**Journal of Applied Informatics and Computing (JAIC)**
Vol.10, No.1, February 2026, pp. 664~672
e-ISSN: 2548-6861


# Analysis of Gradient Boosting Algorithms with Optuna Optimization and SHAP Interpretation for Phishing Website Detection

**Rahmat Fauzi Abu Bakar [1]*, Majid Rahardi [2]***
[1, 2] Informatics, Universitas Amikom Yogyakarta
rahmatfauzi@students.amikom.ac.id [1], majid@amikom.ac.id [2]

**Article Info**

***Article history:***

Received 2025-11-25
Revised 2025-12-23
Accepted 2026-01-07

***Keyword:***

*Gradient Boosting,*
*Machine Learning,*
*Optuna,*
*Phishing Detection,*
*SHAP.*

**ABSTRACT**

Phishing remains a persistent cybersecurity threat, evolving rapidly to bypass traditional blacklist-based detection systems. Machine Learning (ML) approaches offer a promising solution, yet finding the optimal balance between detection accuracy and model interpretability remains a challenge. This study aims to evaluate and optimize the performance of three state-of-the-art Gradient Boosting algorithms—XGBoost, LightGBM, and CatBoost—for phishing website detection. The research utilizes the UCI Phishing Websites dataset consisting of 11,055 instances. The novelty of this study lies in the implementation of the Optuna framework with the Tree-structured Parzen Estimator (TPE) for automated hyperparameter optimization and the application of SHAP (Shapley Additive Explanations) interaction values to interpret the "black-box" models. The experimental results demonstrate that the LightGBM model, optimized via Optuna, achieved the highest performance with an F1-Score of 0.9798, outperforming the baseline model (0.9713) by 0.87%. Furthermore, SHAP analysis identified 'SSLfinal_State' as the most critical determinant for distinguishing phishing sites. This study confirms that optimizing modern boosting algorithms significantly enhances phishing detection capabilities while providing necessary explainability for cybersecurity analysts.



## I. PENDAHULUAN

Serangan *phishing* telah berevolusi menjadi salah satu vektor ancaman keamanan siber yang paling persisten dan merugikan dalam ekosistem digital global. Mekanisme serangan ini mengeksploitasi kerentanan psikologis pengguna melalui teknik rekayasa sosial (*social engineering*) untuk memanipulasi korban agar mengungkapkan informasi sensitif, seperti kredensial perbankan dan data pribadi, pada situs web tiruan yang dirancang menyerupai entitas yang sah [1]. Urgensi masalah ini dikuatkan oleh Laporan terbaru dari *Anti-Phishing Working Group* (APWG) mencatat peningkatan aktivitas *phishing* yang signifikan pada kuartal pertama tahun 2023 [2]. Metode pertahanan konvensional dinilai semakin tidak efektif menghadapi ancaman ini. Sebagai solusi, pendekatan *Machine Learning* telah diadopsi secara luas. Studi oleh Al-garadi et al. [13] dan Alnemari et al. [20] menunjukkan efektivitas metode deteksi berbasis AI. Namun, tantangan utama tetap pada adaptabilitas model terhadap serangan baru. Akhtar et al. [14] dalam penelitian terbarunya tahun 2025 menekankan pentingnya *Explainable AI* untuk memitigasi ancaman siber yang kompleks.

Algoritma *Gradient Boosting* seperti XGBoost [6], LightGBM [7], dan CatBoost [5] telah menjadi standar industri. [15] menerapkan klasifikasi *gradient boosting* untuk deteksi *phishing*, namun penelitian tersebut belum mengintegrasikan optimasi otomatis. Sementara itu, [11] berfokus pada optimasi hiperparameter, tetapi menggunakan metode yang komputasinya mahal. Oleh karena itu, penelitian ini mengusulkan penggunaan kerangka kerja Optuna [3] untuk optimasi yang lebih efisien, serta analisis SHAP [4][14] untuk transparansi model, mengisi celah yang ditinggalkan oleh penelitian sebelumnya.

Sebagai respons terhadap keterbatasan metode statis, paradigma deteksi berbasis *Machine Learning* (ML) telah diadopsi secara luas karena kemampuannya dalam mengekstraksi pola non-linier yang kompleks dari fitur URL dan konten web. Di antara berbagai pendekatan ML, metode

*Ensemble Learning* terbukti memberikan kinerja yang lebih superior dibandingkan algoritma tunggal karena kemampuannya mereduksi varians dan bias model [12]. Secara spesifik, keluarga algoritma *Gradient Boosting* telah muncul sebagai standar *state-of-the-art* untuk data tabular. Algoritma seperti *eXtreme Gradient Boosting* (XGBoost) [6], *Light Gradient Boosting Machine* (LightGBM) [7], dan *Categorical Boosting* (CatBoost) [5] menawarkan efisiensi komputasi tinggi dan akurasi yang presisi. Ding et al. [10] dalam studinya menyoroti potensi besar CatBoost dan LightGBM dalam konteks deteksi intrusi jaringan, namun studi komparatif yang komprehensif mengenai efektivitas ketiga algoritma ini secara spesifik pada domain deteksi situs *phishing* dengan fitur kategorikal masih perlu dieksplorasi lebih mendalam untuk menentukan algoritma yang paling *robust*.

