2)$$

where E_i denotes the exposure proxy score for host i . The first four terms represent the binary metadata indicators, while $x'_{5,i}$ and $x'_{6,i}$ represent the normalized open-port count and normalized risky-port count, respectively.

Higher values of E_i indicate hosts with a larger observable service surface and a greater amount of attribution-related metadata. Lower values correspond to hosts with fewer externally observable services and metadata fields.

It is important to note that this proxy does not directly measure vulnerability or exploitability. Instead, it provides a structured representation of externally observable exposure derived from passive scan data. This formulation enables

comparative analysis across hosts and geographic regions while remaining consistent with the study's non-intrusive measurement design.

3.4. Statistical Analysis Methods

To explore the extent to which scan-derived service-level features contribute to differentiation between hosts exhibiting higher and lower exposure characteristics, an exploratory feature separability analysis was conducted. Hosts were categorized into higher- and lower-exposure groups using a median split of the exposure proxy variable.

Because the exposure groups are derived from the exposure proxy itself, and because service-level features contribute to the construction of that proxy, the resulting analysis should not be interpreted as an independent assessment of security exposure. Rather, it serves as an exploratory evaluation of the extent to which observable service-surface characteristics differentiate between higher- and lower-exposure groups.

The analysis was restricted to service-level features derived from open-port observations. Specifically, the normalized open-port count and normalized risky-port count were used as input variables. These features were selected because they represent externally observable characteristics of the exposed service surface and are available across all sampled hosts.

To examine the degree of separation between higher- and lower-exposure groups, two analytical approaches were evaluated. A logistic regression approach was used as a baseline linear method, while a random forest approach was used to assess potential nonlinear relationships within the service-level feature space.

The ability of service-level features to differentiate between exposure groups was evaluated using stratified k -fold cross-validation ($k = 5$) to preserve group balance across folds. Evaluation metrics included accuracy, precision, recall, F1-score, and balanced accuracy. Class-weighting was applied to mitigate any residual imbalance. Accordingly, the resulting metrics are interpreted solely as measures of feature separability between exposure groups rather than indicators of predictive performance or an independent security outcome.

All statistical analyses and feature separability procedures were implemented in Python using the SciPy and scikit-learn libraries^[19,20].

3.5. Limitations

This study has several limitations that should be considered when interpreting the results. First, the dataset represents a partial sample of the global population of TR-069-exposed hosts. While the initial query identified over 37 million hosts, API constraints limited data collection to approximately 10,000 hosts, yielding a final analytical dataset of 7,634 hosts after preprocessing. Although the sampling strategy prioritizes countries with the highest observed prevalence of exposure, the dataset is not exhaustive and may not fully capture the global distribution of TR-069 deployments.

Second, the analysis is based exclusively on scan-derived, externally observable data. While this enables large-scale measurement without requiring device access, it limits the ability to infer internal system state, configuration correctness, or the presence of unobserved services. As a result, the exposure proxy reflects observable configuration characteristics rather than confirmed security posture.

Third, the definition of exposure relies on a constructed proxy that combines metadata indicators and service-level features. Although this approach enables systematic comparison across hosts, it does not directly measure vulnerability or exploitability. In the present study, all proxy components were assigned equal weight to provide a simple and interpretable representation of externally observable exposure characteristics. Alternative weighting strategies, including data-driven or expert-informed weighting schemes, may influence the relative contribution of metadata and service-level features and could produce different exposure rankings. A formal sensitivity analysis of alternative weighting approaches was beyond the scope of this study but represents an important direction for future research.

Fourth, the feature separability analysis is based on exposure groups derived from the exposure proxy itself. Although care was taken to reduce circularity by restricting the analysis to service-level features, the results should be interpreted as an assessment of the degree to which observable service-surface characteristics distinguish between higher- and lower-exposure groups. The moderate separability observed in this study suggests that port-based features alone capture only a limited portion of the variation represented by the exposure proxy.

