Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

Penerapan Soft Voting Ensemble Berbasis LightGBM, XGBoost, dan CatBoost untuk Deteksi Website Phishing Berdasarkan Fitur URL dan Domain

Application of a Soft Voting Ensemble Based on LightGBM, XGBoost, and CatBoost for Phishing Website Detection Using URL and Domain Features

Rozan Sanura Albary

Fakultas Teknologi Informasi, Program Studi Informatika, Universitas Nusa Mandiri, Jakarta, Indonesia
Jl. Raya Jatiwaringin No.2, Cipinang Melayu, Kec. Makasar, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia
Email: 14240048@nusamandiri.ac.id
Email Penulis Korespondensi: 14240048@nusamandiri.ac.id
Received: **26/08/2026**, Revised: **31/08/2026**, Accepted: **28/10/2026**, Published**: 31/10/2026**

Abstrak−Situs web phishing dirancang menyerupai layanan resmi untuk memperoleh kredensial dan informasi sensitif pengguna. Mekanisme berbasis daftar hitam kerap terlambat mengenali situs yang baru muncul, sementara nilai accuracy yang tinggi belum memadai untuk menggambarkan risiko operasional karena kesalahan mengklasifikasikan situs berbahaya sebagai sah berdampak lebih besar daripada kesalahan sebaliknya. Penelitian ini menerapkan metode Soft Voting Ensemble yang menggabungkan probabilitas dari LightGBM, XGBoost, dan CatBoost, serta membandingkannya dengan model individual, Stacking Ensemble, dan Multilayer Perceptron. Evaluasi dilakukan menggunakan dataset publik yang terdiri atas 11.055 instance dan 30 fitur prediktif yang menggambarkan karakteristik URL, domain, sertifikat, serta indikator konten halaman, dengan phishing ditetapkan sebagai kelas positif. Prosedur imputasi median dan normalisasi MinMaxScaler ditempatkan di dalam pipeline untuk mengurangi risiko kebocoran data, kemudian data dibagi secara stratified dengan rasio 90:10. Kinerja model dinilai pada data uji sekaligus melalui stratified 5-fold cross-validation. Pada data uji, XGBoost memperoleh accuracy tertinggi sebesar 98,01% dengan F1-score phishing 97,73%, sedangkan Soft Voting mencapai accuracy 97,83% dan F1-score phishing 97,52% dengan 472 true positive, 18 false negative, 6 false positive, dan 610 true negative. Pada validasi silang, Soft Voting mencatat rata-rata accuracy tertinggi sebesar 97,25% dan rata-rata F1-score phishing 96,86%. Temuan tersebut menunjukkan bahwa keunggulan suatu model dapat berpindah ketika skema evaluasi berubah. Kontribusi utama penelitian ini terletak pada rancangan evaluasi yang konsisten dan dapat ditelusuri antar model, sehingga selisih kinerja yang kecil tidak serta-merta dimaknai sebagai keunggulan yang berlaku umum.

Kata Kunci: Phishing; Machine Learning; Ensemble Learning; Soft Voting; Validasi Silang

Abstract−Phishing websites are crafted to resemble legitimate services in order to obtain user credentials and other sensitive information. Blacklist-based mechanisms frequently fail to recognise newly emerging sites, while high reported accuracy remains insufficient to characterise operational risk, since misclassifying a malicious site as legitimate carries a greater consequence than the opposite error. This study applies a Soft Voting Ensemble method that combines the class probabilities of LightGBM, XGBoost, and CatBoost, and compares it against individual models, a Stacking Ensemble, and a Multilayer Perceptron. The evaluation employs a public dataset comprising 11,055 instances and 30 predictive features describing URL, domain, certificate, and page content characteristics, with phishing designated as the positive class. Median imputation and MinMax scaling were placed within a pipeline to mitigate data leakage, after which the data were divided using a stratified 90:10 split. Model performance was assessed on the test set as well as through stratified five-fold cross-validation. On the test set, XGBoost attained the highest accuracy of 98.01% with a phishing F1-score of 97.73%, whereas Soft Voting reached 97.83% accuracy and a phishing F1-score of 97.52%, with 472 true positives, 18 false negatives, 6 false positives, and 610 true negatives. Under cross-validation, Soft Voting recorded the highest mean accuracy of 97.25% and mean phishing F1-score of 96.86%. These findings indicate that model superiority may shift as the evaluation scheme changes. The principal contribution lies in a consistent and traceable evaluation design, so that small performance margins are not readily interpreted as generally valid superiority.

Keywords: Phishing; Machine Learning; Ensemble Learning; Soft Voting; Cross-Validation

1. PENDAHULUAN

Phishing masih menjadi salah satu jalur awal yang sering digunakan untuk memperoleh akses ke sistem dan data pengguna. Laporan kuartal pertama 2026 mencatat 971.181 serangan phishing, meningkat 13,8% dibandingkan kuartal sebelumnya [1]. Laporan ancaman siber lainnya juga menunjukkan bahwa phishing menyumbang sekitar 60% vektor intrusi awal selama periode pengamatan [2]. Angka tersebut memperlihatkan bahwa masalah phishing tidak hanya berkaitan dengan banyaknya situs palsu, tetapi juga dengan kecepatan pelaku mengganti domain, memodifikasi struktur URL, dan menyesuaikan tampilan halaman. Akibatnya, mekanisme berbasis daftar hitam sering terlambat mengenali situs baru karena alamat tersebut belum pernah dilaporkan atau dimasukkan ke dalam basis data keamanan.

Pendekatan machine learning digunakan untuk melengkapi mekanisme tersebut dengan mempelajari pola yang terdapat pada URL, sertifikat, domain, pengalihan halaman, dan indikator konten. Tinjauan terhadap metode

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

machine learning untuk deteksi phishing mencatat bahwa perbedaan prosedur evaluasi antar penelitian menyulitkan perbandingan hasil secara langsung [3]. Sejumlah penelitian menunjukkan bahwa pemilihan fitur dan rancangan model berpengaruh langsung terhadap kemampuan klasifikasi [4], [5]. Meskipun nilai accuracy yang dilaporkan umumnya tinggi, angka tersebut belum cukup untuk menggambarkan risiko operasional. Dalam deteksi phishing, kesalahan ketika situs berbahaya diprediksi sebagai legitimate lebih berisiko daripada kesalahan yang memblokir situs legitimate. Oleh karena itu, evaluasi perlu menempatkan phishing sebagai kelas positif dan membaca accuracy bersama precision, recall, F1-score, serta confusion matrix.

Penggunaan ensemble menjadi salah satu arah penelitian karena beberapa pembelajar dapat mengurangi ketergantungan terhadap pola kesalahan satu model. Perbandingan model dasar dan model gabungan menunjukkan bahwa urutan kinerja dapat berubah ketika skema cross-validation atau komposisi data berubah [6]. Penelitian lain membandingkan beberapa algoritma untuk klasifikasi URL [7], sedangkan pendekatan ensemble berbasis pohon juga dilaporkan memberikan kinerja yang kompetitif [8], [9]. Namun, peningkatan yang diperoleh tidak selalu besar. Hasil akhir tetap dipengaruhi oleh karakteristik dataset, fitur yang tersedia, konfigurasi model, dan prosedur validasi yang digunakan.

