EPiC Series in Computing

Volume 111, 2026, Pages 49–65

Proceedings of The 21st Asia Joint Conference on Information Security

A Longitudinal Validation of Cluster-Based Phishing Analysis: Tracking High-Impact Clusters Targeting Japan

Shingo Matsugaya[123], Makoto Shimamura[2], and Masaki Hashimoto[1]

[1] Graduate School of Science for Creative Emergence, Kagawa University
[2] Trend Micro, Inc., Tokyo, Japan
shingo_matsugaya@trendmicro.com
[3] Japan Cybercrime Control Center, Tokyo, Japan

Abstract

Phishing attacks targeting Japan have caused growing financial damage, motivating cluster-based analysis as a framework for prioritizing defensive actions. However, the longitudinal validity of such clusters—whether their definitions remain effective as phishing activity evolves—has not been empirically evaluated. We address this gap by applying cluster definitions and YARA rules established from 2023 data to 100,594 domains observed in 2024, and by comparing fixed 2023 rules against periodically updated 2024 operational rules. We find that 60.6% of 2024 domains are classified into the same clusters using only the 2023 rules, but trackability is non-uniform: high-impact clusters remain stable through Q1–Q2 and then degrade rapidly at the Q2–Q3 boundary, with the rule stable rate of high impact clusters dropping by an order of magnitude within a single quarter. The gap between fixed and updated rules splits into three operationally distinct components: rule degradation (18.0%), newly emerged clusters (18.3%), and classification conflicts (3.1%)—of which the first two dominate at comparable scale. Characterizing clusters along two axes—stability category (High/Medium/Low) and temporal pattern (Stable/Degrading/Volatile)—we provide quantitative guidance for scheduling rule maintenance.

Keywords: Phishing; Cluster-based analysis; Longitudinal validation; YARA.

1 Introduction

Phishing attacks in Japan have been causing increasingly serious damage in recent years. According to a report by the National Police Agency, the total financial damage from unauthorized bank transfers related to internet banking reached a record high of approximately 10.4 billion yen in 2025, with phishing identified as the cause of about 90% of these cases. Losses from unauthorized credit card use also remained high, totaling approximately 51 billion yen. Furthermore, in 2025, fraudulent trading in securities accounts attributed to phishing increased sharply, resulting in serious damage of approximately 740.8 billion yen [1].

Despite the scale of these threats, a structured understanding of which threats should be prioritized is not sufficient. While aggregating data by brand or industry clarifies what is being

T.-T. Tsai, H.K. Kim, Y. Wang and A. Yamada (eds.), AsiaJCIS 2026 (EPiC Series in Computing, vol. 111), pp. 49–65

targeted, it is not always clear whether attacks are supported by a wide variety of independent implementations or by the reuse of a small number of shared technical foundations.

To address this challenge, an earlier study [2] proposed a group-based analysis method for phishing sites targeting Japan. By combining source code characteristics, IP addresses, and WHOIS information for classification, that study showed that a small number of high-impact groups account for the majority of observed phishing activity. Furthermore, through a comprehensive evaluation using multiple metrics such as the number of domains, observation days, and rules, the study identified high-impact and trackable groups, demonstrating their effectiveness for prioritizing defensive actions.

However, this analysis was limited to observation data from 2023, and it has not been verified whether the defined groups and YARA rules remain continuously effective against phishing activity that emerges afterward. Phishing activity changes continuously, and the effectiveness of rules may decrease due to updates in source code or changes in infrastructure. If cluster definitions degrade quickly, their practical value for continuous monitoring would be limited. On the other hand, if they remain stable, the framework could serve as a foundation for sustained defensive operations.

This study addresses this gap by adopting an analytical unit equivalent to the groups defined in [2], which we refer to as a *cluster*, and evaluating its longitudinal validity. Specifically, we apply the cluster definitions and YARA rules established from 2023 data to phishing pages observed in 2024, and verify whether classification into the same clusters can be continuously performed. We further quantitatively assess the practical value of rule maintenance by comparing the results obtained using fixed 2023 rules against those obtained through rule updates in 2024 operational use.

This study is guided by the following research questions.

- RQ1: To what extent can the YARA rules of clusters defined in 2023 classify phishing pages observed in 2024 into the same clusters?
- RQ2: Is the trackability of clusters uniform, or do differences in stability, degradation, and conflict arise depending on the cluster? If degradation occurs, what are the timing and proportion?
- RQ3: What does the difference between fixed 2023 rules and periodically updated 2024 operational rules indicate? What practical implications does this difference have for maintaining a cluster-based defensive framework?

Key Findings Through this longitudinal validation, we surface three findings that have not been reported in prior work and that carry direct operational implications: (i) cluster trackability degrades *non-uniformly*—high-impact clusters remain stable through Q1–Q2 and then lose nearly all trackability at the Q2–Q3 boundary, the rule stable rate of high impact clusters dropping by an order of magnitude within a single quarter; (ii) the gap between fixed and updated rules splits into three operationally distinct components—rule degradation (18.0%), new-cluster emergence (18.3%), and classification conflicts (3.1%)—each requiring a different operational response, with the first two dominating at comparable scale; and (iii) cluster stability inversely correlates with the number of YARA rules defining each cluster, suggesting that signature simplicity is a useful proxy for longevity in cluster-based defense.

The main contributions of this study are as follows.

- Longitudinal validation of a cluster-based defensive framework. We demonstrated to what extent clusters defined in 2023 remain trackable in 2024 data.

- Characterization of clusters along two axes: stability category (High/Medium/Low) and temporal pattern (Stable/Degrading/Volatile). We quantitatively measured degradation timing and the proportion of degraded, new, and conflicting domains.
- Quantitative assessment of the practical value of rule updates. We evaluated this based on a comparison between fixed 2023 rules and 2024 operational results.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the methodology, including data collection, cluster definitions, and the longitudinal validation design. Section 4 presents the analysis results for all three research questions. Section 5 discusses the practical implications of these results, including operational guidance for rule maintenance. Section 6 concludes the paper.

