Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

ANALISIS KARAKTERISTIK URL PHISHING DAN LEGITIMATE SEBAGAI INDIKATOR ANCAMAN KEAMANAN SIBER MENGGUNAKAN DATASET PUBLIK

Nasywa Azzahra[1*], Taufiq [2], Yanita [3]
[1,2,3]Universitas Malikussaleh, Indonesia
[1*]nasywa.240170225@mhs.unimal.ac.id, [2]Taufiq.te@unimal.ac.id, [3]yanita@unimal.ac.id

ABSTRACT

Phishing URLs can imitate legitimate services and expose users to credential theft. This study compared lexical and structural characteristics of phishing and legitimate URLs using the public PhiUSIIL dataset, which documents 100,945 phishing and 134,850 legitimate URLs. URL strings were parsed offline without visiting their destinations. Eight features were examined: URL length, hostname length, digit proportion, hyphen count, hostname dot count, IP-based hosts, HTTPS scheme, and query presence. After removing 425 duplicate records, 235,370 URLs were analyzed using medians, interquartile ranges, proportions, Mann–Whitney tests, rank-biserial effect sizes, and Holm correction. Phishing URLs showed larger digit proportion (r = 0.45), URL length (r = 0.44), hyphen count (r = 0.30), and host length (r = 0.25), whereas hostname dot count differed negligibly (r = −0.03). HTTPS appeared in 48.6% of phishing URLs versus 100% of legitimate URLs; IP-based hosts and query strings occurred only among phishing URLs. Because all legitimate URLs were HTTPS homepage addresses without paths or queries, these contrasts partly reflect how the classes were built. The features are candidate warning indicators, not deterministic rules.

Keywords:
phishing URL, legitimate URL, lexical features, cybersecurity, public dataset

PENDAHULUAN

Perkembangan layanan digital menjadikan tautan (*uniform resource locator*, URL) sebagai pintu masuk utama menuju layanan perbankan, perdagangan elektronik, media sosial, dan sistem akademik. Kondisi ini dimanfaatkan dalam serangan phishing, yaitu upaya menipu pengguna melalui identitas, pesan, atau halaman yang menyerupai pihak tepercaya untuk memperoleh informasi sensitif. Serangan tersebut dapat berlangsung melalui surel, situs web, perangkat bergerak, dan media sosial, serta terus berubah mengikuti kebiasaan pengguna dan perkembangan teknologi (Aleroud & Zhou, 2017; Chiew et al., 2018; Alabdan, 2020). Karena tautan hampir selalu menjadi sarana yang mengarahkan korban ke halaman tiruan, pengamatan terhadap URL menjadi bagian penting dari analisis ancaman siber.

URL memuat skema, host, jalur, kueri, dan fragmen yang dapat dihitung langsung dari string tanpa membuka halaman tujuan. Karakteristik tersebut, misalnya panjang string, jumlah digit, tanda hubung, struktur hostname, dan penggunaan HTTPS, telah dimanfaatkan dalam banyak studi deteksi phishing, baik melalui fitur yang dirancang manual maupun representasi yang dipelajari model (Jain & Gupta, 2018; Sahingoz et al., 2019; Aljofey et al., 2020; Butnaru et al., 2021). Namun, sebagian besar studi tersebut berfokus pada kinerja klasifikasi. Pertanyaan yang lebih mendasar, yaitu seberapa besar dan seberapa konsisten fitur-fitur itu membedakan URL phishing dari URL legitimate, jarang dilaporkan secara deskriptif dan transparan.

Pertanyaan tersebut penting karena hasil analisis sangat bergantung pada cara dataset dibangun. Dataset benchmark publik umumnya menggabungkan URL phishing dari umpan ancaman dengan URL legitimate dari daftar atau sumber terkurasi, sehingga struktur kelas legitimate dapat sangat berbeda dari lalu lintas web yang sebenarnya (Hannousse & Yahiouche, 2021; Prasad & Chandra, 2023). Dataset LegitPhish, misalnya, mengambil URL legitimate dari situs seperti Wikipedia dan Stack Overflow (Potpelwar et al., 2025). Perbedaan fitur antarkelas karenanya perlu dibaca bersama bias pengumpulan datanya, bukan semata sebagai sifat umum URL phishing.

Penelitian ini menggunakan dataset PhiUSIIL Phishing URL (Website) yang memuat 235.795 URL berlabel (Prasad & Chandra, 2024a, 2024b). Tujuannya adalah membandingkan delapan karakteristik leksikal dan struktural URL phishing dan legitimate secara deskriptif dan inferensial, tanpa membangun model klasifikasi. Pertanyaan penelitian adalah: (1) bagaimana perbedaan karakteristik URL phishing dan legitimate berdasarkan delapan fitur yang dipilih; (2) fitur mana yang menunjukkan perbedaan terbesar dan seberapa besar distribusinya tumpang tindih; dan (3) batas apa yang perlu diperhatikan ketika fitur tersebut dijadikan indikator awal ancaman, termasuk akibat bentuk URL pada kelas legitimate.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

Kontribusi penelitian ini ada tiga. Pertama, prosedur analisis komparatif yang transparan dan dapat direproduksi atas string URL. Kedua, pelaporan ukuran efek dan tumpang tindih distribusi, bukan hanya nilai p. Ketiga, pemeriksaan eksplisit terhadap bentuk URL pada kelas legitimate (jalur, awalan www, kueri, dan skema) yang menunjukkan bahwa sebagian perbedaan antarkelas mencerminkan cara dataset dibangun. Bagian selanjutnya memuat kajian literatur, metode, hasil dan pembahasan, serta kesimpulan.