Finally, the analysis is constrained to a single observation window. IoT deployments are dynamic, and exposure

characteristics may change over time due to device churn, network reconfiguration, or firmware updates. As a result, the findings represent a cross-sectional snapshot rather than a longitudinal assessment of exposure dynamics.

Despite these limitations, the study provides a reproducible framework for analyzing IoT exposure using scan-derived data and highlights both the potential and the constraints of externally observable measurement approaches.

3.6. Reproducibility

Data collection was conducted during a fixed observation window in April 2026 using the Shodan Search API and Shodan InternetDB. Host discovery was performed using the query `port:7547`, targeting Internet-exposed devices associated with the TR-069/CWMP protocol. Retrieved records were processed and analyzed using Python within the Google Colab environment.

All preprocessing, feature engineering, statistical analyses, feature separability analyses, and visualization procedures were implemented using reproducible Python workflows. Statistical analyses were performed using the SciPy library, while the separability analyses were implemented using the scikit-learn framework. The study relies exclusively on publicly observable scan-derived information and does not involve active interaction with target devices or exploitation of vulnerabilities.

4. Analysis

4.1. Descriptive Statistics of Exposure

The exposure proxy variable exhibits moderate variability across the analytical dataset of 7,634 hosts. Summary statistics indicate a mean exposure value of 3.66 with a standard deviation of 0.61, and a range from 1.00 to 6.00. The distribution is slightly right-skewed, with a median of 3.25, indicating that a substantial proportion of hosts exhibit relatively similar baseline exposure characteristics.

Service-level features derived from port observations show distinct distributional properties. The open port count has a mean of 2.92 with a standard deviation of 39.03, reflecting the presence of extreme outliers with unusually large service surfaces, as illustrated in **Figure 2**. These outliers were retained in the analysis as they represent genuinely

observable configurations of Internet-exposed hosts rather than measurement artifacts. The median open port count of 1 indicates that most hosts expose a minimal number of services. Risky port count exhibits lower variability, with a mean of 0.80 and a maximum value of 4, suggesting that a limited set of commonly targeted services accounts for most observable exposure.

4.2. Cross-Country Comparison of Exposure

Figure 2 highlights clear cross-country variation in observable exposure characteristics. Countries differ not only in their mean exposure proxy values but also in the composition of their observable service surfaces, as reflected by differences in mean open-port count and mean risky-port count.

The figure summarizes the mean normalized values of the exposure proxy indicators for each sampled country. Because these values represent normalized country-level summaries, they are intended to facilitate relative comparisons across countries rather than represent individual host-level exposure scores. Great Britain, the United States, and Australia exhibit the highest normalized exposure values, while Italy shows the lowest relative exposure among the sampled countries.

The observed variation suggests that exposure is not uniformly distributed across geographic regions. Instead, the results indicate systematic differences in externally observable service configurations and deployment characteristics across the sampled populations. These patterns suggest that scan-derived features may capture geographically differentiated exposure signals, motivating the distributional analysis and statistical testing presented below.

Distributional differences across countries are shown in **Figure 3**. Italy stands out with a substantially wider interquartile range than any other sampled country, with exposure proxy values spanning from approximately 1.0 to 4.5, indicating considerable heterogeneity in observable service configurations among Italian TR-069 hosts. South Korea and Morocco show markedly narrower distributions, suggesting more uniform deployment patterns within those populations. These contrasts indicate that country-level exposure differences are driven not only by differences in mean values but also by structural variation in the underlying configuration landscape.

