P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

PERBANDINGAN KINERJA ALGORITMA RANDOM FOREST DAN LIGHT GRADIENT BOOSTING MACHINE UNTUK DETEKSI DOMAIN PHISING BERBASIS ANALISIS TYPOSQUATTING

Puteri Elisabeth Pakpahan[1)], Nuke Lu'lu Ul Chusna[2)], Harry Dwiyana Kartika[3)]

[1,2,3] *Program Studi Teknik Informatika, Fakultas Teknik, Universitas Krisnadwipayana*

email: [1] puteripkphn@gmail.com, [2] nukelchusna@unkris.ac.id , [3] harry_d@unkris.ac.id

ARTICLE INFO

Article History:
Recieved : 2 Oktober 2026
Accepted : 19 Oktober 2026
Published : 16 November 2026

Keywords:
Phishing;
Typosquatting;
Random Forest;
LightGBM;
SHAP;

IEEE style in citing this article:
P. E. Pakpahan, N. L. U. Chusna, and H. D. Kartika, "Perbandingan Kinerja Algoritma Random Forest dan Light Gradient Boosting Machine untuk Deteksi Domain Phishing Berbasis Analisis Typosquatting", *jurnal.ilmiah.informatika*, vol. 11, no. 2, pp. 221-xx, Des. 2026.

Corresponding Author:
Puteri Elisabeth Pakpahan
Teknik Informatika
Universitas Krisnadwipayana

ABSTRACT

Typosquatting-based phishing registers domains that resemble official brands to deceive users, and blacklist-based protection cannot stop newly registered look-alike domains. This study compares Random Forest and Light Gradient Boosting Machine (LightGBM) for phishing domain detection using typosquatting, lexical, suspicious-keyword, and top-level-domain features, and deploys the best model in a Streamlit simulation with SHAP interpretability. Following CRISP-DM, the dataset contains 7,999 domains after cleaning (5,000 legitimate domains from the Tranco list and 2,999 phishing domains generated for 13 target brands through the HaveIBeenSquatted API), described by 10 numeric features. Class imbalance was handled with class_weight='balanced', and both models were tuned with GridSearchCV on a stratified 70:30 split. With phishing-class recall as the main criterion, the tuned LightGBM reached an accuracy of 94.37%, a phishing precision of 95.16%, and a phishing recall of 89.56%; the tuned Random Forest reached 94.46%, 95.71%, and 89.22%. The difference is small, and LightGBM was selected for its higher recall. A local SHAP explanation showed jaro_winkler_score as the largest contributor for a sample domain. The prototype reports a prediction, a confidence score, and a feature-contribution plot.

© 2026 Jurnal Ilmiah Informatika (Scientific Informatics Journal) with CC BY NC licence

P-ISSN : 2549-7480
P. E. Pakpahan dkk/ JIMI 11 (2) pp. 221-229 E-ISSN : 2549-6301

1. PENDAHULUAN

Pertumbuhan pengguna internet di Indonesia belum diikuti kesiapan keamanan digital. Badan Siber dan Sandi Negara mencatat sejumlah insiden siber serius pada periode 2021–2025, antara lain serangan *ransomware* terhadap Pusat Data Nasional dan kebocoran data BPJS Kesehatan [1]. Pada saat yang sama, penetrasi internet nasional mencapai 81,72% dari populasi, atau sekitar 235 juta jiwa [2], sementara aspek keamanan digital masih menjadi indikator literasi digital dengan pemahaman yang relatif rendah.

Phishing adalah salah satu ancaman yang paling umum. Serangan ini memanfaatkan rekayasa sosial untuk mencuri identitas dan kredensial keuangan dengan menyamar sebagai bank, *platform e-commerce*, atau lembaga pemerintah [3], [4]. Anti-Phishing Working Group (APWG) mencatat 1.003.924 serangan pada kuartal pertama 2025 dan 1.130.393 serangan pada kuartal kedua 2025 [5]. Di Indonesia, PANDI mencatat 34.622 kasus *phishing* dalam lima tahun terakhir [3], dan pengelola domain .id menerima puluhan ribu laporan yang menyasar domain berekstensi Indonesia [6].

