Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Explainable Phishing Website Detection Using Comparative Machine Learning and SHAP

Juni Ismail[1*], Raja Anan Nasution[2], Muhammad Nasri Gea[3]

[1]Teknik Komputer, Politeknik Bisnis Indonesia, Pematangsiantar, Indonesia
[2,3]Manajemen Informatika, Akademi Manajemen Informatika dan Komputer ITMI, Medan, Indonesia
E-mail: [1*]juniismailll@gmail.com, [2]rajanasutionnnn@gmail.com, [3]muhammadnasrigea@gmail.com
(*: *corresponding author*)

Abstract - *Phishing attacks distributed through fraudulent URLs remain one of the most damaging cyber threats, including in Indonesia where malicious links spread widely through messaging applications and e-mail. Blacklist-based approaches cannot recognize newly created phishing URLs, while accurate machine learning models are often difficult to interpret. This study aims to compare six machine learning algorithms, namely XGBoost, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression, and Naive Bayes, for phishing website detection based on lexical URL, page content, and external reputation features, while providing model transparency through Explainable Artificial Intelligence (XAI). Experiments were conducted on the Web Page Phishing Detection Dataset containing 11,430 URLs with 87 features and a balanced class distribution. The research stages include exploratory data analysis, feature selection analysis using Chi-Square, Mutual Information, and Recursive Feature Elimination, an 80:20 data split, model training, hyperparameter optimization using RandomizedSearchCV, and interpretation of the best model using SHapley Additive exPlanations (SHAP). The results show that XGBoost delivers the best performance with 96.50% accuracy, 96.26% precision, 96.76% recall, 96.51% F1-score, and an AUC of 0.9943. SHAP analysis identifies google_index, page_rank, and nb_hyperlinks as the most influential features, dominated by external reputation-based features. Under this experimental setting, the findings indicate that an accurate phishing detection model can be equipped with interpretable explanations of its feature contributions. The main contribution of this study is an integrated comparative evaluation that combines six-algorithm benchmarking, leakage-free hyperparameter optimization, and SHAP-based interpretation on a public phishing dataset, offering practical guidance for security analysts.*

Keywords*: Explainable Artificial Intelligence, Machine Learning, Phishing URL Detection, SHAP, XGboost.*

Abstrak - Serangan phishing melalui penyebaran URL palsu masih menjadi salah satu ancaman siber paling merugikan, termasuk di Indonesia yang marak menghadapi penipuan tautan melalui aplikasi pesan dan surel. Pendekatan berbasis daftar hitam tidak mampu mengenali URL phishing baru, sementara model *machine learning* yang akurat sering kali sulit dijelaskan proses pengambilan keputusannya. Penelitian ini bertujuan membandingkan enam algoritma *machine learning*, yaitu XGBoost, Random Forest, Support Vector Machine, K-Nearest Neighbors, Logistic Regression, dan Naive Bayes, untuk deteksi situs phishing berbasis fitur leksikal URL, konten halaman, dan reputasi eksternal, sekaligus menghadirkan transparansi model melalui *Explainable Artificial Intelligence* (XAI). Eksperimen dilakukan pada *Web Page Phishing Detection Dataset* yang memuat 11.430 URL dengan 87 fitur dan distribusi kelas seimbang. Tahapan penelitian mencakup analisis data eksploratif, analisis seleksi fitur dengan Chi-Square, *Mutual Information*, dan *Recursive Feature Elimination*, pembagian data 80:20, pelatihan model, optimasi hyperparameter menggunakan RandomizedSearchCV, serta interpretasi model terbaik menggunakan SHapley Additive exPlanations (SHAP). Hasil pengujian menunjukkan XGBoost memberikan kinerja terbaik dengan akurasi 96.50%, presisi 96.26%, *recall* 96.76%, F1-score 96.51%, dan AUC 0.9943. Analisis SHAP mengidentifikasi google_index, page_rank, dan nb_hyperlinks sebagai fitur paling berpengaruh, yang didominasi kategori fitur eksternal berbasis reputasi. Pada skenario pengujian ini, temuan tersebut menunjukkan bahwa model deteksi phishing yang akurat dapat dilengkapi dengan penjelasan kontribusi fitur yang dapat diinterpretasikan. Kontribusi utama penelitian ini adalah evaluasi komparatif terintegrasi yang menggabungkan perbandingan enam algoritma, optimasi *hyperparameter* bebas kebocoran data, dan interpretasi SHAP pada dataset phishing publik sebagai panduan praktis bagi analis keamanan.

Kata Kunci: Deteksi URL Phishing, *Explainable Artificial Intelligence*, *Machine Learning*, SHAP, XGboost.

1. PENDAHULUAN

Perkembangan layanan digital yang pesat diiringi oleh meningkatnya kejahatan siber, dan phishing menempati posisi sebagai salah satu vektor serangan yang paling sering digunakan. Pelaku menyebarkan URL palsu yang menyerupai situs resmi untuk mencuri kredensial, data pribadi, hingga informasi finansial korban [1]. Di Indonesia, modus penipuan tautan palsu yang disebarkan melalui WhatsApp, SMS, dan surel telah menimbulkan kerugian finansial yang besar bagi masyarakat maupun institusi. Karakteristik URL phishing yang berumur pendek dan terus berganti membuat deteksi dini menjadi kebutuhan yang mendesak.

Mekanisme perlindungan konvensional berbasis daftar hitam (blacklist) memiliki kelemahan mendasar karena hanya mengenali URL yang sudah pernah dilaporkan, sehingga gagal menghadapi serangan zero-day dengan domain yang baru didaftarkan [2]. Pendekatan machine learning menjadi alternatif yang lebih adaptif karena model belajar dari pola fitur URL dan halaman web untuk mengenali karakteristik phishing yang belum pernah dijumpai sebelumnya [3]. Berbagai algoritma klasifikasi telah diterapkan pada permasalahan ini dengan hasil yang menjanjikan [2], [4].