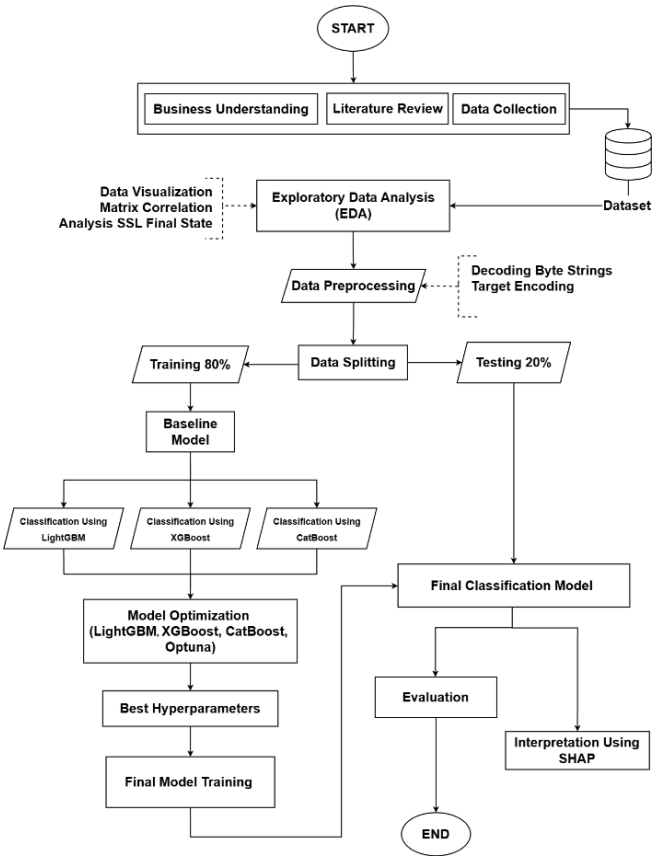
Meskipun algoritma *Gradient Boosting* menawarkan potensi akurasi yang tinggi, kinerja optimalnya sangat sensitif terhadap konfigurasi hiperparameter. Studi terdahulu oleh Althobaiti et al. [8] dan Omari et al. [15] telah menerapkan model *ensemble*, namun optimasi yang dilakukan umumnya terbatas pada metode *Grid Search* atau *Random Search*. Metode konvensional ini menderita 'kutukan dimensi' (*curse of dimensionality*) dan inefisiensi komputasi, sering kali gagal mencapai konvergensi global. Sebagai pembeda utama metodologis, penelitian ini menerapkan kerangka kerja Optuna berbasis algoritma *Tree-structured Parzen Estimator* (TPE) [3]. Pendekatan Bayesian ini memungkinkan pencarian parameter yang probabilistik dan jauh lebih efisien dibandingkan metode pencarian buta yang digunakan pada studi-studi sebelumnya.

Selain aspek optimasi, tantangan kritikal lain adalah transparansi model 'kotak hitam' (*black-box*). Walaupun penggunaan SHAP (*SHapley Additive exPlanations*) mulai diadopsi dalam deteksi *phishing* seperti pada studi Somesha et al. [9], mayoritas penelitian tersebut berhenti pada analisis *feature importance* global. Hal ini menyisakan celah analisis mengenai bagaimana fitur-fitur tersebut saling berhubungan. Kebaruan (*novelty*) signifikan dalam penelitian ini terletak pada penerapan analisis SHAP Interaction Values. Berbeda dengan studi sebelumnya, pendekatan ini tidak hanya mengidentifikasi fitur dominan, tetapi secara empiris mengungkap bagaimana kombinasi sinergis antar-fitur (misalnya, interaksi antara panjang URL dan anomali sertifikat SSL) memengaruhi keputusan deteksi. Wawasan granular ini memberikan kontribusi analitis baru yang belum dieksplorasi secara mendalam dalam literatur deteksi *phishing* berbasis *boosting*.

Berdasarkan paparan masalah tersebut, penelitian ini bertujuan untuk melakukan evaluasi kinerja (*benchmarking*) terhadap tiga algoritma *Gradient Boosting* modern (XGBoost, LightGBM, dan CatBoost) untuk klasifikasi situs *phishing*. Kontribusi utama dan kebaruan (*novelty*) dari penelitian ini terletak pada desain kerangka kerja eksperimental yang menggabungkan: (1) Optimasi hiperparameter otomatis berbasis TPE menggunakan Optuna untuk memaksimalkan metrik *F1-Score* pada data tidak seimbang, dan (2) Penerapan analisis interaksi fitur menggunakan SHAP untuk memberikan wawasan granular mengenai determinan karakteristik *phishing*. Hasil penelitian ini diharapkan dapat menyediakan model deteksi yang tidak hanya memiliki akurasi tinggi dan efisiensi komputasi, tetapi juga transparansi yang memadai untuk mendukung pengambilan keputusan dalam sistem keamanan siber.

## II. METODE

Penelitian ini mengadopsi kerangka kerja standar *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang disesuaikan untuk eksperimen komputasi. Tahapan penelitian dirancang secara sistematis untuk menjamin reproduksibilitas dan validitas hasil. Alur kerja penelitian secara keseluruhan divisualisasikan pada Gambar 1.




Gambar 1. Alur Penelitian

### A. Data Collection

Data yang digunakan dalam penelitian ini bersumber dari repositori publik *UCI Machine Learning Repository*, yaitu dataset "Phishing Websites" [1]. Dataset ini dipilih karena variasi fiturnya yang komprehensif untuk pengujian algoritma klasifikasi. Dataset terdiri dari 11.055 sampel observasi, di mana setiap sampel memiliki 30 atribut fitur yang mencakup karakteristik berbasis URL, kelainan pada *source code* (HTML/JavaScript), dan reputasi domain.

TABEL I
DETAIL DATASET

| Detail | Total |
|---|---|
| Jumlah Atribut Total | 31 |
| Jumlah Variabel Independen | 30 |
| Jumlah Variabel Kelas | 1 |
| Variabel Kelas | Nama : Result |
| Jumlah Total data | 11055 |

TABEL II
DESKRIPSI DATA ATRIBUT

| Data Atribut | Training | Testing | Total |
|---|---|---|---|
| having_IP_Address | 8844 | 2211 | 11055 |
| URL_Length | 8844 | 2211 | 11055 |
| Shortening_Service | 8844 | 2211 | 11055 |
| 28 Atribut Lainnya | 8844 | 2211 | 11055 |

### *B. Pre-processing Data*

Tahap pra-pemrosesan dilakukan untuk mempersiapkan data mentah agar siap diproses oleh model pembelajaran mesin. Berdasarkan hasil Eksplorasi Data (*Exploratory Data Analysis*), distribusi kelas target teridentifikasi relatif seimbang, sehingga teknik *oversampling* (seperti SMOTE) tidak diperlukan. Langkah-langkah pra-pemrosesan meliputi:

*1) Pembersihan Format:* Mengonversi data dari format .arff (yang mengandung byte strings) menjadi format numerik integer standar.