Salah satu tekniknya adalah *typosquatting*, yaitu pendaftaran domain yang menyerupai domain resmi melalui salah ketik, penggantian karakter yang mirip secara visual (*homoglyph*), atau penggantian karakter yang berdekatan pada papan ketik [4], [7]. Contohnya g00gle.com, paypa1.com, dan tokopedla.com. Sistem deteksi berbasis *blacklist* bersifat reaktif karena hanya memblokir domain yang sudah dilaporkan, sehingga domain tiruan yang baru didaftarkan dapat lolos [3], [8].

Pendekatan *machine learning* (ML) dapat mempelajari pola karakteristik situs secara otomatis dan menjadi alternatif bagi *blacklist* [9]. *Random Forest* mencapai akurasi 97,2% pada klasifikasi URL berbasis TF-IDF dan fitur numerik [9], serta mengungguli Naïve Bayes, *Decision Tree*, dan SVM dengan akurasi 90,77% pada studi lain [10]. XGBoost yang dituning mencapai akurasi 96% [4]. Suwarno dan Hardjianto [11] memakai 30 fitur URL pada *Random Forest*, Gabriel [12] mengusulkan model hibrida CNN–XGBoost, dan Perdana [13] membandingkan *Random Forest* dengan XGBoost. Feresa dkk. [14] menguji Naïve Bayes, *Random Forest*, dan LightGBM pada URL *phishing*. Untuk data yang tidak seimbang, Rizky dkk. [15] melaporkan bahwa LightGBM lebih baik dalam memprediksi kelas minoritas.

Penelitian tersebut umumnya memakai fitur leksikal URL yang bersifat umum dan belum menempatkan kemiripan terhadap merek resmi sebagai komponen utama. Akibatnya, domain *typosquatting* yang tampak normal secara leksikal tetap berpotensi lolos. Selain itu, model *ensemble* bersifat *black-box*, sehingga pengguna tidak mengetahui fitur apa yang membuat suatu domain ditandai *phishing*, padahal penjelasan model menjadi perhatian dalam deteksi *phishing* [16].

Penelitian ini membandingkan kinerja *Random Forest* dan *Light Gradient Boosting Machine* (LightGBM) dalam mendeteksi domain *phishing* dengan menggabungkan fitur *typosquatting*, fitur leksikal, kata kunci mencurigakan, dan fitur *Top-Level Domain* (TLD). *Random Forest* dipilih sebagai model *bagging* yang stabil, sedangkan LightGBM sebagai model *gradient boosting* yang efisien dan baik pada kelas minoritas. Model terbaik diimplementasikan dalam simulasi berbasis *Streamlit* yang dilengkapi SHAP (*SHapley Additive exPlanations*) untuk

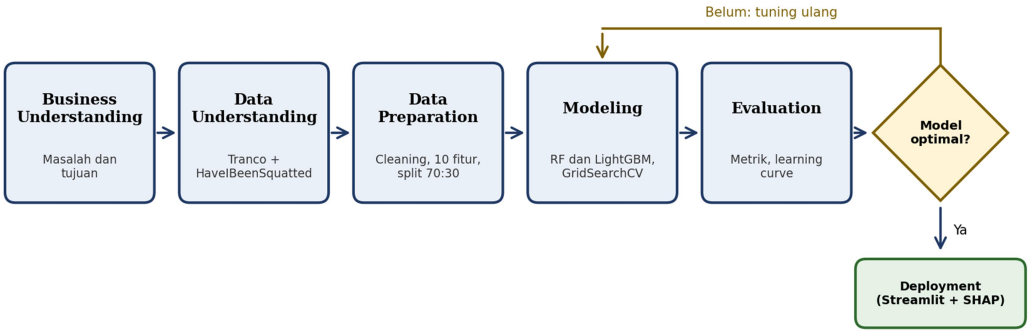
DOI : https://doi.org/10.35316/jimi.v11i2.211-229
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

menjelaskan kontribusi setiap fitur terhadap keputusan klasifikasi.

2. METODE PENELITIAN

Jenis Penelitian dan Tahapan

Penelitian ini adalah penelitian terapan dengan pendekatan kuantitatif komparatif melalui eksperimen komputasi. Kedua algoritma dibandingkan pada data, fitur, dan skema pembagian data yang sama agar perbandingan adil. Tahapan mengikuti kerangka CRISP-DM (*Cross-Industry Standard Process for Data Mining*) seperti pada Gambar 1. Seluruh proses dikerjakan dengan Python 3.12 di Google Colaboratory menggunakan Pandas, NumPy, Scikit-learn, LightGBM, SHAP, dan Streamlit.