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

Sejumlah penelitian terdahulu telah menguji beragam algoritma untuk deteksi phishing. Mahmud dan Wirawan membandingkan beberapa metode klasifikasi pada deteksi situs phishing dan melaporkan keunggulan model berbasis pohon keputusan [2]. Komalasari et al. mengevaluasi algoritma machine learning untuk deteksi domain phishing dan menemukan bahwa metode ensemble memberikan akurasi tertinggi [5]. Pada *dataset* yang sama dengan penelitian ini, Adane et al. menerapkan seleksi fitur univariat terhadap beberapa *ensemble classifier* dan memperoleh akurasi hingga 96.54% menggunakan Random Forest [6], sementara Ponni dan Dhandayudam mengusulkan kombinasi bagging dengan seleksi fitur *ensemble* untuk deteksi URL phishing [7]. Keunggulan algoritma berbasis pohon seperti Random Forest juga dilaporkan pada deteksi situs berbahaya dengan seleksi fitur korelasi Pearson [8]. Studi lain melaporkan bahwa pendekatan machine learning berbasis fitur URL mampu mencapai akurasi tinggi pada skala besar [9]. Kerangka seleksi fitur yang dapat dijelaskan juga terbukti meningkatkan transparansi deteksi phishing [10]. Selain itu, efektivitas perlindungan berbasis daftar hitam terbukti menurun seiring waktu sehingga evaluasi lintas waktu menjadi penting [11].

Meskipun capaian akurasi penelitian-penelitian tersebut tinggi, sebagian besar berhenti pada pelaporan metrik kinerja dan belum menjelaskan alasan di balik keputusan model. Padahal, pada domain keamanan siber, kepercayaan analis terhadap sistem deteksi sangat bergantung pada transparansi model [12]. Explainable Artificial Intelligence (XAI) hadir menjawab kebutuhan tersebut, dan *SHapley Additive exPlanations* (SHAP) menjadi salah satu metode interpretasi yang paling banyak digunakan karena mengkuantifikasi kontribusi setiap fitur terhadap prediksi berbasis teori permainan [13]. Selain itu, tahap optimasi *hyperparameter* kerap dilewatkan padahal berpotensi memperbaiki kinerja model [14]. Kombinasi komparasi algoritma yang luas, optimasi hyperparameter, dan interpretasi SHAP pada deteksi phishing URL, khususnya dalam publikasi berbahasa Indonesia, masih jarang dilakukan. Sebagai pembanding terdekat, Hannousse dan Yahiouche [15] serta Adane et al. [6] menggunakan dataset yang sama, tetapi keduanya berfokus pada seleksi fitur dan *ensemble* tanpa interpretasi keputusan model. Sementara itu, Alotaibi et al. [12] menerapkan XAI untuk klasifikasi phishing pada konteks sistem siber-fisik berbasis IoT dengan cakupan komparasi yang berbeda. Penelitian ini melengkapi ketiganya dengan menyatukan komparasi enam algoritma, optimasi *hyperparameter*, dan interpretasi SHAP dalam satu protokol evaluasi yang konsisten.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk membandingkan kinerja enam algoritma machine learning, yaitu XGBoost, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, dan Naive Bayes, dalam mendeteksi phishing URL pada *Web Page Phishing Detection Dataset*, mengoptimasi model unggulan melalui RandomizedSearchCV, serta menginterpretasikan model terbaik menggunakan SHAP untuk mengidentifikasi fitur-fitur yang paling menentukan keputusan klasifikasi. Kontribusi utama penelitian ini dirumuskan sebagai berikut. Pertama, evaluasi komparatif enam algoritma *machine learning* dengan optimasi *hyperparameter* dilakukan dalam satu protokol eksperimen yang konsisten dan dirancang bebas kebocoran data. Kedua, interpretasi global model terbaik menggunakan SHAP dikaitkan dengan taksonomi fitur dataset, sehingga identifikasi fitur kunci terverifikasi lintas empat metode analisis. Ketiga, implikasi praktis dominasi fitur eksternal terhadap penerapan sistem deteksi dianalisis secara eksplisit, suatu aspek yang jarang dibahas penelitian sebelumnya pada dataset yang sama [6], [15].

2. METODE PENELITIAN

2.1. Tahapan Penelitian

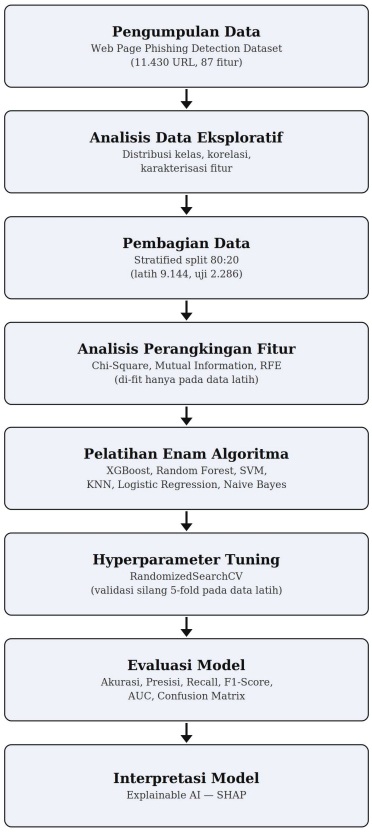
Penelitian ini dirancang sebagai studi komparasi eksperimental yang mengikuti alur kerja *machine learning* terstandar. Rangkaian tahapan dimulai dari pengumpulan data, analisis data eksploratif, pembagian data, analisis perangkingan fitur, pelatihan enam algoritma klasifikasi, optimasi *hyperparameter*, evaluasi kinerja, hingga interpretasi model terbaik menggunakan SHAP. Untuk mencegah kebocoran data (data leakage), pembagian data dilakukan lebih dahulu. Seluruh proses yang memanfaatkan label atau statistik fitur, yaitu standardisasi, perangkingan fitur dengan Chi-Square, *Mutual Information*, dan RFE, serta *hyperparameter tuning*, di-*fit* hanya pada data latih. Keseluruhan tahapan penelitian disajikan pada Gambar 1.