TINJAUAN PUSTAKA

Phishing sebagai Ancaman Keamanan Siber

Phishing adalah kategori serangan rekayasa sosial yang memanfaatkan kepercayaan pengguna terhadap merek, organisasi, atau individu. Aleroud dan Zhou (2017) memetakan lingkungan, teknik, dan tindakan penanggulangan phishing, sedangkan Chiew et al. (2018) dan Alabdan (2020) mensurvei jenis, vektor, dan pendekatan teknis serangan tersebut. Tinjauan sistematis tentang deteksi phishing pada situs web menunjukkan bahwa pendekatan, fitur, dan karakteristik dataset sangat bervariasi sehingga hasil antarstudi perlu dibandingkan dengan hati-hati (Safi & Singh, 2023).

Struktur URL dan Fitur Leksikal

Sebuah URL terdiri atas skema (misalnya http atau https), otoritas yang memuat host, jalur sumber daya, kueri, dan fragmen (Potpelwar et al., 2025). Fitur leksikal dihitung langsung dari string tersebut, misalnya panjang URL, panjang host, jumlah titik, tanda hubung, digit, dan simbol khusus. Fitur semacam ini murah dihitung dan tidak memerlukan akses ke halaman tujuan. PHISH-SAFE mengembangkan deteksi berbasis fitur URL dengan pembelajaran mesin (Jain & Gupta, 2018). Sahingoz et al. (2019) membandingkan tujuh algoritma klasifikasi dengan fitur berbasis pemrosesan bahasa alami dan melaporkan akurasi sekitar 98% dengan Random Forest, sedangkan Butnaru et al. (2021) mengusulkan deteksi ringan berbasis URL. Dalam penelitian ini, istilah host merujuk pada hostname hasil parsing dan tidak disamakan dengan domain terdaftar atau jumlah subdomain.

Pendekatan Berbasis URL dalam Pembelajaran Mesin dan Pembelajaran Mendalam

Selain fitur yang dirancang manual, representasi URL dapat dipelajari langsung dari karakter. URLNet mempelajari representasi karakter dan kata dengan jaringan konvolusional untuk mendeteksi URL berbahaya (Le et al., 2018), dan Aljofey et al. (2020) mengusulkan model CNN tingkat karakter yang melaporkan akurasi 95,02% pada dataset mereka serta 95,22% hingga 98,58% pada tiga dataset pembanding. Pendekatan yang memperluas cakupan ke informasi hyperlink halaman (Jain & Gupta, 2019) atau ke kombinasi fitur keamanan yang beragam seperti kerangka PhiUSIIL (Prasad & Chandra, 2024b) memerlukan akses ke isi halaman atau layanan eksternal. Studi Random Forest yang lebih baru, PhishGuard+, melaporkan akurasi 98,4% pada 10.000 URL berimbang dengan 48 fitur (Srivastava et al., 2026). Evaluasi pada satu dataset berukuran kecil belum menunjukkan ketahanan terhadap pergeseran distribusi, sehingga angka tersebut tidak dapat dibandingkan langsung dengan studi lain. Ringkasan studi yang relevan disajikan pada Tabel 1.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

Tabel 1. Ringkasan studi terdahulu yang relevan

Studi	Fokus dan data	Pendekatan	Relevansi bagi penelitian ini
Jain & Gupta (2018)	Deteksi phishing berbasis fitur URL	Fitur URL dan pembelajaran mesin	Fitur URL dapat dihitung tanpa membuka halaman
Sahingoz et al. (2019)	Deteksi phishing dari URL; tujuh algoritma	Fitur berbasis NLP; Random Forest (±98%)	Landasan pemanfaatan fitur leksikal URL
Le et al. (2018)	Deteksi URL berbahaya	CNN karakter dan kata (URLNet)	Alternatif tanpa fitur manual
Aljofey et al. (2020)	Phishing berbasis URL; dataset sendiri dan tiga dataset pembanding	CNN tingkat karakter (akurasi 95,02%–98,58%)	Membandingkan fitur manual dan *embedding* karakter
Butnaru et al. (2021)	Deteksi phishing ringan berbasis URL	Fitur URL	Mendukung analisis URL yang ringan
Hannousse & Yahiouche (2021)	Skema dataset benchmark yang dapat direproduksi; 87 fitur	Replikasi konseptual; Random Forest paling prediktif	Menekankan kualitas dan komposisi dataset
Prasad & Chandra (2024b)	PhiUSIIL; 235.795 URL	Kerangka *similarity index* dan *incremental learning*	Sumber dataset yang dianalisis
Potpelwar et al. (2025)	LegitPhish; URL berlabel dan fitur terekayasa	Dokumentasi dataset	Pembanding sumber URL legitimate; catatan konsistensi dokumentasi
Srivastava et al. (2026)	PhishGuard+; 10.000 URL, 48 fitur	Random Forest dan SMOTE (akurasi 98,4%)	Contoh studi fitur URL; evaluasi pada satu dataset

Dataset Publik dan Risiko Bias

Kualitas dan komposisi data memengaruhi kesimpulan. Hannousse dan Yahiouche (2021) mengusulkan skema pembuatan dataset yang dapat direproduksi dan diperluas karena situs phishing berumur pendek dan taktiknya terus berkembang. Survei dataset keamanan siber juga menempatkan dataset sebagai topik sentral dalam penerapan pembelajaran mesin untuk menghadapi serangan siber (Prasad & Chandra, 2023). Dataset publik membantu reproduksibilitas, tetapi sumber URL, periode pengumpulan, dan cara pemilihan contoh dapat membentuk pola yang tampak berbeda antarkelas. Sebagai ilustrasi, LegitPhish dibangun dari umpan ancaman untuk kelas phishing dan dari sumber seperti Google Search API, Wikipedia, dan Stack Overflow untuk kelas legitimate. Dokumentasinya memuat angka yang tidak sepenuhnya konsisten: jumlah URL kedua kelas bila dijumlahkan adalah 101.218, sedangkan total yang dinyatakan 101.219, dan jumlah fitur dinyatakan 17 pada satu bagian serta 18 pada bagian lain (Potpelwar et al., 2025). Hal ini menegaskan perlunya pemeriksaan dataset secara mandiri sebelum temuan digeneralisasi.