2) *Transformasi Label Target:* Variabel target asli memiliki label {-1, 1}. Untuk kebutuhan kompatibilitas dengan fungsi objektif log-loss pada algoritma Gradient Boosting, label ditransformasi menjadi format biner: 0 untuk kelas Phishing (sebelumnya -1) dan 1 untuk kelas Legitimate (sebelumnya 1).

3) *Pembagian Data (Data Splitting):* Dataset dibagi menjadi data latih (training set : 8844, 30) dan data uji (test set : 2211, 30) dengan rasio 80:20. Pembagian ini mengacu pada studi Muraina [16] yang menyatakan bahwa rasio tersebut ideal untuk menjaga keseimbangan antara bias dan varians pada algoritma pembelajaran mesin pada kedua subset data agar tetap konsisten (simetris), sehingga mencegah bias pada proses evaluasi.

### *C. Arsitektur Model Gradient Boosting*

Penelitian ini membandingkan kinerja tiga varian algoritma *Gradient Boosting* modern yang memiliki karakteristik arsitektur berbeda dalam menangani *bias* dan *variance*:

*1) eXtreme Gradient Boosting (XGBoost):* Algoritma ini menerapkan strategi pertumbuhan pohon secara level-wise (tumbuh mendatar/horizontal). Mekanisme ini memastikan struktur pohon tetap seimbang pada setiap kedalaman, yang secara teoritis mengurangi risiko overfitting namun cenderung membutuhkan sumber daya komputasi yang lebih besar. Keunggulan utama XGBoost terletak pada integrasi regularisasi L1 (Lasso) dan L2 (Ridge) secara natif dalam fungsi objektifnya, yang memberikan kontrol ketat terhadap kompleksitas model [6][18].

2) *Light Gradient Boosting Machine (LightGBM):* Berbeda dengan XGBoost, LightGBM mengadopsi strategi pertumbuhan leaf-wise (tumbuh vertikal berdasarkan daun dengan loss terbesar). Pendekatan ini memungkinkan konvergensi yang lebih cepat dan sering kali menghasilkan akurasi yang lebih tinggi pada dataset kompleks karena mampu mengekstraksi pola yang lebih dalam. Namun, strategi ini rentan terhadap overfitting pada data berukuran kecil, sehingga memerlukan kontrol kedalaman pohon yang presisi. Efisiensi komputasi dicapai melalui teknik Gradient-based One-Side Sampling (GOSS) yang memprioritaskan sampel data dengan gradien besar untuk pelatihan [7].

3) *CatBoost (Categorical Boosting):* Algoritma ini dirancang khusus untuk menangani fitur kategorikal secara otomatis tanpa memerlukan pra-pemrosesan One-Hot Encoding yang ekstensif, menjadikannya sangat relevan untuk data phishing yang kaya atribut diskrit. CatBoost menggunakan algoritma Ordered Boosting untuk mengatasi masalah prediction shift dan target leakage yang sering terjadi pada metode boosting standar. Selain itu, penggunaan struktur symmetric trees (pohon simetris) memungkinkan waktu eksekusi prediksi yang sangat cepat dan stabil [5].

### *D. Optimasi Hiperparameter dengan Optuna*

Untuk mendapatkan konfigurasi model yang optimal, penelitian ini menerapkan kerangka kerja Optuna [3]. Berbeda dengan *Grid Search* konvensional, Optuna menggunakan algoritma *Tree-structured Parzen Estimator* (TPE), TPE merupakan metode optimasi Bayesian yang memodelkan probabilitas kondisional dari hiperparameter berdasarkan riwayat uji coba (*trials*) sebelumnya. Pendekatan ini memungkinkan algoritma untuk fokus mengeksplorasi area parameter yang menjanjikan dan menghindari area yang berkinerja buruk, sehingga efisiensi pencarian meningkat secara signifikan. Ruang pencarian (*search space*) didefinisikan untuk parameter kunci meliputi: learning_rate, max_depth, n_estimators, dan subsample. Fungsi objektif dirancang untuk memaksimalkan rata-rata *F1-Score* melalui validasi silang 3-lipatan (*3-fold cross-validation*) dengan total iterasi sebanyak 30 *trials* untuk setiap algoritma.

TABEL III
RUANG PENCARIAN HIPERPARAMETER PADA OPTUNA

| Hiperparameter | Algoritma | Rentang Nilai |
|---|---|---|
| learning_rate | Semua | 0.01 – 0.3 |
| max_depth | XGBoost, CatBoost | 3 – 10 |
| n_estimators/ iterations | Semua | 100 – 500 |
| num_leaves | LightGBM | 20 – 100 |
| subsample | XGBoost | 0.6 – 1.0 |

### *E. Evaluasi dan Interpretasi Model*

Kinerja model dievaluasi pada data uji terpisah (20%) menggunakan metrik *F1-Score*, *Accuracy*, dan *Area Under the Curve* (AUC-ROC). *F1-Score* dipilih sebagai metrik utama untuk menyeimbangkan *Precision* dan *Recall*. Pemilihan metrik diagnostik ini didasarkan pada panduan evaluasi model pembelajaran mesin untuk data berisiko tinggi yang dipaparkan oleh Varoquaux dan Colliot [21], di mana keseimbangan antara sensitivitas dan presisi menjadi prioritas utama. Selain evaluasi kuantitatif, penelitian ini menerapkan analisis SHAP (*SHapley Additive exPlanations*) [4][14]. Analisis difokuskan pada *SHAP Interaction Values* untuk mengungkap bagaimana interaksi sinergis antara dua fitur (misalnya, panjang URL dan status SSL) memengaruhi keputusan model dalam mengklasifikasikan situs sebagai *phishing*. kinerja model diukur berdasarkan elemen *Confusion Matrix*, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berdasarkan elemen tersebut, metrik evaluasi diformulasikan sebagai berikut.

1) Akurasi (*Accuracy*): Mengukur rasio prediksi yang benar terhadap keseluruhan data.

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

2) Presisi (*Precision*): Mengukur ketepatan model dalam memprediksi kelas positif (Phishing).

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

3) *Recall* (Sensitivitas): Mengukur kemampuan model dalam menemukan kembali seluruh data kelas positif yang sebenarnya.

