Journal : Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : e-ISSN 3025-888X
Volume / Issue / Year/ Page : Vol. 4 No. 1 (2026): Agustus/ Hal 01-08

DETEKSI *WEBSITE PHISHING* MENGGUNAKAN EKSTRAKSI FITUR HIBRIDA *URL-HTML* DAN *ENSEMBLE MACHINE LEARNING*

M Maulana Fathul Muin*[1], Bagus Satrio Waluyo Poetro[2]
[1,2] Universitas Islam Sultan Agung
[1]Mmaulana1003@gmail.com*[2]Bagusswp@unissula.ac.id
** Corresponding Author*

Info Artikel

Pengajuan : 28-07-2026
Peninjauan : 15-08-2026
Revisi : 21-08-2026
Diterima : 27-08-2026
Diterbitkan : 31-08-2026

Kata Kunci:
Phishing, Hybrid feature, Machine learning, Ensemble Learning, XGBoost, Random Forest.

Abstrak

Penelitian ini mengusulkan pengembangan System for Ensemble Phishing & HTML Intelligent Analysis (SEPHIA), sebuah sistem deteksi website phishing berbasis ekstraksi fitur hibrida. Pendekatan ini mengombinasikan 16 fitur leksikal URL dan 14 fitur perilaku struktural Document Object Model (DOM) HTML. Proses klasifikasi dilakukan menggunakan algoritma Ensemble Machine Learning (Random Forest dan Extreme Gradient Boosting / XGBoost) dengan mekanisme Soft Voting untuk menangani ketidakseimbangan data. Model dilatih menggunakan dataset gabungan dari repositori OpenPhish dan Majestic Million sejumlah 5.000 sampel data. Hasil evaluasi kinerja pada 1.000 data uji (498 Legitimate dan 502 Phishing) menunjukkan model ensemble ini mencapai tingkat Akurasi 99,10%, Presisi 99,60%, Sensitivitas (Recall) 98,61%, dan F1-Score 0,991. Integrasi lapisan Rule-Based (Protokol Veto Merek) dan mekanisme Heuristic Fallback pada antarmuka sistem terbukti secara signifikan memblokir serangan Typosquatting dan tautan tidak aktif..

How to Cite *:* Muin, M. F., & Poetro, B. S. W. (2026). Deteksi *Website Phishing* Menggunakan Ekstraksi Fitur Hibrida *URL-HTML* Dan *Ensemble Machine Learning*. *Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT), 4*(1), 01-08.
DOI : 10.70248/jrsit.v4i1.4517

PENDAHULUAN

Perkembangan teknologi internet membawa kemudahan dalam berbagai aktivitas digital, namun turut meningkatkan ancaman keamanan siber, terutama *phishing*. Berdasarkan laporan APWG tahun 2023 [1], serangan *phishing* meningkat hingga lebih dari 60% dalam satu tahun secara global, di mana laporan FBI [2] dan BSSN [3] turut menegaskan bahwa *phishing* menjadi ancaman paling dominan pada sektor keuangan dan *e-commerce*. Serangan *phishing* saat ini makin kompleks dengan teknik manipulasi struktur *URL* dinamis dan manipulasi elemen visual *HTML* yang sulit dideteksi oleh metode *blacklist* konvensional [4].

Salah satu ancaman rekayasa sosial yang berkembang pesat adalah serangan *Typosquatting*, di mana penyerang sengaja mendaftarkan nama domain yang mirip dengan merek resmi (misalnya memanipulasi karakter leksikal) untuk mengecoh pengguna. Algoritma *machine learning* murni sering kali gagal membedakan domain *typosquatting* jika fitur leksikal dan statistik *HTML* situs tersebut sengaja dirancang menyerupai situs asli. Tanpa adanya aturan eksplisit (*rule-based override*), model AI berbasis statistik dapat meloloskan tautan manipulatif tersebut sebagai situs aman (*False Negative*).