Posisi Penelitian

Berbeda dari studi pada Tabel 1 yang berorientasi pada kinerja klasifikasi, penelitian ini tidak melatih model dan tidak melaporkan akurasi, *precision*, *recall*, atau *F1-score*. Fokusnya adalah menjelaskan distribusi fitur yang dapat ditafsirkan langsung pada dua kelas berlabel, mengukur seberapa besar perbedaannya, dan memeriksa bagaimana struktur kelas legitimate memengaruhi tafsirannya. Dengan demikian temuan dapat menjadi dasar pemilihan indikator awal sekaligus bahan kehati-hatian bagi studi klasifikasi yang memakai dataset yang sama.

METODE PENELITIAN

Desain dan Sumber Data

Penelitian ini merupakan analisis kuantitatif komparatif atas data sekunder publik yang dilakukan secara offline terhadap string URL tanpa mengunjungi situs tujuan. Dataset yang digunakan adalah PhiUSIIL Phishing URL (Website) dari UCI Machine Learning Repository (Prasad & Chandra, 2024a). Dokumentasi dataset melaporkan 235.795 record, terdiri atas 100.945 phishing dan 134.850 legitimate, dengan label 0 untuk phishing dan 1 untuk legitimate. Berkas yang dianalisis adalah PhiUSIIL_Phishing_URL_Dataset.csv (56 kolom; berkas bertanggal 29 September 2024 pada arsip ZIP yang diperoleh melalui Kaggle). Berkas dianalisis pada 21 Oktober 2026. Hanya kolom URL dan label yang digunakan agar hasil dapat ditafsirkan sebagai karakteristik string URL, bukan sebagai evaluasi seluruh fitur bawaan dataset.

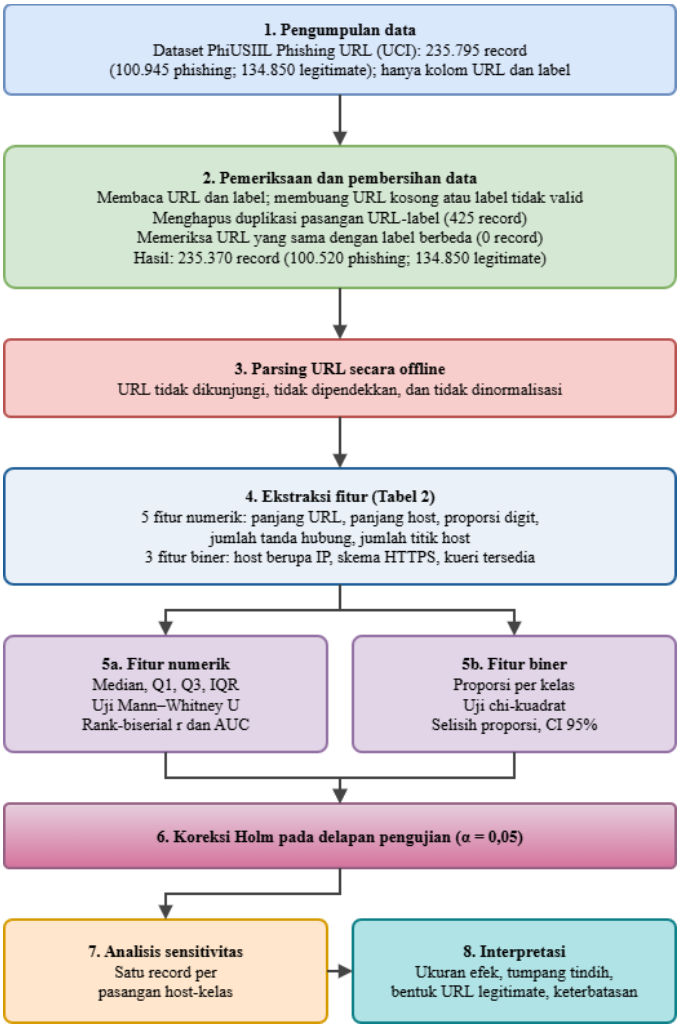
Checksum SHA-256 berkas: a236549cd369cd80bd478ff8e1779cbf44c58d5c3f79f7a51a1adbed7d06d1c6

Tahapan penelitian secara keseluruhan, mulai dari pengumpulan data hingga interpretasi hasil, diringkas dalam diagram alur pada Gambar 1. Rincian setiap tahap diuraikan pada subbagian berikut.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370



Gambar 1. Diagram alur tahapan penelitian

Pemeriksaan dan Pembersihan Data

Pengolahan dilakukan berurutan: (1) membaca kolom URL dan label; (2) membuang record dengan URL kosong atau label selain 0 dan 1 setelah spasi tepi dihapus; (3) menghapus duplikasi pasangan URL-label yang identik dan menyisakan satu record; (4) memeriksa URL yang sama dengan label berbeda; dan (5) mem-parse URL. URL asli dipertahankan, tidak dipendekkan, tidak dinormalisasi, dan tidak dikunjungi. Hostname diambil dengan modul *urllib.parse* dan alamat IP diperiksa dengan modul *ipaddress*. Kegagalan parsing, jika terjadi, dikeluarkan dan dihitung secara eksplisit. URL tanpa skema diparse dengan menambahkan awalan // dan skemanya diperlakukan sebagai tidak tersedia, bukan otomatis dianggap HTTP.

