Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index

ANALISIS PERFORMA ALGORITMA *LIGHTGBM* DENGAN TEKNIK *SMOTE* UNTUK KLASIFIKASI *WEBSITE PHISHING*

Angelyn Indriaty[1*)], Rizka Hafsari[2)]
[1,2]Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
210402040@umri.ac.id[1*], rizkahafsari@umri.ac.id[2]

Abstract

Phishing website attacks are a severe cyber security threat that manipulates users into surrendering sensitive credentials. Machine learning detection based on URL lexical features offers a proactive, real-time alternative to reactive blacklists, but large-scale datasets such as URL-Phish 2025 (89.56% benign vs. 10.44% phishing across 111,660 records) exhibit extreme class imbalance that biases classifiers toward the majority class. This study analyzes the performance of combining the Synthetic Minority Over-sampling Technique (SMOTE) with the Light Gradient Boosting Machine (LightGBM) classifier, following the experimental baseline of Dam and Tran (2025). The dataset was divided using a stratified 75:10:15 split into training (83,745), validation (11,166), and testing (16,749) sets. SMOTE was applied only to the training set, balancing it to 75,000:75,000 samples in 0.5339 seconds. On 16,749 test samples, the optimal LightGBM + SMOTE model (Scenario II) achieved Accuracy of 97.94% (+4.29% over baseline), Recall of 98.23% (+56.83%), F1-Score of 90.88% (+33.25%), and ROC-AUC of 99.32%, while reducing False Negative threat leakage by 96.98% (from 1,025 to 31 cases), at an inference latency of only 0.0039 ms per URL. These results confirm that SMOTE and LightGBM together form an effective, real-time-ready cybersecurity detection solution.

Keywords: *Phishing Website, LightGBM, SMOTE, Class Imbalance, Lexical URL Features*

Abstrak

Serangan *website phishing* merupakan ancaman keamanan siber serius yang memanipulasi pengguna untuk menyerahkan kredensial sensitif. Deteksi berbasis *machine learning* pada fitur leksikal URL menawarkan solusi proaktif secara real-time dibanding daftar hitam yang reaktif, namun dataset berskala besar seperti URL-*Phish* 2025 (89,56% benign vs 10,44% phishing dari 111.660 records) mengalami ketidakseimbangan kelas ekstrem yang membuat model bias ke kelas mayoritas. Penelitian ini menganalisis performa kombinasi Synthetic Minority Over-sampling Technique (SMOTE) dan algoritma *Light Gradient Boosting Machine* (LightGBM), mengacu pada metodologi eksperimen Dam dan Tran (2025). Dataset dibagi menggunakan stratified split 75:10:15 menjadi data latih (83.745), validasi (11.166), dan uji (16.749). SMOTE diterapkan hanya pada data latih, menyeimbangkan rasio menjadi 75.000:75.000 dalam 0,5339 detik. Pada 16.749 data uji, model optimal *Light*GBM + SMOTE (Skenario II) mencapai Akurasi 97,94% (+4,29% dari baseline), *Recall* 98,23% (+56,83%), F1-*Score* 90,88% (+33,25%), dan ROC-AUC 99,32%, serta menekan kebocoran *False Negative* sebesar 96,98% (dari 1.025 menjadi 31 kasus), dengan latensi inferensi hanya 0,0039 ms per URL. Hasil ini membuktikan bahwa kombinasi SMOTE dan LightGBM merupakan solusi deteksi keamanan siber yang efektif dan siap real-time.

Keywords: *Website Phishing*, *Light*GBM, SMOTE, Ketidakseimbangan Kelas, Fitur Leksikal URL

PENDAHULUAN

Keamanan siber menjadi isu yang sangat krusial seiring pesatnya transformasi digital dan tingginya ketergantungan masyarakat terhadap layanan daring. Salah satu ancaman yang paling marak adalah serangan *website phishing*, yaitu bentuk rekayasa sosial di mana penyerang merancang situs tiruan untuk memanipulasi pengguna agar menyerahkan data sensitif seperti kredensial akun dan kode OTP (Blake, 2025; Latif, 2025), dengan lebih dari satu juta insiden baru tercatat hanya pada kuartal pertama tahun 2025 (APWG, 2025). Metode deteksi konvensional berbasis blacklist bersifat reaktif dan gagal menangkal serangan *zero-day* yang secara dinamis mengubah struktur URL, sehingga pendekatan machine learning berbasis fitur leksikal

URL menjadi solusi proaktif yang mampu bekerja secara *real-time* tanpa perlu mengunduh konten *website* (Latif, 2025; Mahmud & Wirawan, 2024).