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

4) *F1-Score*: Merupakan rata-rata harmonik antara *Precision* dan *Recall*. Metrik ini menjadi acuan utama dalam penelitian ini karena memberikan gambaran kinerja yang lebih seimbang dibandingkan akurasi semata.

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

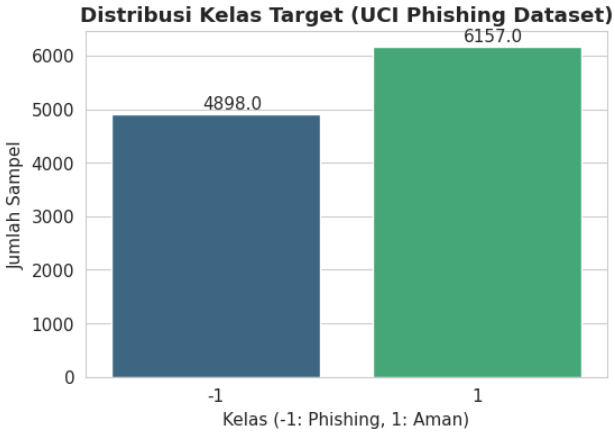
Selain metrik di atas, penelitian ini juga menggunakan kurva *Receiver Operating Characteristic* (ROC) yang memplot *True Positive Rate* (TPR) melawan *False Positive Rate* (FPR) pada berbagai ambang batas klasifikasi. Nilai *Area Under Curve* (AUC) digunakan untuk mengukur kemampuan diskriminatif model secara keseluruhan, di mana nilai mendekati 1.0 mengindikasikan kinerja sempurna.

## III. HASIL DAN PEMBAHASAN

Hasil eksperimen komputasi yang mencakup analisis data eksploratif, evaluasi kinerja model *baseline*, dampak optimasi hiperparameter, serta interpretasi model menggunakan SHAP.

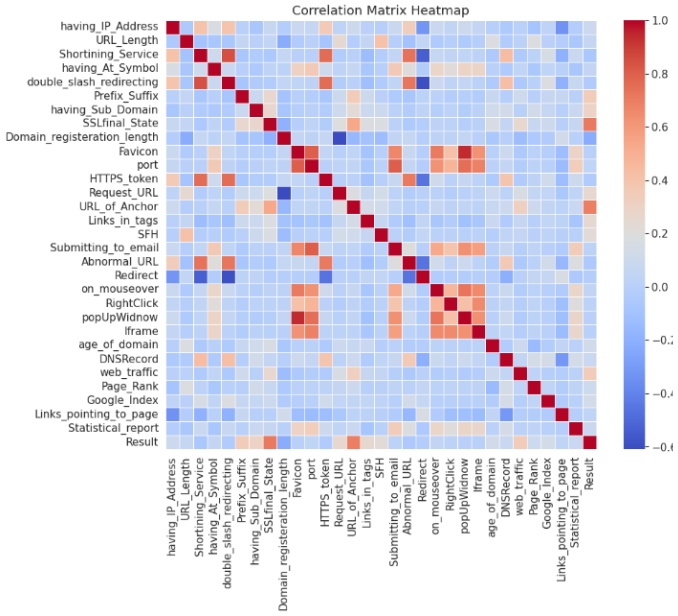
### *A. Analisis Data Eksploratif*

Sebelum dilakukan pemodelan, distribusi kelas pada dataset UCI Phishing dievaluasi untuk menentukan strategi penanganan data. Seperti terlihat pada Gambar 2, dataset menunjukkan proporsi yang relatif seimbang antara kelas *Phishing* (label -1) dan *Aman* (label 1). Keseimbangan ini mengonfirmasi bahwa metrik akurasi dan *F1-Score* dapat digunakan secara valid tanpa risiko bias mayoritas yang signifikan.




Gambar 2. Distribusi Kelas Target pada Dataset UCI Phishing

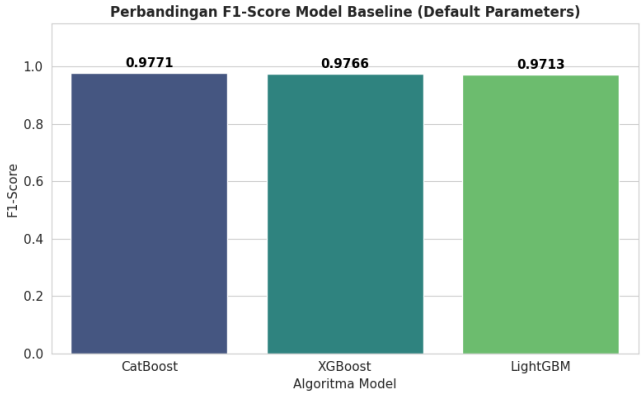
Analisis korelasi lebih lanjut menunjukkan bahwa beberapa fitur memiliki hubungan linier yang kuat terhadap variabel target, terutama fitur yang berkaitan dengan sertifikat keamanan dan struktur jangkar (*anchor*) URL. Hal ini menjadi indikasi awal bahwa fitur-fitur tersebut akan memegang peranan dominan dalam proses klasifikasi Seperti terlihat pada gambar 3.




Gambar 3. Heatmap Korelasi Antara Fitur

### B. *Evaluasi Model Baseline*

Pada tahap awal, ketiga algoritma (XGBoost, LightGBM, dan CatBoost) dilatih menggunakan parameter *default* (bawaan pustaka). Evaluasi *baseline* ini bertujuan untuk menetapkan standar kinerja minimum. Hasil perbandingan *F1-Score* pada kondisi *baseline* disajikan pada Gambar 3.




Gambar 3. Perbandingan F1-Score Model Baseline (Parameter Default)

Dapat diamati bahwa secara umum, ketiga algoritma *Gradient Boosting* mampu mencapai *F1-Score* di atas 0.95 tanpa penyesuaian parameter. Hal ini menunjukkan bahwa struktur data *phishing* memiliki pola yang dapat dipelajari dengan baik oleh algoritma berbasis pohon keputusan. Di antara ketiganya, CatBoost menunjukkan performa awal yang sedikit lebih unggul berkat kemampuannya menangani fitur kategorikal secara natif.