2.2. Dataset Penelitian

Data penelitian bersumber dari *Web Page Phishing Detection Dataset* yang disusun oleh Hannousse dan Yahiouche dan tersedia secara publik melalui repositori Mendeley Data serta Kaggle [15]. Data diunduh melalui Kaggle (shashwatwork/web-page-phishing-detection-dataset) yang merupakan salinan Mendeley Data V3 (DOI: 10.17632/c2gw7fy2j4.3), dipublikasikan 25 Juni 2021 dengan lisensi CC BY 4.0, dan diakses pada 16 Juli 2026. Dataset ini memuat 11.430 sampel URL dengan komposisi kelas yang seimbang, yaitu 5.715 URL phishing dan 5.715 URL *legitimate*. Setiap sampel dideskripsikan oleh 87 fitur yang terbagi ke dalam tiga kategori [15]. Kategori pertama adalah 56 fitur leksikal berbasis struktur dan sintaks URL, misalnya panjang URL, jumlah titik, dan jumlah tanda hubung. Kategori kedua adalah 24 fitur berbasis konten halaman, misalnya jumlah *hyperlink* dan

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

kesesuaian judul halaman. Kategori ketiga adalah 7 fitur eksternal berbasis reputasi dari layanan pihak ketiga, misalnya google_index, page_rank, dan web_traffic. Distribusi kelas yang seimbang menjadikan akurasi dan F1-score dapat digunakan secara langsung tanpa penanganan ketidakseimbangan data. Ringkasan karakteristik dataset disajikan pada Tabel 1.



Gambar 1. Tahapan Penelitian

Tabel 1. Karakteristik *Web Page Phishing Detection Dataset*

Atribut	Keterangan
Sumber data	Mendeley Data / Kaggle [15]
Jumlah sampel	11.430 URL
Jumlah fitur	87 (leksikal URL, konten halaman, eksternal)
Distribusi kelas	5.715 phishing : 5.715 legitimate (50:50)
Data latih	9.144 sampel (80%)
Data uji	2.286 sampel (20%)
Atribut	Keterangan

2.3. Pra-pemrosesan dan Pembagian Data

Tahap pra-pemrosesan mencakup pemeriksaan nilai hilang, penghapusan kolom URL mentah yang tidak digunakan dalam pemodelan, serta pengodean label kelas ke dalam nilai biner, yaitu 0 untuk *legitimate* dan 1 untuk phishing. Standardisasi fitur diterapkan pada algoritma yang sensitif terhadap skala seperti SVM, KNN, dan Logistic Regression. Data kemudian dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan stratified split agar proporsi kelas tetap terjaga, sehingga diperoleh 9.144 sampel latih dan 2.286 sampel uji dengan masing-masing kelas berjumlah sama pada setiap subset. Pembagian data menggunakan fungsi train_test_split dengan random_state = 42 agar hasil dapat direplikasi. Standardisasi di-fit hanya pada data latih melalui mekanisme *pipeline*, kemudian transformasi yang sama diterapkan pada data uji. Dengan demikian, tidak ada statistik data uji yang bocor ke proses pelatihan.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

2.4. Analisis Perangkingan Fitur

Analisis perangkingan fitur dilakukan secara deskriptif untuk mengkarakterisasi fitur yang paling diskriminatif, bukan untuk mereduksi masukan model. Tiga metode yang saling melengkapi digunakan. Chi-Square mengukur dependensi statistik antara fitur dan label dengan prasyarat nilai fitur non-negatif. *Mutual Information* menangkap keterkaitan non-linear antara fitur dan label. *Recursive Feature Elimination* (RFE) mengeliminasi fitur secara iteratif menggunakan estimator Logistic Regression (max_iter = 1000, random_state = 42) hingga tersisa 30 fitur terpilih. Seluruh analisis dihitung hanya pada data latih setelah pembagian data untuk mencegah kebocoran informasi. Hasilnya menjadi pembanding temuan SHAP, sedangkan pelatihan model utama tetap menggunakan 87 fitur. Sebagai tindak lanjut, perbandingan kinerja antara seluruh fitur, fitur terpilih, dan kategori fitur disajikan sebagai eksperimen ablation pada Subbab 3.3.

2.5. Algoritma Klasifikasi

Enam algoritma klasifikasi dengan karakteristik yang berbeda dipilih agar perbandingan mencakup beragam paradigma pembelajaran. XGBoost merupakan implementasi gradient boosting berbasis pohon dengan regularisasi bawaan yang efisien untuk data tabular [16]. Random Forest membangun banyak pohon keputusan melalui mekanisme bagging dan agregasi suara mayoritas [8]. SVM memisahkan kelas dengan mencari hyperplane bermargin maksimum, KNN mengklasifikasikan berdasarkan kedekatan jarak antar sampel, Logistic Regression memodelkan probabilitas kelas secara linear, dan Naive Bayes menerapkan teorema Bayes dengan asumsi independensi antar fitur [9]. Keenam algoritma diimplementasikan menggunakan pustaka scikit-learn dan XGBoost pada lingkungan Python.

2.6. *Hyperparameter Tuning*

Optimasi hyperparameter dilakukan pada tiga model dengan kinerja awal terbaik, yaitu Random Forest, XGBoost, dan SVM, menggunakan RandomizedSearchCV. Metode ini mengambil sampel kombinasi parameter secara acak dari ruang pencarian sehingga lebih efisien dibandingkan pencarian menyeluruh [14]. Ketiga model tersebut menempati tiga peringkat teratas pada evaluasi awal dengan margin yang jelas terhadap model peringkat keempat. Agar dasar pemilihan tidak bergantung pada data uji, peringkat keenam model diverifikasi ulang menggunakan F1-*score* validasi silang 5-fold pada data latih dan menghasilkan tiga model teratas yang sama (F1 masing-masing 0.9653, 0.9647, dan 0.9573, dengan margin jelas terhadap peringkat keempat 0.9467). Pencarian menggunakan n_iter = 20 kombinasi untuk Random Forest dan XGBoost serta 10 kombinasi untuk SVM, dengan random_state = 42. Konfigurasi terbaik dipilih berdasarkan F1-score validasi silang 5-fold pada data latih, kemudian model dengan parameter terbaik dilatih ulang pada keseluruhan data latih sebelum diuji pada data uji. Ruang pencarian tiap hyperparameter dan konfigurasi terbaik dirangkum pada Tabel 2.

