MENGOPTIMALKAN DETEKSI METADATA URL PHISHING MENGGUNAKAN FEATURE SUBSET RANKING DAN LIGHTWEIGHT MODELS

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

disusun oleh

AKBAR HAMONANGAN

21.61.0209

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

MENGOPTIMALKAN DETEKSI METADATA URL PHISHING MENGGUNAKAN FEATURE SUBSET RANKING DAN LIGHTWEIGHT MODELS

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

disusun oleh

AKBAR HAMONANGAN

21.61.0209

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2025

HALAMAN PERSETUJUAN

SKRIPSI

MENGOPTIMALKAN DETEKSI METADATA URL PHISHING MENGGUNAKAN FEATURE SUBSET RANKING DAN LIGHTWEIGHT MODELS

yang disusun dan diajukan oleh

Akbar Hamonangan

21.61.0209

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 Juni 2025

Dosen Pembimbing,

Subektiningsih, M.Kom

NIK. 190302413

HALAMAN PENGESAHAN

SKRIPSI

MENGOPTIMALKAN DETEKSI METADATA URL PHISHING MENGGUNAKAN FEATURE SUBSET RANKING DAN LIGHTWEIGHT MODELS

yang disusun dan diajukan oleh

Akbar Hamonangan

21.61.0209

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 Juni 2025

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Ainul Yaqin, S.Kom., M.Kom **NIK. 190302255**	
Ali Mustopa, S.Kom., M.Kom **NIK. 190302192**	
Subektiningsih, M.Kom **NIK. 190302413**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 Juni 2025

DEKAN FAKULTAS ILMU KOMPUTER

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Akbar Hamonangan
NIM : 21.61.0209

Menyatakan bahwa Skripsi dengan judul berikut:

MENGOPTIMALKAN DETEKSI METADATA URL PHISHING MENGGUNAKAN FEATURE SUBSET RANKING DAN LIGHTWEIGHT MODELS

Dosen Pembimbing : Subektiningsih, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 24 Juni 2025

Yang Menyatakan,

Akbar Hamonangan

HALAMAN PERSEMBAHAN

Skripsi ini dipersembahkan kepada Allah SWT atas rahmat dan petunjuknya yang tiada henti, kedua orang tua tercinta atas doa, kasih sayang, dan dukungannya yang tidak ternilai. Serta kepada para dosen dan pembimbing yang telah membagikan ilmu dan bimbingan selama masa studi, dan juga seluruh pihak yang tekah memberikan dukungan moral maupun material dalam proses penyusunan skripsi ini.

KATA PENGANTAR

Puji syukur saya panjatkan ke hadirat Allah SWT atas segala rahmat dan karunia-nya sehingga saya dapat menyelesaikan skripsi ini sebagai salah satu syarat untuk memperoleh gelar sarjana di program studi informatika, fakultas ilmu komputer, Universitas Amikom Yogyakarta.

Penyusunan skripsi ini tidak lepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, saya menyampaikan ucapan terimakasih yang sebesar-besarnya kepada:

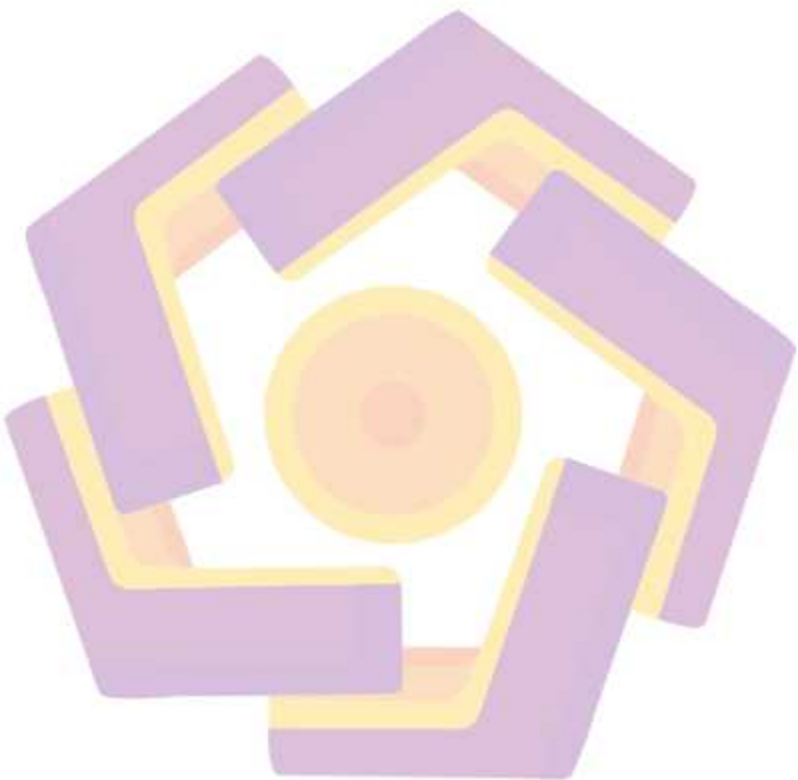
1. Subektiningsih, M. Kom, selaku dosen pembimbing yang telah meluangkan waktu, memberikan bimbingan, arahan, dan motivasi kepada saya selama proses penyusunan skripsi ini.
2. Eli Pujastuti, M.Kom, selaku ketua program studi informatika yang telah memberikan dukungan selama masa studi.
3. Tim Dosen penguji, yang telah memberikan masukan dan kritik yang membangun demi kesempurnaan skripsi ini.
4. Bapak dan ibu tercinta, atas doa, kasih sayang, serta dukungan moral dan material yang tiada henti.
5. Seluruh pihak yang telah membantu secara langsung maupun tidak langsung dalam penyusunan skripsi ini.