Gambar 1. Tahapan penelitian berdasarkan CRISP-DM

Data dan Praproses

Data yang dipakai adalah data sekunder dari dua sumber. Domain *phishing* diperoleh melalui HaveIBeenSquatted API untuk 13 merek target (Amazon, Apple, Microsoft, Facebook, Google, Dana, BCA, Shopee, Traveloka, Gojek, Tokopedia, Bank Mandiri, dan Direktorat Jenderal Pajak), menghasilkan 7.638 domain dengan 18 teknik *squatting*. Domain legitim diambil dari Tranco List, yaitu daftar 1.000.000 domain berdasarkan peringkat popularitas. Dari kedua sumber diambil 3.000 domain *phishing* (label 1) dan 5.000 domain legitim (label 0) dengan *random_state*=42, sehingga rasio kelas 1,67:1. Label 1 ditetapkan sebagai kelas positif karena ancaman adalah kejadian yang ingin dideteksi [19], [20], dan label 0 sebagai kelas negatif [21].

Pembersihan data mencakup penghapusan domain kosong dan duplikat, penyeragaman huruf kecil, serta pengecekan domain yang tumpang tindih antarkelas. Satu domain yang tumpang tindih (facebook.pl) tersaring pada deduplikasi, sehingga dataset akhir berjumlah 7.999 baris (5.000 legitim dan 2.999 *phishing*) tanpa nilai kosong. Dari setiap domain diekstraksi 10 fitur numerik dalam empat kelompok:

a. Fitur leksikal: *digit_count*, *hyphen_count*, dan *entropy*. Entropi Shannon dihitung pada *second-level domain* dengan Persamaan (1), dengan p_i adalah proporsi kemunculan karakter ke-i dan k jumlah karakter unik.
b. Fitur *typosquatting*: *homoglyph_count* (karakter yang menyerupai huruf lain, misalnya 0, 1, 3, 4, 5, 7, dan @), *levenshtein_distance*, *damerau_levenshtein_distance* (menambahkan operasi transposisi), dan *jaro_winkler_score* terhadap merek referensi terdekat.

P-ISSN : 2549-7480
E-ISSN : 2549-6301

c. Fitur kata kunci: *suspicious_keywords*, yaitu jumlah kata pemicu seperti login, secure, dan verify.
d. Fitur TLD: *is_common_tld* (indikator biner) dan *tld_abuse_score* (skor risiko penyalahgunaan TLD, dengan nilai bawaan 0,2 untuk TLD yang tidak terdaftar).

$$H(s) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad (1)$$

Pemodelan dan Evaluasi

Data dibagi dengan *stratified split* 70:30 (data latih 5.599 dan data uji 2.400). Seluruh fitur dinormalisasi dengan *StandardScaler* yang di-*fit* hanya pada data latih untuk mencegah *data leakage*. Ketidakseimbangan kelas ditangani dengan *class_weight*='balanced' pada kedua algoritma tanpa *oversampling* (SMOTE). Model *baseline Random Forest* memakai *n_estimators*=100, dan LightGBM memakai *objective*='binary'.

Optimasi *hyperparameter* memakai *GridSearchCV* dengan *StratifiedKFold* 5 lipatan dan skor F1 agar kedua kelas seimbang [17]. Ruang pencarian *Random Forest* adalah *n_estimators* {100, 200}, *max_depth* {5, 8, 10}, *min_samples_split* {10, 20}, dan *min_samples_leaf* {5, 10}. Ruang pencarian LightGBM adalah *n_estimators* {300}, *learning_rate* {0,01; 0,05}, *num_leaves* {15, 31}, *max_depth* {5, 8}, *min_child_samples* {20, 40}, *reg_alpha* {0,1; 0,5}, dan *reg_lambda* {0,1; 0,5}. Kompleksitas ruang pencarian sengaja dibatasi untuk menekan *overfitting*. Selisih skor F1 latih dan validasi silang di atas 0,05 dianggap peringatan *overfitting*, dan model LightGBM final dilatih ulang dengan *early stopping*.