Operasionalisasi Fitur

Delapan fitur dipilih untuk mewakili panjang, komposisi karakter, struktur host, serta elemen protokol dan kueri. Definisi operasional ditetapkan sebelum membandingkan kelas agar pemilihan indikator tidak mengikuti pola yang menguntungkan hasil, dan disajikan pada Tabel 2.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

Tabel 2. Fitur yang digunakan dan definisi operasional

Fitur	Definisi operasional	Jenis
Panjang URL	Jumlah karakter URL asli setelah spasi tepi dihapus	Numerik
Panjang host	Jumlah karakter hostname hasil parsing	Numerik
Proporsi digit	Jumlah karakter ASCII 0–9 dibagi panjang URL	Numerik
Jumlah tanda hubung	Jumlah karakter - pada URL asli	Numerik
Jumlah titik host	Jumlah titik pada hostname; bukan jumlah subdomain	Numerik
Host berupa IP	1 jika hostname adalah IPv4/IPv6 valid; selain itu 0	Biner
Skema HTTPS	1 jika skema eksplisit https; 0 untuk skema lain yang tersedia	Biner
Kueri tersedia	1 jika komponen kueri tidak kosong; selain itu 0	Biner

Tabel 2 mengelompokkan delapan fitur ke dalam empat aspek. Aspek panjang diwakili panjang URL dan panjang host, yang dihitung sebagai jumlah karakter sehingga URL dengan lebih banyak komponen atau karakter bernilai lebih besar. Aspek komposisi karakter diwakili proporsi digit, yaitu jumlah digit dibagi panjang URL sehingga bernilai 0 sampai 1 dan tidak semata-mata mengikuti panjang string, serta jumlah tanda hubung. Aspek struktur host diwakili jumlah titik pada hostname dan host berupa IP, yang bernilai 1 apabila hostname merupakan alamat IPv4 atau IPv6 yang valid. Aspek protokol dan kueri diwakili skema HTTPS, yang bernilai 1 apabila skema eksplisit adalah https, dan kueri tersedia, yang bernilai 1 apabila komponen kueri tidak kosong. Kolom "Jenis" menunjukkan perlakuan statistiknya: lima fitur pertama bersifat numerik sehingga dibandingkan antarkelas dengan uji Mann–Whitney U, sedangkan tiga fitur terakhir bersifat biner sehingga dibandingkan menurut proporsinya dengan uji chi-kuadrat. Seluruh nilai dihitung langsung dari string URL tanpa membuka halaman tujuan.

Jumlah titik host tidak disebut jumlah subdomain karena penghitungan subdomain yang benar memerlukan daftar akhiran publik (*public suffix list*). Selain delapan fitur utama, empat deskriptor dihitung hanya secara deskriptif untuk memeriksa bentuk URL tiap kelas: keberadaan jalur (path selain kosong atau /), awalan www. pada host, skema http, serta keberadaan sedikitnya satu digit atau satu tanda hubung. Deskriptor tersebut tidak dimasukkan ke dalam pengujian inferensial dan koreksi Holm.

Teknik Analisis

Analisis dijalankan dengan Python 3.12.3, pandas 3.0.2, NumPy 2.4.4, dan SciPy 1.17.1. Fitur numerik diringkas dengan jumlah observasi valid, median, kuartil pertama (Q1), kuartil ketiga (Q3), dan rentang antarkuartil (IQR); fitur biner diringkas dengan proporsi per kelas. Fitur numerik dibandingkan menggunakan uji Mann–Whitney U dua arah (aproksimasi asimtotik) dan ukuran efek rank-biserial $r = 2U/(n_1 n_2) - 1$, dengan nilai positif berarti nilai fitur cenderung lebih besar pada phishing. Secara konvensional $|r|$ sekitar 0,1 dianggap kecil, 0,3 sedang, dan 0,5 besar. Karena $r = 2\text{AUC} - 1$, probabilitas bahwa URL phishing yang diambil acak bernilai lebih besar daripada URL legitimate yang diambil acak (AUC; nilai sama dihitung setengah) dilaporkan sebagai ukuran tumpang tindih.

Fitur biner dibandingkan dengan uji chi-kuadrat tanpa koreksi kontinuitas, selisih proporsi dalam poin persentase, dan interval kepercayaan (CI) 95% pendekatan Wald. Koreksi Holm diterapkan pada delapan pengujian dengan taraf 0,05, dan nilai p yang tercetak nol oleh perangkat lunak dilaporkan sebagai $p < 0{,}001$. Tumpang tindih distribusi juga digambarkan lewat persentase URL phishing yang melampaui persentil ke-95 dan nilai maksimum kelas legitimate; ukuran ini bersifat deskriptif dan bukan aturan klasifikasi. Analisis sensitivitas mengulang perbandingan numerik dengan satu record per pasangan host-kelas karena record dari host yang sama tidak independen.

Keamanan, Etika, dan Reproduksibilitas

Seluruh analisis dilakukan offline atas string URL. Tidak ada kunjungan ke halaman, login, pengiriman formulir, pemindaian aktif, maupun pengambilan konten situs, dan URL berbahaya tidak ditampilkan sebagai tautan aktif. Kode analisis, berkas keluaran, versi pustaka, dan checksum berkas disimpan untuk reproduksibilitas. Dataset bukan sampel acak dari seluruh internet; oleh karena itu signifikansi statistik dibaca bersama ukuran efek, sebaran data, dan kemungkinan bias pengumpulan data.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

HASIL DAN PEMBAHASAN

Kualitas Data dan Komposisi Kelas

Tabel 3. Rekap pemeriksaan kualitas data

Tahap	Jumlah record
Record awal yang dibaca	235.795
URL atau label kosong atau tidak valid	0
Duplikasi pasangan URL-label yang dihapus	425
Record dengan konflik label	0
Parsing gagal	0
Record akhir untuk analisis	235.370
Phishing pada data akhir	100.520
Legitimate pada data akhir	134.850
Host unik	220.077

Dari 235.795 record awal tidak ditemukan URL atau label kosong maupun tidak valid, tidak ada URL dengan label konflik, dan seluruh URL berhasil diparse. Satu-satunya pengecualian adalah 425 pasangan URL-label duplikat yang seluruhnya berasal dari kelas phishing, sehingga record akhir berjumlah 235.370: 100.520 phishing (42,7%) dan 134.850 legitimate (57,3%). Jumlah phishing akhir karenanya berbeda dari metadata sumber (100.945). Seluruh URL memiliki skema eksplisit dan hostname, sehingga tidak ada nilai tidak tersedia pada delapan fitur dan denominator tiap fitur sama dengan jumlah record kelasnya. Terdapat 220.077 host unik dan 54 host muncul pada kedua kelas.