2 Related Work

2.1 Phishing Activity Patterns: Lifecycle and Infrastructure

Oest et al. analyzed the end-to-end life cycle of phishing attacks using large-scale real-world data and revealed a concentrated structure in which a small number of campaigns account for approximately 89% of phishing damage [3]. Lee et al. statistically analyzed the lifespan of phishing sites after detection using approximately 280,000 URLs and showed that while many sites are short-lived, a certain number continue to operate for a relatively long period even after detection, suggesting that detection-based countermeasures alone may not be sufficient [4].

From an infrastructure perspective, Luo et al. analyzed phishing emails reaching enterprise environments from the perspective of sending networks and showed that a small number of networks send large volumes of phishing emails over extended periods [5]. Phishing sites have similarly been reported to concentrate in specific hosting providers or autonomous systems [6].

Together, these studies establish that phishing activity exhibits concentration along temporal, volumetric, and infrastructural dimensions. However, they primarily rely on time-based information, URL-level observations, or network-level metadata, and do not treat the implementation structure or code-level characteristics of phishing pages themselves as analytical units. Our study addresses this gap by analyzing phishing pages at the level of structural source-code features, complementing these external indicators with an implementation-centric view.

2.2 Structural and Signature-based Clustering of Phishing Pages

Feng et al. proposed a method for clustering phishing pages by vectorizing and comparing structural similarities, focusing on the DOM structure of phishing web pages [7]. The method demonstrated that pages originating from the same phishing kit can be structurally identified even when they target different URLs or brands.

On the signature side, YARA, originally developed for malware analysis, has been increasingly applied to phishing countermeasures. Leitold presented implementation examples of phishing email and phishing site detection using YARA and reported its effectiveness as a flexible detection method based on pattern descriptions [8]. A publicly available collection of YARA rules covering known phishing kits also provides an example of signature management at the kit level [9].

While both lines of work are effective for detection, most existing studies primarily aim to improve detection accuracy or automated clustering performance, and do not address how the resulting clusters or signatures can be used as reusable analytical units for prioritization and

continuous monitoring in defensive operations. A framework [2] treats YARA-based analytical units, expressing structural features extracted from rendered HTML and JavaScript implementation patterns rather than relying on specific text content or known kits. Our contribution is to evaluate whether such YARA-based analytical units remain valid over time, an aspect not addressed by these prior studies.

2.3 Temporal Validity of Detection Models

Phishing activity continues to evolve, and the temporal validity of detection methods is an important practical challenge. Ejaz et al. visualized through PCA analysis that the feature distribution of phishing site HTML content changes across years, using data collected over three years from 2018 to 2020, and quantitatively demonstrated that the detection accuracy of conventional ML classifiers degrades by up to 20.65% [10]. This result shows that the problem of temporal validity cannot be avoided in phishing detection.

However, this study focuses on the accuracy degradation of ML classifiers as monolithic detectors, and it has not been verified whether specific cluster definitions continue to function as effective analytical units over time. To our knowledge, no prior work has empirically evaluated the longitudinal validity of YARA-based phishing clusters. Our study fills this gap by applying clusters and YARA rules defined in 2023 to 2024 data, and quantifying not only *whether* clusters remain trackable but also *when* and *how* their trackability degrades—providing operationally actionable insights about rule-maintenance scheduling.

3 Methodology

3.1 Data Collection

The dataset used in this study was collected through a phishing site monitoring platform operated in cooperation with a Japanese cybersecurity organization. This platform is a system that automatically crawls target sites based on phishing URLs reported by volunteers and participating organizations, and collects relevant information. The collected information includes observation date and time, URL, redirect URL, IP address, domain, WHOIS information (registrar, registrant, contact email address, etc.), screenshots, site source code, JavaScript files specific to phishing sites, target brand, industry of the target brand, source feature name (rule name), and cluster name.

The 2023 dataset covers phishing site information collected from January to December 2023. After removing duplicates, it consists of 359,009 URLs and 255,929 domains. The 2024 dataset covers phishing site information collected from January to December 2024 in the same manner. After removing duplicates, it consists of 144,687 URLs and 100,594 domains. An overview of both datasets is shown in Table 1. Although the collection period and method are the same for both years, the 2024 dataset is approximately 40% smaller than the 2023 dataset. The impact of this difference on the analysis results is discussed in Section 5.

Table 1: Comparison of 2023 and 2024 Datasets

Dataset	Period	URLs	Domains
2023	Jan.-Dec. 2023	359,009	255,929
2024	Jan.-Dec. 2024	144,687	100,594

Table 2: Correlation Factors Used for Cluster Classification

Factor	Evaluation Method	Primary	Special Handling
IP Address	Hosting IP matches an existing cluster.	✓	CDN IPs excluded; 30-day observation window.
Registrant Email	WHOIS registrant e-mail matches an existing cluster.		Registrar-provided addresses excluded; non-unique values ignored.
Registrant Name	WHOIS registrant name matches an existing cluster.		Registrar names excluded; non-unique values ignored.
Domain	Domain (excluding subdomains) matches an existing cluster.	✓	Dynamic DNS and shared domains excluded.
URL Path	Most URL-path features and transition paths match.		Non-unique values ignored.
Source Code Feature	YARA rules match characteristic source-code strings.	✓	Evaluated with other factors when linked to multiple clusters.
Temporal Observation	Source-code changes observed over time match cluster rules.	✓	Evaluated with other factors when linked to multiple clusters.

3.2 Cluster Definition and Analytical Framework

This study builds on the group-based analysis method proposed in [2] and adopts an equivalent analytical unit, referred to as a *cluster*, whose longitudinal validity is evaluated in this work. The remainder of this subsection (i) defines the cluster, (ii) describes the systematic process for classifying phishing pages into clusters, and (iii) details the procedure for generating the YARA rules that operationalize each cluster.

3.2.1 Cluster Definition

In this study, a cluster is defined as a collection of phishing pages that share common characteristics in source code features, infrastructure information, and operational behavior. Clusters do not assert the identity of a single attacker, but are positioned as reusable analytical units that support prioritization and continuous monitoring in defensive operations.