Kinerja diukur dengan *confusion matrix* dan metrik pada Persamaan (2)–(5) [18], dengan TP, TN, FP, dan FN berturut-turut adalah *true positive*, *true negative*, *false positive*, dan *false negative*. *Recall* kelas *phishing* menjadi kriteria utama karena *false negative* (domain *phishing* dianggap aman) lebih berisiko daripada *false positive*. Stabilitas model diperiksa dengan membandingkan skor data latih dan data uji serta dengan *learning curve* (*StratifiedKFold* 5 lipatan, metrik akurasi).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Re} \quad (5)$$

Model terbaik diimplementasikan sebagai prototipe *Streamlit* yang menerima nama domain, mengekstraksi fitur secara otomatis, dan menampilkan prediksi, skor kepercayaan, serta grafik kontribusi fitur dari SHAP *TreeExplainer*.

3. HASIL DAN PEMBAHASAN

Hasil Tuning Hyperparameter

Pada model *baseline*, akurasi data latih *Random Forest* adalah 0,9718 dan LightGBM 0,9645. *Learning curve baseline* menunjukkan selisih yang menetap antara skor latih dan validasi silang, sekitar 3–4 poin persentase pada *Random Forest* dan sekitar 2,3 poin persentase pada LightGBM. Pola ini mengindikasikan kecenderungan *overfitting* tanpa pembatasan kompleksitas, sehingga *tuning* dengan regularisasi dilakukan. Kombinasi parameter terbaik disajikan pada Tabel 1.

DOI : https://doi.org/10.35316/jimi.v11i2.211-229
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

P-ISSN: 2549-7480
E-ISSN: 2549-6301

Tabel 1. Parameter terbaik hasil GridSearchCV (split 70:30)

Parameter	*Random Forest*	LightGBM
n_estimators	200	300
max_depth	10	5
min_samples_split	10	–
min_samples_leaf	5	–
max_features	*sqrt*	–
learning_rate	–	0,05
num_leaves	–	15
min_child_samples	–	20
reg_alpha; *reg_lambda*	–	0,5; 0,1

Skor F1 validasi silang terbaik: *Random Forest* 0,9226 dan LightGBM 0,9278.

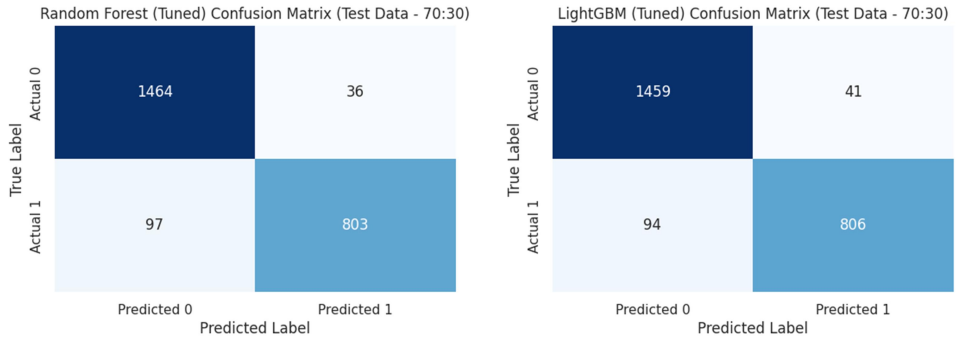
Setelah *tuning*, akurasi data latih turun menjadi 0,9505 pada *Random Forest* dan 0,9539 pada LightGBM. Penurunan ini merupakan pembatasan kompleksitas model saat tuning mencegah model menghafal data latih, sehingga skor latih turun sedikit tetapi model lebih mampu mengenali data baru. Selisih skor F1 latih dan validasi silang adalah 0,0096 (0,9322 dan 0,9226) pada *Random Forest* dan 0,0127 (0,9406 dan 0,9278) pada LightGBM, keduanya jauh di bawah ambang 0,05.

Kinerja pada Data Uji

Kinerja pada data uji disajikan pada Tabel 2 dan Gambar 2. Dari 900 domain *phishing* pada data uji, *Random Forest* mendeteksi 803 domain dan melewatkan 97 (*false negative*), sedangkan LightGBM mendeteksi 806 domain dan melewatkan 94. Pada 1.500 domain legitim, *Random Forest* salah menandai 36 domain dan LightGBM salah menandai 41 domain sebagai *phishing* (*false positive*).