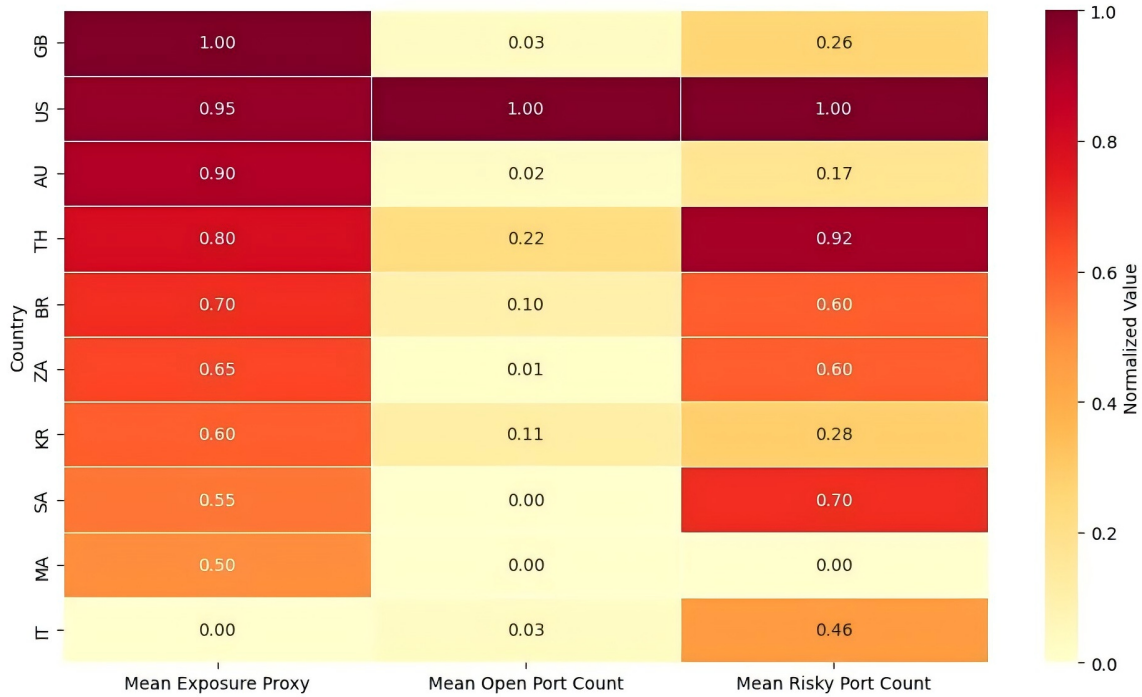


Figure 2. Comparison of mean normalized exposure indicators across the ten sampled countries.

Note: The first column represents the normalized mean exposure proxy value for each country, while the remaining columns represent the normalized mean open-port count and normalized mean risky-port count. These normalized country-level summaries are intended for relative comparison across countries and should not be interpreted as individual host-level exposure scores. Higher values indicate greater observable exposure for the corresponding indicator.

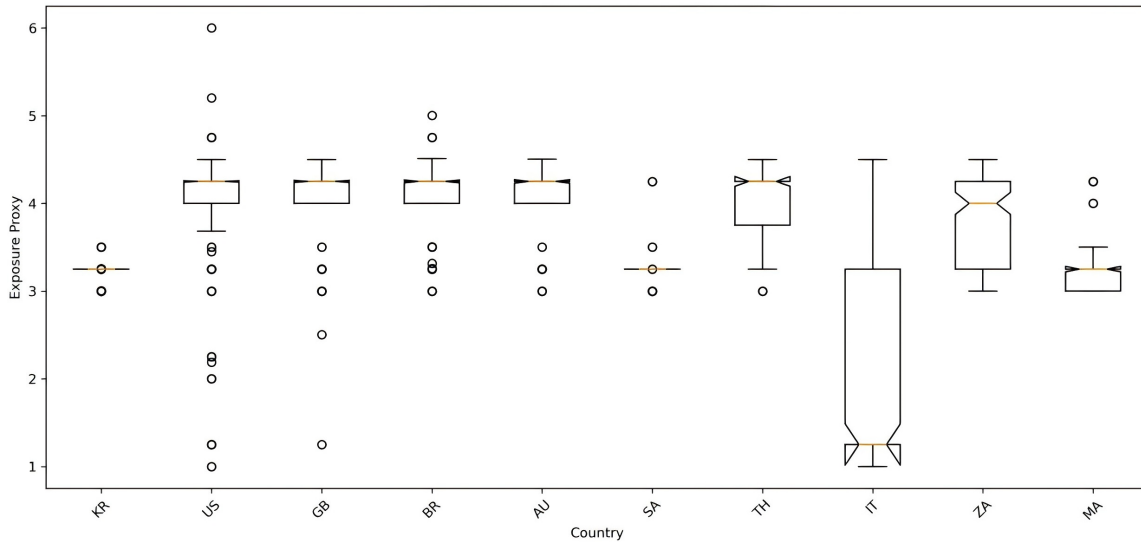


Figure 3. Distribution of raw exposure proxy scores across the ten sampled countries.