3.2.2 Cluster Classification Process

Cluster classification follows a systematic evaluation process that evaluates the following seven correlation factors in Table 2.

None of these factors alone serves as a decisive indicator. Each factor is interpreted as either a primary or secondary indicator based on operational experience, and the final classification decision is made based on the overall consistency across multiple factors. If sufficient consistency cannot be confirmed, the page is not forcibly assigned to an existing cluster but is instead treated as a new cluster. This classification process incorporates human judgment and emphasizes explainability and robustness. It is designed to be suitable for large-scale operational analysis in continuously evolving threat environments.

3.2.3 YARA Rule Generation

YARA rules are generated through the following procedure. In this study, common structure extraction refers to the process of identifying implementation-level features that repeatedly

appear across multiple phishing pages, such as HTML structures and JavaScript implementation patterns, while excluding variable elements such as brand names and text originating from legitimate websites.

1. Obtain the source code of the top page of the phishing site in the form of rendered HTML.
2. Selectively retrieve additional JavaScript files loaded by the page. The files to be retrieved are determined based on their file names.
3. Compare the obtained source code with the source code of the legitimate site of the target brand, and extract phishing-specific common structures.
4. From the extracted common structures, select structural features that have high reproducibility as cluster-specific characteristics and do not appear in legitimate sites or other clusters.
5. Construct YARA rules to detect content containing these features. To avoid dependence on a single feature and preserve structural similarity, the default condition requires at least four features to match. However, this requirement may be relaxed when the number of features is limited or when the features are clearly cluster-specific.
6. Apply the generated YARA rules to the source code of the legitimate site of the target brand and confirm that no false positives occur.
7. Confirm that the newly created YARA rules do not misclassify the source code of existing clusters.
8. Deploy the rules as new cluster rules.

Source code was obtained using a WebDriver, and the rendered HTML—produced by serializing the DOM internally constructed by the browser after JavaScript execution into an HTML string—was used as the analysis target. Since many recent phishing sites dynamically generate content using JavaScript, it is necessary to analyze the HTML source code as interpreted by the browser in order to extract phishing-specific common structures. Examples of structural elements targeted by YARA rules are shown below.

- **Resource Layout and Directory Structure**: Load paths and file naming conventions for JavaScript files, CSS, images, and PHP files.
- **URL Path Definitions, Function Names, and APIs**: Code sections defining page transitions, site-specific function names, and API call paths.
- **Development and Testing Artifacts**: Download traces from the target brand site, debug log output code, and code comments.
- **Unique File Names and Display Messages**: Message text intended to make users feel anxious or threatened, and unique resource file names.
- **Frameworks and Services Used**: Names of access analytics services, JavaScript framework names, and library hosting service names.
- **Analysis-Hindering Features**: Empty title tags, anti-crawler service file names, and access control code based on User-Agent or IP address.

3.3 Longitudinal Validation Design

In this study, the following design is used to evaluate the longitudinal validity of clusters defined in 2023.

Validation Design Two types of classification results are used for comparative analysis.

- **Result-A (Fixed 2023 Rules)**: Classification results obtained by applying the YARA rules and cluster definitions established from 2023 data to 2024 data without any modification. No rule updates are made.
- **Result-B (Periodically Updated 2024 Rules)**: Classification results obtained by periodically updating YARA rules during actual 2024 operations. This serves as a baseline reflecting real-world defensive operations.

RQ1 and RQ2 are evaluated primarily based on the results of Result-A, while RQ3 is evaluated through the difference between Result-A and Result-B.

Based on the results of applying Result-A to data classified by Result-B, the stability of each cluster is categorized into the following four types.

- **Type-S (Stable)**: The page was classified into the same cluster by both Result-A and Result-B. This indicates that the rules defined in 2023 remain effective in 2024.
- **Type-D (Degraded)**: The page could not be classified by Result-A, although it was classified by Result-B. This indicates that the effectiveness of the 2023-defined rules has decreased.
- **Type-C (Conflicting)**: The page was classified into a different cluster by Result-A, or was simultaneously classified into multiple clusters. Classification into a single different cluster suggests the possibility of imitation or reuse of implementation by another cluster, but this cannot be confirmed. Identifying the cause of simultaneous classification into multiple clusters requires source code inspection, which is outside the scope of this study. A detailed discussion of Type-C is provided in Section 5.
- **Type-N (New)**: Result-A cluster that newly emerged in 2024. It is naturally expected that such clusters cannot be captured by 2023 rules, and this is distinguished from rule degradation. These are positioned as activities that require the creation of new rules.

The analysis is limited to data classified by Result-B. Data that could not be classified by B is described only at the level of domain counts. This limitation is discussed in Section 5.

The following metrics are calculated for each cluster.

- **Type-S Rate (Stability Rate)**: The proportion of data classified by Result-B that was also classified into the same cluster by Result-A (the proportion of Type-S). Used to answer RQ1.
- **Type Distribution**: The distribution of Type-S/D/C/N for each cluster. Used to answer RQ2.
- **Gap between Result-A and Result-B**: The number of domains and their proportion corresponding to Type-D, Type-C, and Type-N. Used to answer RQ3. Note that Type-N is attributed to newly emerged clusters rather than rule degradation, and is therefore interpreted separately from Type-D and Type-C.

- **Monthly Type-S Rate Trend**: The monthly Type-S rate is introduced to track temporal changes in rule degradation. Used as a supplementary indicator for RQ3.

4 Analysis Results

4.1 Overview of 2024 Dataset

Table 1 provides an overview of the 2024 dataset. The 2024 dataset consists of 144,687 URLs and 100,594 domains, which is approximately 40% of the size of the 2023 dataset (359,009 URLs and 255,929 domains). The collection period and method are the same for both years, and the impact of this difference in scale on the analysis results is discussed in Section 5. In this study, 100,594 domains classified by B are used as the analysis target.