Saya menyadari bahwa skripsi ini masih jauh dari kata sempurna. Oleh karena itu, kritik dan saran yang membangun sangat saya harapkan untuk penyempurnaan di masa mendatang.

Yogyakarta, 19 Juni 2025

Penulis

DAFTAR ISI



DAFTAR TABEL

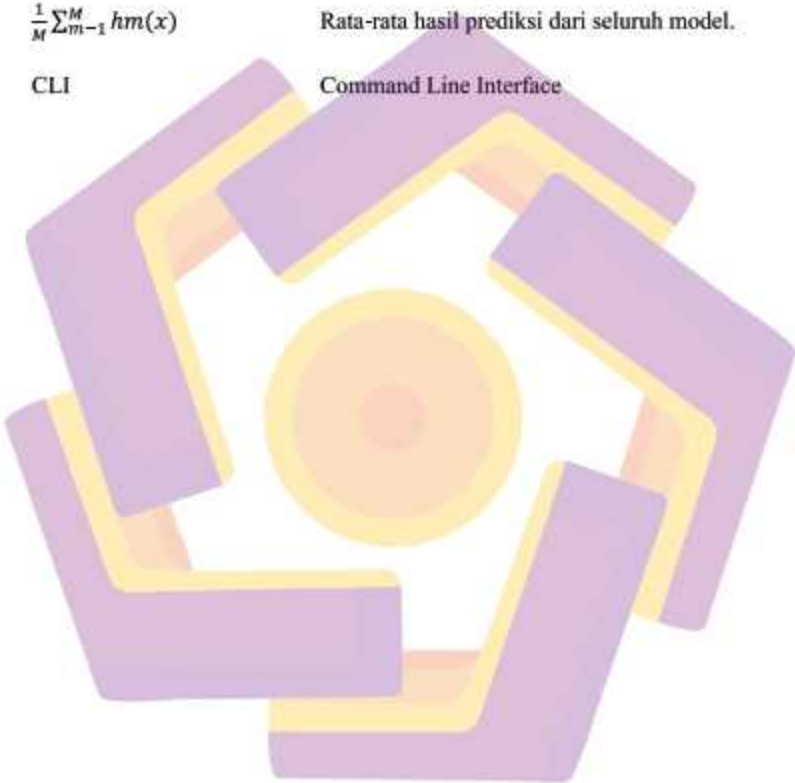
DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

RFE	Recursive Feature Elimination.
Bagging	Bootstrap Aggregating.
AI	Artificial Intelligence.
SVM	Support Vector Machine.
NLP	Natural Language Processing.
RF	Random Forest.
$\hat{y}$	Nilai prediksi (Output dari model).
$ft(x)$	Fungsi prediktor dari model ke-t (terutama decision tree).
T	Jumlah total model.
$(f1(x), f2(x), \ldots, f_T(x))$	Nilai modus.
η	Learning rate atau shrinkage faktor yang mengontrol kontribusi model.
$L^{(t)}$	Fungsi objektif (loss function) pada iterasi ke-t.
$l(yi, \hat{y}i)$	Loss Function untuk prediksi ke-I, misalnya log loss atau squared error.
$\Omega(f_t)$	Fungsi regularisasi dari model f_t, digunakan untuk menghindari overfitting.
γ	Parameter regularisasi jumlah daun (leaf) dalam pohon.
λ	Parameter regularisai L2 untuk bobot daun.
ω_j	Bobot dari daun ke-j dalam pohon keputusan.

$\sum$	Simbol penjumlahan.
$hm(x)$	Fungsi prediksi ke-m, digunakan pada metode averaging seperti bagging.
M	Jumlah model pada metode ensemble bagging.
$\frac{1}{M}\sum_{m-1}^{M} hm(x)$	Rata-rata hasil prediksi dari seluruh model.
CLI	Command Line Interface