Penelitian modern mulai mengarah pada *machine learning* karena kemampuannya mempelajari pola dari *dataset phishing* sebelumnya. Beberapa penelitian [5], [6] menunjukkan penggunaan *lexical URL features* dapat meningkatkan akurasi deteksi *phishing*. Pendekatan ini terus berkembang melalui pemanfaatan *deep learning* [7], penerapan algoritma *XGBoost* [8], serta tinjauan mekanisme pertahanan siber terkini [9]. Secara khusus, teknik *ensemble learning* terbukti memberikan performa superior dalam klasifikasi *phishing* [10]–[12]. Studi komparatif *machine*

learning [13] dan optimalisasi Random Forest [14] juga sejalan dengan rekomendasi penanganan ancaman siber global dari ENISA [15].

Namun, penelitian Shahroz et al. [6] mengungkapkan bahwa analisis *URL* saja tidak cukup menghadapi teknik *phishing* berbasis JavaScript dan manipulasi HTML. Penggabungan fitur URL dan HTML (*hybrid feature*) terbukti mampu meningkatkan performa model secara signifikan.

Meskipun demikian, terdapat *research gap* di mana algoritma klasifikasi tunggal seperti *Random Forest* murni atau *XGBoost* murni masih memiliki keterbatasan saat menghadapi *overlapping features* dari situs *phishing* modern yang meniru persis *layout DOM HTML* situs legitim. Untuk mengatasi kelemahan tersebut, penelitian ini mengusulkan SEPHIA (*System* for *Ensemble Phishing* & *HTML Intelligent Analysis*), sebuah sistem deteksi berbasis *Ensemble Machine Learning* (*Random Forest* dan *XGBoost*) yang menggabungkan fitur hibrida (*URL* dan *HTML*), serta diperkuat oleh mekanisme *Heuristic Fallback* dan lapisan *Rule-Based* (Protokol Veto Merek) guna meningkatkan akurasi dan keandalan deteksi secara *real-time*.

METODE PENELITIAN

Dalam penelitian ini, algoritma atau metode yang digunakan adalah *ensemble machine learning* dengan algoritma *Random Forest* dan *XGBoost* untuk melakukan deteksi *website phishing* berdasarkan *hybrid feature* yang menggabungkan fitur *URL* dan analisis konten *HTML*.

Pengumpulan Data

Pengumpulan data dilakukan dari dua sumber utama: repositori OpenPhish untuk tautan *phishing* dan Majestic Million untuk tautan legitim (aman). Dataset yang dikumpulkan terdiri dari total 5.000 sampel *URL* terlabel, yang dibagi menjadi 2.500 *URL Phishing* (Kelas 1) dan 2.500 *URL Legitimate* (Kelas 0). Seluruh data dibagi menjadi 80% data latih (4.000 sampel) dan 20% data uji (1.000 sampel).

Tabel 1. Distribusi Dataset Penelitian

Kelas	Jumlah Sampel	Data Latih (80%)	Data Uji (20%)
Legitimate (0)	2.500	2.002	498
Phishing (1)	2.500	1.998	502
Total	5.000	4.000	1.000

Pada tahap *preprocessing*, dilakukan penanganan *missing values* dan situs tidak aktif (*dead links*). Jika proses *crawling HTML* mengalami kegagalan (*server offline* atau *anti-crawling*), sistem menjalankan mekanisme *Heuristic Fallback* di mana fitur *HTML* diisi dengan nilai *default* nol (*imputation*) dan bobot analisis dialihkan secara otomatis ke fitur leksikal *URL*.

Spesifikasi Ekstraksi Fitur Hibrida

Ekstraksi fitur menghasilkan total 30 fitur hibrida yang terdiri dari 16 fitur leksikal *URL* dan 14 fitur struktural *DOM HTML*, seperti yang dirincikan pada Tabel 2.