Table 3: Summary of Evaluation Metrics

Metric	Definition	RQ
Type-S Rate	Type-S rate among B-classified domains	RQ1
Type Distribution	Distribution of Type-S/D/C/N for each cluster	RQ2
Gap between Result-A & B	Type-D/C/N total proportion	RQ3
Monthly Type-S Rate Trend	Time-series change of monthly Type-S rate	RQ3 (Supp.)

4.2 RQ1: Stability of Cluster Classification

The analysis results for RQ1 are presented below. Table 4 shows the results of applying Result-A to the 100,594 domains classified by Result-B.

Type-S accounts for 60,912 domains (60.6%), confirming that the YARA rules defined in 2023 can classify approximately 60% of the 2024 data into the same clusters. Type-D accounts for 18,094 domains (18.0%), representing activity that belongs to known 2023 clusters but could not be classified using the fixed 2023 rules. This suggests that the implementation structure of phishing sites has changed and the effectiveness of existing rules has decreased. Type-C accounts for 3,153 domains (3.1%), representing cases where the classification results of Result-A and Result-B did not match; this is discussed in detail in Section 5. Type-N accounts for 18,435 domains (18.3%), attributed to clusters that newly emerged in 2024. This differs in nature from rule degradation and indicates that phishing activity that did not exist in 2023 has appeared.

These results confirm that the YARA rules defined in 2023 retain a certain level of reproducibility when applied to 2024 data. On the other hand, the presence of Type-D (18.0%) and Type-C (3.1%) suggests the need for periodic rule updates. Furthermore, Type-N (18.3%) indicates that new rules need to be created to address newly emerging phishing activity.

Table 4: Distribution of the Four Categories

Category	Definition	Number of Domains	Proportion (%)
Type-S	Result-A = Result-B match	60,912	60.6%
Type-D	Result-A: no match, Result-B: matched	18,094	18.0%
Type-C	Result-A $\neq$ Result-B	3,153	3.1%
Type-N	Newly emerged cluster in 2024	18,435	18.3%
Total		100,594	100%

4.3 RQ2: Stability Differences Across Clusters

Cluster-level Type Distribution Table 5 shows the Type-S/D/C distribution for each cluster. The analysis targets the major clusters reported in [2]. The results show that cluster stability is not uniform, and clusters can be classified into the following three categories.

High Stability (Type-S 80% or above): BP1_(MASKED), BP2, BP6, CP31, IP4, and MP5 are classified into this category. These clusters demonstrate that the YARA rules defined in 2023 remain largely effective in 2024, functioning as stable analytical units for continuous monitoring. Note that BP1_(MASKED) has been masked because its name contains the target brand name. Furthermore, based on the analysis results, this cluster was determined to be unrelated to BP1, which belongs to the medium stability category.

Medium Stability (Type-S 55–80%): BP1 and CP13 are classified into this category. BP1, the largest cluster, shows a certain level of stability with a Type-S rate of 70.7%, but the presence of Type-D at 27.4% indicates that partial rule degradation has occurred. CP13 is a small-scale cluster and is less stable than BP1, with a Type-S rate of 56.3%. In contrast to BP1, no degradation was observed, but rule conflicts were prominent, with a Type-C rate of 43.7%.

Low Stability (Type-S below 55%): CP20, CP29, and CP30 are classified into this category. All clusters show instability with degradation rates of 34% or higher. CP30 shows the most prominent degradation with a Type-D rate of 31.2%, while conflicts remained below 1%. Only CP29 showed some degree of conflict, with a Type-C rate of 10.7%.

On the other hand, among the other known clusters defined in [2] (totaling 16,521 domains, shown in the aggregated row of Table 5), approximately 14.0% of domains fall into the low-stability category, while the remainder are mostly stable. Additionally, no phishing activity corresponding to UKN344 was observed during the 2024 observation period.

Table 5: Cluster-level Type-S/D/C Distribution

Cluster	Total Domains	Type-S Rate(%)	Type-D Rate(%)	Type-C Rate(%)
BP1	42,769	70.7%	27.4%	1.9%
BP6	8,205	93.4%	6.6%	0.0%
CP30	3,088	31.2%	68.8%	0.1%
IP4	3,129	92.3%	7.7%	0.0%
CP29	2,358	54.9%	34.4%	10.7%
CP20	1,170	51.0%	48.1%	0.9%
MP5	1,441	93.7%	6.2%	0.1%
CP31	602	93.7%	0.0%	6.3%
CP13	494	56.3%	0.0%	43.7%
BP2	187	98.4%	1.6%	0.0%
BP1_(MASKED)	2,195	95.9%	1.3%	2.8%
Other known (aggregated)	16,521	77.3%	12.1%	10.7%

† Aggregates other clusters from [2] not individually analyzed; UKN344 had no 2024 observations.

‡ Newly emerged 2024 clusters (18,435 domains) appear as Type-N in Table 4.