Note: Boxes represent interquartile ranges, orange lines indicate medians, whiskers extend to 1.5 times the interquartile range, and circles denote outliers. Unlike the normalized country-level summaries presented in Figure 2, this figure illustrates the distribution of individual host-level exposure proxy scores within each country. Italy exhibits notably higher distributional variability compared to the other sampled countries.

Statistical testing using the Kruskal–Wallis H test confirms that these differences are statistically significant. The test yields a large H-statistic ($H = 5602.35$) with a p -value effectively equal to zero ($p < 0.001$), leading to rejection of

the null hypothesis that all countries share the same exposure distribution. To assess the practical magnitude of this difference, epsilon-squared (ϵ^2) was computed as an effect-size measure for the Kruskal–Wallis test. The resulting value

($\epsilon^2 = 0.734$) indicates a large practical effect, suggesting that a substantial proportion of the observed variation in exposure proxy values is associated with country-level differences rather than random variation alone.

The observed variation is reflected not only in mean exposure values but also in the composition of service surfaces, including differences in open port counts and the presence of risky ports. These findings suggest that exposure structure is influenced by regional deployment practices, network configurations, and operational environments.

4.3. Feature Separability Analysis of Exposure Groups

To assess the degree to which service-level features contribute to differentiation between exposure groups, an exploratory feature separability analysis was performed using normalized open port count and normalized risky port count as input variables. Because the exposure-group labels are derived from the exposure proxy itself, the analysis should not be interpreted as an independent assessment of security exposure. Instead, the objective is to evaluate the extent to which service-level characteristics contribute to observable

separation between higher- and lower-exposure hosts.

Exposure groups were defined using a median split of the exposure proxy variable, resulting in a balanced dataset containing 3,820 lower-exposure hosts and 3,814 higher-exposure hosts. This near-equal group distribution supports the use of stratified cross-validation and enables a balanced assessment of feature separability without requiring extensive resampling techniques^[21].

The logistic regression approach achieved an accuracy of 0.51 and a balanced accuracy of 0.51, indicating limited separation between the two exposure groups. However, the approach achieved a recall of 0.79 and an F1-score of 0.62, suggesting that observable service-level features contain some information associated with higher-exposure hosts, despite limited overall discriminatory capability.

The random forest approach yielded improved results, achieving an accuracy of 0.58 and a balanced accuracy of 0.58. The approach achieved a recall of 0.83 and an F1-score of 0.67, indicating that nonlinear relationships among service-level features contribute additional information relevant to differentiating between exposure groups. A comparison of performance metrics is shown in **Table 4**.

Table 4. Performance metrics for the feature separability analysis using normalized open port count and normalized risky port count as service-level indicators.

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1
Logistic Regression	0.510	0.510	0.506	0.788	0.616
Random Forest	0.585	0.585	0.557	0.829	0.666

The improvement observed with Random Forest suggests that nonlinear interactions among service-level features contain additional information not captured by the linear decision boundary of logistic regression.

Despite this improvement, overall separability remains moderate, suggesting that port-based features alone capture only a limited portion of the variation represented by the exposure proxy. This behavior is further illustrated in **Figure 4**, which reveals an asymmetric error structure in which higher-exposure hosts are identified with greater sensitivity, while lower-exposure hosts exhibit greater overlap within the service-level feature space. These findings indicate that observable port-based characteristics provide only a partial representation of the broader exposure patterns captured by the composite exposure proxy. The substantially higher recall than precision indicates that the model is more effective

at identifying potentially higher-exposure hosts than at excluding lower-exposure hosts, resulting in a relatively large number of false-positive classifications.