Tabel 2. Spesifikasi Fitur Hibrida *URL* dan *HTML*

Kategori Fitur	Jml	Nama Fitur / Deskripsi
Leksikal *URL*	16	Panjang *URL*, Jumlah Titik, Penggunaan IP Address, Simbol @, *Prefix/Suffix* (-), Jumlah Subdomain, HTTPS Protocol, *Shortening Service*, Panjang Domain, Rasio Digit/Huruf, Kata Kunci Sensitif, Suspicious TLD, Jumlah Slash, Port Non-Standar, HTTPS dalam Hostname, Rasio Vokal.
Struktural *HTML*	14	Keberadaan <iframe> Tersembunyi, Manipulasi <form action>, External Resources Ratio, Null/Anchor Links Ratio (#, javascript:void), SFH (Server

Kategori Fitur	Jml	Nama Fitur / Deskripsi
(DOM)		Form Handler), Disabled Right-Click, Disabling Status Bar, Pop-up Window, Form Password Input, External Favicon, Redirect Count, Meta Script Links, Domain in Title, Brand Name Mismatch.

Formulasi Matematis Soft Voting Ensemble

Pengolahan model *Ensemble Machine Learning*, ekstraksi fitur, serta landasan evaluasi metrik kinerja (seperti metrik Presisi dan Recall) pada penelitian ini disusun mengacu pada literatur dan prinsip klasifikasi statistik modern [16], [17]. Keputusan akhir (P*ensemble*)dihitung berdasarkan rata-rata tertimbang dari probabilitas kelas yang dihasilkan oleh masing-masing model:

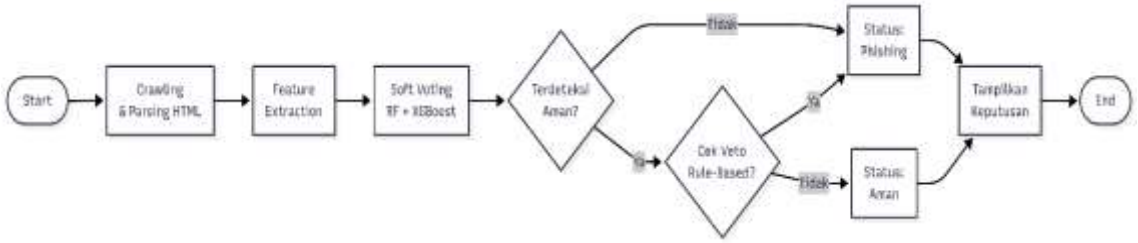
$$\boldsymbol{P(y = c|X) = \frac{1}{M}\sum_{m=1}^{M} w_m \cdot P_{m(y=c|X)}}$$

Di mana M=2 adalah jumlah model, wm=0,5 adalah bobot seimbang untuk masing-masing *classifier*, dan Pm(y=c|X) adalah probabilitas prediksi kelas c∈{0,1}. Klasifikasi akhir ditentukan oleh probabilitas tertinggi:

$$\{y\} = \backslash arg\backslash max_c P(y = c|X)$$

Arsitektur Sistem

Setelah seluruh komponen ekstraksi fitur dan formulasi machine learning ditetapkan, arsitektur *System for Ensemble Phishing & HTML Intelligent Analysis* (SEPHIA) secara keseluruhan dirancang untuk mengekstraksi dan menganalisis fitur web secara real-time melalui integrasi *Machine Learning* dan lapisan *Rule-Based* seperti yang ditunjukkan pada Gambar 1