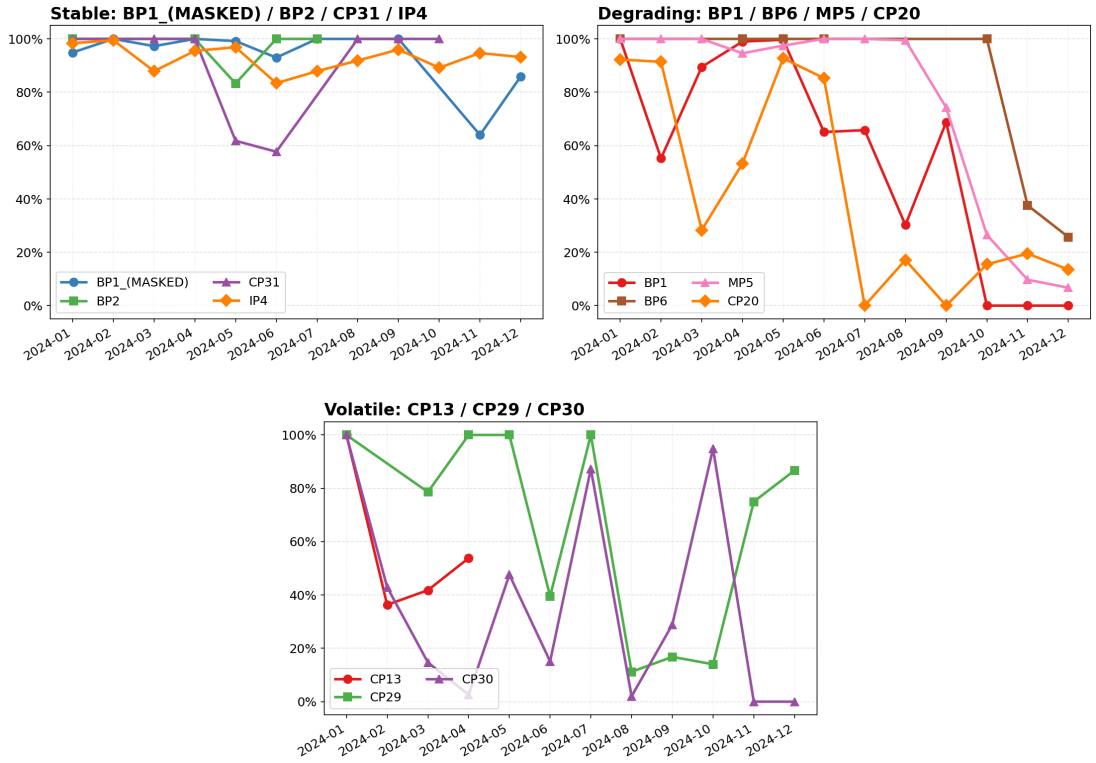
Analysis of Monthly Type-S Rate Trends Figure 1 shows the monthly Type-S rate trends for the analyzed clusters. Data points with fewer than 10 observed domains in a given month are excluded to focus on reliable trends. The results show that cluster stability varies considerably over time, and the clusters can be classified into the following three patterns based on their trends.

Stable: BP1_(MASKED), BP2, CP31, and IP4 are classified into this pattern. These clusters maintained a high Type-S rate throughout the year without a sustained downward

trend. CP31 is a notable case in this category: it experienced temporary drops in May (61.8%) and June (57.7%) accompanied by elevated Type-C rates (38.2% and 42.3%), but recovered in subsequent months. We retain CP31 in the Stable pattern because no sustained degradation or volatility trend was observed; this case nevertheless illustrates that even Stable clusters can experience transient conflict-driven dips, and the Stable label captures the absence of *long-term* trends rather than the absence of short-term variability.

Degrading: BP1, BP6, MP5, and CP20 are classified into this pattern. Like the Stable pattern, these clusters are initially stable, but their Type-S rate drops sharply after a certain period. MP5 began to decline in September, and BP6 in November. In both cases, the Type-C rate remained at 0% while the Type-D rate increased, indicating progressive rule degradation. BP1 shows greater monthly variation compared to BP6 and MP5, but its Type-S rate generally began to decline from June, reaching 0% from October onward and becoming untrackable. Similar to BP6 and MP5, an increase in Type-D rate was observed, except in December, where the Type-C rate was exceptionally 26.9%, indicating a conflicting state. CP20 experienced a temporary drop in March but remained largely stable through the first half of the year. From July onward, it declined to near 0% and showed a degradation trend with no recovery.

Volatile: CP13, CP29, and CP30 exhibit this pattern, with the nature of instability differing across clusters. CP13 and CP29 experienced periods where the conflict rate exceeded 45%, whereas CP30 showed intermittently high degradation rates throughout the year.



† Type-S rate is calculated excluding data points with fewer than 10 observed domains.

Figure 1: Monthly Trend of the Type-S Rate

As shown in Table 6, which summarizes the correspondence between stability category and temporal pattern for each cluster, cluster stability varies significantly depending on the

cluster. Stability changes over the observation period, and even clusters belonging to the high stability category showed degradation and instability in the second half of the year, while some clusters exhibited large monthly variations in stability. The cause of instability differs by cluster, whether it is due to rule degradation or rule conflicts.

Table 6: Cluster Category Mapping

Cluster	Type-S	Category	Trend
BP2	98.4%	High	Stable
BP1_(MASKED)	95.9%	High	Stable
CP31	93.7%	High	Stable
IP4	92.3%	High	Stable
MP5	93.7%	High	Degrading
BP6	93.4%	High	Degrading
BP1	70.7%	Medium	Degrading
CP13	56.3%	Medium	Volatile
CP29	54.9%	Low	Volatile
CP20	51.0%	Low	Degrading
CP30	31.2%	Low	Volatile

Degradation Timing Among clusters classified as Degrading, the high-impact clusters BP1 and CP20 showed degradation from Q3 onward, while BP6 and MP5 showed degradation from Q4 onward (no observations were recorded for BP6 in Q3). As shown in Table 7, most of these clusters became untrackable by Q4. The implications of this timing difference for maintenance scheduling are discussed in Section 5.1.

Table 7: Quarterly Type-S Rate of Degrading Clusters

Cluster	Q1	Q2	Q3	Q4
BP1	85.2%	86.7%	51.5%	0%
CP20	80.3%	86.0%	6.8%	16.2%
BP6	100%	100%	N/A	44.4%
MP5	100%	97.8%	97.9%	16.3%

Cluster Conflicts The 3.1% of domains classified as Type-C reflect cases where Result-A and Result-B disagree on cluster assignment. Among the analyzed clusters, CP13, CP29, and CP31 exhibit the most prominent conflicts. Figure 2 visualizes these conflicts between Result-A and Result-B. CP30, CP31, and CP13 share many common features and frequently conflict with each other; MP5, also included in the analysis, shows frequent conflicts with CP31 and CP30; and CP29 frequently conflicts with BP1. The operational implications of these conflicts are discussed in Section 5.2.

4.4 RQ3: Gap between Fixed and Operational Rules

This subsection quantifies the gap between Result-A (fixed 2023 rules) and Result-B (periodically updated 2024 rules), thereby characterizing where rule maintenance contributes value in operational defense.