Variasi metode juga terlihat pada penelitian yang membandingkan teknik machine learning untuk mendeteksi URL phishing [10] serta penelitian yang menilai model ensemble berdasarkan kinerjanya [11]. Pendekatan berbasis fitur dapat mencapai akurasi tinggi ketika indikator yang digunakan sesuai dengan pola pada dataset [12]. Pada konteks phishing email, penggabungan probabilitas melalui Soft Voting juga telah digunakan bersama fitur hibrida [13]. Meskipun objek email tidak identik dengan website, mekanisme tersebut memberi gambaran bahwa keputusan gabungan dapat memanfaatkan kekuatan beberapa model. Perbedaan sumber data tersebut sekaligus menunjukkan bahwa hasil dari satu jenis phishing tidak dapat langsung digeneralisasi ke jenis lainnya.

Di sisi lain, pendekatan deep learning mulai memanfaatkan URL mentah, HTML, dan representasi karakter. Kombinasi karakteristik URL dan HTML digunakan untuk menangkap pola yang tidak selalu terlihat pada fitur tabular [14]. Pendekatan lain memanfaatkan indeks kemiripan dan pembelajaran inkremental [15], beberapa varian convolutional network [16], [17], serta penggabungan pengklasifikasi deep learning dengan pre-trained transformer [17]. Investigasi terhadap pemilihan fitur yang dipadukan dengan model deep learning juga melaporkan accuracy sebesar 98,94% menggunakan stratified k-fold cross-validation[18]. Metode tersebut menawarkan ekstraksi fitur yang lebih otomatis, tetapi biasanya membutuhkan data, waktu pelatihan, dan sumber daya komputasi yang lebih besar. Untuk dataset tabular dengan jumlah fitur yang telah ditentukan, model berbasis pohon masih relevan karena lebih ringan dan relatif mudah dibandingkan dalam skenario eksperimen yang sama.

Penelitian ensemble terbaru menggunakan komposisi pembelajar dan objek phishing yang beragam [19]. Model voting lain menggabungkan Decision Tree, Random Forest, dan k-Nearest Neighbors melalui Hard Voting [20]. Perbedaan tersebut meninggalkan ruang untuk menguji penggabungan probabilitas dari tiga model boosting modern, yaitu LightGBM, XGBoost, dan CatBoost, pada data website phishing tabular. Selain itu, sebagian perbandingan hanya menonjolkan satu hasil pengujian, padahal nilai yang tinggi pada satu pembagian data belum tentu tetap berada pada urutan yang sama ketika data latih dan validasi berganti.

Beberapa penelitian dalam lima tahun terakhir telah menguji penggabungan model boosting pada objek yang serupa. Penelitian dengan tiga base classifier, yaitu CatBoost, XGBoost, dan LightGBM, yang digabungkan melalui teknik voting melaporkan accuracy sebesar 93,0% pada data URL phishing [21]. Penelitian lain mengevaluasi sembilan base classifier, termasuk ketiga model boosting tersebut, pada empat dataset dengan tingkat keseimbangan kelas yang berbeda [22]. Kerangka yang memadukan strategi stacking dan voting atas enam algoritma boosting juga dilaporkan memberikan klasifikasi URL berbahaya yang stabil [23]. Sementara itu, pendekatan super learner menunjukkan bahwa model gabungan lebih unggul dibandingkan model tunggal dari sisi stabilitas dan accuracy lintas dataset [24]. Keempat penelitian tersebut memperlihatkan bahwa kombinasi model boosting memang menjanjikan, tetapi sebagian besar berhenti pada satu skema penggabungan tanpa membandingkan Soft Voting dan Stacking secara langsung pada pipeline, data latih, dan metrik yang sama.

Berdasarkan kondisi tersebut, research gap penelitian ini terletak pada kebutuhan evaluasi yang konsisten antara model individual, Stacking Ensemble, dan Soft Voting dengan definisi kelas positif, pipeline, data latih, data uji, serta metrik yang sama. Penelitian ini menggunakan dua sudut evaluasi. Pengujian pada satu data uji tetap digunakan untuk melihat perilaku model terhadap 1.106 data yang tidak digunakan saat pelatihan, sedangkan stratified 5-fold cross-validation digunakan untuk melihat rata-rata kinerja dan variasi antar lipatan. Pemisahan dua sudut pandang tersebut penting agar selisih kecil antar model tidak langsung ditafsirkan sebagai keunggulan yang berlaku umum.

Tujuan penelitian adalah membandingkan tujuh model dan menilai posisi Soft Voting dalam deteksi website phishing. Kontribusi penelitian bukan berupa algoritma baru, melainkan rancangan evaluasi yang dapat ditelusuri: phishing diperlakukan sebagai kelas positif, preprocessing ditempatkan di dalam pipeline untuk mengurangi risiko kebocoran data, dan kesalahan klasifikasi dianalisis melalui confusion matrix. Hasil pengujian selanjutnya digunakan untuk menjelaskan apakah model terbaik pada data uji juga menghasilkan nilai rata-rata terbaik pada validasi silang, sekaligus menunjukkan batas interpretasi dari perbedaan kinerja yang relatif kecil.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*.
This work is licensed under a Creative Commons Attribution 4.0 International License.

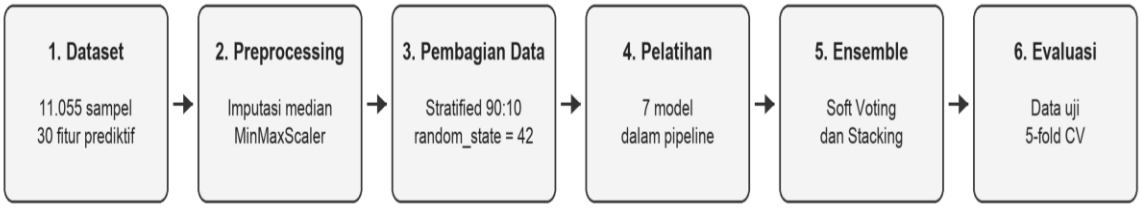
Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian dilakukan melalui enam tahap, yaitu pemeriksaan dataset, penyiapan fitur dan label, pembagian data, pembentukan pipeline, pelatihan model, serta evaluasi. Pemeriksaan awal mencakup jumlah data, jumlah fitur, distribusi kelas, tipe data, dan nilai yang hilang. Tahapan tersebut dirancang agar seluruh model menerima data latih dan data uji yang sama, sehingga perbedaan kinerja lebih mencerminkan perilaku model daripada perbedaan masukan atau perlakuan data.