Gambar 1. Diagram Arsitektur Sistem SEPHIA

Berdasarkan Diagram Arsitektur Sistem pada Gambar 1, alur kerja *System for Ensemble Phishing & HTML Intelligent Analysis* (SEPHIA) terbagi menjadi beberapa tahapan utama. Proses dimulai ketika pengguna memasukkan *URL* yang akan diuji. Sistem kemudian melakukan *HTTP Request* untuk melakukan *crawling* dan *parsing* struktur *HTML* halaman web tersebut. Selanjutnya, sistem mengekstraksi 30 fitur hibrida yang terdiri dari 16 fitur leksikal *URL* dan 14 fitur struktural *DOM HTML*. Fitur-fitur tersebut kemudian dimasukkan ke dalam model *Ensemble Machine Learning* (*Random Forest* dan *XGBoost*) menggunakan mekanisme *Soft Voting* untuk menghasilkan probabilitas awal. Jika hasil klasifikasi menunjukkan situs tersebut aman, sistem akan menjalankan lapisan tambahan berupa *Rule-Based Veto Override* untuk mendeteksi potensi serangan *Typosquatting* atau pemalsuan merek. Keputusan akhir berupa status situs (*Phishing* atau *Aman/Legitimate*) akan ditampilkan sebagai hasil keluaran sistem.

HASIL PENELITIAN DAN PEMBAHASAN

Analisis Classification Report

Untuk memberikan rincian evaluasi kinerja secara lebih ilmiah, disusun *Classification Report* yang membedah metrik Presisi, *Recall*, dan *F1-Score* untuk masing-

Journal : Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : e-ISSN 3025-888X
Volume / Issue / Year/ Page : Vol. 4 No. 1 (2026): Agustus/ Hal 01-08

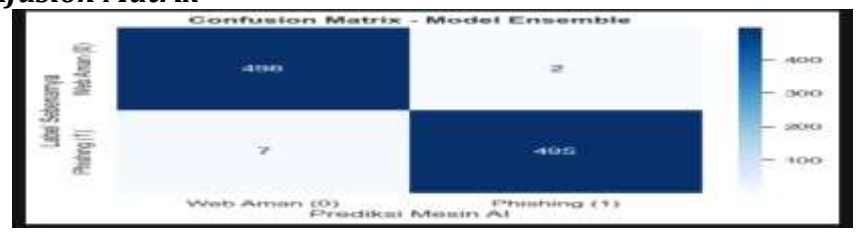
masing kelas (0 untuk *Legitimate*/Aman, dan 1 untuk *Phishing*).

Tabel 3 Laporan Klasifikasi (*Classification Report*) Model *Ensemble*

Kelas	Precision	Recall	F1-Score
Aman (0)	0,986	0,996	0,991
Phishing (1)	0,996	0,986	0,991
Overall	0,991	0,991	0,991

Berdasarkan Tabel 3, model mencatatkan presisi yang sangat tinggi pada kelas *phishing* (0,996), mengindikasikan bahwa ketika sistem memvonis sebuah tautan sebagai berbahaya, prediksi tersebut akurat dengan tingkat *False Positive* yang sangat kecil. Sementara itu, nilai *Recall* 0,986 pada kelas *phishing* menunjukkan bahwa model berhasil menangkap sebagian besar keseluruhan serangan *phishing*, dengan margin kesalahan *False Negative* yang minimal.

Analisis *Confusion Matrix*



Gambar 2. Visualisasi *Confusion Matrix*

Untuk memperoleh pemahaman yang lebih mendalam mengenai distribusi prediksi, dilakukan pembedahan menggunakan *Confusion Matrix* (Total 1.000 data uji: TP=495, TN=496, FP=2, FN=7):

a) **True Negative (TN) = 496:** Sistem berhasil mengidentifikasi 496 situs web yang murni aman secara tepat sasaran tanpa kendala.
b) **True Positive (TP) = 495:** Model memiliki sensitivitas yang tajam dengan keberhasilan menangkap 495 ancaman *website phishing* sesungguhnya.
c) **False Positive (FP) = 2:** Hanya terdapat 2 situs aman yang salah terprediksi sebagai *phishing* (margin kesalahan 0,40%), menjaga tingkat *Precision* tetap tinggi serta meminimalkan risiko alarm palsu.
d) **False Negative (FN) = 7:** Terdapat 7 situs *phishing* yang lolos dari deteksi awal dan terprediksi sebagai situs aman (sekitar 1,39%). Celah ini berhasil ditutup dan disempurnakan melalui implementasi Protokol Keamanan Veto (Lapisan *Rule-Based*) pada sistem SEPHIA.