As shown in Table 4, the gap between Result-A and Result-B totals approximately 39% of all observed domains. This gap consists of three qualitatively different components, each calling for a distinct operational response. The first is the *degradation gap*, comprising Type-D (18.0%): domains that belong to known 2023 clusters but cannot be classified by Result-A because the characteristic values of phishing sites have shifted. Recovering trackability for these domains requires updating the signatures of existing rules. The second is the *new-emergence*

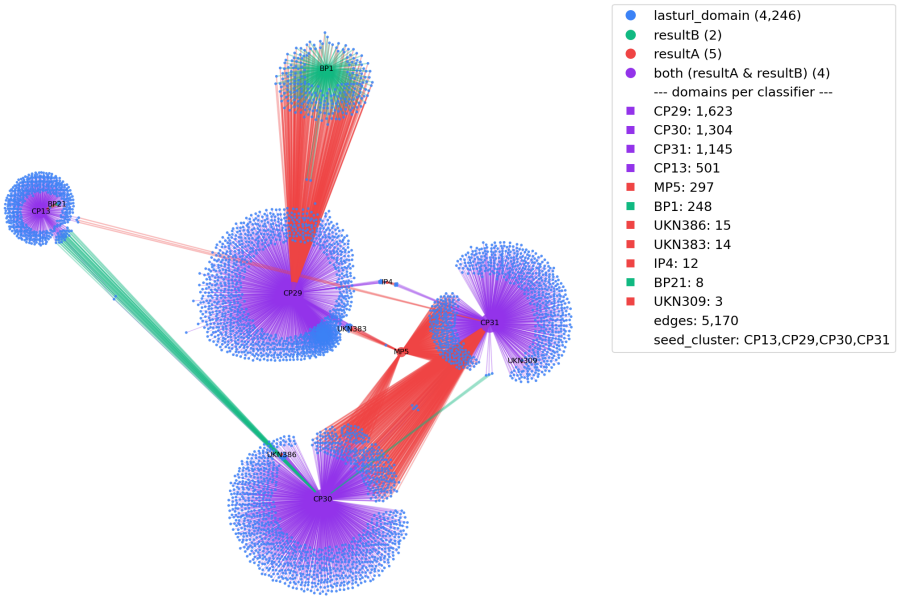


Figure 2: Conflict Clusters Similarity

gap, comprising Type-N (18.3%): domains that belong to clusters that did not exist in 2023 and therefore cannot be captured by Result-A at all. Recovering trackability here requires the creation of new rules rather than updates to existing ones. The third is the *conflict gap*, comprising Type-C (3.1%): although smaller in proportion, this component is qualitatively distinct from the previous two. Detection itself succeeds, but the classification is inconsistent—a domain is assigned to a different cluster or to multiple clusters simultaneously by Result-A. The operational response is therefore neither signature update nor new rule creation, but *cluster-level review*: deciding whether two rules should be merged, whether one should be removed, or whether the cluster boundaries themselves should be redrawn.

It is noteworthy that Type-D and Type-N are of approximately the same scale, meaning that the detection-recovery burden distributes almost equally between maintaining existing rules and authoring new ones. Type-C is smaller in scale but requires distinct attention because it reflects classification ambiguity rather than detection failure. Combined with the fact that approximately 60% of domains can be stably tracked without any rule maintenance, this indicates that the cluster-based defensive framework maintains a certain level of effectiveness as an analytical foundation, provided that ongoing maintenance addresses all three operational streams.

Trackability of High-Impact Clusters To characterize how the gap manifests for the most operationally significant clusters, we examine the high-impact clusters defined in [2], where clusters are classified into four impact ratings, S, A, B, and C (S: BP1/CP20, A: CP31, B: CP29/IP4, C: BP2/CP13/CP30/MP5/BP6). BP1 and CP20, classified as the highest-impact rating S, achieved Type-S rates of 70.7% and 51.0% respectively, but were not in a consistently stable state throughout the year. The quarterly Type-S rates in Table 8 reveal that both clusters remained stably trackable through Q2, but degraded rapidly thereafter: CP20 dropped from 86.0% in Q2 to 6.8% in Q3, and BP1 dropped from 86.7% to 51.5% over the same transition. CP31, classified as rating A, remained stably trackable throughout 2024.

Table 8: Quarterly Type-S Rate of High-Impact Clusters

	Q1	Q2	Q3	Q4
BP1 Domains	7295	19054	14856	1717
BP1 Type-S	6127	16524	7655	0
BP1 Type-S(%)	85.2%	86.7%	51.5%	0.0%
CP20 Domains	295	351	294	235
CP20 Type-S	237	302	20	38
CP20 Type-S(%)	80.3%	86.0%	6.8%	16.2%

These observations indicate that S- and A-rated high-impact clusters can be tracked using fixed 2023 rules through the first half of the year; once Q3 begins, however, S-rated clusters can lose trackability rapidly within a single quarter. The practical implications of this Q2–Q3 transition for operational rule maintenance are discussed in Section 5.1.

5 Discussion

5.1 Practical Implications for Rule Maintenance

Section 4.4 shows that the cluster-based defensive framework requires rule maintenance on roughly a semiannual schedule. The Q2–Q3 boundary observed for S-rated clusters should be interpreted not as a gradual decay but as a relatively sharp transition: until the end of Q2 the framework operates with high accuracy on its most critical targets, and after this point the most impactful clusters can lose nearly all trackability within a single quarter. This pattern suggests that rule updates targeting S-rated clusters should be scheduled before the end of Q2 to avoid sudden capability gaps, rather than being performed reactively after degradation has been observed.

Tiered Maintenance Scheduling The Q2–Q3 boundary observed for S-rated clusters does not apply uniformly. As shown in Section 4.3, lower-impact clusters with a Degrading pattern (BP6, MP5) tend to lose trackability later in the year, from Q4 onward. This staggered timing supports a tiered maintenance schedule: S-rated clusters require focused updates before the Q2–Q3 boundary, while medium- and lower-impact Degrading clusters can be addressed in a consolidated end-of-year update cycle. Aligning maintenance effort with both impact and observed degradation timing reduces overall operational burden without compromising coverage of the most consequential clusters.