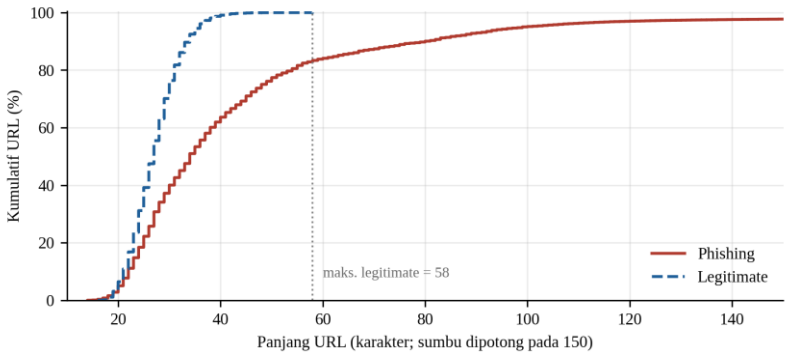
Perbandingan Fitur Numerik

Tabel 4. Ringkasan fitur numerik menurut kelas

Fitur	Phishing: median (IQR)	Legitimate: median (IQR)	*r* rank-biserial	*p* (Holm)
Panjang URL	34 (22)	27 (6)	0,44	< 0,001
Panjang host	22 (13)	19 (6)	0,25	< 0,001
Proporsi digit	0,000 (0,102)	0,000 (0,000)	0,45	< 0,001
Jumlah tanda hubung	0 (1)	0 (0)	0,30	< 0,001
Jumlah titik host	2 (0)	2 (0)	−0,03	< 0,001

Seluruh fitur numerik berbeda signifikan menurut uji Mann–Whitney U setelah koreksi Holm (*n* valid 100.520 phishing dan 134.850 legitimate), tetapi besar efeknya sangat bervariasi. Proporsi digit ($r = 0{,}45$) dan panjang URL ($r = 0{,}44$) memberi perbedaan terbesar. Median panjang URL phishing adalah 34 karakter (Q1–Q3: 26–48) dibandingkan 27 karakter (24–30) pada legitimate, dan sebaran phishing jauh lebih lebar dengan maksimum 6.097 karakter versus 58. Nilai ekstrem memengaruhi rata-rata (46,2 versus 27,2 karakter) sehingga median dan IQR lebih informatif untuk data yang tidak simetris.

Median proporsi digit adalah nol pada kedua kelas, sehingga perbedaannya bersumber dari ekor distribusi: Q3 phishing mencapai 0,102 sedangkan Q3 legitimate tetap nol, dan rata-rata proporsi digit 6,4% pada phishing berbanding 0,2% pada legitimate. Jumlah tanda hubung ($r = 0{,}30$) berefek sedang dengan median nol pada kedua kelas dan maksimum 286 pada phishing. Panjang host ($r = 0{,}25$; median 22 versus 19 karakter) berefek kecil hingga sedang. Jumlah titik host hampir tidak membedakan kelas (median dua pada keduanya; $r = -0{,}03$); signifikansinya semata akibat ukuran sampel yang besar dan menunjukkan pentingnya membaca ukuran efek, bukan hanya nilai *p*.



Gambar 2. Distribusi kumulatif panjang URL (sumbu dipotong pada 150 karakter; maksimum phishing 6.097 karakter)

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

Gambar 2 memperlihatkan tumpang tindih yang cukup besar. Seluruh URL legitimate berukuran paling panjang 58 karakter, dan 83,4% URL phishing berada pada rentang yang sama. Perbedaan baru tampak jelas pada ekor kanan: sekitar 10% URL phishing lebih panjang dari 80 karakter, sedangkan persentil ke-99 kelas legitimate hanya 40 karakter. Tabel 5 meringkas tumpang tindih ini untuk kelima fitur numerik.

Tabel 5. Tumpang tindih distribusi fitur numerik (deskriptif)

Fitur	AUC	Persentil 95 legitimate	Phishing > persentil 95 legitimate	Phishing > maksimum legitimate
Panjang URL	0,72	36	44,3%	16,6%
Panjang host	0,62	28	26,5%	3,3%
Proporsi digit	0,73	0,000	47,5%	2,4%
Jumlah tanda hubung	0,65	1	16,9%	0,5%
Jumlah titik host	0,49	3	5,3%	0,4%

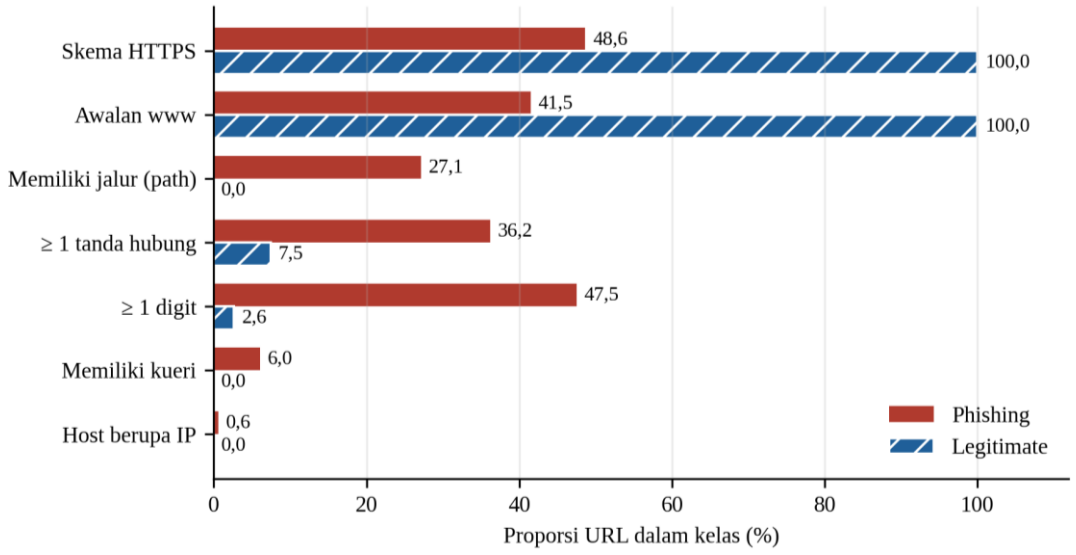
AUC tertinggi dimiliki proporsi digit (0,73) dan panjang URL (0,72), artinya URL phishing yang dipilih acak memiliki nilai lebih besar daripada URL legitimate yang dipilih acak pada sekitar tiga dari empat pasangan. Nilai AUC jumlah titik host (0,49) berada di sekitar 0,5, yaitu tidak membedakan kelas. Hanya 16,6% URL phishing yang lebih panjang dari URL legitimate terpanjang dan 2,4% yang memiliki proporsi digit di atas maksimum legitimate. Artinya, aturan ambang tunggal pada fitur numerik akan melewatkan sebagian besar URL phishing, dan nilai ambang itu pun bergantung pada dataset.