DAFTAR ISTILAH

Phishing	Teknik penipuan siber yang bertujuan mencuri informasi sensitif.
Machine Learning	Cabang dari AI yang memungkinkan sistem belajar dari data.
Ensemble Learning	Teknik menggabungkan beberapa model untuk meningkatkan akurasi.
Boosting	Metode ensemble untuk memperbaiki kesalahan model terdahulu.
Bagging	Metode ensemble dengan pelatihan model secara paralel pada subset acak data.
Feature Selection	Proses pemilihan fitur paling relevan untuk efisiensi model.
Feature Subset Ranking	Teknik mengurutkan fitur berdasarkan kontribusi.
RFE	Teknik seleksi fitur secara iteratif dengan menghapus fitur paling tidak penting.
Random Forest	Metode ensemble berbasis ecision tree dengan metode bagging.
RF Feature Importance	Penilaian pentingnya fitur berdasarkan seberapa sering digunakan dalam pohon keputusan.
XGBoost	Model boosting yang fokus pada efisiensi dan akurasi.
LightGBM	Model boosting efisien yang menggunakan leaf-wise growth.
CatBoost	Model boosting yang dioptimalkan untuk data kategorikal.
Decision Tree	Struktur pohon yang digunakan dalam klasifikasi dan regresi.
Accuracy	Rasio prediksi yang benar terhadap total prediksi.
Precision	Rasio prediksi benar terhadap semua prediksi positif.

Recall	Rasio deteksi benar terhadap semua kasus positif sebenarnya.
F1-Score	Rata-rata harmonik dari precision dan recall.
Confusion Matrix	Matriksyang menggambarkan kinerja klasifikasi.
StandarScaler	Metode normalisasai data agar memiliki mean = 0 dan std=1.
Missing Data	Data yang kosong/tidak tersedia dalam dataset.
Encoding	Proses konversi data kategorikal menjadi numerik.
Train-Test Split	Pemisahan data menjadi data pelatihan dan pengujian.
Google Colab	Platform cloud untuk eksperimen python.
Python 3.x	Bahasa pemrograman utama untuk eksperimen ini.
Scikit-learn	Library python untuk machine learning.
Docker	Platform untuk membuat container aplikasi yang terisolasi.
CLI	Aplikasi berbasis teks untuk menjalankan model.

INTISARI

Phishing adalah jenis serangan siber yang semakin umum yang menggunakan situs web berbahaya untuk mengumpulkan informasi penting pengguna. Mendeteksi *URL phishing* merupakan hal yang rumit karena seiring berkembangnya teknologi, phishing akan semakin sulit diatasi. Penelitian ini mencoba untuk meningkatkan identifikasi *URL* phishing dengan menggunakan strategi pemeringkatan subset fitur dan model pembelajaran mesin yang ringan seperti *XGBoost, LightGBM, CatBoost, dan Random Forest*.

Strategi yang digunakan dalam penelitian ini terdiri dari pemilihan atribut berdasarkan kepentingannya dalam mengidentifikasi *URL phishing* dari *URL* yang sah. Setelah itu, eksperimen dijalankan dengan menggunakan beberapa model pembelajaran mesin untuk menilai kegunaan fitur yang dipilih dalam meningkatkan akurasi klasifikasi. Dataset diperoleh dari sumber yang telah dikurasi dan dapat diandalkan. Untuk mencapai kinerja deteksi yang optimal, model yang diimplementasikan dinilai dengan menggunakan ukuran seperti *akurasi*, *presisi*, *recall*, dan *F1-score*.

Temuan ini menunjukkan bahwa memilih subset karakteristik yang tepat dapat meningkatkan akurasi model secara signifikan dengan meminimalkan fitur-fitur yang tidak relevan. Dari empat model yang diuji, *LightGBM, XGBoost, CatBoost, dan Random Forest* menunjukkan performa terbaik dalam mendeteksi *URL* phishing dengan akurasi lebih dari 80%. Penelitian ini menunjukkan bahwa menggabungkan pemeringkatan subset fitur dan model yang ringan dapat menjadi pilihan yang efektif untuk mendeteksi phishing secara *real-time*. Penelitian ini dirancang untuk menjadi sumber daya bagi pembuat sistem keamanan siber yang ingin meningkatkan pertahanan terhadap upaya phishing.

Kata kunci: *Phishing, URL Detection, Machine Learning, Feature Subset Ranking, Lightweight Models.*

ABSTRACT

Phishing is an increasingly common type of cyber attack that uses malicious websites to collect important user information. Detecting phishing URLs is complicated because as technology develops, phishing will become more difficult to overcome. This research attempts to improve phishing URL identification by using feature subset ranking strategies and lightweight machine learning models such as XGBoost, LightGBM, CatBoost, and Random Forest.

The strategy used in this research consists of selecting attributes based on their importance in identifying phishing URLs from Legitimate URLs. Afterward, experiments were run using several machine learning models to assess the usefulness of the selected features in improving classification accuracy. Datasets were obtained from curated and reliable sources. To achieve optimal detection performance, the implemented models were assessed using measures such as accuracy, precision, recall, and F1 score.

The findings indicate that picking the appropriate subset of characteristics can considerably enhance model accuracy by minimizing irrelevant features. Of the four models tested, LightGBM, XGBoost, CatBoost, and Random Forest showed the best performance in detecting phishing URLs with more than 80% accuracy. This research demonstrates that combining feature subset ranking and a lightweight model can be an effective option for real-time phishing detection. This study is designed to serve as a resource for cybersecurity system makers seeking to increase defense against phishing attempts.

Keyword: Phising, URL Detection, Machine Learning, Feature Subset Ranking, Lightweight Models.