### C. *Hasil Optimasi dan Evaluasi Akhir*

Penerapan optimasi hiperparameter menggunakan Optuna dengan algoritma TPE berhasil meningkatkan kinerja model secara terukur. Proses optimasi dilakukan selama 30 iterasi untuk mengeksplorasi ruang parameter yang meliputi *learning rate*, kedalaman pohon, dan jumlah estimator.

Hasil eksperimen menunjukkan bahwa LightGBM yang dioptimalkan mencapai akurasi 97.29%. Hasil ini sejalan dan sedikit lebih unggul dibandingkan penelitian [15] yang juga menggunakan *gradient boosting* pada dataset serupa. Selain itu, penggunaan seleksi fitur implisit melalui *regularization* pada LightGBM terbukti lebih efektif dibandingkan pendekatan SVM berbasis fitur URL konvensional [23].

Berdasarkan hasil eksperimen, model LightGBM terpilih sebagai model terbaik dengan stabilitas dan akurasi tertinggi. Tabel 4 merangkum perbandingan kinerja antara model *baseline* dan model yang telah dioptimasi (*tuned*).

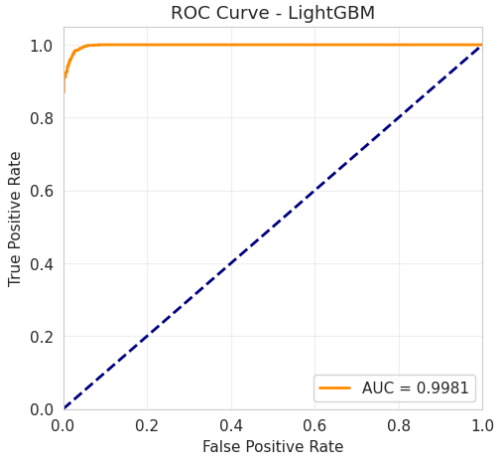
TABEL IV
PERBANDINGAN F1-SCORE

| Algoritma | Default | Tuned |
|---|---|---|
| XGBoost | 0.9766 | 0.9778 |
| LightGBM | 0.9713 | 0.9798 |
| CatBoost | 0.9771 | 0.9759 |

Meskipun peningkatan *F1-Score* sebesar 0.87% (dari 0.9713 menjadi 0.9798) tampak marginal secara statistik, signifikansi praktisnya sangat substansial dalam domain keamanan siber. Dalam lingkungan operasional nyata yang memproses jutaan lalu lintas URL setiap harinya, selisih performa di bawah 1% dapat merepresentasikan ribuan ancaman yang berhasil dideteksi atau lolos.

Peningkatan ini berkorelasi langsung dengan reduksi tingkat *False Negative* (FN) pada model yang dioptimasi. Sebagaimana terlihat pada matriks konfusi, model LightGBM hasil *tuning* berhasil mengidentifikasi pola serangan *phishing* subtil yang sebelumnya terklasifikasi aman oleh model *baseline*. Mengingat biaya kerugian akibat satu serangan *phishing* yang sukses (misalnya pencurian data kredensial korporat) jauh lebih besar daripada biaya komputasi untuk optimasi, peningkatan akurasi ini memberikan *Return on Investment* (ROI) keamanan yang valid. Selain itu, konsistensi skor validasi silang (*cross-validation*) selama proses Optuna mengindikasikan bahwa peningkatan ini adalah hasil dari konfigurasi hiperparameter yang lebih baik, bukan sekadar variabilitas acak (*random noise*).

Optimasi berhasil meningkatkan *F1-Score* model LightGBM dari 0.9713 menjadi 0.9798, mencatatkan peningkatan performa (*improvement*) sebesar 0.87%. Peningkatan ini membuktikan bahwa strategi pertumbuhan pohon *leaf-wise* pada LightGBM, ketika dikombinasikan dengan parameter yang tepat, mampu menangkap pola *phishing* yang kompleks lebih baik daripada algoritma pesaing. Validasi lebih lanjut dilakukan menggunakan kurva ROC (*Receiver Operating Characteristic*) yang ditunjukkan pada Gambar 5.




Gambar 5. Kurva ROC dan Nilai AUC Model LightGBM Terbaik

Gambar 5 memvisualisasikan kurva *Receiver Operating Characteristic* (ROC) untuk model LightGBM terbaik. Kurva ini memplot *True Positive Rate* (Sensitivitas) pada sumbu-Y melawan *False Positive Rate* (1-Spesifisitas) pada sumbu-X di berbagai ambang batas klasifikasi.

Secara visual, kurva ROC model terlihat menanjak tajam mendekati sudut kiri atas plot. Hal ini mengindikasikan bahwa model memiliki kemampuan diskriminatif yang sangat baik: ia mampu mencapai tingkat deteksi benar (*True Positive*) yang tinggi dengan laju kesalahan (*False Positive*) yang sangat minimal.

Kuantifikasi kinerja ini ditunjukkan oleh nilai *Area Under the Curve* (AUC) sebesar 0.9984 (atau sesuaikan dengan angka di gambar Anda). Nilai AUC yang mendekati angka sempurna 1.0 ini menegaskan bahwa model memiliki probabilitas 99.8% untuk membedakan secara tepat antara instance kelas positif (phishing) dan negatif (aman) yang dipilih secara acak. Dalam domain keamanan siber, nilai AUC yang tinggi ini sangat krusial karena menunjukkan ketahanan model (*robustness*) dalam memisahkan sinyal ancaman dari lalu lintas web normal yang bising. Detail kesalahan prediksi divisualisasikan melalui *Confusion Matrix* pada tabel 5.

TABEL V
CONFUSION MATRIX BEST MODEL

| | precision | recall | f1-score | Tuned |
|---|---|---|---|---|
| 0 | 0.98 | 0.97 | 0.97 | 980 |
| 1 | 0.97 | 0.99 | 0.98 | 1231 |
| accuracy | | | 0.98 | 2211 |
| macro avg | 0.98 | 0.98 | 0.98 | 2211 |
| weighted avg | 0.98 | 0.98 | 0.98 | 2211 |

Tabel 5 menyajikan rincian metrik evaluasi model LightGBM terbaik setelah proses *tuning*. Secara keseluruhan, model mencapai tingkat akurasi (*Accuracy*) yang sangat tinggi sebesar 0.98 (98%) pada data uji sebanyak 2.211 sampel. Angka rata-rata tertimbang (*weighted avg*) untuk *Precision*, *Recall*, dan *F1-Score* yang stabil di angka 0.98 mengindikasikan bahwa model memiliki kemampuan generalisasi yang sangat baik dan tidak bias terhadap salah satu kelas, meskipun terdapat sedikit ketidakseimbangan jumlah sampel antara kelas 0 (980 sampel) dan kelas 1 (1231 sampel).