The confusion matrix suggests substantial overlap between the two exposure groups in the feature space, indicating that normalized open-port and risky-port counts alone are insufficient to cleanly distinguish exposure levels. To further evaluate the extent to which service-level features differentiate between exposure groups across varying decision thresholds, **Figure 5** presents the receiver operating characteristic (ROC) curve obtained from the Random Forest analysis, yielding an area under the curve (AUC) of 0.67. This result indicates limited but non-random discriminative capability, consistent with the moderate balanced accuracy observed in the confusion matrix analysis. The ROC curve lies above the diagonal baseline across most threshold values,

confirming that scan-derived service-level features encode measurable separation between higher- and lower-exposure hosts. However, the relatively modest AUC highlights that this separability remains weak, reflecting the constrained expressive power of port-based features. Taken together with

the confusion matrix results and feature importance analysis, these findings suggest that while externally observable service configurations provide useful signals for exposure characterization, they do not fully capture the underlying complexity of exposure structure across IoT populations.

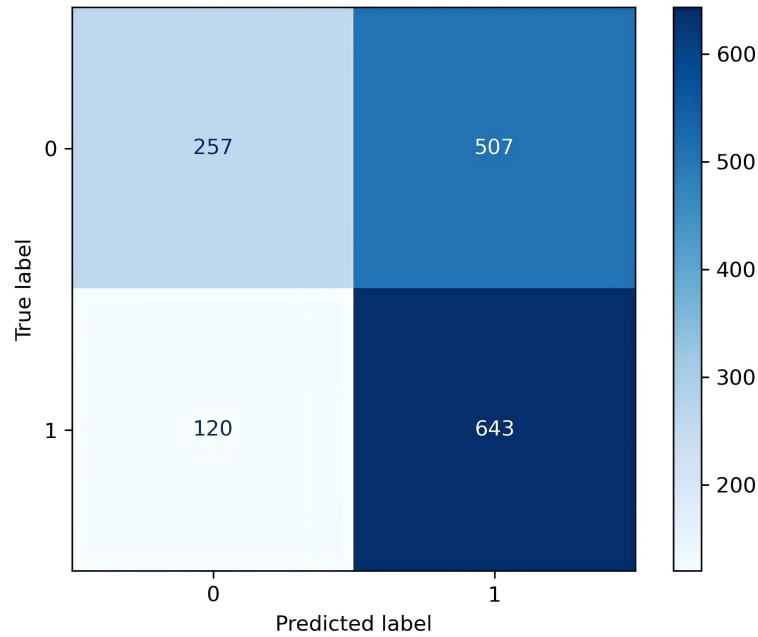


Figure 4. Confusion matrix for the Random Forest approach.

Note: The analysis demonstrates stronger identification of higher-exposure hosts, while lower-exposure hosts exhibit greater overlap within the service-level feature space. This pattern is consistent with the moderate degree of separation observed between the two exposure groups.