Setelah label dipetakan, data dibagi secara stratified dengan rasio 90:10 dan random_state 42. Imputer dan scaler ditempatkan di dalam pipeline agar parameter preprocessing hanya dipelajari dari data latih. Model kemudian dievaluasi pada data uji dan melalui stratified 5-fold cross-validation pada data latih. Keseluruhan alur penelitian tersebut ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian dan Evaluasi

2.2 Objek Penelitian

Objek utama penelitian ini adalah website phishing, yaitu situs yang dirancang menyerupai layanan resmi untuk memperoleh kredensial dan informasi sensitif pengguna. Setiap website direpresentasikan sebagai data tabular yang terdiri atas 30 fitur prediktif beserta satu label kelas, sehingga permasalahan deteksi dapat diperlakukan sebagai klasifikasi biner. Terhadap objek tersebut diterapkan sejumlah model machine learning, dengan fokus pada skema ensemble learning berupa Soft Voting yang menggabungkan probabilitas kelas dari LightGBM, XGBoost, dan CatBoost, serta Stacking Ensemble sebagai pembanding. Kinerja seluruh model kemudian dinilai melalui dua sudut evaluasi, yaitu pengujian pada data uji dan stratified 5-fold cross-validation, untuk melihat apakah keunggulan suatu model tetap konsisten ketika skema evaluasi berubah.

2.3 Dataset Penelitian

Dataset yang digunakan merupakan dataset publik untuk deteksi website phishing yang terdiri atas 11.055 instance, 30 fitur prediktif, dan satu label kelas. Dataset yang sama telah digunakan pada penelitian deteksi phishing [20]. Fitur tersebut menggambarkan karakteristik URL, domain, sertifikat, pengalihan halaman, perilaku konten, dan indikator eksternal. Daftar lengkap fitur disajikan pada Tabel 1.

Tabel 1. Daftar 30 Fitur Prediktif

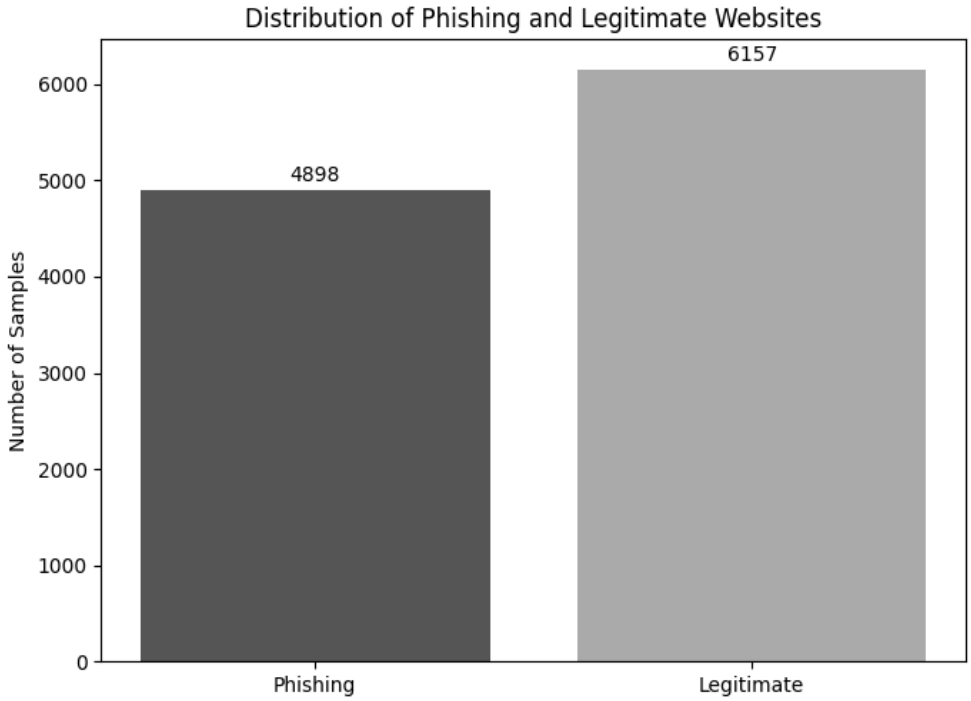
No.	Fitur	No.	Fitur
1	having_IPhaving_IP_Address	16	SFH
2	URLURL_Length	17	Submitting_to_email
3	Shortining_Service	18	Abnormal_URL
4	having_At_Symbol	19	Redirect
5	double_slash_redirecting	20	on_mouseover
6	Prefix_Suffix	21	RightClick
7	having_Sub_Domain	22	popUpWidnow
8	SSLfinal_State	23	Iframe
9	Domain_registeration_length	24	age_of_domain
10	Favicon	25	DNSRecord
11	Port	26	web_traffic
12	HTTPS_token	27	Page_Rank
13	Request_URL	28	Google_Index
14	URL_of_Anchor	29	Links_pointing_to_page
15	Links_in_tags	30	Statistical_report

Seluruh fitur pada Tabel 1 bertipe numerik dengan nilai diskret, sehingga tidak memerlukan proses encoding tambahan. Fitur nomor 1 sampai 15 berkaitan dengan struktur URL dan properti domain, sedangkan nomor 16 sampai 30 mencakup perilaku konten halaman serta indikator eksternal seperti web traffic, Page Rank, dan status pengindeksan Google.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 215
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

Distribusi kelas pada Gambar 2 terdiri atas 4.898 website phishing dan 6.157 website legitimate. Proporsi kelas tidak sepenuhnya seimbang, tetapi setiap kelas masih memiliki jumlah data yang memadai. Pembagian stratified digunakan agar proporsi kedua kelas tetap terjaga pada data latih, data uji, dan setiap lipatan validasi. Penelitian tidak melakukan oversampling maupun undersampling. Kinerja model tetap dibaca menggunakan metrik yang berfokus pada kelas phishing dan Macro-F1 agar evaluasi tidak hanya bergantung pada accuracy.



Gambar 2. Distribusi Kelas Dataset

2.4 Preprocessing Data

Label asli menggunakan nilai -1 untuk phishing dan 1 untuk legitimate. Pada penelitian ini, label dipetakan menjadi 1 untuk phishing dan 0 untuk legitimate sehingga phishing diperlakukan sebagai kelas positif. Dataset tidak memiliki nilai yang hilang. Meskipun demikian, SimpleImputer dengan strategi median tetap ditempatkan di dalam pipeline sebagai pengaman apabila model digunakan pada data lain yang mengandung nilai kosong. MinMaxScaler diterapkan setelah imputasi agar seluruh model, terutama MLP, menerima rentang fitur yang konsisten.

Pembagian data menghasilkan 9.949 instance pelatihan dan 1.106 instance pengujian. Data latih mencakup 4.408 phishing dan 5.541 legitimate, sedangkan data uji terdiri atas 490 phishing dan 616 legitimate. Parameter random_state ditetapkan sebesar 42 agar pembagian dapat direproduksi. Seluruh transformasi di dalam pipeline dipelajari dari data latih, sehingga informasi dari data uji tidak ikut menentukan nilai imputasi maupun skala fitur.

2.5 Implementasi Model

Tujuh model yang dievaluasi adalah Random Forest, XGBoost, LightGBM, CatBoost, MLP, Stacking Ensemble, dan Soft Voting. Random Forest mewakili bagging, sedangkan XGBoost, LightGBM, dan CatBoost mewakili keluarga gradient boosting. MLP digunakan sebagai pembanding jaringan saraf sederhana. Pemilihan model mempertimbangkan kemampuannya menangani klasifikasi data tabular dan praktik evaluasi model yang dijelaskan dalam literatur machine learning dan keamanan [25], [26].