Tabel 2. Ruang Pencarian Hyperparameter dan Konfigurasi Terbaik

Model	*Hyperparameter*	Ruang Pencarian	Nilai Terbaik
Random Forest	n_estimators; max_depth; min_samples_split; min_samples_leaf; max_features	{100, 200, 300, 500}; {None, 10, 20, 30, 50}; {2, 5, 10}; {1, 2, 4}; {sqrt, log2, None}	200; 50; 2; 1; log2
XGBoost	n_estimators; learning_rate; max_depth; subsample; colsample_bytree; gamma	{100, 200, 300}; {0.01, 0.05, 0.1, 0.2}; {3, 5, 7, 9}; {0.7, 0.8, 1.0}; {0.7, 0.8, 1.0}; {0, 0.1, 0.2}	200; 0.05; 7; 0.8; 0.7; 0.2
SVM	C; gamma; kernel	{0.1, 1, 10, 100}; {scale, auto, 0.01, 0.1}; {rbf, poly}	1.0; scale; rbf

2.7. Evaluasi Model

Kinerja setiap model diukur pada data uji menggunakan *confusion matrix* beserta metrik turunannya. *True Positive* (TP) menyatakan URL phishing yang terdeteksi benar, *True Negative* (TN) menyatakan URL *legitimate* yang terklasifikasi benar, sedangkan *False Positive* (FP) dan *False Negative* (FN) menyatakan kesalahan klasifikasi pada masing-masing kelas. Akurasi mengukur proporsi keseluruhan prediksi yang benar. Presisi mengukur ketepatan model ketika menandai sebuah URL sebagai phishing, sehingga penting untuk menekan alarm palsu yang mengganggu pengguna. Recall mengukur kemampuan model menemukan seluruh URL phishing, sehingga penting untuk mencegah serangan yang lolos deteksi. F1-*score* merangkum keseimbangan presisi dan recall dalam satu nilai tunggal. Keempat metrik digunakan bersama karena masing-masing menyoroti aspek kesalahan yang berbeda dan sama-sama relevan pada konteks keamanan. Metrik akurasi, presisi, *recall*, dan F1-score dihitung berdasarkan Persamaan (1) sampai dengan Persamaan (4).

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Sebagai pelengkap, kurva *Receiver Operating Characteristic* (ROC) beserta nilai *Area Under the Curve* (AUC) digunakan untuk menilai kemampuan model memisahkan kedua kelas pada berbagai ambang keputusan. Nilai AUC yang mendekati 1 menandakan kemampuan diskriminasi yang semakin baik. Skor untuk ROC dan AUC diperoleh dari probabilitas kelas positif melalui predict_proba pada setiap model, dengan opsi *probability=True* diaktifkan pada SVM.

2.8. Interpretasi Model dengan SHAP

Interpretasi model terbaik dilakukan menggunakan SHAP yang menghitung kontribusi setiap fitur terhadap prediksi berdasarkan nilai Shapley dari teori permainan kooperatif [13]. Implementasi TreeExplainer digunakan karena efisien untuk model berbasis pohon seperti XGBoost. Analisis mencakup SHAP *summary* plot untuk melihat arah dan besaran pengaruh nilai fitur terhadap keluaran model serta peringkat mean absolute SHAP *value* untuk mengidentifikasi fitur paling berpengaruh secara global. Nilai SHAP dihitung pada 500 sampel data uji yang diambil secara acak (random_state = 42) menggunakan pustaka SHAP versi 0.52.0 dengan pengaturan baku TreeExplainer tanpa *background* dataset khusus. Keluaran TreeExplainer untuk XGBoost berada pada skala log-odds, sehingga arah pengaruh fitur diinterpretasikan terhadap skor log-odds model, bukan probabilitas secara langsung.

3. HASIL DAN PEMBAHASAN

3.1. Karakterisasi Fitur Dataset

Analisis eksploratif mengonfirmasi distribusi kelas yang seimbang sehingga tidak diperlukan teknik penyeimbangan data. Analisis korelasi pada tahap eksploratif menunjukkan google_index memiliki korelasi positif tertinggi terhadap label sebesar 0.73, diikuti page_rank sebesar -0.51 dan nb_www sebesar -0.44. Hasil perangkingan fitur pada data latih memperlihatkan konsistensi yang tinggi antar metode, yaitu google_index menempati peringkat teratas baik pada Chi-Square maupun Mutual Information, disusul fitur reputasi lain seperti page_rank, web_traffic, dan domain_age, serta fitur leksikal seperti nb_www dan ratio_digits_url. Konsistensi lintas metode ini mengindikasikan bahwa fitur-fitur kunci tersebut bersifat robust dan bukan artefak dari satu teknik perangkingan tertentu.

3.2. Hasil Evaluasi Algoritma

Hasil evaluasi keenam algoritma pada 2.286 sampel data uji dirangkum pada Tabel 3. Perbandingan visual akurasi antar algoritma ditunjukkan pada Gambar 2(a), sedangkan perbandingan F1-*score* antar algoritma ditampilkan pada Gambar 2(b). XGBoost memperoleh nilai tertinggi pada pembagian data ini dengan akurasi 96.50%, presisi 96.26%, recall 96.76%, F1-score 96.51%, dan AUC 0.9943, diikuti secara ketat oleh Random Forest dengan F1-score 96.39%. SVM menempati posisi ketiga dengan F1-score 95.39%, sementara KNN dan Logistic Regression berada pada kisaran 93%. Naive Bayes menjadi model dengan kinerja terendah, yaitu akurasi 74.28% dan F1-score 77.28%.