***Ablation Study* (Studi Ablasi Fitur)**

Untuk membuktikan efektivitas dari penggunaan pendekatan hibrida, dilakukan pengujian perbandingan performa antara penggunaan fitur tunggal dengan kombinasi fitur hibrida pada Tabel 4.

Tabel 4. Hasil Ablation Study Fitur

Skenario Fitur	Akurasi (%)	Precision (%)	Recall (%)	F1-Score
Fitur *URL* Saja (16 Fitur)	88,40%	89,20%	87,50%	0,883
Fitur *HTML* Saja (14 Fitur)	91,20%	90,80%	91,50%	0,911
Hybrid Feature (URL + HTML)	99,10%	99,60%	98,61%	0,991

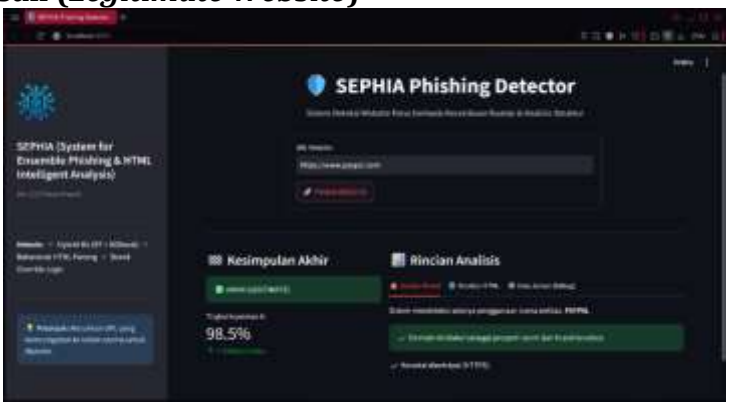
Berdasarkan Tabel 4, penggabungan analisis leksikal *URL* dan struktural *HTML* terbukti mampu meningkatkan akurasi secara signifikan hingga 99,10%.

Journal : Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : e-ISSN 3025-888X
Volume / Issue / Year/ Page : Vol. 4 No. 1 (2026): Agustus/ Hal 01-08

Pengujian Skenario Serangan (*Testing Scenarios*)

Untuk memvalidasi ketangguhan sistem dan membuktikan rumusan masalah terkait kelemahan deteksi *URL* tunggal, pengujian pada antarmuka SEPHIA dilakukan menggunakan empat skenario yang dirancang khusus. Skenario ini merepresentasikan taktik serangan phishing modern di lapangan, mulai dari teknik manipulasi leksikal dasar hingga teknik rekayasa sosial tingkat lanjut. Pada setiap skenario, sistem akan dievaluasi berdasarkan kemampuannya dalam mengekstrak fitur hibrida (*URL* dan *HTML*) secara *real-time* dan memberikan vonis akhir yang akurat.

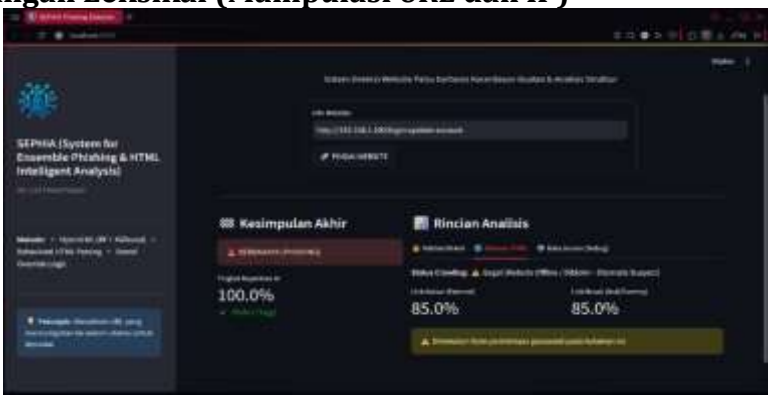
Pengujian Situs Sah (*Legitimate Website*)