Tabel 2. Kinerja model *tuned* pada data uji (2.400 domain)

Model	Kelas	*Precision*	*Recall*	*F1-score*	Akurasi
Random Forest	Aman (0)	0,9379	0,9760	0,9566	0,9446
	Phishing (1)	0,9571	0,8922	0,9235	
LightGBM	Aman (0)	0,9395	0,9727	0,9558	0,9437
	Phishing (1)	0,9516	0,8956	0,9227	



Gambar 2. *Confusion matrix* data uji: *Random Forest* (kiri) dan LightGBM (kanan)

P-ISSN : 2549-7480
P. E. Pakpahan dkk/ JIMI 11 (2) pp. 221-229 E-ISSN : 2549-6301

Perbandingan Model dan Stabilitas

Pada Tabel 2, *Random Forest* sedikit lebih tinggi pada *precision* kelas *phishing* (0,9571 dibanding 0,9516) dan akurasi (0,9446 dibanding 0,9437), sedangkan LightGBM lebih tinggi pada *recall* kelas *phishing* (0,8956 dibanding 0,8922). Selisih *recall* tersebut setara tiga domain dari 900 domain *phishing* pada data uji, sehingga perbedaan kedua model kecil. Karena kriteria utama penelitian ini adalah *recall* kelas *phishing*, LightGBM *tuned* dipilih sebagai model akhir. Pemilihan ini perlu dibaca bersama konsekuensinya: LightGBM menghasilkan lebih banyak *false positive* (41 dibanding 36). Karakter kedua algoritma sejalan dengan dugaan awal. *Random Forest* berbasis *bagging* stabil dan memberi *precision* tinggi, sedangkan LightGBM berbasis *boosting* memperbaiki kesalahan secara bertahap dan sedikit lebih peka terhadap kelas *phishing*.

Stabilitas model diringkas pada Tabel 3. Setelah *tuning*, selisih akurasi data latih dan data uji hanya sekitar 0,6 poin persentase pada *Random Forest* dan 1,0 poin persentase pada LightGBM, lebih kecil dibandingkan model *baseline* yang sekitar 2–3 poin persentase. *Learning curve* pada Gambar 3 mendukung hal ini. Pada *Random Forest tuned*, skor latih turun dari sekitar 0,956 dan stabil di kisaran 0,950, sedangkan skor validasi silang naik dari sekitar 0,938 menjadi sekitar 0,944 pada 4.500 sampel latih, sehingga selisih akhirnya sekitar 0,6 poin persentase. Kurva validasi yang masih naik menunjukkan bahwa penambahan data masih dapat membantu.

Tabel 3. Stabilitas model setelah *tuning*

Model	F1 latih	F1 validasi silang	Selisih F1	Akurasi latih	Akurasi uji
Random Forest	0,9322	0,9226	0,0096	0,9505	0,9446
LightGBM	0,9406	0,9278	0,0127	0,9539	0,9437



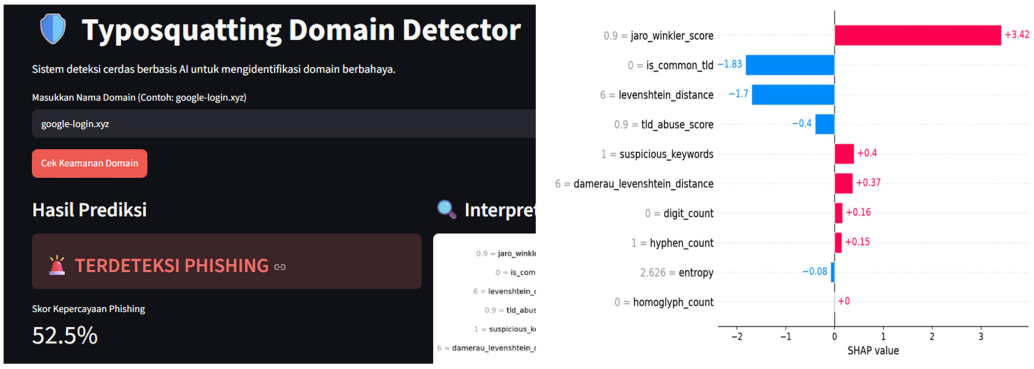
Gambar 3. *Learning curve* model *tuned*: *Random Forest* (kiri) dan LightGBM (kanan)

Implementasi dan Interpretasi SHAP

Model LightGBM *tuned* diimplementasikan pada prototipe *Streamlit* (Gambar 4, kiri). Sebagai contoh, domain google-login.xyz diklasifikasikan sebagai *phishing* dengan skor kepercayaan 52,5%. Skor ini adalah probabilitas kelas *phishing* untuk satu domain, sedikit di atas ambang 0,5, sehingga keyakinan model pada domain ini tergolong tipis. Skor tersebut berbeda dari akurasi 94,37% yang merupakan ukuran agregat pada seluruh data uji.