LightGBM, XGBoost, dan CatBoost merupakan varian gradient boosting berbasis pohon yang menambahkan pohon keputusan secara bertahap untuk memperbaiki kesalahan model sebelumnya [25]. Ketiganya banyak digunakan pada data tabular karena efisien dan mampu menangkap interaksi antar fitur, serta telah dilaporkan memberikan kinerja kompetitif sebagai base classifier pada deteksi website phishing [21], [23]. Mekanisme Soft Voting menggabungkan rata-rata probabilitas kelas dari beberapa model, sedangkan Stacking melatih meta-model di atas keluaran base learner; kedua skema tersebut telah diterapkan pada deteksi phishing dengan hasil yang saling bersaing [11] , [13].

Soft Voting menggabungkan probabilitas kelas dari LightGBM, XGBoost, dan CatBoost dengan bobot yang sama. Kelas akhir ditentukan dari rata-rata probabilitas terbesar. Stacking Ensemble menggunakan ketiga model tersebut sebagai base learner dan Random Forest sebagai meta-model dengan lima lipatan internal. Konfigurasi MLP menggunakan solver lbfgs, max_iter 5000, dan max_fun 50000 agar proses optimasi mencapai kondisi konvergen pada eksperimen yang dijalankan. Konfigurasi setiap model disajikan pada Tabel 2.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 216
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

Tabel 2. Konfigurasi Model Machine Learning

Model	Parameter	Nilai
Random Forest	n_estimators; criterion; max_depth; random_state	100; gini; None; 42
XGBoost	learning_rate; n_estimators; max_depth; eval_metric	0,1; 500; 6; logloss
LightGBM	n_estimators; learning_rate; max_depth	200; 0,1; -1
CatBoost	iterations; learning_rate; depth	500; 0,1; 6
MLP	solver; hidden_layer_sizes; activation; max_iter; max_fun	lbfgs; (100,); relu; 5000; 50000
Soft Voting	estimators; voting; weights	LGBM, XGB, CatBoost; soft; sama
Stacking Ensemble	base_estimators; final_estimator; cv	LGBM, XGB, CatBoost; Random Forest; 5

Konfigurasi pada Tabel 2 disusun agar setiap model diuji pada kondisi yang setara. Parameter random_state 42 diterapkan pada seluruh model untuk menjaga reprodusibilitas, sedangkan LightGBM, XGBoost, dan CatBoost menggunakan learning rate yang sama sebesar 0,1 agar perbedaan kinerja tidak berasal dari laju pembelajaran yang berbeda.

2.6 Evaluasi Model

Evaluasi menggunakan accuracy, precision phishing, recall phishing, F1-score phishing, Macro-F1, Weighted-F1, dan confusion matrix. Precision phishing menunjukkan proporsi prediksi phishing yang benar, sedangkan recall phishing menunjukkan proporsi seluruh website phishing yang berhasil dideteksi. F1-score merangkum keseimbangan precision dan recall. Macro-F1 memberi bobot yang sama kepada kedua kelas, sedangkan Weighted-F1 mempertimbangkan jumlah data pada setiap kelas. Dalam konteks keamanan, false negative perlu mendapat perhatian karena menunjukkan website phishing yang diprediksi sebagai legitimate.

Pengujian pertama dilakukan pada data uji yang dipisahkan dengan rasio 10% dari keseluruhan data. Pengujian kedua menggunakan stratified 5-fold cross-validation pada data latih. Pada prosedur ini, data dibagi menjadi lima lipatan dengan proporsi kelas yang dipertahankan; empat lipatan digunakan untuk pelatihan dan satu lipatan untuk validasi secara bergantian. Hasil validasi silang dilaporkan sebagai rata-rata dan simpangan baku dari lima nilai, sehingga tabel menunjukkan tingkat kinerja sekaligus variasi antar lipatan.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian Model

Ketujuh model dilatih menggunakan data latih yang sama dan dievaluasi pada 1.106 data uji. Hasil pada Tabel 3 menunjukkan bahwa seluruh model mencapai accuracy antara 0,9756 dan 0,9801. Rentang yang sempit tersebut menandakan bahwa sebagian besar model mampu mempelajari pola pada 30 fitur prediktif. Meskipun demikian, urutan kinerja tetap perlu dibaca bersama precision, recall, dan F1-score phishing karena selisih beberapa prediksi dapat mengubah risiko kesalahan pada kelas positif.

Tabel 3. Hasil Pengujian pada Data Uji

Model	Accuracy	Precision	Recall	F1 Phishing	Macro-F1	Weighted-F1
LightGBM	0,9765	0,9854	0,9612	0,9731	0,9761	0,9765
Stacking Ensemble	0,9792	0,9875	0,9653	0,9763	0,9789	0,9792
Random Forest	0,9756	0,9793	0,9653	0,9723	0,9752	0,9756
XGBoost	0,9801	0,9875	0,9673	0,9773	0,9798	0,9801
MLP	0,9774	0,9854	0,9633	0,9742	0,977	0,9774
CatBoost	0,9774	0,9834	0,9653	0,9743	0,9771	0,9774
Soft Voting	0,9783	0,9874	0,9633	0,9752	0,978	0,9783

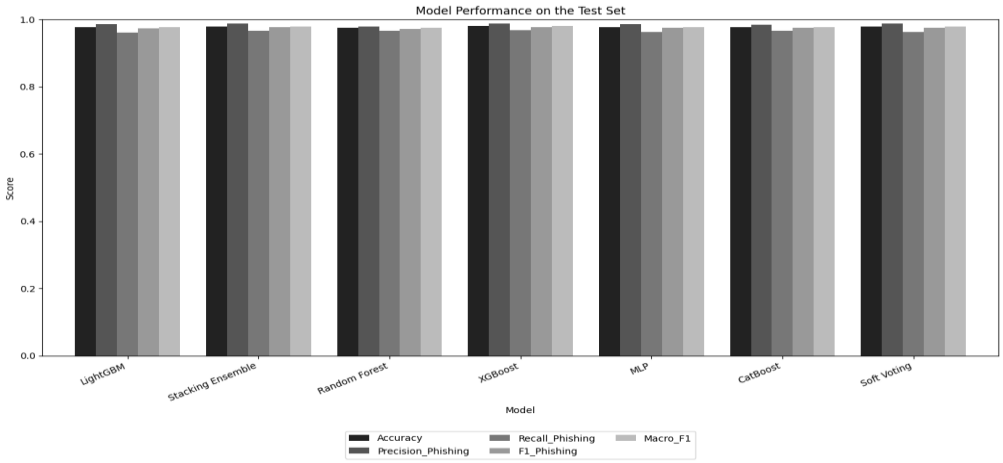
XGBoost memperoleh hasil tertinggi pada data uji dengan accuracy 0,9801, precision phishing 0,9875, recall phishing 0,9673, dan F1-score phishing 0,9773. Stacking Ensemble berada pada urutan berikutnya dengan accuracy 0,9792 dan F1-score phishing 0,9763. Soft Voting menghasilkan accuracy 0,9783 dan F1-score phishing 0,9752. Dengan demikian, Soft Voting bukan model terbaik pada pembagian data uji tetap. Namun, selisih accuracy antara XGBoost dan Soft Voting hanya 0,0018 atau 0,18 poin persentase, sehingga perbedaannya perlu ditafsirkan secara proporsional.