Tantangan utama penerapan *machine learning* pada deteksi phishing skala industri adalah ketidakseimbangan kelas (*class imbalance*) yang ekstrem (Talukder et al., 2024). Pada dataset URL-Phish 2025 yang terdiri atas 111.660 records, proporsi kelas legitimate mencapai 89,56% (100.000 sampel) sedangkan kelas phishing hanya 10,44% (11.660 sampel) (Dam & Tran, 2025). Dam dan Tran (2025) melaporkan bahwa model *baseline* seperti *Random Forest* hanya mencapai *Recall* 62,50% terhadap kelas phishing meskipun akurasi nominalnya tinggi (94,10%) — fenomena yang dikenal sebagai accuracy paradox, di mana model menjadi bias terhadap kelas mayoritas (Woods et al., 2025; Wilk-Jakubowski et al., 2025). Dalam konteks keamanan siber, angka false negative yang tinggi sangat berbahaya karena meloloskan ancaman aktif kepada pengguna.

Untuk mengatasi bias tersebut, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) yang dipelopori Chawla et al. (2002) terbukti menyeimbangkan distribusi kelas melalui interpolasi sampel sintetis tanpa risiko *overfitting* seperti pada duplikasi acak, serta terbukti mendongkrak *Recall* dan F1-Score deteksi phishing pada berbagai studi (Rahman & Maslan, 2021; Saputra, 2024), termasuk optimasi akurasi hingga 97,10% saat dipadukan dengan gradient boosting (Zheng et al., 2024). Di sisi algoritma klasifikasi, Light Gradient Boosting Machine (LightGBM) menawarkan efisiensi komputasi unggul dibanding ensemble tree konvensional melalui arsitektur *leaf-wise tree growth, Gradient-based One-Side Sampling* (GOSS), dan *Exclusive Feature Bundling*/EFB (Al Subaiey et al., 2024; Tjahjono et al., 2025).

Kendati demikian, penelitian Tjahjono et al. (2025) belum menerapkan SMOTE, sedangkan Tyas et al. (2026) dan Mahmud dan Wirawan (2024) masih terbatas pada dataset kecil (<30.000 records) atau tanpa penanganan class imbalance sehingga menyisakan kebocoran siber yang signifikan. Pendekatan berbasis citra seperti VGG16 pada Gabriel et al. (2026) membutuhkan komputasi grafis besar sehingga kurang adaptif untuk deteksi real-time. Kondisi ini menunjukkan adanya celah penelitian (research gap): belum tersedia kajian empiris yang menguji secara khusus integrasi SMOTE dan efisiensi *Light*GBM pada fitur leksikal tabular berskala di atas 100.000 data. Ringkasan posisi penelitian ini terhadap sejumlah studi terdahulu disajikan pada Tabel 1.

Tabel 1 Penelitian Terdahulu

Peneliti & Tahun	Metode	Hasil Utama	Keterbatasan
Dam & Tran (2025)	*Random Forest*	Akurasi 94,10%	*Recall* kelas *phishing* rendah (62,50%)
Zheng et al. (2024)	SMOTE + *Light*GBM	Akurasi 97,10%	Fokus pada IDS jaringan, bukan fitur leksikal URL
Mahmud & Wirawan (2024)	*Light*GBM *Baseline*	*Recall* 58,20%	Kebocoran *False Negative* tinggi
Tjahjono et al. (2025)	XGBoost + *Light*GBM	Akurasi 96,50%	Tanpa *oversampling* data latih
Tyas et al. (2026)	*Light*GBM + SMOTE	Akurasi 96,80%	Ukuran dataset terbatas (<30.000 data)
Gabriel et al. (2026)	VGG16 CNN	Akurasi 95,80%	Komputasi grafis besar, kurang real-time
Rahman & Maslan (2021)	SMOTE + *Decision Tree*	F1-*Score* 88,40%	Waktu komputasi pelatihan lambat