Penjelasan SHAP untuk domain tersebut ditampilkan pada Gambar 4 (kanan). Nilai positif mendorong prediksi ke arah *phishing* dan nilai negatif ke arah aman. *jaro_winkler_score* bernilai 0,9 dan menjadi kontributor terbesar ke arah

phishing (+3,42), yang sesuai dengan kemiripan ejaan yang tinggi terhadap "google". Fitur *suspicious_keywords* (+0,40), *damerau_levenshtein_distance* (+0,37), *digit_count* (+0,16), dan *hyphen_count* (+0,15) juga mendorong ke arah *phishing* dengan kontribusi yang lebih kecil. Sebaliknya, *is_common_tld* (−1,83), *levenshtein_distance* (−1,70), *tld_abuse_score* (−0,40), dan *entropy* (−0,08) mendorong ke arah aman, dan *homoglyph_count* tidak berkontribusi. Arah kontribusi *is_common_tld* dan *levenshtein_distance* pada contoh ini berlawanan dengan dugaan awal bahwa TLD tidak umum dan jarak edit yang kecil mengarah ke *phishing*, sehingga perlu dikaji lebih lanjut. Perlu ditekankan bahwa SHAP pada Gambar 4 adalah penjelasan lokal untuk satu domain dan merupakan kontribusi statistik terhadap prediksi model, bukan penjelasan yang mutlak.



Gambar 4. Simulasi pada *Streamlit* (kiri) dan penjelasan SHAP untuk domain google-login.xyz (kanan)

Keterbatasan

Domain *phishing* pada penelitian ini adalah kandidat *typosquatting* yang diperoleh dari HaveIBeenSquatted API untuk 13 merek, bukan domain yang terbukti dipakai dalam serangan, dan domain legitim berasal dari satu daftar peringkat popularitas. Hasil karena itu belum tentu berlaku pada distribusi domain di dunia nyata. Perbandingan hanya memakai satu pembagian data 70:30 tanpa uji signifikansi, sehingga selisih kecil antara kedua model sebaiknya tidak ditafsirkan sebagai keunggulan yang pasti. *Recall* kelas *phishing* yang masih sekitar 90% juga berarti masih ada domain *phishing* yang lolos.

4. KESIMPULAN

Penelitian ini membandingkan *Random Forest* dan LightGBM untuk mendeteksi domain *phishing* berbasis *typosquatting* dengan fitur *typosquatting*, leksikal, kata kunci mencurigakan, dan TLD. Kedua algoritma menghasilkan kinerja yang berdekatan. LightGBM *tuned* dipilih sebagai model terbaik karena *recall* kelas *phishing* yang sedikit lebih tinggi, sedangkan *Random Forest tuned* sedikit lebih baik pada *precision* dan akurasi. *Tuning* dengan regularisasi mengurangi kecenderungan *overfitting* pada kedua model.

Model terbaik berhasil diimplementasikan pada simulasi *Streamlit* yang menampilkan prediksi, skor kepercayaan, dan penjelasan kontribusi fitur dengan SHAP, sehingga keputusan model dapat ditelusuri oleh pengguna. Penelitian selanjutnya dapat memperluas dataset dengan domain *phishing* aktual, menambah fitur berbasis WHOIS atau konten halaman, membandingkan algoritma lain seperti XGBoost dan CatBoost, menerapkan

P-ISSN : 2549-7480
E-ISSN : 2549-6301

penyesuaian ambang klasifikasi atau *cost-sensitive learning* untuk menekan *false negative*, dan mengembangkan layanan deteksi *real-time*. Uji signifikansi dan analisis SHAP global juga dapat ditambahkan.

5. REFERENSI

[1] Badan Siber dan Sandi Negara, "Lanskap keamanan siber Indonesia," *Id-SIRTII/CC*, vol. 70, pp. 1–107, 2024.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia, "Profil Internet Indonesia 2026," APJII, Jakarta, 2026.