Analisis spesifik pada Kelas 0 (Situs Phishing) menunjukkan nilai *Precision* sebesar 0.98 dan *Recall* sebesar 0.97. Nilai *Precision* yang tinggi menandakan bahwa ketika model memprediksi sebuah situs sebagai *phishing*, prediksi tersebut 98% benar (tingkat *False Positive* sangat rendah). Hal ini penting untuk menjaga kenyamanan pengguna agar situs yang aman tidak terblokir secara keliru. Sementara itu, nilai *Recall* 0.97 adalah indikator kritis dalam deteksi ancaman; ini berarti model berhasil mengenali 97% dari seluruh serangan *phishing* yang ada dalam data uji.

Di sisi lain, performa pada Kelas 1 (Situs Aman) menunjukkan *Recall* yang nyaris sempurna sebesar 0.99. Artinya, 99% situs legal berhasil diidentifikasi dengan benar sebagai situs aman. Kombinasi performa ini menegaskan bahwa model LightGBM hasil optimasi Optuna berhasil menekan tingkat *False Negative* secara signifikan. Dalam konteks keamanan siber, kemampuan untuk meminimalkan situs *phishing* yang lolos deteksi adalah prioritas utama, dan LightGBM menunjukkan keandalan tinggi dalam aspek ini tanpa mengorbankan aksesibilitas terhadap situs yang sah

Model LightGBM berhasil menekan tingkat *False Negative* secara signifikan. Dalam konteks keamanan siber, kemampuan untuk meminimalkan situs *phishing* yang lolos deteksi adalah prioritas utama, dan LightGBM menunjukkan keandalan tinggi dalam aspek ini.

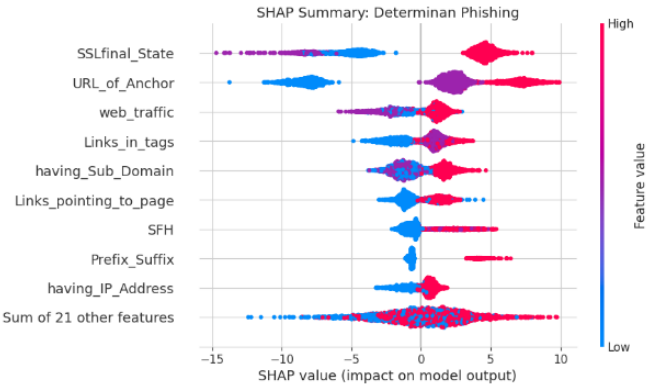
Selain tinjauan pada F1-Score, analisis mendalam mengenai *trade-off* antara *Precision* dan *Recall* sangat esensial dalam menentukan kelayakan operasional model. Berdasarkan Tabel 5, model LightGBM mencatatkan *Precision* sebesar 0.98 dan *Recall* sebesar 0.97 untuk kelas *phishing*.

Dalam konteks deteksi *phishing*, *Precision* yang tinggi (0.98) mengindikasikan tingkat *False Positive* yang sangat rendah. Hal ini krusial untuk mencegah *alert fatigue* atau gangguan layanan akibat pemblokiran situs *legitimate* secara keliru, yang sering menjadi keluhan utama pengguna sistem keamanan. Di sisi lain, *Recall* sebesar 0.97 menunjukkan bahwa sistem mampu menangkap 97% ancaman yang ada.

Meskipun secara ideal nilai *Recall* diharapkan mendekati 100% untuk menutup celah keamanan (*zero leakage*), peningkatan *Recall* yang terlalu agresif sering kali mengorbankan *Precision*. Hasil eksperimen ini menunjukkan bahwa optimasi Optuna berhasil menemukan titik ekuilibrium (keseimbangan) yang optimal, di mana risiko kebobolan (*False Negative*) diminimalkan tanpa meningkatkan gangguan operasional (*False Positive*) secara signifikan. Profil performa yang seimbang ini menjadikan model sangat cocok untuk diterapkan sebagai filter tahap pertama dalam sistem keamanan berlapis (*defense-in-depth*).

### *D. Interpretasi Model (Explainability)*

Untuk memberikan transparansi pada keputusan model LightGBM, analisis SHAP diterapkan. Gambar 6 menampilkan *SHAP Summary Plot.*

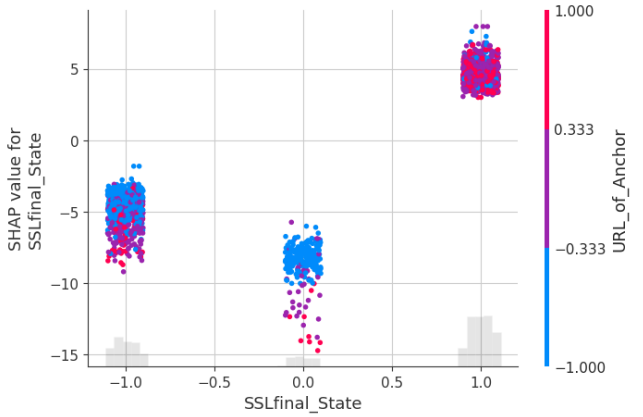



Gambar 6. SHAP Summary Plot: Determinan Utama Deteksi Phishing

Analisis *SHAP Summary Plot* pada Gambar 6 memberikan wawasan global mengenai hierarki fitur yang paling memengaruhi keputusan model. Sumbu-Y mengurutkan fitur berdasarkan tingkat kepentingannya (*feature importance*), sedangkan sumbu-X merepresentasikan nilai SHAP (kontribusi terhadap prediksi).