Case of BP1: The Concrete Cost of Delayed Updates BP1, the largest cluster in the analyzed dataset, was reported to account for 80.3% of observed domains in 2023 [2]. In 2024, this proportion decreased by 37.8 percentage points to 42.5%, yet BP1 remained the largest cluster. As shown in Table 5, approximately 30% of its domains could not be captured using the fixed 2023 rules. The case of BP1 makes the aggregate Type-D figure concrete: without rule updates, the largest single source of phishing activity targeting Japan would have lost roughly one third of its coverage. This illustrates both the necessity of rule updates to close such gaps and the practical value that updates provide on the most consequential clusters.

Operational Priorities Combined with the gap decomposition in Section 4.4, these observations suggest a three-tier maintenance strategy. First, periodic updates targeting S-rated

clusters should be performed before the Q2–Q3 boundary to maintain coverage of the highest-impact phishing activity. Second, monitoring for newly emerged clusters should be continuous, since Type-N accounts for a comparable share of the gap and some new clusters may grow into high-impact ones. Third, conflicts (Type-C) call for a qualitatively different response: since detection itself succeeds but classification is inconsistent, the appropriate action is cluster-level review—deciding whether conflicting rules should be merged, removed, or refined to eliminate overlap between clusters—rather than signature update or new rule authoring. This prioritization identifies clusters requiring maintenance; given that approximately 60% domains are correctly tracked without additional effort, it reduces maintenance costs and enables resources to be focused on areas with the highest operational impact.

The Six-Month Boundary as a Working Hypothesis Our observations support, but do not prove, the following working hypothesis: *for phishing clusters defined by structural features, the half-life of a fixed cluster definition is on the order of six months, with degradation tending to manifest as a sharp transition rather than a gradual decline.* This hypothesis is consistent with the Q2–Q3 inflection observed for S-rated clusters, the Q3–Q4 inflection observed for the Degrading clusters BP6 and MP5, and the inverse relationship between rule count and longitudinal stability discussed in Section 5.2. Validating this hypothesis at greater scale—across additional years of data and against phishing activity targeting other regions—would clarify whether the six-month boundary is specific to the clusters analyzed here or a more general regularity of cluster-based phishing defense.

5.2 Cluster Characteristics and Stability

Correlation Between Rule Count and Stability The longitudinal stability observed in this study varies markedly across clusters, and one factor that appears to correlate with this variation is the number of YARA rules defining a cluster. Using standardized rule-count scores reported in [2] (where higher scores indicate clusters defined by more YARA rules), high-stability clusters have an average score of 53.3, whereas medium- and low-stability clusters have an average of 61.7—a difference of 8.4 points. This indicates a possible inverse relationship between the number of rules defining a cluster and its longitudinal stability: clusters described by fewer, more focused signatures appear to remain trackable longer than those defined by larger rule sets. One possible explanation is that clusters with more uniform implementation structures require fewer signatures to capture, and these uniform structures change less frequently over time.

Implications of Cluster Conflicts The 3.1% of domains classified as Type-C, although small in aggregate, raise operational questions that go beyond their numerical share. Several conflict pairs identified in Section 4.3 link high-impact and lower-impact clusters (notably CP29 with BP1, and CP30/CP31/CP13 with each other), suggesting that some lower-impact clusters may represent imitative or derivative versions of high-impact ones—or vice versa. Whether such conflicts arise from genuine structural similarity (suggesting cluster consolidation) or from rule overspecification (suggesting rule revision) cannot be determined without source-code-level inspection, which lies outside the scope of this study. Nonetheless, the conflict patterns observed here identify candidate cluster pairs for which targeted re-examination would have the highest operational value, and inform prioritization in cluster-level audits.

The Two-Axis Characterization Framework Combining the stability category (High/Medium/Low) and the temporal pattern (Stable/Degrading/Volatile) yields a two-axis characterization of cluster behavior, summarized in Table 6. This framework is more informative than either dimension alone: a cluster's annual Type-S rate captures its overall trackability, while its temporal pattern captures whether and when that trackability changes. For instance, BP1 and CP20 occupy different stability categories yet share a Degrading temporal pattern with the same Q2–Q3 inflection point, which pure stability scoring would obscure. We therefore offer this two-axis view as a generalizable lens for evaluating cluster-based detection schemes against longitudinal data, applicable beyond the specific clusters analyzed here.

5.3 Limitations and Constraints

Limitations of Data Collection The 2024 dataset used in this study was collected using the same method and over the same period as the 2023 dataset, but is approximately 40% smaller in scale (2023: 255,929 domains, 2024: 100,594 domains). This difference in scale may be due to changes in phishing activity itself or biases in data collection, but identifying the cause is outside the scope of this study. As an impact of this scale difference on the analysis, the reliability of the monthly Type-S rate may decrease for clusters with a small number of observations. In this study, this impact was partially mitigated by excluding months with fewer than 10 observed domains from the plots.

Methodological Constraints This study verified whether data classified into any cluster by Result-B could be tracked using fixed 2023 rules. Therefore, data that was unclassified during 2024 operations is outside the scope of this study. It is possible that some of this unclassified data may actually belong to the clusters analyzed in this study. However, since the focus of this study is to analyze the operational value of the cluster-based defensive framework by comparing results as Type-S/D/C/N against those obtained using 2024 operational rules, the impact of excluding unclassified data on the direction of the analysis is considered limited.

Scope of Analysis The purpose of this study is to verify the trackability of high-impact clusters. Therefore, identifying clusters newly observed in 2024 and analyzing their impact are outside the scope of this study. In addition, the analysis targets only the clusters reported in [2] in order to enable comparison with 2023. In this study, newly emerged clusters and other known clusters are presented collectively to suggest the extent of their impact. The findings of this study are positioned as an initial empirical validation of the trackability of high-impact clusters in phishing activity targeting Japan. Verifying the applicability to other regions and analyzing the impact of newly emerged clusters are left as future work.