[3] A. F. Mahmud and S. Wirawan, "Deteksi *phishing* website menggunakan *machine learning* metode klasifikasi," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 4, 2024.

[4] Lukito and W. B. T. Handaya, "Deteksi website *phishing* menggunakan teknik *machine learning*," *Jurnal Informatika Atma Jogja*, vol. 6, no. 1, pp. 69–80, 2025, doi: 10.24002/jiaj.v6i1.11538.

[5] Anti-Phishing Working Group, "Phishing activity trends report: 1st quarter 2025," 2025. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2025.pdf

[6] Y. Y. Santika, R. Rianto, and E. Ujianto, "Studi komprehensif keamanan siber: perbandingan teknologi AI dengan sistem non-AI dalam deteksi dan pencegahan ancaman," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 9, no. 1, pp. 45–64, 2025, doi: 10.31603/komtika.v9i1.13149.

[7] A. M. Almuhaideb *et al.*, "Homoglyph attack detection model using machine learning and hash function," 2022.

[8] S. Agarwal and M. Vasek, "Examining newly registered phishing domains at scale," *Journal of Cybersecurity*, vol. 12, doi: 10.1093/cybsec/tyag020.

[9] R. Fauzan, A. V. Vitianingsih, D. Cahyono, A. L. Maukar, and Y. A. B. Suprio, "Application of classification algorithms in machine learning for phishing detection," *Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 2, pp. 531–540, 2025.

[10] N. B. Putri and A. W. Wijayanto, "Analisis komparasi algoritma klasifikasi data mining dalam klasifikasi website *phishing*," *Komputika: Jurnal Sistem Komputer*, vol. 11, no. 1, pp. 59–66, 2022, doi: 10.34010/komputika.v11i1.4350.

[11] D. B. Suwarno and M. Hardjianto, "Deteksi website *phishing* dari analisis URL menggunakan algoritma *Random Forest*," vol. 21, no. 2, pp. 145–152, 2024.

[12] P. R. E. Gabriel, "Deteksi URL *phishing* menggunakan hybrid deep learning CNN dan XGBoost dengan teknik balancing SMOTE-ENN," vol. 5, no. 2, pp. 70–79, 2025.

[13] Y. Perdana, "Comparative analysis of Random Forest and XGBoost for detecting phishing websites: a machine learning approach," vol. 7, no. 2, pp. 906–921, 2025.

[14] C. Feresa, M. Foozy, M. Amir, I. Anuar, A. Maslan, and H. A. Mohd, "Phishing URLs detection using Naïve Bayes, Random Forest and LightGBM algorithms," vol. 5, no. 1, pp. 56–63, 2024.

[15] P. S. Rizky, R. H. Hirzi, and U. Hidayaturrohman, "Perbandingan metode LightGBM dan XGBoost

DOI : https://doi.org/10.35316/jimi.v11i2.211-229
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence

dalam menangani data dengan kelas tidak seimbang," 2022.

[16] M. C. Calzarossa and R. Zieni, "Explainable machine learning for phishing feature detection," pp. 362–373, 2024, doi: 10.1002/qre.3411.

[17] M. Sipper, "High per parameter: a large-scale study of hyperparameter tuning for machine learning algorithms," pp. 1–13, 2022.

[18] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[19] S. Remya, M. J. Pillai, K. K. Nair, S. R. Subbareddy, and Y. Y. Cho, "An effective detection approach for phishing URL using ResMLP," *IEEE Access*, vol. 12, pp. 79367–79382, 2024, doi: 10.1109/ACCESS.2024.3409049.

[20] E.-S. M. El-kenawy, M. M. Eid, H. L. Hussein, A. M. Osman, and A. M. Elshewey, "Optimized deep learning model using binary particle swarm optimization for phishing attack detection: a comparative study," vol. 5, no. 2, pp. 685–702, 2025.

[21] R. M. Aziz, R. Mahto, K. Goel, A. Das, and P. Kumar, "Modified genetic algorithm with deep learning for fraud transactions of Ethereum smart contract," *Applied Sciences*, 2023.

DOI : https://doi.org/10.35316/jimi.v11i2.211-229
© 2026 Jurnal Ilmiah Informatika with CC BY NC licence