Nilai precision tertinggi sebesar 0,9875 diperoleh XGBoost dan Stacking Ensemble, sedangkan Soft Voting berada sangat dekat pada 0,9874. Perbedaan lebih terlihat pada recall XGBoost mencapai recall 0,9673, sementara Soft Voting mencapai 0,9633. Artinya, pada data uji yang sama, XGBoost mengenali dua website phishing lebih banyak daripada Soft Voting. Random Forest dan CatBoost juga memperoleh recall 0,9653, tetapi nilai precision dan F1-score keduanya berada sedikit di bawah XGBoost. Pola ini memperlihatkan bahwa accuracy yang hampir

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 217
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

sama dapat terbentuk dari susunan kesalahan klasifikasi yang berbeda. Perbandingan kinerja ketujuh model pada data uji divisualisasikan dalam bentuk diagram batang pada Gambar 3.



Gambar 3. Perbandingan Kinerja Model pada Data Uji

Gambar 3 memperlihatkan bahwa batang kinerja antar model berada pada rentang yang berdekatan. Visual tersebut membantu menunjukkan bahwa tidak ada satu model yang terpisah jauh dari model lainnya. Selisih yang kecil belum cukup untuk menyatakan keunggulan secara umum jika kesimpulan hanya didasarkan pada satu pembagian data. Komposisi data uji dapat lebih sesuai dengan pola yang dipelajari model tertentu, sehingga urutan model perlu dibandingkan dengan hasil validasi silang.

Setiap model selanjutnya dievaluasi menggunakan stratified 5-fold cross-validation pada data latih. Nilai pada Tabel 4 disajikan sebagai rata-rata dan simpangan baku dari lima lipatan. Rata-rata menggambarkan tingkat kinerja model secara keseluruhan, sedangkan simpangan baku menunjukkan seberapa besar nilai berubah ketika komposisi data pelatihan dan validasi berganti. Penyajian kedua angka tersebut mencegah nilai rata-rata dibaca tanpa mempertimbangkan kestabilannya.

Tabel 4. Hasil Stratified 5-Fold Cross-Validation

Model	Accuracy	Precision	Recall	F1 Phishing	Macro-F1
LightGBM	0,9702 ± 0,0064	0,9724 ± 0,0062	0,9601 ± 0,0100	0,9662 ± 0,0073	0,9698 ± 0,0065
Stacking Ensemble	0,9707 ± 0,0063	0,9740 ± 0,0050	0,9594 ± 0,0098	0,9666 ± 0,0073	0,9702 ± 0,0064
Random Forest	0,9705 ± 0,0046	0,9762 ± 0,0031	0,9569 ± 0,0102	0,9664 ± 0,0054	0,9701 ± 0,0047
XGBoost	0,9720 ± 0,0055	0,9767 ± 0,0072	0,9596 ± 0,0081	0,9681 ± 0,0063	0,9715 ± 0,0056
MLP	0,9626 ± 0,0029	0,9658 ± 0,0027	0,9492 ± 0,0066	0,9574 ± 0,0035	0,9620 ± 0,0030
CatBoost	0,9712 ± 0,0045	0,9775 ± 0,0052	0,9569 ± 0,0069	0,9671 ± 0,0051	0,9707 ± 0,0045
Soft Voting	0,9725 ± 0,0063	0,9778 ± 0,0066	0,9596 ± 0,0102	0,9686 ± 0,0073	0,9720 ± 0,0065

Catatan: nilai disajikan sebagai rata-rata ± simpangan baku dari lima lipatan.

Nilai pada Tabel 4 memperlihatkan bahwa Soft Voting memperoleh rata-rata accuracy tertinggi sebesar 0,9725 ± 0,0063, diikuti XGBoost sebesar 0,9720 ± 0,0055 dan CatBoost sebesar 0,9712 ± 0,0045. Rata-rata F1 phishing Soft Voting juga tertinggi, yaitu 0,9686 ± 0,0073, sedangkan XGBoost memperoleh 0,9681 ± 0,0063. Selisih keduanya hanya 0,0005 pada accuracy dan 0,0005 pada F1 phishing. Nilai tersebut menunjukkan perubahan urutan dibandingkan pengujian data uji, tetapi belum membuktikan dominasi Soft Voting.

Simpangan baku memberikan konteks tambahan. Variasi accuracy Soft Voting sebesar 0,0063 lebih besar daripada XGBoost sebesar 0,0055 dan CatBoost sebesar 0,0045. Dengan demikian, rata-rata Soft Voting memang sedikit lebih tinggi, tetapi variasi antar lipatannya bukan yang paling kecil. Karena penelitian ini belum menerapkan uji signifikansi berpasangan pada skor setiap lipatan, selisih rata-rata yang sangat kecil tidak ditafsirkan sebagai perbedaan yang signifikan secara statistik.

Keunggulan rata-rata Soft Voting juga tidak muncul secara seragam pada seluruh metrik. Precision phishing Soft Voting sebesar 0,9778 ± 0,0066 hanya sedikit di atas CatBoost sebesar 0,9775 ± 0,0052 dan XGBoost sebesar 0,9767 ± 0,0072. Pada recall phishing, Soft Voting dan XGBoost memperoleh rata-rata yang sama, yaitu 0,9596, meskipun simpangan bakunya berbeda. Sementara itu, Macro-F1 Soft Voting sebesar 0,9720 berada dekat dengan

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 218
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

XGBoost sebesar 0,9715. Pola tersebut menunjukkan bahwa pemilihan model sebaiknya mengikuti metrik yang paling sesuai dengan kebutuhan sistem, bukan hanya peringkat accuracy.

3.2 Analisis Model Soft Voting

Soft Voting membentuk keputusan dari rata-rata probabilitas yang dihasilkan LightGBM, XGBoost, dan CatBoost. Ketiga model tersebut sama-sama berasal dari keluarga boosting, tetapi menggunakan strategi pembentukan pohon dan pengendalian bias yang berbeda. Penggabungan probabilitas dapat mengurangi pengaruh prediksi yang terlalu yakin dari satu model apabila dua model lainnya memberikan probabilitas yang lebih moderat. Mekanisme inilah yang dapat membantu Soft Voting mempertahankan kinerja rata-rata ketika komposisi data berubah pada setiap lipatan.

Pada pembagian data uji tetap, karakteristik data lebih sesuai dengan batas keputusan XGBoost sehingga model tersebut menghasilkan recall dan F1-score phishing yang lebih tinggi. Pada validasi silang, setiap model menerima kombinasi data pelatihan dan validasi yang berbeda. Soft Voting tidak bergantung pada satu pola prediksi karena keputusan akhir dibentuk dari tiga keluaran probabilitas. Hasil yang diperoleh menunjukkan manfaat kecil pada rata-rata kinerja, bukan peningkatan besar. Oleh sebab itu, Soft Voting lebih tepat disebut kompetitif daripada selalu lebih unggul.