Sebagai kelanjutan dari kajian penerapan machine learning pada berbagai domain yang telah dilakukan sebelumnya, termasuk pada klasterisasi lahan pertanian (Mukhtar et al., 2024) dan prediksi curah hujan (Hendra et al., 2023), penelitian ini memperluas cakupan penerapannya ke ranah keamanan siber. Berdasarkan tabel 1, penelitian ini bertujuan (1) mengevaluasi dan membandingkan performa komprehensif (*Accuracy, Precision, Recall*, F1-*Score*, ROC-AUC, dan waktu komputasi) algoritma LightGBM sebelum (Skenario I/*baseline*) dan sesudah (Skenario II/optimal) penyeimbangan data latih dengan SMOTE, mengacu pada metodologi eksperimen Dam dan Tran (2025); serta (2) membuktikan efektivitas kombinasi SMOTE dan *Light*GBM sebagai solusi deteksi *website phishing* yang akurat, sensitif terhadap kelas minoritas, dan hemat sumber daya komputasi.

METODOLOGI PENELITIAN

Penelitian ini menerapkan pendekatan eksperimen kuantitatif dengan delapan tahapan utama sebagaimana diilustrasikan pada Gambar 1.

Gambar 1 Alur Penelitian

Dataset

Dataset yang digunakan adalah URL-Phish 2025, dataset sekunder publik yang dipublikasikan oleh Dam dan Tran (2025) melalui repositori Mendeley Data, terdiri atas 111.660 records dan 26 kolom, dengan distribusi kelas timpang: 100.000 sampel (89,56%) legitimate/benign dan 11.660 sampel (10,44%) phishing.

Pra-pemrosesan Data

Tahap pembersihan data mengeliminasi 4 kolom bertipe string non-fitur (url, domain, tld, dan scheme metadata), memangkas dimensi dataset dari 26 menjadi 22 kolom fitur leksikal numerik murni tanpa ditemukan nilai kosong (null). Label kelas dikodekan secara biner (0 = legitimate, 1 = phishing), dan seluruh fitur numerik dinormalisasi menggunakan *StandardScaler* ($z = (x - \mu) / \sigma$) untuk menyamakan skala nilai antar fitur. Delapan dari 22 fitur leksikal tersebut disajikan pada tabel 2.

Tabel 2 Contoh Fitur Leksikal Dataset URL-Phish 2025

Nama Fitur	Tipe	Definisi Sintaksis	Karakteristik Diskriminatif
url_len	*Integer*	*Len*(*string*_URL)	URL *phishing* cenderung sangat panjang
domain_len	*Integer*	*Len*(*string_domain*)	Domain tiruan seringkali panjang
tld_len	*Integer*	*Len*(*string*_TLD)	Panjang Top Level Domain (.com, .info, dsb)
count_dot	*Integer*	Jumlah karakter '.'	Banyaknya titik mencerminkan subdomain bertingkat
count_hyphen	*Integer*	Jumlah tanda '-'	Tanda hubung sering meniru nama brand
count_slash	*Integer*	Jumlah karakter '/'	Mengukur kedalaman struktur path direktori
count_at	*Integer*	Jumlah simbol '@'	Simbol '@' memicu pengalihan kredensial
count_equal	*Integer*	Jumlah tanda '='	Indikator banyaknya variabel parameter URL

Selengkapnya, seluruh 22 fitur numerik tersebut menjadi masukan langsung bagi algoritma *Light*GBM tanpa reduksi dimensi lebih lanjut, mengingat *Light*GBM secara inheren mampu menyeleksi fitur paling informatif melalui mekanisme split gain pada setiap node pohon keputusan.

Pemisahan Data

Dataset dibagi menggunakan *Stratified Split* dengan rasio 75% data latih, 10% data validasi, dan 15% data uji sebagaimana dirinci pada Tabel 3, guna menjaga proporsi kelas yang konsisten pada setiap subset.