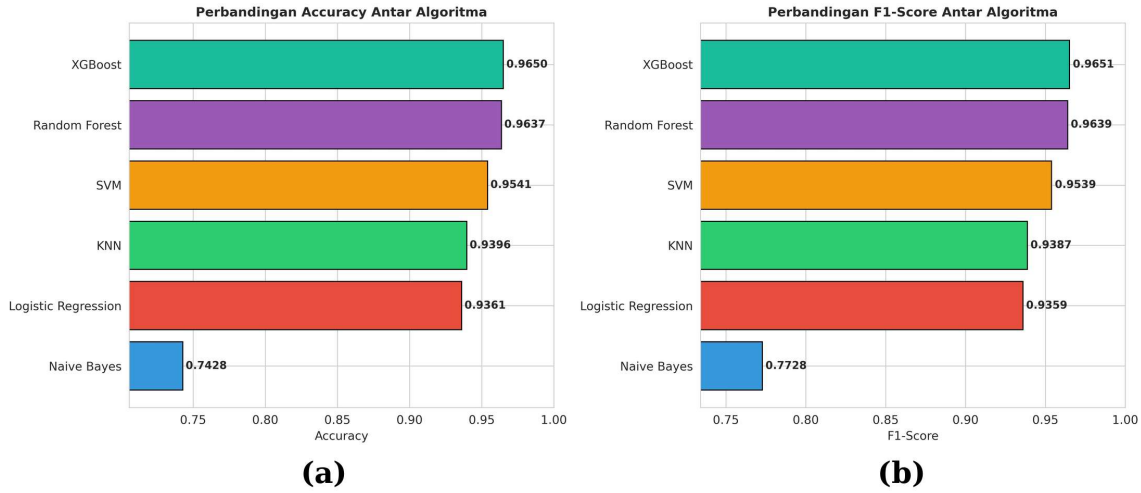
Tabel 3. Hasil Evaluasi Enam Algoritma pada Data Uji

Algoritma	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)	AUC
XGBoost	96.50	96.26	96.76	96.51	0.9943
Random Forest	96.37	95.93	96.85	96.39	0.9933
SVM	95.41	95.77	95.01	95.39	0.9883
KNN	93.96	95.39	92.39	93.87	0.9788
Logistic Regression	93.61	93.92	93.26	93.59	0.9814
Naive Bayes	74.28	69.20	87.49	77.28	0.8051

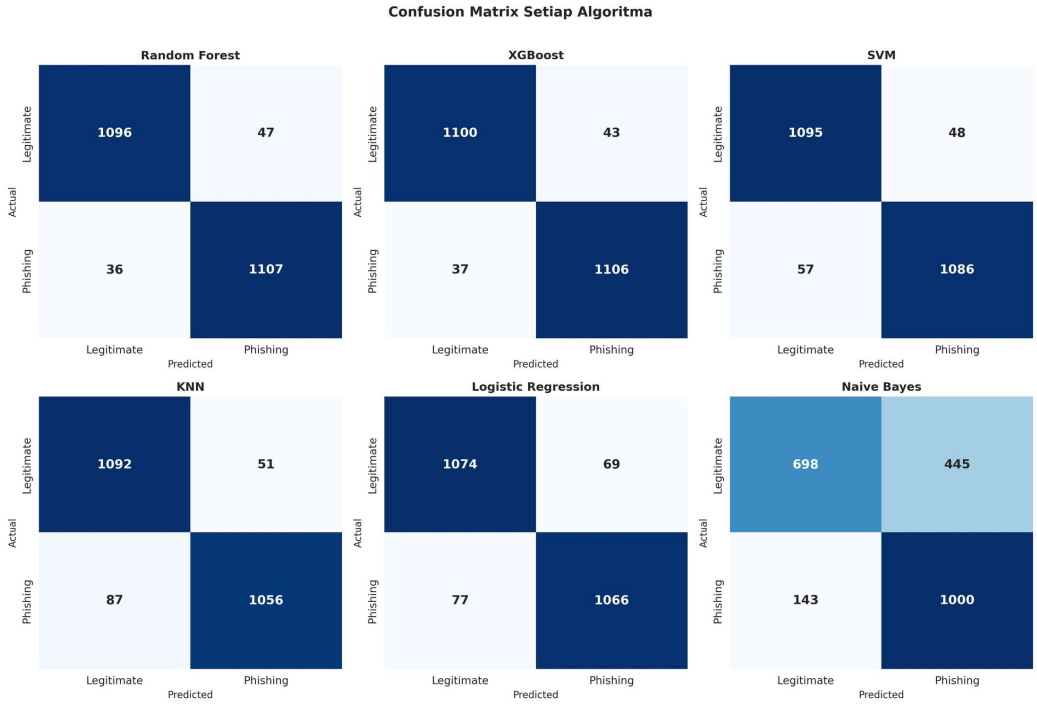
Rincian kesalahan klasifikasi setiap algoritma dapat diamati pada confusion matrix yang disajikan pada Gambar 3. XGBoost hanya menghasilkan 43 *false positive* dan 37 *false negative* dari 2.286 sampel uji. Jumlah *false negative* yang rendah penting pada konteks keamanan karena URL phishing yang lolos deteksi berdampak

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

langsung pada potensi kerugian pengguna. Sebaliknya, Naive Bayes menghasilkan 445 *false positive*, yang menunjukkan kecenderungan berlebihan menandai URL *legitimate* sebagai phishing. Pola ini diduga berkaitan dengan pelanggaran asumsi independensi antar fitur pada Naive Bayes, mengingat banyak fitur URL saling berkorelasi, misalnya ratio_digits_url dengan ip yang berkorelasi 0.77. Fitur yang saling berkorelasi membuat Naive Bayes menghitung bukti yang redundan secara berulang, sehingga model cenderung berlebihan menandai URL *legitimate* sebagai phishing. Dugaan ini masih bersifat hipotesis dan memerlukan analisis lanjutan untuk dikonfirmasi.