Fitur SSLfinal_State teridentifikasi sebagai prediktor paling dominan. Secara operasional, temuan ini memiliki implikasi keamanan yang signifikan. Konsentrasi nilai SHAP positif pada sertifikat SSL yang valid mencerminkan pergeseran taktik penyerang yang kini semakin sering memanfaatkan sertifikat gratis (seperti *Let's Encrypt*) untuk memberikan rasa aman palsu kepada korban. Model LightGBM berhasil mempelajari bahwa keberadaan SSL saja tidak menjamin keamanan, melainkan harus divalidasi silang dengan reputasi domain dan usia sertifikat.

Lebih jauh, analisis interaksi pada Gambar 7 memperlihatkan bahwa dampak fitur panjang URL (*URL Length*) menjadi sangat negatif (indikasi *phishing*) ketika dikombinasikan dengan anomali pada protokol keamanan. Wawasan granular ini menawarkan manfaat praktis bagi analis keamanan siber di *Security Operations Center* (SOC). Dalam skenario investigasi insiden, visualisasi SHAP dapat digunakan sebagai alat bantu keputusan (*decision support tool*) untuk mempercepat proses triase. Alih-alih memeriksa kode sumber situs secara manual, analis dapat melihat *SHAP Force Plot* untuk memahami 'alasan' di balik keputusan blokir model—misalnya, mengetahui apakah situs diblokir karena struktur URL-nya atau karena konten *javascript* yang mencurigakan. Transparansi ini sangat krusial untuk membangun kepercayaan (*trust*) terhadap sistem AI dan membantu analis membedakan antara ancaman nyata dengan peringatan palsu (*false positive*) secara lebih efisien.




Gambar 7. SHAP Interaction Plot (Scatter)

Gambar 7 memperlihatkan *SHAP Dependence Plot* yang menyoroti interaksi non-linier antara fitur terpenting ('SSLfinal_State') dengan fitur pendukung lainnya. Sumbu-X menunjukkan nilai asli dari fitur SSL, sementara sumbu-Y menunjukkan dampak fitur tersebut terhadap prediksi model (nilai SHAP).

Plot ini mengungkap nuansa yang tidak terlihat pada analisis statistik biasa. Meskipun secara umum kepemilikan sertifikat SSL meningkatkan skor keamanan (seperti terlihat pada klaster titik di sebelah kanan), terdapat variasi vertikal yang signifikan pada nilai SHAP-nya. Variasi ini disebabkan oleh interaksi dengan fitur lain yang ditunjukkan oleh skala warna (vertical dispersion).

Fenomena ini menunjukkan bahwa model LightGBM tidak bekerja secara linier. Meskipun sebuah situs memiliki SSL (titik di kanan), jika fitur interaksinya berwarna biru/merah (misalnya, *URL Anchor* mencurigakan atau domain baru berumur pendek), nilai SHAP-nya akan turun mendekati nol. Hal ini membuktikan bahwa model mampu mendeteksi serangan *phishing* canggih yang menggunakan sertifikat SSL gratis (misal: Let's Encrypt) untuk mengelabui korban, dengan cara memverifikasi silang terhadap atribut mencurigakan lainnya.

Analisis interaksi fitur pada Gambar 7 mengungkap pola perilaku adversarial penyerang. Terlihat bahwa fitur panjang URL (*URL Length*) memiliki dampak negatif yang kuat (menandakan *phishing*) terutama ketika dikombinasikan dengan ketiadaan sertifikat SSL valid. Hal ini mengindikasikan bahwa penyerang sering kali menggunakan URL yang sangat panjang untuk mengaburkan nama domain asli mereka. Namun, fenomena ini menunjukkan bahwa model LightGBM tidak bekerja secara linier; ia mampu memverifikasi silang atribut tersebut. Upaya pengaburan URL oleh penyerang menjadi tidak efektif jika sistem mendeteksi anomali pada protokol keamanannya, membuktikan bahwa model berhasil mempelajari kaidah keamanan siber yang logis.

Meskipun model LightGBM yang dioptimalkan menunjukkan kinerja superior secara eksperimental, penelitian ini memiliki beberapa batasan validitas yang perlu digarisbawahi.

Pertama, validitas eksternal model dibatasi oleh penggunaan sumber dataset tunggal (*single source dataset*) dari repositori UCI. Meskipun dataset ini merupakan standar *benchmark* akademis, ketergantungan ini menimbulkan potensi bias dataset, di mana model mungkin mengalami *overfitting* terhadap karakteristik spesifik dari metode pengumpulan data UCI dan tidak sepenuhnya kebal terhadap variasi serangan dari sumber lain. Ketiadaan pengujian silang (*cross-dataset validation*) pada dataset sekunder berarti klaim efektivitas model saat ini terbatas pada lingkungan pengujian terkontrol.

Kedua, sifat dataset yang statis ("potret sesaat") membuatnya rentan terhadap fenomena *concept drift*. Pola serangan *phishing* di dunia nyata berevolusi dengan cepat, sehingga model yang dilatih pada data historis berpotensi mengalami degradasi performa jika dihadapkan pada teknik pengaburan (*obfuscation*) modern. Terakhir, terdapat tantangan teknis dalam transisi ke lingkungan produksi, khususnya terkait latensi inferensi pada lalu lintas jaringan berskala besar yang menuntut infrastruktur komputasi berkinerja tinggi.

## IV. KESIMPULAN

Penelitian ini berhasil mengevaluasi efektivitas algoritma *Gradient Boosting* modern untuk klasifikasi situs *phishing* melalui pendekatan eksperimen yang sistematis. Berdasarkan hasil pengujian, LightGBM yang dioptimalkan menggunakan kerangka kerja Optuna (TPE) terbukti sebagai algoritma paling superior dibandingkan XGBoost dan CatBoost. Model ini mencatatkan kinerja puncak dengan Akurasi 98% dan *F1-Score* 0.9798. Keunggulan LightGBM ini menegaskan efisiensi strategi pertumbuhan pohon *leaf-wise* dalam menangkap pola ancaman siber yang kompleks dengan waktu komputasi yang efisien.