Perbandingan Fitur Biner

Tabel 6. Proporsi fitur biner menurut kelas

Fitur	Phishing: n/N (%)	Legitimate: n/N (%)	Selisih (poin)	CI 95%
Host berupa IP	606/100.520 (0,60%)	0/134.850 (0,00%)	+0,60	[0,56; 0,65]
Skema HTTPS	48.885/100.520 (48,63%)	134.850/134.850 (100,00%)	−51,37	[−51,68; −51,06]
Kueri tersedia	6.071/100.520 (6,04%)	0/134.850 (0,00%)	+6,04	[5,89; 6,19]

Ketiga perbedaan proporsi signifikan (p < 0,001 setelah koreksi Holm), tetapi harus dibaca bersama struktur kelas legitimate. HTTPS terdapat pada 48,6% URL phishing dan pada seluruh URL legitimate; selisihnya −51,37 poin persentase. Hampir separuh URL phishing memakai HTTPS, sehingga HTTPS tidak dapat dijadikan penanda legitimasi. HTTPS hanya menunjukkan bahwa koneksi ke server terenkripsi, bukan bahwa pemilik situs tepercaya, dan situs phishing dapat memperoleh sertifikat TLS. Sebaliknya, skema HTTP hanya ditemukan pada phishing dalam dataset ini (51,4%). Host berupa IP ditemukan pada 0,6% phishing (606 URL) dan kueri pada 6,0% phishing (6.071 URL), tetapi tidak ada pada legitimate. Keduanya sangat spesifik namun jarang: lebih dari 99% URL phishing tidak memakai host IP dan sekitar 94% tidak memiliki kueri, sehingga keduanya tidak layak berdiri sebagai indikator tunggal.



Gambar 3. Proporsi URL yang memiliki ciri tertentu menurut kelas

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

Gambar 3 menampilkan persentase URL dalam tiap kelas yang memiliki ciri tertentu. Setiap ciri diwakili sepasang batang, yaitu merah untuk phishing (n = 100.520) dan biru berarsir untuk legitimate (n = 134.850), dengan angka di ujung batang menunjukkan persentasenya dalam kelas tersebut. Skema HTTPS (48,6% pada phishing berbanding 100,0% pada legitimate) dan awalan www (41,5% berbanding 100,0%) justru lebih sering ditemukan pada URL legitimate karena seluruh URL legitimate berbentuk https://www. Ciri memiliki jalur (27,1% berbanding 0,0%), kueri (6,0% berbanding 0,0%), dan host berupa IP (0,6% berbanding 0,0%) hanya ditemukan pada phishing, tetapi batang kueri dan host berupa IP sangat pendek sehingga kedua ciri ini jarang. Sedikitnya satu tanda hubung (36,2% berbanding 7,5%) dan sedikitnya satu digit (47,5% berbanding 2,6%) lebih sering ditemukan pada phishing, meskipun masing-masing belum mencapai separuh URL phishing. Dengan demikian tidak ada satu pun ciri yang dimiliki seluruh URL phishing, dan selisih pada HTTPS, awalan www, serta jalur sebagian mencerminkan bentuk beranda pada URL legitimate yang dibahas pada subbagian berikut (Tabel 7).

Bentuk URL pada Kelas Legitimate

Tabel 7. Deskriptor struktur URL menurut kelas (deskriptif)

Ciri	Phishing: n (%)	Legitimate: n (%)
Host berawalan www.	41.720 (41,5%)	134.850 (100,0%)
Memiliki jalur (selain kosong atau /)	27.284 (27,1%)	0 (0,0%)
Skema http	51.635 (51,4%)	0 (0,0%)
Memuat sedikitnya satu digit	47.741 (47,5%)	3.560 (2,6%)
Memuat sedikitnya satu tanda hubung	36.358 (36,2%)	10.064 (7,5%)

Pemeriksaan deskriptif menunjukkan bahwa seluruh 134.850 URL legitimate berbentuk beranda dengan pola yang seragam: berskema https, berawalan www., tanpa jalur, tanpa kueri, dan tanpa host berupa IP. Sebaliknya, 27,1% URL phishing memiliki jalur dan 51,4% memakai HTTP. Bentuk gabungan https://www.host tanpa jalur dan tanpa kueri hanya ditemukan pada 1,0% URL phishing (1.052 URL). Pola ini kemungkinan mencerminkan prosedur pengumpulan data, yakni URL legitimate diambil sebagai alamat beranda situs; kesimpulan ini merupakan inferensi dari data yang dianalisis, bukan keterangan resmi dataset.