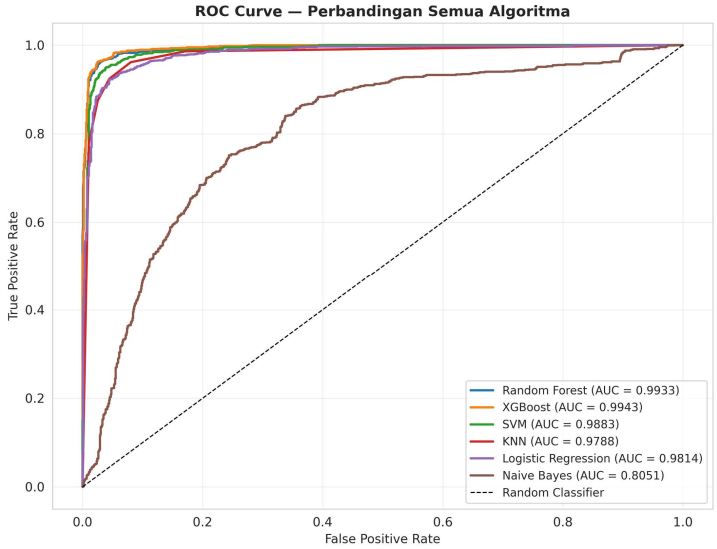
Gambar 2. Perbandingan Kinerja Antar Algoritma: (a) Akurasi; (b) F1-*Score*



Gambar 3. *Confusion Matrix* Setiap Algoritma

Kurva ROC pada Gambar 4 memperkuat temuan tersebut. Seluruh model berbasis pohon dan model diskriminatif mencapai AUC di atas 0.97, dengan XGBoost tertinggi sebesar 0.9943, sedangkan Naive Bayes hanya mencapai 0.8051. Keunggulan XGBoost pada seluruh metrik dalam skema pengujian ini selaras dengan temuan penelitian terdahulu mengenai superioritas algoritma gradient boosting untuk klasifikasi data tabular pada domain keamanan siber [5], [7], [16].

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index



Gambar 4. Kurva ROC Perbandingan Seluruh Algoritma

3.3. Pengaruh *Hyperparameter Tuning*

Optimasi *hyperparameter* menggunakan RandomizedSearchCV pada Random Forest, XGBoost, dan SVM menghasilkan peningkatan F1-score dibandingkan konfigurasi baseline masing-masing model, sebagaimana dirangkum pada Tabel 4. Peningkatan yang relatif terbatas ini mengindikasikan bahwa konfigurasi default ketiga algoritma telah mendekati optimal pada ruang pencarian dan dataset yang digunakan, sehingga klaim tersebut dibatasi pada kondisi eksperimen ini. Meskipun demikian, tahap optimasi tetap berperan penting sebagai prosedur verifikasi bahwa kinerja yang dilaporkan bukan akibat pemilihan parameter yang kurang tepat [14], sekaligus memastikan perbandingan antar algoritma dilakukan pada konfigurasi terbaik masing-masing.

Tabel 4. Perbandingan F1-Score Sebelum dan Sesudah *Hyperparameter Tuning*

Model	F1 *Baseline* (%)	F1 Setelah Tuning (%)	Selisih (pp)	Skor F1 CV Terbaik (%)
Random Forest	95.95	96.39	+0.44	96.73
XGBoost	96.25	96.51	+0.26	96.90
SVM	95.39	95.39	0.00	95.73

Untuk menilai stabilitas kinerja di luar satu pembagian data, evaluasi tambahan dilakukan menggunakan 10-fold cross validation berstratifikasi. XGBoost memperoleh F1-score sebesar 96.75% ± 0.48 dan Random Forest sebesar 96.86% ± 0.34 antar-fold. Uji Wilcoxon signed-rank terhadap skor antar-fold menunjukkan bahwa perbedaan keduanya tidak signifikan secara statistik ($p = 0.3223$). Oleh karena itu, keunggulan XGBoost dinyatakan sebagai capaian tertinggi pada pembagian data penelitian ini, sedangkan secara statistik kinerja XGBoost dan Random Forest dapat dianggap setara.

Eksperimen ablation dilakukan untuk menguji peran perangkingan fitur dan kategori fitur, dengan hasil pada Tabel 5. Konfigurasi yang dibandingkan meliputi seluruh 87 fitur, subset fitur terpilih hasil perangkingan, fitur leksikal saja, fitur eksternal saja, serta gabungan fitur leksikal dan konten tanpa fitur eksternal. Perbandingan ini sekaligus menjawab kebutuhan praktis penerapan model ketika layanan eksternal tidak tersedia. Hasilnya menunjukkan konfigurasi tanpa fitur eksternal menurunkan F1-score sekitar 1.9 poin persentase menjadi 94.60%, sedangkan tujuh fitur eksternal saja telah mencapai F1-score 93.39%, yang menegaskan besarnya peran fitur reputasi. Subset 30 fitur hasil RFE mempertahankan kinerja mendekati konfigurasi penuh dengan F1-score 95.86%, sementara model leksikal murni mencapai 91.31% sebagai batas bawah skenario tanpa konektivitas eksternal.

Tabel 5. Hasil Eksperimen Ablation Konfigurasi Fitur pada XGBoost

Konfigurasi Fitur	Jumlah Fitur	Akurasi (%)	F1-Score (%)
Seluruh fitur	87	96.46	96.47
Fitur terpilih RFE	30	95.84	95.86
Fitur leksikal saja	56	91.29	91.31
Fitur eksternal saja	7	93.39	93.39
Leksikal + konten (tanpa eksternal)	80	94.62	94.60