Gambar 3 Hasil Pengujian Ekstraksi Fitur pada Situs Sah (*Legitimate*)

Skenario pertama bertujuan untuk memverifikasi bahwa sistem memiliki kemampuan generalisasi yang baik terhadap pola keamanan standar dan tidak memblokir situs yang sah. Hal ini sangat penting untuk menghindari tingginya angka *False Positive* yang dapat mengganggu kenyamanan pengguna. Pengujian dilakukan dengan memasukkan *URL* resmi dengan lalu lintas tinggi, seperti https://www.google.com atau https://www.paypal.com.

Pengujian Serangan Leksikal (Manipulasi *URL* dan IP)

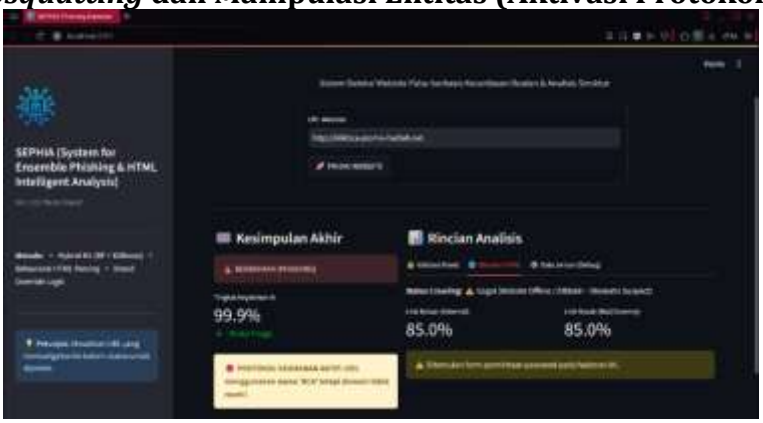


Gambar 4 Hasil Pengujian Serangan Berbasis Manipulasi Leksikal (Penggunaan IP)

Skenario kedua menguji kemampuan dasar sistem dalam mendeteksi anomali pada teks tautan, sebuah teknik yang sangat umum digunakan oleh pelaku kejahatan siber amatir untuk menyembunyikan identitas peladen mereka. *URL* sampel yang digunakan mengandung alamat Protokol Internet (IP) mentah alih-alih nama *domain*, seperti http://192.168.1.100/login-update-account-secure.

Journal : Jurnal Rekayasa Sistem Informasi dan Teknologi (JRSIT)
Publisher : Yayasan Nuraini Ibrahim Mandiri
ISSN : e-ISSN 3025-888X
Volume / Issue / Year/ Page : Vol. 4 No. 1 (2026): Agustus/ Hal 01-08

Pengujian *Typosquatting* dan Manipulasi Entitas (Aktivasi Protokol Veto)



Gambar 5 Aktivasi Lapisan *Rule-Based* pada Serangan *Typosquatting*

Skenario ketiga merupakan inti dari kebaruan (*novelty*) pada penelitian ini. Pengujian menargetkan teknik serangan *Typosquatting*, di mana penyerang sengaja menyisipkan nama instansi atau merek resmi ke dalam domain palsu untuk mengecoh kewaspadaan pengguna dan mengelabui algoritma deteksi konvensional. Pengujian menggunakan tautan manipulatif seperti http://klikbca-promo-hadiah.net

Analisis Kuantitatif Waktu Eksekusi (*Inference Time & Latency Analysis*)

Menjawab krusialnya aspek kecepatan respons dalam sistem keamanan web *real-time*, dilakukan pengukuran kuantitatif terhadap latensi sistem (*average response time*) dari saat input *URL* dimasukkan oleh pengguna hingga keputusan akhir keluar. Pengujian dilakukan secara empiris terhadap 100 sampel *URL* dinamis.