Perbandingan dengan Stacking Ensemble memperjelas pengaruh rancangan penggabungan. Stacking mencapai accuracy 0,9792 pada data uji, tetapi rata-rata accuracy validasi silangnya turun menjadi 0,9707. Model ini melibatkan pembelajaran dua tingkat karena keluaran base learner diproses kembali oleh Random Forest sebagai meta-model. Tanpa penyetelan khusus, proses tersebut dapat menghasilkan variasi tambahan. MLP menunjukkan penurunan yang lebih besar, dari accuracy 0,9774 pada data uji menjadi rata-rata 0,9626 pada validasi silang. Kondisi ini menunjukkan bahwa satu hasil pengujian tidak selalu mewakili perilaku model pada beberapa pembagian data.

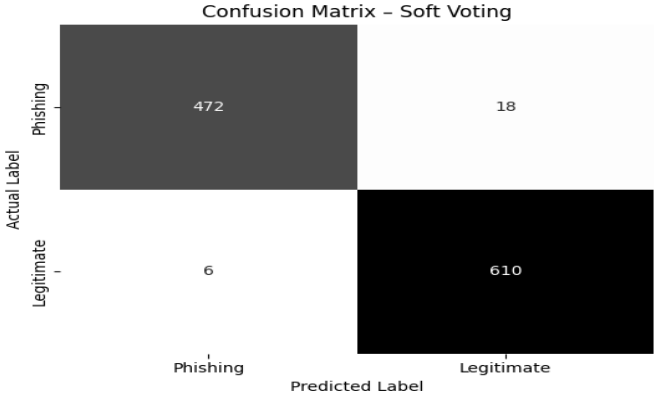
Dari sisi implementasi, Soft Voting memiliki konsekuensi komputasi karena tiga model harus tersedia saat inference. XGBoost lebih sederhana untuk digunakan karena hanya memerlukan satu model dan pada penelitian ini memberikan hasil terbaik pada data uji. Soft Voting dapat dipertimbangkan ketika kinerja rata-rata pada beberapa pembagian data lebih diprioritaskan. Pemilihan keduanya sebaiknya mempertimbangkan waktu prediksi, penggunaan memori, kemudahan pemeliharaan, serta toleransi sistem terhadap false negative. Penelitian ini belum mengukur waktu inference, sehingga perbandingan efisiensi tersebut belum dapat dibuktikan secara empiris.

3.3 Analisis Confusion Matrix

Confusion matrix Soft Voting dianalisis dengan phishing sebagai kelas positif. Urutan kelas pada Gambar 4 adalah phishing dan legitimate untuk label aktual maupun prediksi. Penetapan urutan tersebut penting karena perubahan urutan label dapat menyebabkan true positive dan true negative tertukar saat dibaca. Dengan definisi ini, false negative menunjukkan website phishing yang lolos sebagai legitimate, sedangkan false positive menunjukkan website legitimate yang diprediksi sebagai phishing.

Soft Voting menghasilkan 472 true positive, 18 false negative, 6 false positive, dan 610 true negative. Dari 1.106 data uji, sebanyak 1.082 instance diklasifikasikan dengan benar dan 24 instance salah diklasifikasikan. Susunan tersebut menghasilkan accuracy 0,9783, precision phishing 0,9874, recall phishing 0,9633, dan F1-score phishing 0,9752. Kesesuaian antara jumlah pada confusion matrix dan metrik pada Tabel 3 juga menjadi pemeriksaan bahwa kelas positif telah ditetapkan secara konsisten.

Sebanyak 18 dari 490 website phishing diprediksi sebagai legitimate, sehingga false negative rate mencapai 3,67%. Kesalahan ini memiliki dampak keamanan yang lebih tinggi karena pengguna dapat mengakses situs berbahaya tanpa menerima peringatan. Sebaliknya, 6 dari 616 website legitimate diprediksi sebagai phishing, sehingga false positive rate sebesar 0,97%. Kesalahan tersebut dapat mengganggu akses pengguna, tetapi umumnya lebih mudah ditangani melalui proses verifikasi atau pengecualian dibandingkan phishing yang tidak terdeteksi.



Gambar 4. Confusion Matrix Soft Voting

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 219
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

Confusion matrix tersebut merepresentasikan satu titik operasi yang dibentuk oleh ambang keputusan bawaan model dan bobot Soft Voting yang sama. Ambang yang lebih rendah untuk kelas phishing berpotensi meningkatkan recall dan mengurangi false negative, tetapi dapat menambah false positive. Sebaliknya, ambang yang lebih tinggi dapat mempertahankan precision dengan risiko lebih banyak situs phishing terlewat. Penentuan ambang dan bobot seharusnya dilakukan hanya menggunakan data pelatihan atau validasi agar data uji tetap independen. Karena tahap tersebut belum dilakukan, hasil penelitian ini dibaca sebagai kinerja konfigurasi awal, bukan sebagai batas terbaik yang dapat dicapai oleh Soft Voting.

Sebagai pembanding, XGBoost mendeteksi 474 website phishing dengan 16 false negative dan 6 false positive. Perbedaan dua true positive menjelaskan mengapa recall dan F1-score XGBoost lebih tinggi pada data uji. Dalam penerapan nyata, perbedaan kecil tersebut tetap relevan apabila volume URL yang diperiksa sangat besar. Namun, keputusan memilih model tidak cukup hanya berdasarkan dua prediksi karena kestabilan, biaya komputasi, kalibrasi probabilitas, serta perubahan pola phishing juga perlu dipertimbangkan.

3.4 Pembahasan

Hasil penelitian mendukung temuan bahwa model ensemble dapat memberikan kinerja yang kompetitif, tetapi besarnya peningkatan bergantung pada dataset, komposisi pembelajar, dan prosedur evaluasi [5], [7], [10], [17]. Model voting dengan komposisi pembelajar yang berbeda juga menunjukkan bahwa penggabungan keputusan dapat menjadi pilihan yang efektif [18]. Pada penelitian ini, keuntungan Soft Voting terlihat pada rata-rata lima lipatan, sedangkan XGBoost tetap unggul pada data uji tetap. Perbedaan tersebut memperlihatkan pentingnya menyatakan konteks evaluasi ketika melaporkan model terbaik.

Perbandingan langsung dengan nilai accuracy penelitian lain perlu dilakukan secara hati-hati. Studi terdahulu menggunakan jumlah fitur, sumber data, bentuk URL, periode pengambilan data, dan prosedur validasi yang berbeda [9], [11], [13]-[16]. Pendekatan berbasis URL mentah dan HTML dapat mempelajari representasi yang lebih kompleks, sedangkan penelitian ini menggunakan fitur tabular yang telah tersedia. Oleh sebab itu, nilai yang lebih tinggi pada studi lain tidak otomatis menunjukkan bahwa model tersebut lebih sesuai untuk dataset penelitian ini. Perbandingan yang paling dapat dipertanggungjawabkan adalah perbandingan internal karena seluruh model diuji menggunakan data dan metrik yang sama.