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

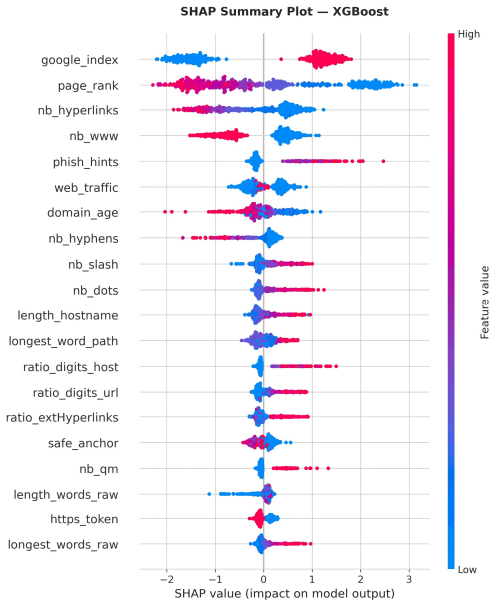
3.4. Interpretasi Model dengan SHAP

Interpretasi model XGBoost menggunakan SHAP disajikan pada Gambar 5 dan Gambar 6. Berdasarkan nilai mean absolute SHAP, tiga fitur paling berpengaruh secara berurutan adalah google_index, page_rank, dan nb_hyperlinks, diikuti nb_www, phish_hints, web_traffic, dan domain_age. Pada dataset ini, fitur google_index dikodekan bernilai 1 apabila halaman tidak terindeks oleh Google dan 0 apabila terindeks [15], sehingga arah pengaruhnya sejalan dengan intuisi. *Summary plot* pada Gambar 5 menunjukkan nilai google_index yang tinggi, yang berarti halaman tidak terindeks, mendorong prediksi ke arah kelas phishing. Sebaliknya, nilai page_rank, nb_hyperlinks, dan nb_www yang tinggi menurunkan skor log-odds ke arah kelas *legitimate*. Kehadiran kata kunci mencurigakan yang ditangkap fitur phish_hints serta rasio digit yang tinggi pada URL juga menaikkan skor log-odds model ke arah kelas phishing.

Apabila ditinjau dari taksonomi fitur dataset [15], fitur-fitur dominan tersebut sebagian besar termasuk kategori fitur eksternal berbasis reputasi (google_index, page_rank, web_traffic, domain_age), disusul fitur berbasis konten halaman (nb_hyperlinks) dan fitur leksikal URL (nb_www, nb_hyphens, nb_dots). Temuan ini menegaskan bahwa reputasi domain pada ekosistem mesin pencari merupakan diskriminator terkuat antara URL phishing dan *legitimate*, konsisten dengan nilai korelasi google_index terhadap label sebesar 0.73 serta hasil perangkingan Chi-Square dan *Mutual Information* pada tahap analisis perangkingan fitur. Konsistensi temuan lintas empat metode analisis yang berbeda memperkuat validitas identifikasi fitur kunci tersebut [12].

3.5. Pembahasan

Keunggulan XGBoost dapat dijelaskan dari beberapa aspek. Pertama, mekanisme boosting yang membangun pohon secara sekuensial memungkinkan model menangkap interaksi non-linear antar fitur yang kompleks pada data phishing [16]. Kedua, regularisasi bawaan menjaga model tetap robust terhadap overfitting meskipun jumlah fitur mencapai 87. Ketiga, dari sisi efisiensi, XGBoost membutuhkan waktu pelatihan 1.21 detik dan waktu prediksi 0.014 detik untuk keseluruhan data uji. Namun, angka tersebut belum mencakup waktu ekstraksi fitur dan kueri ke layanan eksternal, sehingga kelayakan penerapan waktu nyata masih memerlukan pengujian latensi end-to-end. Hasil ini juga sebanding dengan studi pada dataset yang sama, yaitu akurasi 96.83% oleh Hannousse dan Yahiouche menggunakan Random Forest dengan seleksi fitur [15] dan 96.54% oleh Adane et al. [6], yang menempatkan capaian penelitian ini pada kisaran kinerja terbaik dataset tersebut.

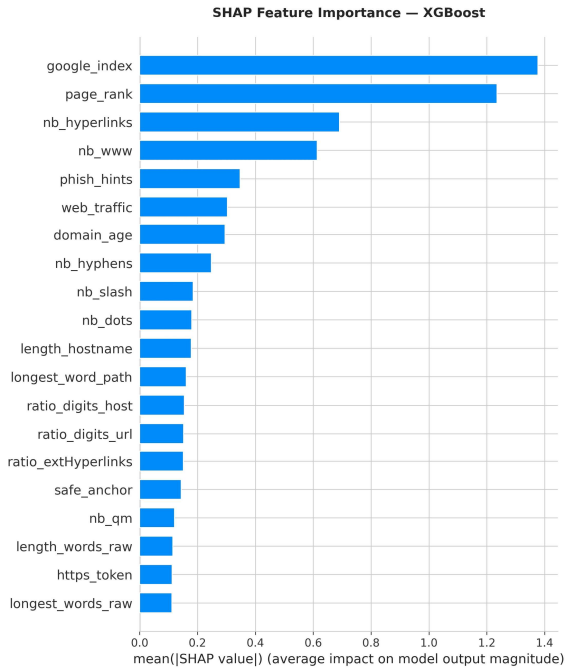


Gambar 5. SHAP *Summary Plot* Model XGBoost

Dominasi fitur eksternal membawa konsekuensi praktis yang perlu dicermati. Fitur seperti google_index, page_rank, dan web_traffic menuntut kueri ke layanan pihak ketiga pada saat inferensi, sehingga menimbulkan latensi tambahan, ketergantungan pada ketersediaan layanan eksternal, serta potensi biaya akses. Implementasi waktu nyata berbasis ekstensi peramban atau gateway surel perlu mempertimbangkan strategi caching, batas waktu kueri, atau model cadangan yang hanya memanfaatkan fitur leksikal ketika layanan eksternal tidak terjangkau. Di sisi lain, transparansi yang dihasilkan SHAP memberi nilai praktis bagi analis keamanan, karena fitur-fitur kunci

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

dapat dijadikan aturan penyaring awal sebelum tahap deteksi berbasis model [12], sekaligus menjadi dasar audit ketika terjadi kesalahan klasifikasi.



Gambar 6. Peringkat *Mean Absolute SHAP Value* Model XGBoost

Penelitian ini memiliki sejumlah keterbatasan. Pertama, evaluasi akhir dilakukan pada satu skema pembagian data sehingga estimasi kinerja belum disertai ukuran variabilitas antar pembagian. Kedua, dataset dikumpulkan pada tahun 2021 sehingga pergeseran pola serangan (concept drift) berpotensi menurunkan kinerja model pada URL phishing terkini. Ketiga, ketergantungan pada fitur eksternal membatasi penerapan langsung pada lingkungan dengan konektivitas terbatas. Keempat, model belum divalidasi pada URL phishing berkonteks Indonesia yang mungkin memiliki karakteristik tersendiri. Kelima, validasi eksternal maupun temporal pada sumber data lain belum dilakukan, sehingga ketergantungan hasil pada satu sumber data belum dapat dikesampingkan. Keenam, model produksi yang hanya mengandalkan fitur leksikal belum dikembangkan dan diuji secara *end-to-end*, padahal fitur eksternal dapat berubah nilainya atau tidak tersedia pada saat *deployment*.