Tabel 3 Pembagian Dataset

Subset Data	Persentase	*Legitimate*	*Phishing*	Total Sampel
Data Latih (*Training*)	75%	75.000	8.745	83.745
Data Validasi (*Validation*)	10%	10.000	1.166	11.166
Data Uji (*Testing*)	15%	15.000	1.749	16.749

Penyeimbangan Kelas (SMOTE)

Ketidakseimbangan kelas pada subset data latih (75.000 legitimate: 8.745 phishing) ditangani menggunakan SMOTE (k_neighbors = 5), yang mensintesis sampel minoritas baru melalui interpolasi linier antar-tetangga terdekat. SMOTE hanya diterapkan pada data latih — data validasi dan data uji tetap pada distribusi aslinya agar evaluasi tetap merepresentasikan kondisi dunia nyata.

Pemodelan Klasifikasi LightGBM

Algoritma *Light Gradient Boosting Machine* (LightGBM) dipilih karena efisiensi arsitektur leaf-wise tree growth yang dipadukan dengan Gradient-based *One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling*/EFB (Ke et al., 2017).: Skenario I (Baseline) dilatih pada data latih asli tanpa SMOTE dengan *hyperparameter* *n_estimators*=25, *max_depth*=3, *learning_rate*=0,05, *random_state*=42; sedangkan Skenario II (Optimal) dilatih pada data latih hasil SMOTE dengan konfigurasi lebih kompleks *n_estimators*=100, *max_depth*=10, *learning_rate*=0,10, *random_state*=42, dilengkapi pemantauan *binary_logloss* pada data validasi dan early stopping 10 iterasi untuk mencegah overfitting.

Konfigurasi Skenario I dibuat ringan dan minim tuning untuk merepresentasikan praktik umum penerapan LightGBM tanpa penanganan khusus terhadap class imbalance, sehingga berfungsi sebagai pembanding baseline yang wajar. Sebaliknya, konfigurasi Skenario II diperluas jumlah pohon dan kedalaman maksimum ditingkatkan serta learning rate dinaikkan untuk mengakomodasi kompleksitas data latih hasil SMOTE yang meningkat dua kali lipat (150.000 sampel), sementara early stopping pada data validasi ditambahkan guna mencegah overfitting akibat kapasitas model yang lebih besar tersebut.

Lingkungan Eksperimen

Seluruh eksperimen dijalankan pada lingkungan komputasi sebagaimana dirinci pada Tabel 4. Karena dataset URL-Phish 2025 tersedia terbuka melalui repositori Mendeley Data (Dam & Tran, 2025).

Tabel 4 Spesifikasi Lingkungan Komputasi

Komponen Sistem	Spesifikasi/Versi	Peran dalam Eksperimen
Bahasa Pemrograman	Python 3.10.0 (64-bit)	Penerjemah kode utama eksperimen
Algoritma Klasifikasi	LightGBM Classifier v4.0.0	Model boosting pohon keputusan utama
Teknik Oversampling	SMOTE (imbalanced-learn v0.11)	Sintesis data latih kelas minoritas
Pustaka Evaluasi	Scikit-Learn v1.3.0	Perhitungan metrik evaluasi & pembagian data
Memori Komputer	16 GB DDR4 RAM	Media eksekusi komputasi data berskala besar

Evaluasi Performa

Performa kedua skenario diuji secara independen pada 16.749 sampel data uji (15.000 legitimate, 1.749 phishing) menggunakan metrik Confusion Matrix, *Accuracy, Precision, Recall*, F1-*Score*, ROC-AUC, serta waktu komputasi (pelatihan dan inferensi). Kelima metrik utama dihitung berdasarkan komponen *Confusion Matrix* — *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) — menggunakan formulasi berikut:

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1-*Score* = 2 x (*Precision* x *Recall*) / (*Precision* + *Recall*)

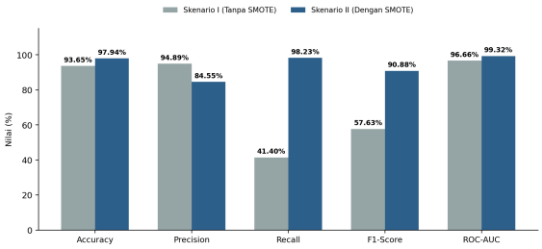
Recall menjadi metrik paling krusial dalam konteks keamanan siber karena secara langsung mengukur proporsi ancaman phishing aktif yang berhasil terdeteksi, sedangkan ROC-AUC digunakan untuk menilai kemampuan model memisahkan kedua kelas pada keseluruhan ambang batas (threshold) klasifikasi.