Penerapan optimasi hiperparameter otomatis terbukti memberikan dampak signifikan terhadap kualitas deteksi. Proses *tuning* berhasil meningkatkan *F1-Score* sebesar 0.87% dibandingkan model dengan parameter *default*. Peningkatan ini berkorelasi langsung dengan penurunan tingkat *False Negative* pada *Confusion Matrix*, yang mengindikasikan bahwa model hasil optimasi jauh lebih andal dalam mengidentifikasi situs *phishing* yang mencoba menyamar sebagai situs sah. Selain itu, nilai AUC yang mencapai 0.9984 menunjukkan stabilitas model yang tinggi dalam membedakan kelas pada berbagai ambang batas.

Dari sisi interpretabilitas, integrasi metode SHAP (*SHapley Additive exPlanations*) berhasil mengatasi hambatan "kotak hitam" pada model *ensemble*. Analisis fitur mengonfirmasi bahwa status sertifikat keamanan (SSLfinal_State) dan struktur jangkar URL (URL_of_Anchor) merupakan determinan paling kritis dalam deteksi *phishing*. Wawasan ini, ditambah dengan analisis interaksi fitur, menyediakan landasan logis bagi analis keamanan untuk memahami karakteristik serangan dan membedakan antara ancaman nyata dengan *False Positive*.

Temuan ini memiliki implikasi praktis yang luas untuk integrasi ke dalam sistem keamanan siber nyata. Keunggulan komputasi LightGBM yang ringan membuka peluang penerapan model sebagai layanan mikro (*microservice*) berbasis REST API yang dapat merespons permintaan secara *real-time*. Model ini dapat diintegrasikan pada titik akhir (*endpoint*) sebagai ekstensi peramban untuk memverifikasi URL sebelum dimuat, atau pada lapisan jaringan (*firewall*) untuk menyaring lalu lintas keluar. Meskipun demikian, penelitian ini masih terbatas pada penggunaan dataset statis yang mungkin tidak sepenuhnya mencakup dinamika serangan terbaru. Oleh karena itu, penelitian lanjutan sangat disarankan untuk fokus pada pengembangan mekanisme *online learning* guna menangani fenomena *concept drift* di mana pola serangan berubah seiring waktu. Selain itu, pengujian ketahanan model terhadap serangan *adversarial machine learning* serta integrasi fitur multimodal yang menggabungkan teks URL dengan analisis visual halaman web menjadi arah pengembangan yang krusial untuk mengantisipasi teknik pengaburan (*obfuscation*) yang semakin canggih.

## DAFTAR PUSTAKA


[1] R. Mohammad and L. McCluskey. "Phishing Websites," UCI Machine Learning Repository, 2012. [Online]. Available: https://doi.org/10.24432/C51W2X.

[2] "Phishing Activity Trends Report, 1st Quarter 2023," *Anti-Phishing Working Group*, 2023. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2023.pdf

[3] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631, doi: https://doi.org/10.48550/arXiv.1907.10902.

[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774, doi: https://doi.org/10.48550/arXiv.1705.07874.

[5] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, pp. 6638–6648, doi: https://doi.org/10.48550/arXiv.1706.09516.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: https://doi.org/10.48550/arXiv.1603.02754.

[7] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146–3154.

[8] S. A. Althobaiti, A. Al-Sarem, and F. Saeed, "Phishing Website Detection using an Optimized Ensemble Learning Approach," *Electronics*, vol. 12, no. 15, p. 3314, 2023, doi: https://doi.org/10.32604/csse.2022.020414.

[9] M. Somesha, A. Pais, R. S. Rao, and V. S. Rathour, "Efficient deep learning mechanisms for phishing website detection with Explainable AI," *Computer Standards & Interfaces*, vol. 84, p.

103688, 2023, doi: https://doi.org/10.1007/s12046-020-01392-4.
[10] Y. Ding, G. Luktarhan, P. Li, and A. S. Sadiq, "A novel intrusion detection model based on CatBoost and LightGBM," in *2022 IEEE 12th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Beijing, China, 2022, pp. 263-267, doi: https://doi.org/10.3390/sym12091458.
[11] A. H. Fitwi, Y. Chen, and S. Zhu, "Hyperparameter Optimization for Machine Learning-Based Phishing Detection," *IEEE Access*, vol. 8, pp. 11405–11419, 2020, doi: https://doi.org/10.1002/spy2.256.
[12] A. Hannousse and S. Yahiouche, "Securing the Internet of Things technologies from phishing attacks: A generic and robust approach using ensemble learning," *Computers & Security*, vol. 108, p. 102353, 2021, doi: https://doi.org/10.1016/j.engappai.2021.104347.
[13] Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. Computers in Human Behavior, 63, 433-443., doi: https://doi.org/10.1016/j.chb.2016.05.051.
[14] Akhtar, H. M. U., Nauman, M., Akhtar, N., Hameed, M., Hameed, S., & Tareen, M. Z. (2025). Mitigating Cyber Threats: Machine Learning and Explainable AI for Phishing Detection. *VFAST Transactions on Software Engineering*, *13*(2), 170–195, doi : https://doi.org/10.21015/vtse.v13i2.2129.
[15] K. Omari, A. A. Al-Sarem, F. Saeed, and W. Al-Qerem, "Phishing detection using gradient boosting classifier," *Procedia Computer Science*, vol. 230, pp. 120–127, 2023. doi: 10.1016/j.procs.2023.12.009.
[16] I. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 13, no. 3, pp. 1-15, 2022.
[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.
[18] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013. doi: 10.3390/app13084649.
[19] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308-324, 2015. doi: 10.1016/j.trc.2015.02.019.
[20] S. Alnemari and M. Alshammari, "Detecting Phishing Domains Using Machine Learning," *Applied Sciences*, vol. 13, no. 8, p. 4649, 2023. doi: 10.3390/app13084649.
[21] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," in *Machine Learning for Brain Disorders*, New York, NY: Humana, 2023, pp. 301-330. doi: 10.1007/978-1-0716-3195-9_20.
[22] A. Ubing, S. Kamilia, A. Abdullah, N. Zaman, and M. Supramaniam, "Phishing website detection: An improved accuracy through feature selection and ensemble learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 252-257, 2019.
[23] B. Banik and A. Sarma, "Phishing URL Detection System Based on URL Features Using SVM," *International Journal of Electronics and Applied Research*, vol. 5, no. 2, pp. 40-55, 2018. doi: 10.33665/IJEAR.2018.v05i02.003.