Threats to Validity Beyond the methodological constraints noted above, three further threats deserve mention. First, clusters labeled as newly emerged in 2024 (Type-N) may include clusters that existed in 2023 but were too small to be defined as analytical units at the time; we cannot fully distinguish genuine emergence from delayed recognition without finer-grained 2023 data. Second, the observed rule-degradation patterns could in principle reflect either real changes in phishing implementations or systematic shifts in upstream data collection; the consistent direction of the Q2–Q3 transition across clusters argues against the latter, but we cannot fully exclude it. Third, our analysis treats Result-B—which is itself produced through human-in-the-loop rule maintenance—as the reference baseline, so any classification errors in Result-B propagate into the Type-D and Type-C categories. We have mitigated this risk by constraining

the analysis to clusters already defined in 2023, but a fully ground-truthed evaluation remains future work.

6 Conclusion

The following findings were obtained through the research questions based on the results above.

- RQ1: Applying the YARA rules defined in 2023 to 2024 data, 60.6% of domains were classified into the same clusters, confirming a certain level of reproducibility.
- RQ2: Cluster stability was not uniform. Clusters fell into three Type-S–rate categories (High: ≥80%, Medium: 55–80%, Low: <55%) and three temporal patterns (Stable, Degrading, Volatile). Furthermore, high-stability clusters were defined by relatively fewer YARA rules than medium- and low-stability clusters (a difference of 8.4 points in their relative rule-count score), suggesting that more focused signatures decay more slowly.
- RQ3: High-impact clusters can be tracked stably for approximately six months (Q1–Q2) without rule updates; thereafter, S-rated clusters can lose trackability rapidly, dropping by an order of magnitude within a single quarter. Across the full year, 18.0% of domains exhibit degradation requiring rule updates, with degradation concentrated in the second half. Capturing newly emerged high-impact clusters requires creating new rules covering an additional 18.3% of domains, and resolving potential conflicts between high- and low-impact clusters requires reviewing a further 3.1%.

This study demonstrated that the cluster-based defensive framework is capable of tracking high-impact clusters to a certain degree. In addition, quantitative findings were obtained regarding the timing of rule degradation and the considerations for rule updates.

- Longitudinal validation of the cluster-based defensive framework. We demonstrated to what extent clusters defined in 2023 remain trackable in 2024 data.
- We propose a two-axis cluster characterization framework—combining annual stability category (High/Medium/Low) with temporal pattern (Stable/Degrading/Volatile)—that captures both overall trackability and the timing of degradation in a single view. The framework reveals that clusters with similar annual stability can exhibit qualitatively different temporal behaviors, a distinction obscured by single-axis evaluation. We argue that this framework offers a generalizable lens for evaluating cluster-based detection schemes against longitudinal data, applicable beyond the specific clusters analyzed here.
- Quantitative assessment of the practical value of rule maintenance, decomposing the gap between fixed and operational rules into three operationally distinct components—signature degradation (Type-D), new-cluster emergence (Type-N), and classification conflicts (Type-C)—each calling for a different response (rule update, new rule authoring, and cluster-level review respectively). This decomposition translates directly into a tiered maintenance schedule keyed to the observed Q2–Q3 boundary.

The operational methods for stably maintaining the trackability of high-impact clusters identified in this study are left as future work.

- Analysis of the impact of newly emerged clusters and quantitative evaluation of their transition into high-impact clusters over time.

- Verification of the six-month boundary as a working hypothesis through large-scale data analysis over multiple years.
- Verification of applicability to other regions.

Acknowledgments

We acknowledge the use of Claude AI assistant (Anthropic) for the English translation of this paper. This work was supported in part by the JSPS/MEXT KAKENHI under Grant 24K14956.

References

[1] National Police Agency of Japan, Cyber Police Bureau. Cybersecurity Threat Landscape in Japan, 2025. https://www.npa.go.jp/publications/statistics/cybersecurity/data/R7/R07_cyber_jousei.pdf, March 2026. Published March 2026 (Japanese Government Report).

[2] Shingo Matsugaya, Makoto Shimamura, Kosuke Takeshige, Keisuke Sakai, and Masaki Hashimoto. Characteristic analysis and attack group identification of phishing sites targeting japan. In Leonard Barolli, editor, *Advances in Internet, Data and Web Technologies*, pages 65–76, Cham, 2025. Springer Nature Switzerland.

[3] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 361–377. USENIX Association, August 2020.

[4] Kiho Lee, Kyungchan Lim, Hyoungshick Kim, Yonghwi Kwon, and Doowon Kim. 7 days later: Analyzing phishing-site lifespan after detected. *Proceedings of the ACM on Web Conference 2025*, 2025.

[5] Elisa Luo, Liane Young, Grant Ho, M. H. Afifi, Marco Schweighauser, Ethan Katz-Bassett, and Asaf Cidon. Characterizing the networks sending enterprise phishing emails. In *Passive and Active Measurement*. Springer Nature Switzerland, 2025.

[6] Cybercrime Information Center. Phishing activity in hosting networks: May–april 2025. https://www.cybercrimeinfocenter.org/phishing-activity-in-hosting-networks-may-april-2025, 2025. Accessed May 2026.

[7] Jian Feng, Yuqiang Qiao, Ou Ye, and Ying Zhang. Detecting phishing webpages via homology analysis of webpage structure. *PeerJ Computer Science*, 8:e868, February 2022.

[8] Ferenc Leitold. Utilizing yara: An effective method for phishing attack response. In *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media (CyberSciTech 2024)*, Springer Proceedings in Complexity, pages 195–205. Springer, 2025.

[9] t4d. Phishingkit-yara-rules: Repository of yara rules dedicated to phishing kits. https://github.com/t4d/PhishingKit-Yara-Rules, 2019. Accessed May 2026.

[10] Asif Ejaz, Adnan Noor Mian, and Sanaullah Manzoor. Life-long phishing attack detection using continual learning. *Scientific Reports*, 13:11488, jul 2023.