Konsekuensinya ada tiga. Pertama, perbedaan pada panjang URL, kueri, dan HTTPS sebagian berasal dari perbedaan bentuk URL (beranda versus halaman dalam), sehingga daya pembeda fitur-fitur itu pada URL dunia nyata mungkin lebih kecil daripada yang tampak di sini. Kedua, memisahkan kedua kelas pada dataset dengan kelas legitimate seperti ini kemungkinan jauh lebih mudah daripada mendeteksi phishing pada lalu lintas web nyata, sehingga akurasi tinggi pada dataset serupa perlu ditafsirkan hati-hati. Ketiga, ciri yang jelas hanya muncul pada sebagian phishing: 34,1% URL phishing tidak memuat digit, tanda hubung, host IP, kueri, maupun jalur, sehingga tidak satu pun ciri tersebut dapat mendeteksi URL-URL itu.

Analisis Sensitivitas

Tabel 8. Ukuran efek rank-biserial pada seluruh record dan satu record per host-kelas

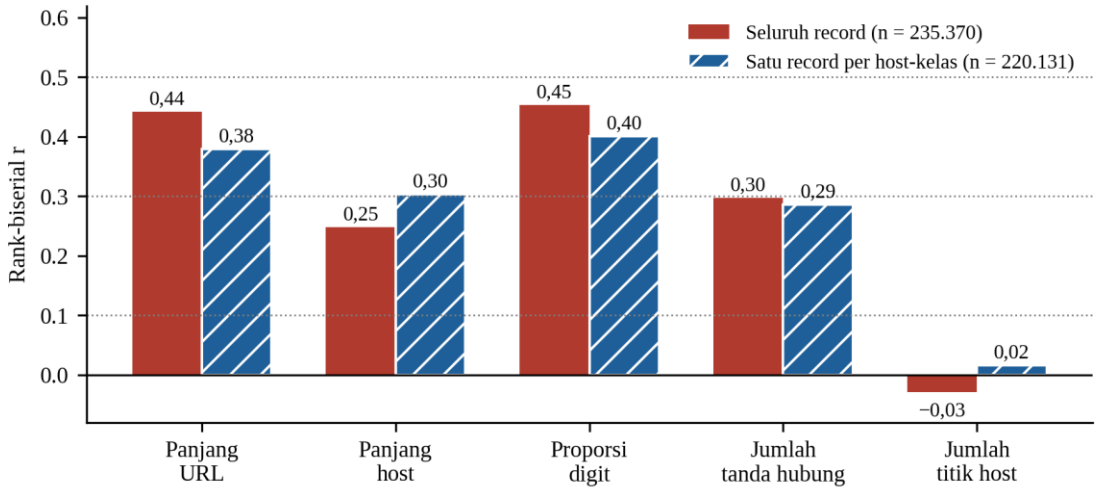
Fitur	r seluruh record (n = 235.370)	r satu record per host-kelas (n = 220.131)	Perubahan
Panjang URL	0,44	0,38	−0,06
Panjang host	0,25	0,30	+0,06
Proporsi digit	0,45	0,40	−0,05
Jumlah tanda hubung	0,30	0,29	−0,01
Jumlah titik host	−0,03	0,02	+0,05

Dengan satu record per pasangan host-kelas (220.131 record: 85.281 phishing dan 134.850 legitimate), arah efek tetap sama dan besarnya serupa, kecuali jumlah titik host yang tetap mendekati nol (Tabel 8 dan Gambar 4). Proporsi digit dan panjang URL tetap menjadi fitur numerik dengan efek terbesar (r = 0,40 dan 0,38), sedangkan efek jumlah titik host tetap dapat diabaikan ($|r| < 0{,}03$). Perubahan terbesar, sekitar 0,06, terjadi pada panjang URL (turun) dan panjang host (naik). Dengan demikian temuan utama tidak bergantung pada keberulangan host.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370



Gambar 4. Perbandingan ukuran efek pada seluruh record dan pada satu record per host-kelas (garis putus-putus: 0,1; 0,3; 0,5)

Gambar 4 membandingkan ukuran efek rank-biserial (*r*) kelima fitur numerik pada dua keadaan data, yaitu seluruh record (batang merah; n = 235.370) dan satu record per pasangan host-kelas (batang biru berarsir; n = 220.131). Sumbu vertikal menunjukkan nilai *r*, dan garis putus-putus menandai batas konvensional efek kecil (0,1), sedang (0,3), dan besar (0,5). Pada kedua keadaan, proporsi digit (0,45 dan 0,40) dan panjang URL (0,44 dan 0,38) tetap memiliki batang tertinggi, keduanya di atas batas efek sedang. Jumlah tanda hubung hampir tidak berubah (0,30 dan 0,29), sedangkan panjang host naik dari 0,25 menjadi 0,30. Jumlah titik host berada di sekitar nol pada kedua keadaan (−0,03 dan 0,02), sehingga tidak memiliki arah efek yang konsisten. Urutan fitur dan besar efeknya yang serupa menunjukkan bahwa temuan tidak bergantung pada keberulangan host.

Interpretasi dan Implikasi

Berdasarkan ukuran efek, AUC, dan proporsi, kandidat indikator awal yang paling informatif pada dataset ini adalah proporsi digit (r = 0,45; AUC 0,73), panjang URL (r = 0,44; AUC 0,72), jumlah tanda hubung (r = 0,30; AUC 0,65), dan panjang host (r = 0,25; AUC 0,62). Jumlah titik host tidak informatif. Pada fitur biner, penggunaan HTTP dan keberadaan kueri merupakan petunjuk pendukung, sedangkan host berupa IP jarang. Arah temuan, yaitu URL phishing lebih panjang serta memuat lebih banyak digit dan tanda hubung, sejalan dengan fitur leksikal yang lazim dipakai pada studi terdahulu (Jain & Gupta, 2018; Sahingoz et al., 2019; Srivastava et al., 2026). Studi-studi tersebut mengukur kinerja klasifikasi dengan dataset dan tujuan berbeda, sehingga nilai numeriknya tidak dibandingkan langsung.