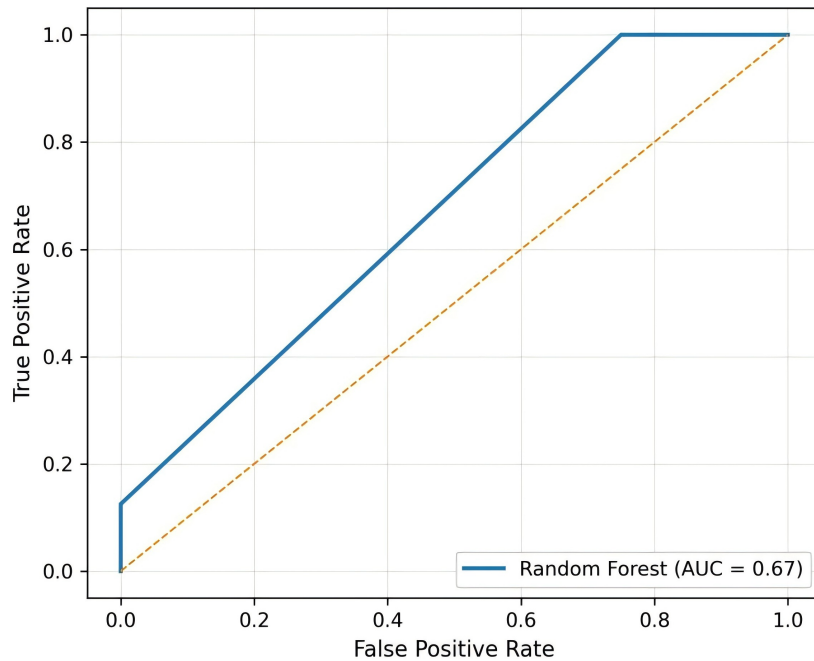


Figure 5. Receiver operating characteristic (ROC) curve for the Random Forest approach.

Note: The model achieves an area under the curve (AUC) of 0.67, indicating limited but non-random discriminative capability. The curve remains above the diagonal baseline, confirming that scan-derived service-level features capture measurable separation between higher- and lower-exposure hosts.

4.4. Interpretation of Findings

The feature separability analysis indicates that scan-derived service-level features capture measurable but limited information related to the exposure proxy. While both models achieve relatively high recall (0.79 for logistic regression and 0.83 for random forest), indicating effective identification of higher-exposure hosts, the overall balanced accuracy remains modest (0.51 and 0.58, respectively). This disparity suggests that the models are biased toward detecting higher-exposure cases while struggling to reliably distinguish lower-exposure hosts.

The results of the feature separability analysis should be interpreted with caution because the exposure-group labels are derived from the exposure proxy used elsewhere in the study. Although service-level features constitute only a subset of the proxy components, the resulting analysis does not represent an independent evaluation of security exposure. Consequently, the analysis is best viewed as an exploratory assessment of the extent to which observable service-surface characteristics differentiate between higher- and lower-exposure groups rather than as an independent validation of exposure structure^[22,23].

The moderate performance of the Random Forest approach, despite its ability to capture nonlinear feature interactions, suggests that port-based features alone capture only a limited portion of the information represented by the exposure proxy^[24,25]. The modest improvement over the baseline approach further indicates that a substantial share of exposure-related variation is not reflected in externally observable service configurations.

These findings imply that exposure is influenced by additional factors beyond scan-derived features, including device-specific implementation differences, network architecture, firmware behavior, and configuration policies that are not directly observable through Internet-wide scanning. This interpretation is consistent with the observed cross-country variation in exposure structure, where statistically significant differences are present despite the limited degree of separation achieved using port-based features alone.

As a result, scan-derived indicators should be interpreted as partial and probabilistic signals of exposure rather than definitive measures of security risk.

Overall, the results support the value of Internet-wide scan data for large-scale exposure measurement and comparative analysis, while highlighting the inherent limitations of relying exclusively on observable network characteristics to capture the full complexity of exposure structure.

5. Conclusions

This study investigated whether scan-observable network configurations can capture measurable differences in Internet-exposed IoT deployment characteristics across geographically distributed host populations. Using a dataset of 7,634 TR-069-exposed hosts collected from ten countries during a fixed observation period, the analysis combined descriptive statistics, nonparametric hypothesis testing, and exploratory feature separability analysis to examine variation in externally observable exposure characteristics.

The results demonstrate that observable exposure characteristics are not uniformly distributed across the sampled countries. Statistical analysis revealed significant differences in exposure distributions, indicating that externally visible service configurations vary systematically across geographic regions. These findings suggest that regional deployment practices, network management strategies, and operational environments contribute to measurable differences in observable exposure structure. The large effect size observed in the country-level analysis further indicates that these differences are practically meaningful rather than solely the result of random variation.