HASIL DAN PEMBAHASAN

Tabel 5 dan Gambar 2 menyajikan perbandingan metrik evaluasi kedua skenario pada 16.749 sampel data uji.

Tabel 5 Hasil Evaluasi Metrik Komparatif Skenario I vs Skenario II

Metrik Evaluasi	Skenario I (Tanpa SMOTE)	Skenario II (Dengan SMOTE)	Peningkatan
Accuracy	93,65%	97,94%	+4,29%
Precision	94,89%	84,55%	-10,34%
Recall	41,40%	98,23%	+56,83%
F1-Score	57,63%	90,88%	+33,25%
ROC-AUC	96,66%	99,32%	+2,66%

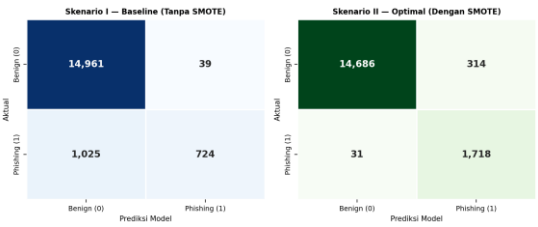


Gambar 2 Perbandingan Metrik Evaluasi Skenario I vs Skenario II

Model Skenario II (LightGBM + SMOTE) mencapai Akurasi 97,94% (naik +4,29% dari baseline 93,65%), Recall 98,23% (melonjak +56,83% dari baseline 41,40%), F1-Score 90,88% (naik +33,25% dari baseline 57,63%), dan ROC-AUC 99,32% (naik +2,66% dari baseline 96,66%). Nilai Precision menurun dari 94,89% menjadi 84,55% (-10,34%) sebagai konsekuensi pergeseran decision boundary akibat SMOTE — trade-off yang wajar dan menguntungkan dalam konteks pertahanan siber, karena kenaikan sedikit false alarm jauh lebih murah ditangani dibandingkan lolosnya ancaman phishing aktif.

Secara teknis, penurunan Precision ini terjadi karena SMOTE menggeser decision boundary model agar lebih sensitif terhadap pola minoritas, sehingga sejumlah URL legitimate dengan karakteristik leksikal mirip phishing (mis. URL panjang atau mengandung tanda hubung) ikut terklasifikasi sebagai phishing. Secara operasional, konsekuensi ini tergolong dapat dikelola, false positive dapat disaring melalui mekanisme verifikasi tambahan seperti whitelist domain resmi tanpa risiko kerugian finansial, berbeda dengan false negative yang berpotensi langsung merugikan pengguna melalui pencurian kredensial.

Perbandingan rinci confusion matrix kedua skenario disajikan pada Gambar 3 dan Tabel 6.



Gambar 3 Confusion Matrix Skenario I vs Skenario II

Tabel 6 Analisis Penurunan Kasus False Negative (Kebocoran Siber)

Skenario Model	Jumlah FN (Phishing Lolos)	Persentase Kebocoran	Tingkat Pencegahan
Skenario I (Tanpa SMOTE)	1.025 kasus	58,60%	41,40% (Baseline)
Skenario II (Dengan SMOTE)	31 kasus	1,77%	98,23% (Penurunan FN 96,98%)

Pada Skenario I, model gagal mendeteksi 1.025 dari 1.749 sampel phishing pada data uji (False Negative), merepresentasikan tingkat kebocoran siber sebesar 58,60% dan tingkat pencegahan hanya 41,40%. Setelah penerapan SMOTE pada Skenario II, jumlah False Negative dipangkas drastis menjadi hanya 31 kasus — penurunan sebesar 96,98% — sehingga tingkat pencegahan ancaman siber melonjak menjadi 98,23%. Sebagai konsekuensinya, False Positive meningkat dari 39 menjadi 314 sampel, namun peningkatan ini jauh lebih dapat ditoleransi dibandingkan risiko lolosnya 1.025 ancaman phishing aktif pada Skenario I.