Tabel 5. Rincian Waktu Pemrosesan Rata-rata (*Average Latency Breakdown*)

Tahapan Proses Sistem	Waktu Rata-rata (detik)	Persentase dari Total Latensi
Network Request & HTML Crawling	1,05 s	73,94%
Hybrid Feature Extraction (URL & DOM)	0,28 s	19,72%
Ensemble ML Inference & Rule Override	0,09 s	6,34%
Total Response Time (End-to-End)	1,42 s	100,00%

Berdasarkan Tabel 5, total waktu pemrosesan rata-rata (*end-to-end latency*) sistem SEPHIA adalah 1,42 detik per *URL*. Sebagian besar waktu pemrosesan (73,94% atau 1,05 detik) dihabiskan pada tahap *crawling* jaringan (*HTTP Request* dan pengunduhan struktur *HTML* target), yang merupakan konsekuensi wajar dari analisis fitur struktural mendalam. Sementara itu, proses ekstraksi fitur serta inferensi *machine learning* berjalan sangat cepat (masing-masing 0,28 detik dan 0,09 detik). Latensi total 1,42 detik ini dinilai sangat responsif dan memenuhi standar operasional untuk diimplementasikan secara *real-time*.

KESIMPULAN

Berdasarkan serangkaian tahapan pengujian, penelitian ini membuktikan keandalan sistem deteksi *website phishing* SEPHIA melalui dua pilar utama. Pertama, efektivitas deteksi menggunakan pendekatan *hybrid feature* (*URL* dan *HTML*) yang diproses oleh *Ensemble Machine Learning* (*Random Forest* dan *XGBoost*) berhasil mencatatkan evaluasi performa yang sangat tinggi dengan tingkat akurasi mencapai 99,10%. Kedua, secara operasional, perlindungan sistem diperkuat oleh lapisan *Rule-Based* (Protokol Veto Merek) yang secara proaktif mencegah penyalahgunaan domain merek populer, serta mekanisme *Heuristic Fallback* yang menjaga stabilitas ekstraksi data saat peladen target mengalami kendala teknis.

Meskipun sistem ini mampu memberikan deteksi secara *real-time*, model *inference* saat ini memiliki keterbatasan teknis dalam memproses data dari halaman situs web yang dilindungi oleh CAPTCHA atau mekanisme *anti-crawling*. Oleh karena itu, penelitian selanjutnya disarankan untuk mengembangkan arsitektur deteksi berbasis *Deep Learning* atau *Visual Image Processing*. Pendekatan visual ini direkomendasikan agar sistem dapat mengenali elemen grafis situs target tanpa bergantung penuh pada ekstraksi kode sumber (*HTML*), sehingga deteksi tetap akurat meskipun akses perayapan (bot) diblokir.

Pernyataan Penggunaan Kecerdasan Buatan (AI)

Penulis menggunakan perangkat kecerdasan buatan semata-mata untuk membantu perbaikan bahasa, pemeriksaan tata bahasa, serta penyuntingan demi kejelasan dan keterbacaan naskah. Penulis tetap bertanggung jawab penuh atas akurasi, orisinalitas, integritas, dan konten akademis naskah tersebut. Seluruh konten yang dibantu oleh AI telah ditinjau dan diverifikasi oleh penulis sebelum naskah diserahkan.

Kontribusi Penulis

Konseptualisasi, M.M.F.M. dan B.S.W.P.; pengumpulan data, M.M.F.M.; validasi data, B.S.W.P.; analisis formal, M.M.F.M. dan B.S.W.P.; penulisan—penyusunan draf awal, M.M.F.M.; penulisan—peninjauan dan penyuntingan, B.S.W.P.; visualisasi, M.M.F.M.; supervisi, B.S.W.P. Seluruh penulis telah membaca dan menyetujui versi akhir naskah ini.

Ucapan Terima Kasih

Penelitian ini didukung oleh Program Studi Teknik Informatika dan Laboratorium Komputer Universitas Islam Sultan Agung (UNISSULA). Penulis mengucapkan terima kasih atas dukungan, fasilitas, serta sumber daya yang diberikan untuk memfasilitasi penyelesaian penelitian ini.