The feature separability analysis showed that service-level characteristics derived from scan data contain measurable information related to exposure structure but provide only limited discriminatory capability. While logistic regression achieved near-random performance, the Random Forest model demonstrated modest improvement, suggesting that nonlinear interactions among service-level features contribute additional information relevant to exposure differentiation. However, the overall classification performance remained moderate, with a balanced accuracy of 0.585 and an AUC of 0.67. These results indicate that observable port-based characteristics capture only a portion of the variation represented by the exposure proxy and are insufficient to fully characterize exposure levels on their own.

Taken together, the findings support the use of Internet-wide scan data as a scalable and reproducible resource for population-level measurement and comparative analysis of IoT exposure characteristics. At the same time, the results highlight important limitations associated with scan-derived indicators. Exposure was assessed using externally observable network characteristics and a proxy-based exposure measure rather than validated vulnerability data or direct security assessments. Consequently, scan-derived exposure metrics should be interpreted as observational indicators of relative exposure rather than definitive measures of device vulnerability or security risk.

This work contributes a reproducible framework for analyzing Internet-exposed IoT populations using passive scan data and provides empirical evidence of geographic variation in observable exposure characteristics. Future research should incorporate additional sources of information, including protocol-level measurements, device fingerprinting techniques, vulnerability intelligence, and longitudinal observations, to improve exposure characterization and better capture the complex factors that influence security posture. An important next step is to validate the proposed exposure proxy against independently verified vulnerability datasets or devices with confirmed security assessments to determine how well scan-derived exposure characteristics correspond to actual security outcomes. Integrating these complementary data sources may provide a more comprehensive understanding of Internet-scale IoT exposure and strengthen both the interpretation and empirical validation of scan-derived exposure metrics.

Author Contributions

Conceptualization, R.W.; methodology, R.W.; software, R.W.; validation, R.W.; formal analysis, R.W.; investigation, R.W.; resources, R.W.; data curation, R.W.; writing—original draft preparation, R.W.; writing—review and editing, R.W. and F.K.; visualization, R.W.; supervision, F.K.; project administration, R.W. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived for this study because it relied exclusively on publicly available Internet scan data and did not involve human participants, animals, identifiable personal information, or direct interaction with networked devices.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data analyzed in this study were obtained from the Shodan Search API and Shodan InternetDB during a fixed observation window in April 2026. Due to the dynamic nature of Internet-wide scan data and the terms of service governing Shodan data access, the complete dataset cannot be publicly redistributed. The methods, data collection procedures, and analytical workflow are described in sufficient detail within the manuscript to support reproducibility.

Conflict of Interest

The authors declare no conflict of interest.

AI Use Statement

Artificial intelligence (AI) tools were used to assist with language editing and improve the clarity and readability of the manuscript. The authors reviewed, verified, and take full responsibility for all content, analyses, interpretations, and conclusions presented in this work.

References

- [1] Bakhshi, T., Ghita, B., Kuzminykh, I., 2024. A review of IoT firmware vulnerabilities and auditing techniques. *Sensors*. 24(2), 708.
- [2] Alshammari, B.M., 2025. A machine learning-based framework for measuring attack surfaces of IoT systems. *IEEE Access*. 13, 134297–134311.
- [3] Rajasekar, V.R., Rajkumar, S., 2023. A study on internet of things devices vulnerabilities using shodan. *International Journal of Computing*. 22(2), 149–158.
- [4] Yaben, R., Vasilomanolakis, E., 2025. Digital ghost