Kontribusi penelitian terletak pada konsistensi rancangan eksperimen. Phishing ditetapkan sebagai kelas positif sejak pemetaan label, sehingga precision, recall, F1-score, dan confusion matrix memiliki arah interpretasi yang sama. Imputasi dan normalisasi ditempatkan di dalam pipeline agar parameter transformasi tidak dipelajari dari data uji. Pembagian stratified menjaga proporsi kelas, sedangkan penyajian data uji dan validasi silang memberikan dua sudut pandang yang saling melengkapi. Rancangan ini membuat perbedaan antarmodel lebih mudah ditelusuri dan mengurangi risiko kesimpulan yang hanya bergantung pada satu angka accuracy.

Komposisi dataset terdiri atas sekitar 44,31% phishing dan 55,69% legitimate. Ketimpangan tersebut tidak ekstrem, tetapi tetap dapat membuat accuracy terlihat baik apabila model lebih sering mengikuti kelas yang jumlahnya lebih besar. Karena itu, Macro-F1 digunakan untuk memberi bobot yang sama kepada kedua kelas, sedangkan Weighted-F1 menunjukkan kinerja dengan mempertimbangkan proporsi data. Kedekatan nilai accuracy, Macro-F1, dan Weighted-F1 pada Tabel 3 menunjukkan bahwa kinerja tinggi tidak hanya berasal dari kelas legitimate. Hasil recall phishing dan confusion matrix tetap diperlukan untuk memastikan bahwa situs berbahaya benar-benar terdeteksi. Penggunaan beberapa metrik juga mengurangi risiko pemilihan model hanya karena selisih accuracy yang sangat kecil. Dengan cara ini, konsekuensi kesalahan pada kelas phishing dapat dinilai lebih langsung dan tidak tertutup oleh jumlah prediksi benar pada kelas legitimate.

Temuan penelitian juga memiliki implikasi praktis. Jika sistem mengutamakan model tunggal dengan hasil data uji terbaik, XGBoost menjadi pilihan yang lebih langsung. Jika pengembang lebih mempertimbangkan rata-rata kinerja pada beberapa pembagian data, Soft Voting layak dipertimbangkan. Meskipun demikian, selisih keduanya sangat kecil dan belum diuji signifikansinya. Penggunaan Soft Voting juga berarti tiga model harus dilatih, disimpan, dan dijalankan. Oleh karena itu, peningkatan rata-rata sebesar 0,05 poin persentase perlu dibandingkan dengan tambahan biaya komputasi sebelum model diterapkan.

Penelitian masih memiliki beberapa keterbatasan. Eksperimen hanya menggunakan satu dataset publik dengan fitur yang telah ditentukan sebelumnya, sehingga kemampuan generalisasi terhadap pola phishing baru belum dapat dipastikan. Penyetelan hyperparameter belum dilakukan secara menyeluruh, probabilitas ketiga model belum dikalibrasi secara khusus, dan perbedaan antarmodel belum diuji secara statistik. Penelitian juga belum mencakup feature importance, waktu inference, pengujian lintas dataset, serta validasi temporal. Aspek tersebut perlu ditambahkan agar ketahanan model terhadap perubahan pola serangan dan kelayakannya dalam sistem dunia nyata dapat dinilai lebih lengkap.

4. KESIMPULAN

Penelitian ini membandingkan Random Forest, XGBoost, LightGBM, CatBoost, MLP, Stacking Ensemble, dan Soft Voting untuk mendeteksi website phishing pada dataset yang terdiri atas 11.055 instance dan 30 fitur prediktif. Seluruh model diuji dengan phishing sebagai kelas positif, preprocessing di dalam pipeline, pembagian data

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 220
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

stratified 90:10, dan metrik yang sama. Pada data uji, XGBoost memberikan hasil tertinggi dengan accuracy 98,01% dan F1-score phishing 97,73%. Soft Voting memperoleh accuracy 97,83% dan F1-score phishing 97,52%, dengan 472 true positive, 18 false negative, 6 false positive, dan 610 true negative. Perbedaan kedua model pada data uji terutama berasal dari dua website phishing tambahan yang berhasil dikenali XGBoost. Pada stratified 5-fold cross-validation, urutan kinerja berubah. Soft Voting menghasilkan rata-rata accuracy tertinggi sebesar 97,25% dan rata-rata F1-score phishing 96,86%, sedikit di atas XGBoost yang masing-masing memperoleh 97,20% dan 96,81%. Hasil tersebut menjawab tujuan penelitian bahwa penggabungan probabilitas LightGBM, XGBoost, dan CatBoost mampu memberikan kinerja rata-rata yang kompetitif, tetapi tidak selalu menjadi pilihan terbaik pada satu pembagian data. Selisih rata-rata antar model hanya 0,05 poin persentase dan belum diuji melalui uji signifikansi berpasangan. Karena itu, temuan ini tidak digunakan untuk menyatakan keunggulan mutlak Soft Voting. Untuk penerapan praktis, XGBoost lebih sederhana dan memberikan hasil data uji terbaik, sedangkan Soft Voting dapat dipertimbangkan ketika rata-rata kinerja pada beberapa pembagian data lebih diprioritaskan. Kesimpulan tersebut masih terbatas pada satu dataset dan fitur tabular yang tersedia. Penelitian berikutnya perlu melakukan penyetelan hyperparameter, kalibrasi probabilitas, uji statistik, analisis feature importance, serta pengukuran waktu inference dan penggunaan memori. Pengujian lintas dataset dan validasi temporal juga diperlukan untuk mengetahui apakah peringkat model tetap konsisten ketika pola URL dan teknik phishing berubah. Dengan pengujian tersebut, pemilihan model dapat didasarkan tidak hanya pada accuracy, tetapi juga pada kestabilan, risiko false negative, efisiensi, dan kemampuan generalisasi.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi, metodologi, perangkat lunak, validasi, analisis formal, investigasi, sumber daya, kurasi data, penulisan draf awal, peninjauan dan penyuntingan, serta visualisasi dilakukan oleh Rozan Sanura Albary. Penulis telah membaca dan menyetujui versi akhir manuskrip.

5.2 Ketersediaan Data

Dataset yang digunakan tersedia secara publik. Notebook eksperimen dan hasil pengolahan dapat diberikan oleh penulis korespondensi berdasarkan permintaan yang wajar.

5.3 Pendanaan

Penelitian ini tidak menerima pendanaan eksternal.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak berlaku karena penelitian menggunakan dataset publik dan tidak melibatkan manusia maupun hewan sebagai subjek penelitian.

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak berlaku karena penelitian tidak melibatkan partisipan manusia.

5.6 Pernyataan Konflik Kepentingan

Penulis menyatakan tidak memiliki konflik kepentingan yang dapat memengaruhi penelitian ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Teknologi AI generatif digunakan untuk membantu penyuntingan bahasa dan penyesuaian format, dan penelusuran kendala teknis pada source code. Seluruh hasil eksperimen, isi ilmiah, dan keputusan akhir manuskrip ditinjau kembali serta menjadi tanggung jawab penuh penulis.