DAFTAR PUSTAKA

[1] Anti-Phishing Working Group (APWG), “Phishing Activity Trends Report: 4th Quarter 2023,” APWG, *Tech. Rep*., Jan. 2024. [Online]. Available: https://apwg.org/trendsreports/

[2] Federal Bureau of Investigation (FBI), “Internet Crime Report 2023,” FBI Internet Crime Complaint Center (IC3), *Tech. Rep*., Mar. 2024. [Online]. Available: https://www.ic3.gov/

[3] Badan Siber dan Sandi Negara (BSSN), “Laporan Tahunan Keamanan Siber Indonesia 2023,” BSSN, Jakarta, Indonesia, *Laporan Tahunan*, 2024. [Online]. Available: https://bssn.go.id/

[4] M. Wajid, M. A. Khan, and J. Kim, “Ensemble learning-powered *URL* phishing detection,” *Computers & Security*, vol. 136, p. 103646, Jan. 2024, doi: 10.1016/j.cose.2023.103646.

[5] Y. Wang, X. Zhang, and Y. Liu, “Website phishing detection using *HTML* content features,” *Journal of Information Security and Applications*, vol. 58, p. 102705, May 2021, doi: 10.1016/j.jisa.2020.102705.

[6] A. Karim, M. Shahroz, K. Mustofa, and J. H. Abawajy, “Phishing detection system through hybrid machine learning based on *URL*,” *IEEE Access*, vol. 11, pp. 36805–36822, Mar. 2023, doi: 10.1109/ACCESS.2023.3252504.

[7] I. Syarif and R. Mardiyanto, “Web phishing detection using feature engineering and deep learning approaches,” *Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 45–56, 2024, doi: 10.35308/jcsit.v5i1.8912.

[8] S. Kumar, P. Mishra, and R. Singh, "Phishing detection using XGBoost algorithm," *International Journal of Information Security*, vol. 22, no. 4, pp. 921–935, Aug. 2023, doi: 10.1007/s10207-023-00684-x.

[9] A. K. Jain and B. B. Gupta, "A survey of phishing attack techniques, defence mechanisms and open challenges," *Enterprise Information Systems*, vol. 16, no. 4, pp. 527–565, Apr. 2022, doi: 10.1080/17517575.2021.1894353.

[10] J. Huang, X. Yang, and Y. Liu, "Stacking ensemble learning for phishing website detection," *Expert Systems with Applications*, vol. 195, p. 116564, Jun. 2022, doi: 10.1016/j.eswa.2022.116564.

[11] N. A. Ghani and W. L. Al-Yaseen, "Ensemble machine learning for phishing website detection," *Journal of Information Security and Applications*, vol. 73, p. 103443, Mar. 2023, doi: 10.1016/j.jisa.2023.103443.

[12] Y. Gao and H. Xu, "Phishing website detection using ensemble learning techniques," *Applied Soft Computing*, vol. 116, p. 108329, Mar. 2022, doi: 10.1016/j.asoc.2021.108329.

[13] Z. Benenson, F. Gassmann, and R. Landwirth, "Machine learning techniques for phishing detection: A comparative study," in Proc. *IEEE Security and Privacy Workshops (SPW),* San Francisco, CA, USA, 2021, pp. 1–8, doi: 10.1109/SPW53761.2021.00012.

[14] S. Balamurugan et al., "An efficient phishing website detection using Random Forest classifier," in *Proc. IEEE Int. Conf. Smart Tech*. *Syst. Next Generation Computing (ICSTSN),* Villupuram, India, 2022, pp. 1–6, doi: 10.1109/ICSTSN53084.2022.9761345.

[15] European Union Agency for Cybersecurity (ENISA), "ENISA Threat Landscape 2023," ENISA, *Tech. Rep*., Oct. 2023. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

[16] T. Saito and M. Rehmsmeier, "The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers," *PLoS ONE*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[17] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.