4. KESIMPULAN

Pada skenario pengujian penelitian ini, XGBoost memperoleh kinerja tertinggi untuk deteksi situs phishing pada *Web Page* Phishing Detection Dataset dengan akurasi 96.50%, F1-score 96.51%, dan AUC 0.9943, diikuti lima algoritma pembanding dengan Naive Bayes sebagai baseline terendah pada F1-score 77.28%. Optimasi hyperparameter memberikan peningkatan F1-*score* yang terbatas pada tiga model unggulan, yang mengindikasikan konfigurasi default ketiganya telah mendekati optimal pada ruang pencarian dan dataset ini, sekaligus memastikan perbandingan antar model dilakukan pada konfigurasi terbaik masing-masing. Analisis SHAP berhasil mengidentifikasi google_index, page_rank, dan nb_hyperlinks sebagai fitur paling menentukan, yang didominasi fitur eksternal berbasis reputasi, sehingga keputusan model dapat dijelaskan secara lebih transparan kepada analis keamanan.

Kontribusi utama penelitian ini adalah kerangka evaluasi komparatif terintegrasi yang menggabungkan benchmarking enam algoritma, optimasi hyperparameter bebas kebocoran data, dan interpretasi SHAP yang dikaitkan dengan taksonomi fitur, beserta analisis implikasi praktis dominasi fitur eksternal. Keterbatasan utama penelitian mencakup penggunaan satu dataset publik yang dikumpulkan pada tahun 2021, evaluasi akhir pada satu skema pembagian data, serta ketergantungan pada fitur eksternal yang menuntut kueri ke layanan pihak ketiga. Penelitian selanjutnya disarankan mengeksplorasi arsitektur deep learning seperti CNN, LSTM, atau Transformer yang memproses URL langsung sebagai barisan karakter, membangun dataset phishing URL berkonteks Indonesia, menerapkan model pada ekstensi peramban atau API waktu nyata dengan mempertimbangkan ketergantungan fitur eksternal, serta mengkaji adversarial training untuk meningkatkan ketahanan model terhadap serangan penghindaran.

Idealis: Indonesia Journal Information System
Volume 9, Nomor 2, Juli 2026
ISSN 2684-7280 (*online*)
Halaman 268-277
available online at http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index

DAFTAR PUSTAKA

[1] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 590-611, 2023, doi: 10.1016/j.jksuci.2023.01.004.

[2] A. F. Mahmud and S. Wirawan, "Deteksi Phishing Website Menggunakan Machine Learning Metode Klasifikasi," *Sistemasi: J. Sist. Inf.*, vol. 13, no. 4, pp. 1368-1380, 2024, doi: 10.32520/stmsi.v13i4.3456.

[3] A. S. Y. Irawan, N. Heryana, H. S. Hopipah, and D. Rahma, "Identifikasi Website Phishing dengan Perbandingan Algoritma Klasifikasi," *Syntax: J. Inform.*, vol. 10, no. 1, pp. 57-67, 2021, doi: 10.35706/syji.v10i01.5292.

[4] Y. Muliono, M. A. Ma'ruf, and Z. M. Azzahra, "Phishing Site Detection Classification Model Using Machine Learning Approach," *J. EMACS (Eng. Math. Comput. Sci.)*, vol. 5, no. 2, pp. 63-67, 2023, doi: 10.21512/emacsjournal.v5i2.9951.

[5] D. Komalasari, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria, Z. Abdullah, and A. Alanda, "Phishing Domain Detection Using Machine Learning Algorithms," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 15, no. 1, pp. 318-327, 2025, doi: 10.18517/ijaseit.15.1.12553.

[6] K. Adane, B. Beyene, and M. Abebe, "ML and DL-based Phishing Website Detection: The Effects of Varied Size Datasets and Informative Feature Selection Techniques," *J. Artif. Intell. Technol.*, vol. 4, no. 1, pp. 18-30, 2024, doi: 10.37965/jait.2023.0269.

[7] P. Ponni and P. Dhandayudam, "An Optimized Bagging Learning with Ensemble Feature Selection Method for URL Phishing Detection," *J. Electr. Eng. Technol.*, vol. 19, no. 3, pp. 1881-1889, 2023, doi: 10.1007/s42835-023-01680-z.

[8] E. Sangra, R. Agrawal, P. R. Gundalwar, K. Sharma, D. Bangri, and D. Nandi, "Malicious Website Detection Using Random Forest and Pearson Correlation for Effective Feature Selection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 772-780, 2024, doi: 10.14569/IJACSA.2024.0150876.

[9] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345-357, 2019, doi: 10.1016/j.eswa.2018.09.029.

[10] S. S. Shafin, "An explainable feature selection framework for web phishing detection with machine learning," *Data Sci. Manag.*, vol. 8, no. 2, pp. 127-136, 2025, doi: 10.1016/j.dsm.2024.08.004.

[11] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, A. Doupe, and G.-J. Ahn, "PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," in *Proc. 29th USENIX Security Symposium*, 2020, pp. 379-396.

[12] S. R. Alotaibi et al., "Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems," *Alexandria Eng. J.*, vol. 110, pp. 490-505, 2025, doi: 10.1016/j.aej.2024.09.115.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765-4774.

[14] S. Y. Nailendra, W. Witanti, and G. Abdillah, "Optimasi Prediksi Penjualan Retail Online Menggunakan LightGBM dan Hyperparameter Tuning," *J. Algoritm.*, vol. 22, no. 2, pp. 1931-1942, 2025, doi: 10.33364/algoritma/v.22-2.2551.

[15] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Eng. Appl. Artif. Intell.*, vol. 104, p. 104347, 2021, doi: 10.1016/j.engappai.2021.104347.

[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.