Hasil ini dapat dimanfaatkan sebagai bahan edukasi bahwa pemeriksaan URL perlu melihat beberapa karakteristik sekaligus. Aturan sederhana seperti "URL panjang pasti phishing" akan keliru karena distribusi kedua kelas tumpang tindih, dan "URL berawalan https pasti aman" akan keliru karena hampir separuh URL phishing memakai HTTPS. Untuk penggunaan praktis, sinyal dari string sebaiknya dipadukan dengan konteks, misalnya kesesuaian nama domain dengan organisasi yang diklaim, apakah tautan diterima melalui pesan yang tidak diharapkan, dan apakah halaman meminta kredensial atau informasi pembayaran secara tidak wajar. Pemeriksaan kontekstual tersebut berada di luar pengukuran penelitian ini, tetapi menunjukkan mengapa indikator URL tidak boleh menjadi satu-satunya dasar keputusan.

Keterbatasan

Penelitian ini memiliki beberapa keterbatasan. Pertama, hasil terikat pada kualitas label dan komposisi dataset PhiUSIIL, terutama bentuk seragam URL legitimate. Kedua, dataset merupakan potret pada satu periode, sedangkan pola serangan berubah dari waktu ke waktu. Ketiga, isi halaman, reputasi domain, dan fitur berbasis layanan eksternal tidak digunakan. Keempat, beberapa URL berasal dari host yang sama sehingga tidak sepenuhnya independen, meskipun analisis sensitivitas menunjukkan arah temuan tetap. Kelima, penelitian tidak melatih model prediksi sehingga tidak menghasilkan metrik akurasi, *precision*, *recall*, *F1-score*, maupun AUC klasifikasi. Penelitian lanjutan dapat memvalidasi temuan pada dataset independen seperti LegitPhish, menggunakan URL legitimate yang memiliki jalur dan kueri, menambahkan analisis subdomain berbasis *public suffix list*, serta menguji stabilitas pola pada pembagian data berbasis waktu atau domain. Bila model klasifikasi dikembangkan, pemisahan data harus mencegah kebocoran antar-URL dari host yang sama.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.370

KESIMPULAN

Penelitian ini membandingkan delapan karakteristik leksikal dan struktural URL phishing dan legitimate pada dataset PhiUSIIL. Setelah 425 duplikasi dihapus, 235.370 URL (100.520 phishing dan 134.850 legitimate) dianalisis. Menjawab pertanyaan penelitian, URL phishing cenderung lebih panjang, memuat lebih banyak digit dan tanda hubung, serta memiliki host yang lebih panjang. Perbedaan terbesar terdapat pada proporsi digit (r = 0,45) dan panjang URL (r = 0,44), sedangkan jumlah titik host hampir tidak berbeda (r = −0,03). Tumpang tindih distribusi tetap besar: hanya 16,6% URL phishing yang lebih panjang daripada URL legitimate terpanjang.

Pada fitur biner, HTTPS ditemukan pada 48,6% URL phishing dan 100% URL legitimate, kueri hanya pada phishing (6,0%), dan host berupa IP hanya pada phishing (0,6%). Seluruh URL legitimate berbentuk beranda https://www tanpa jalur dan kueri, sehingga sebagian perbedaan mencerminkan cara kelas legitimate dibentuk. Karena itu fitur-fitur tersebut sebaiknya dipandang sebagai kandidat indikator awal untuk dataset ini, bukan penentu bahwa suatu URL pasti phishing, dan tidak ada fitur tunggal yang cukup untuk memastikan suatu URL berbahaya atau aman. Pengembangan berikutnya dapat diarahkan pada validasi lintas dataset dengan URL legitimate yang lebih beragam, analisis per host, dan evaluasi klasifikasi dengan pemisahan data yang mencegah kebocoran.

REFERENSI

Alabdan, R. (2020). Phishing attacks survey: Types, vectors, and technical approaches. *Future Internet*, *12*(10), 168. https://doi.org/10.3390/fi12100168

Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, *68*, 160–196. https://doi.org/10.1016/j.cose.2017.04.006

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*, *9*(9), 1514. https://doi.org/10.3390/electronics9091514

Butnaru, A., Mylonas, A., & Pitropakis, N. (2021). Towards lightweight URL-based phishing detection. *Future Internet*, *13*(6), 154. https://doi.org/10.3390/fi13060154

Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, *106*, 1–20. https://doi.org/10.1016/j.eswa.2018.03.050

Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence*, *104*, 104347. https://doi.org/10.1016/j.engappai.2021.104347

Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security* (Advances in Intelligent Systems and Computing, Vol. 729). Springer. https://doi.org/10.1007/978-981-10-8536-9_44

Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, *10*, 2015–2028. https://doi.org/10.1007/s12652-018-0798-z

Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. H. (2018). *URLNet: Learning a URL representation with deep learning for malicious URL detection* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1802.03162

Potpelwar, R. S., Kulkarni, U. V., & Waghmare, J. M. (2025). LegitPhish: A large-scale annotated dataset for URL-based phishing detection. *Data in Brief*, *63*, 111972. https://doi.org/10.1016/j.dib.2025.111972

Prasad, A., & Chandra, S. (2023). Machine learning to combat cyberattack: A survey of datasets and challenges. *The Journal of Defense Modeling and Simulation*, *20*(4), 577–588. https://doi.org/10.1177/15485129221094881

Prasad, A., & Chandra, S. (2024a). *PhiUSIIL phishing URL (website)* [Data set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset

Prasad, A., & Chandra, S. (2024b). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Computers & Security*, *136*, 103545. https://doi.org/10.1016/j.cose.2023.103545

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University – Computer and Information Sciences*, *35*(2), 590–611. https://doi.org/10.1016/j.jksuci.2023.01.004

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Srivastava, P., Singh, A., Virmani, A., Shukla, H., Mishra, R. K., Agarwal, R., & Mishra, A. K. (2026). A framework for real-time phishing resource locator detection using structural and lexical feature analysis based on machine learning. *International Journal of Science and Engineering Science Research*, *2*(4), 24–33. https://doi.org/10.63671/ijsesr.v2i4.171

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.