REFERENCES

[1] Anti-Phishing Working Group, “Phishing Activity Trends Report: 1st Quarter 2026,” Cambridge, MA, USA, May 2026. Accessed: Aug. 26, 2026. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2026.pdf
[2] ENISA Threat Landscape 2025, “European Union Agency for Cybersecurity,” Athens, Greece, Oct. 2025. doi: 10.2824/2445233.
[3] E. Kytidou, T. Tsikriki, G. Drosatos, and K. Rantos, “Machine learning techniques for phishing detection: A review of methods, challenges, and future directions,” Intelligent Decision Technologies, vol. 19, no. 6, pp. 4356–4379, Nov. 2025, doi: 10.1177/18724981251366763.
[4] G. S. Nayak, B. Muniyal, and M. C. Belavagi, “Enhancing Phishing Detection: A Machine Learning Approach With Feature Selection and Deep Learning Models,” IEEE Access, vol. 13, pp. 33308–33320, 2025, doi: 10.1109/ACCESS.2025.3543738.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 221
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Information System Research (JOSH)
ISSN 2686-228X (media online)
Volume 8, No. 1, Oktober 2026, pp 213−222
DOI 10.47065/josh.v8i1.11080
https://ejurnal.seminar-id.com/index.php/josh/

[5] N. Yadav and S. P. Panda, "Feature Selection for Email Phishing Detection Using Machine Learning," 2022, pp. 365–378. doi: 10.1007/978-981-16-2597-8_31.
[6] A. Awasthi and N. Goel, "Phishing website prediction using base and ensemble classifier techniques with cross-validation," Cybersecurity, vol. 5, no. 1, p. 22, Nov. 2022, doi: 10.1186/s42400-022-00126-9.
[7] T. M. Bao, K. Shashvat, N. G. Nhu, and D.-N. Le, "A Comparative Analysis of Machine Learning Algorithms for Spam and Phishing URL Classification," Computers, Materials & Continua, vol. 87, no. 2, pp. 1–10, 2026, doi: 10.32604/cmc.2025.075161.
[8] N. Innab et al., "Phishing Attacks Detection Using Ensemble Machine Learning Algorithms," Computers, Materials & Continua, vol. 80, no. 1, pp. 1325–1345, 2024, doi: 10.32604/cmc.2024.051778.
[9] H. Fadhilah, D. R. Maulana, and R. Utari, "Tree-based Ensemble Machine Learning for Phishing Website Detection," Komputika : Jurnal Sistem Komputer, vol. 13, no. 2, pp. 233–243, Oct. 2024, doi: 10.34010/komputika.v13i2.12495.
[10] A. A. Albishri and M. M. Dessouky, "A Comparative Analysis of Machine Learning Techniques for URL Phishing Detection," Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 18495–18501, Dec. 2024, doi: 10.48084/etasr.8920.
[11] A. Jain et al., "PhishNet 1.0: optuna-optimized stacking ensemble with Boruta-based feature selection for phishing URL detection," Sci. Rep., vol. 16, no. 1, p. 1480, Dec. 2025, doi: 10.1038/s41598-025-31447-7.
[12] M. Bahaghighat, M. Ghasemi, and F. Ozen, "A high-accuracy phishing website detection method based on machine learning," Journal of Information Security and Applications, vol. 77, p. 103553, Sep. 2023, doi: 10.1016/j.jisa.2023.103553.
[13] P. Bountakas and C. Xenakis, "HELPHED: Hybrid Ensemble Learning PHishing Email Detection," Journal of Network and Computer Applications, vol. 210, p. 103545, Jan. 2023, doi: 10.1016/j.jnca.2022.103545.
[14] C. Opara, Y. Chen, and B. Wei, "Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics," Expert Syst. Appl., vol. 236, p. 121183, Feb. 2024, doi: 10.1016/j.eswa.2023.121183.
[15] A. Prasad and S. Chandra, "PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning," Comput. Secur., vol. 136, p. 103545, Jan. 2024, doi: 10.1016/j.cose.2023.103545.
[16] M. Hussain, C. Cheng, R. Xu, and M. Afzal, "CNN-Fusion: An effective and lightweight phishing detection method based on multi-variant ConvNet," Inf. Sci. (N. Y)., vol. 631, pp. 328–345, Jun. 2023, doi: 10.1016/j.ins.2023.02.039.
[17] N. Q. Do, A. Selamat, H. Fujita, and O. Krejcar, "An integrated model based on deep learning classifiers and pre-trained transformer for phishing URL detection," Future Generation Computer Systems, vol. 161, pp. 269–285, Dec. 2024, doi: 10.1016/j.future.2024.06.031.
[18] S. Mousavi and M. Bahaghighat, "Phishing Website Detection: An In-Depth Investigation of Feature Selection and Deep Learning," Expert Syst., vol. 42, no. 3, Mar. 2025, doi: 10.1111/exsy.13824.
[19] Z. Salah, H. Abu Owida, E. Abu Elsoud, E. Alhenawi, S. Abuowaida, and N. Alshdaifat, "An Effective Ensemble Approach for Preventing and Detecting Phishing Attacks in Textual Form," Future Internet, vol. 16, no. 11, p. 414, Nov. 2024, doi: 10.3390/fi16110414.
[20] M. N. Yeasmin, M. A. R. Refat, B. C. Singh, Z. Alom, Z. Aung, and M. Azim, "EnLeM: ensemble learning-based model to detect phishing websites," EURASIP J. Inf. Secur., vol. 2026, no. 1, p. 1, Dec. 2025, doi: 10.1186/s13635-025-00219-1.
[21] M. Agagu, I. A. Ogunbiyi, A. Lasisi, and O. Omorogiuwa, "Detection of Phishing Websites from URLs Using Hybrid Ensemble-Based Machine Learning Technique," 2024, pp. 11–22. doi: 10.1007/978-3-031-66965-1_2.
[22] Z. Alamri, A. Alhuzali, B. Alsulami, and D. Alghazzawi, "Enhanced Phishing Website Detection Using Optimized Ensemble Stacking Models," International Journal of Advanced Computer Science and Applications, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.01608100.
[23] D. R. Patil, T. M. Pattewar, T. S. Shinde, K. S. Kumavat, and S. N. Deshpande, "Stacking and Voting-Based Boosting Ensembles for Robust Malicious URL Classification," Informatica, vol. 49, no. 35, Dec. 2025, doi: 10.31449/inf.v49i35.7762.
[24] R. S. Rao, C. Kondaiah, A. R. Pais, and B. Lee, "A hybrid super learner ensemble for phishing detection on mobile devices," Sci. Rep., vol. 15, no. 1, p. 16839, May 2025, doi: 10.1038/s41598-025-02009-8.
[25] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
[26] C. Chio and D. Freeman, Machine Learning and Security: Protecting Systems with Data and Algorithms. Sebastopol, CA, USA: O'Reilly Media, 2018.

Copyright © 2026 The Author(s). Published by *Forum Kerjasama Pendidikan Tinggi*. Page 222
This work is licensed under a Creative Commons Attribution 4.0 International License.