- ships: Abandoned, neglected, and obsolete IoT and OT devices exposed to the internet. TechRxiv. DOI: <https://doi.org/10.36227/techrxiv.174838052.25233921/v1>
- [5] Li, R., Li, Q., Lin, T., et al., 2024. Deviceradar: Online iot device fingerprinting in isps using programmable switches. *IEEE/ACM Transactions on Networking*. 32(5), 3854–3869.
- [6] Liang, B.-H., Hwang, R.-H., Lin, J.-Y., et al., 2025. Comprehensive vulnerability detection and malware infection testing strategies for IoT devices. *IEEE Internet of Things Journal*. 12(12), 20556–20571.
- [7] Haseeb, J., Mansoori, M., Welch, I., 2020. A measurement study of IoT-based attacks using IoT kill chain. In *Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Guangzhou, China, 29 December 2020–1 January 2021; pp. 557–567.
- [8] Srinivasa, S., Pedersen, J.M., Vasilomanolakis, E., 2021. Open for hire: Attack trends and misconfiguration pitfalls of IoT devices. In *Proceedings of the 21st ACM Internet Measurement Conference*, Online, 2–4 November 2021; pp. 195–215.
- [9] Mishra, R., Mishra, A., 2025. Current research on internet of things (IoT) security protocols: A survey. *Computers & Security*. 151, 104310.
- [10] Dritsas, E., Trigka, M., 2025. A survey on cybersecurity in IoT. *Future Internet*. 17(1), 30.
- [11] Durumeric, Z., Adrian, D., Stephens, P., et al., 2024. Ten years of ZMap. In *Proceedings of the 2024 ACM Internet Measurement Conference*, Madrid, Spain, 4–6 November 2024; pp. 139–148.
- [12] Kayas, G., Hossain, M., Payton, J., et al., 2021. Supnp: Secure access and service registration for upnp-enabled internet of things. *IEEE Internet of Things Journal*. 8(14), 11561–11580.
- [13] Antonakakis, M., April, T., Bailey, M., et al., 2017. Understanding the Mirai Botnet. In *Proceedings of the USENIX Security Symposium*, Vancouver, BC, Canada, 16–18 August 2017.
- [14] Alfahaid, A., Alalwany, E., Almars, A.M., et al., 2025. Machine learning-based security solutions for IoT networks: A comprehensive survey. *Sensors*. 25(11), 3341.
- [15] Raja, S., Manikandasaran, S.S., Doss, R., 2022. Threat modeling and IoT attack surfaces. In: Aurelia, S., Paiva, S. (Eds.). *Immersive Technology in Smart Cities: Augmented and Virtual Reality in IoT*. Springer: Cham, Switzerland. pp. 229–258.
- [16] Coston, I., Plotnizky, E., Nojournian, M., 2025. Comprehensive study of IoT vulnerabilities and countermeasures. *Applied Sciences*. 15(6), 3036.
- [17] Lu, Y., Xu, L.D., 2019. Internet of things (IoT) cybersecurity research: A review of current research topics. *IEEE Internet of Things Journal*. 6(2), 2103–2115.
- [18] Ali, R.F., Muneer, A., Dominic, P.D.D., et al., 2021. Internet of things (IoT) security challenges and solutions: A systematic literature review. In: Abdullah, N., Manickam, S., Anbar, M. (Eds.). *Advances in Cyber Security*. Springer: Singapore. pp. 128–154.
- [19] Virtanen, P., Gommers, R., Oliphant, T.E., et al., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*. 17(3), 261–272.
- [20] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12, 2825–2830.
- [21] Owusu-Adjei, M., Hayfron-Acquah, J.B., Frimpong, T., et al., 2023. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*. 2(11), e0000290.
- [22] Waqdan, M., Louafi, H., Mouhoub, M., 2025. Security risk assessment in IoT environments: A taxonomy and survey. *Computers & Security*. 154, 104456.
- [23] AlSalem, T.S., Almaiah, M.A., Lutfi, A., 2023. Cybersecurity risk analysis in the IoT: A systematic review. *Electronics*. 12(18). DOI: <https://doi.org/10.3390/electronics12183958>
- [24] Kouassi, B.M., Ballo, A.B., Ayikpa, K.J., et al., 2025. Top-K feature selection for IoT intrusion detection: Contributions of XGBoost, Lightgbm, and Random Forest. *Future Internet*. 17(11). DOI: <https://doi.org/10.3390/fi17110529>
- [25] Chalichalamala, S., Govindan, N., Kasarapu, R., 2023. Logistic regression ensemble classifier for intrusion detection system in internet of things. *Sensors*. 23(23), 9583.