Efisiensi Komputasi

Meskipun dilatih pada populasi data 150.000 sampel (hampir dua kali lipat Skenario I), model Skenario II hanya membutuhkan waktu pelatihan 0,8452 detik, mengonfirmasi keunggulan arsitektur leaf-wise tree growth dan histogram-based feature binning LightGBM. Latensi inferensi tercatat 0,0727 detik untuk mengklasifikasikan seluruh 16.749 sampel data uji secara batch, atau setara 0,0039 milidetik (3,9 mikrodetik) per URL — kecepatan yang mengindikasikan potensi kelayakan model diintegrasikan sebagai inline filter pada Web Application Firewall (WAF) atau Intrusion Detection/Prevention System (IDS/IPS). Perlu dicatat bahwa pengukuran ini dilakukan pada lingkungan eksperimen offline (Tabel 4) dan belum memperhitungkan overhead jaringan produksi sesungguhnya, sehingga validasi lebih lanjut pada lingkungan real-time tetap diperlukan sebelum implementasi produksi.

Hasil ini konsisten dan melampaui sejumlah studi acuan: Recall 98,23% pada penelitian ini jauh melampaui baseline Random Forest pada dataset yang sama oleh Dam dan Tran (2025) yang hanya mencapai 62,50%, serta melampaui akurasi kombinasi SMOTE+LightGBM pada domain intrusion detection oleh Zheng et al. (2024) sebesar 97,10%. Temuan ini juga menjembatani celah penelitian Tjahjono et al. (2025) yang belum menerapkan oversampling, sekaligus membuktikan pendekatan ini tetap efisien pada skala data (111.660 records) yang jauh lebih besar dibandingkan batasan dataset pada studi Tyas et al. (2026) (<30.000 records). Dengan demikian, kombinasi SMOTE dan LightGBM terbukti menjadi solusi yang akurat, sensitif terhadap kelas minoritas, dan efisien secara komputasi untuk deteksi website phishing berskala besar.

Ditinjau dari struktur fitur pada Tabel 2, karakteristik leksikal seperti url_len, count_hyphen, dan count_at secara konseptual berperan penting dalam pembentukan decision boundary LightGBM, karena URL phishing secara umum dirancang lebih panjang dan menyisipkan tanda hubung atau simbol '@' untuk meniru identitas domain resmi guna mengelabui korban (Blake, 2025; Latif, 2025). Interpretasi ini konsisten dengan mekanisme split gain pada LightGBM, yang secara otomatis memprioritaskan fitur dengan daya pisah kelas tertinggi pada setiap percabangan pohon keputusan tanpa memerlukan seleksi fitur manual.

SIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa kombinasi teknik oversampling SMOTE dan algoritma LightGBM secara signifikan meningkatkan performa klasifikasi website phishing pada dataset besar dan timpang. Model optimal (Skenario II) mencapai Akurasi 97,94%, Recall 98,23%, F1-Score 90,88%, dan ROC-AUC 99,32%, dengan penurunan kasus False Negative sebesar 96,98% (dari 1.025 menjadi 31 kasus) dibandingkan model baseline tanpa SMOTE. Trade-off penurunan Precision (94,89% menjadi 84,55%) terbukti sepadan mengingat signifikansi penurunan risiko lolosnya ancaman phishing aktif. Dengan latensi inferensi hanya 0,0039 milidetik per URL, model ini menunjukkan indikasi kesiapan untuk diintegrasikan pada sistem keamanan siber real-time seperti WAF maupun IDS/IPS, meskipun validasi lebih lanjut pada lingkungan produksi tetap diperlukan..

Penelitian selanjutnya disarankan untuk: (1) menggabungkan fitur leksikal dengan pendekatan Natural Language Processing/Transformer (mis. BERT, URLTran) serta analisis konten HTML/DOM guna mendeteksi teknik obfuscation tingkat lanjut; (2) menguji model pada lalu lintas jaringan real-time serta mengevaluasi ketahanannya terhadap serangan adversarial; dan (3) mengeksplorasi teknik oversampling hybrid seperti SMOTE-Tomek atau SMOTE-ENN untuk membersihkan sampel sintetis di area perbatasan kelas, guna menekan False Positive tanpa mengorbankan Recall.

DAFTAR PUSTAKA

Al Subaiey, A., Al-Thani, M., & Erbad, A. (2024). Interpretable Web AI Platform for Real-Time Lexical URL Phishing Detection. IEEE Access, 12, 45120-45135.

Anti-Phishing Working Group (APWG). (2025). Phishing Activity Trends Report, 1st Quarter 2025. APWG Technical Report.

Blake, R. (2025). PhishSense: 1B Model for URL Obfuscation and Phishing Detection. Journal of Information Security, 16(2), 89-104.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Dam, M. L., & Tran, C. H. (2025). URL-Phish: A comprehensive dataset for URL-based phishing detection. Mendeley Data, V1. https://doi.org/10.17632/v6234j4b29.1

Gabriel, P. R. E., Sari, A. P., & Junaidi, A. (2026). Penerapan VGG16 pada identifikasi website phishing berbasis gambar screenshot. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),

13(1), 45-56.

Hendra, Y., Mukhtar, H., Baidarus, & Hafsari, R. (2023). Prediksi Curah Hujan di Kota Pekanbaru Menggunakan LSTM (Long Short Term Memory). Journal of Software Engineering and Information System (SEIS), 3(2), 74–81. https://doi.org/10.37859/seis.v3i2.5606

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (NeurIPS 30), 3146-3154.

Latif, S. (2025). Detecting Phishing Attacks in Cybersecurity Using Machine Learning Algorithms. Cybersecurity Journal, 8(1), 77-92.

Mahmud, A. F., & Wirawan, S. (2024). Deteksi Phishing Website menggunakan Machine Learning LightGBM dan SMOTE. Jurnal Systemic, 10(2), 101-112.

Mukhtar, H., Syafutri, T. M., Rahman, R. A., Putra, A., & Hafsari, R. (2024). Analisis Kesuburan Pertanian Melalui Irigasi dengan Menggunakan Metode K-Means Clustering. Journal of Software Engineering and Information System (SEIS), 4(2), 102–107. https://doi.org/10.37859/seis.v4i2.7599

Rahman, A., & Maslan, A. (2021). Imbalanced Dataset Handling in Phishing Detection Using SMOTE. International Journal of Advanced Computer Science and Applications, 12(8), 340-348.

Saputra, R. (2024). Implementation of Synthetic Minority Oversampling Technique (SMOTE) on Phishing Website Classification. Jurnal Ilmu Komputer dan Informasi, 17(1), 23-31.

Talukder, M. A., Islam, M. M., & Uddin, M. A. (2024). Evaluation Metrics for Imbalanced Datasets in Cybersecurity. Data & Knowledge Engineering, 150, 102280.

Tjahjono, A. F., Hasan, Putera, R. P., Indranto, D. M. P., & Hermawan, A. T. (2025). Machine Learning-Based Web Phishing Detection Model using LightGBM and Extreme Gradient Boosting. International Journal of Computer Applications, 186(12), 1-9.

Tyas, W. S., Rafrastara, F. A., & Ghozi, W. (2026). Optimizing URL-Based Phishing Detection Using LightGBM and SMOTE. Jurnal Informatika dan Komputer (JIK), 11(1), 15-26.

Wilk-Jakubowski, J. L., Pawlik, L., Wilk-Jakubowski, G., & Sikora, A. (2025). Machine Learning Classifiers for Detection of Phishing URLs: Performance Analysis. Applied Sciences, 15(3), 1205.

Woods, J., Patel, K., & Smith, M. (2025). Predicting Phishing Attacks Using Support Vector Machine and Random Forest. Information Security Journal, 34(1), 45-59.

Zheng, Q., Zhang, X., & Liu, Y. (2024). LightGBM and SMOTE-Based Imbalanced Data Classification for Network Intrusion. IEEE Access, 12, 18920-18931.