IMPLEMENTASI DETEKSI PENIPUAN TAUTAN PALSU SECARA LANGSUNG MENGGUNAKAN LOGISTIC REGRESSION PADA CHROME EXTENSION

PROPOSAL PENELITIAN

SKRIPSI

Dosen Pembimbing	**:**	**Dr. Evta Indra, S.Kom., M.Kom.**
Ketua Peneliti	**:**	**Wahyu Soekanta Ginting (223303040368)**
Anggota Peneliti 1	**:**	**Katrin Wijaya (223303040395)**
Anggota Peneliti 2	**:**	**Nuragustyani br Bangun (223303040619)**

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS PRIMA INDONESIA

MEDAN

2025

Road map : Eksperimen, *Machine Learning*

Penelitian ini bertujuan mengembangkan sistem deteksi penipuan tautan palsu *(phishing)* berbasis Tautan *(URL Only)* langsung yang mampu bekerja secara langsung *(real-time)* dan memberikan peringatan instan kepada pengguna ketika situs yang diakses terindikasi berbahaya. Dengan memanfaatkan algoritma *Logistic Regression*, model yang dikembangkan diharapkan mencapai tingkat akurasi tinggi dalam membedakan tautan phishing dan tautan sahih, sekaligus memiliki efisiensi komputasi tinggi sehingga dapat dijalankan dengan ringan di sisi klien. Karakteristik ini menjadikan sistem sangat relevan untuk diimplementasikan dalam bentuk ekstensi peramban, khususnya *Chrome Extension*, sebagai lapisan perlindungan tambahan bagi pengguna internet. Sistem ini juga dirancang agar mudah diintegrasikan dan dioperasikan tanpa memerlukan sumber daya besar, menjadikannya solusi praktis dan aplikatif untuk deteksi tautan palsu secara langsung.

Tahun	2024	2025	2026	2027
Machine Learning				
Penelitian	Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks [5]	Phishing URL Attack Detection using Logistic Regression and Convolutional Neural Network [7]	Implementasi Real-Time Deteksi Phishing URL Only Menggunakan Logistic Regression Pada Chrome Extension	Sistem Deteksi Phishing Adaptif Berbasis Hybrid Machine Learning dan Data Real-Time URLhaus pada Ekstensi Multi-Browser

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kemajuan teknologi digital telah membawa perubahan signifikan dalam berbagai aspek kehidupan, mulai dari komunikasi, transaksi keuangan, hingga layanan publik. Internet kini menjadi kebutuhan utama yang mendukung efisiensi tinggi dalam aktivitas sehari-hari. Namun, di balik kemudahan tersebut, terdapat ancaman keamanan siber yang dapat menimbulkan kerugian finansial maupun kebocoran privasi. Salah satu ancaman yang paling sering terjadi adalah penipuan tautan palsu *(phising)*, yaitu upaya untuk meniru laman web resmi untuk mencuri informasi sensitif seperti kredensial akun atau data keuangan[1].

Kasus *phising* terus meningkat seiring dengan pertumbuhan pengguna internet. Dampaknya tidak hanya dirasakan oleh individu, tetapi juga institusi bisnis dan pemerintahan. Berbagai laporan menunjukkan peningkatan signifikan dalam jumlah kerugian akibat serangan siber setiap tahun, termasuk kebocoran data pribadi dan penyalahgunaan identitas. Metode deteksi tradisional seperti *blacklist* sering kali gagal dalam menghadapi serangan terbaru *(zero-day attack)*, karena teknik penyerangan laman web berkembang cepat dan menghasilkan variasi tautan berbahaya yang belum tercatat dalam basis data keamanan [2], [3]. Oleh karena itu, diperlukan sistem deteksi yang adaptif, efisien, dan mampu bekerja secara langsung untuk memberikan perlindungan proaktif bagi pengguna [4].

Berbagai penelitian sebelumnya telah menerapkan pendekatan *machine learning* untuk mendeteksi *phising*. Model berbasis algoritma kompleks seperti *Convolutional Neural Network (CNN)* dan *Deep Neural Network (DNN)* terbukti memiliki akurasi tinggi [5], [6]. Tetapi memerlukan sumber daya komputasi besar dan waktu inferensi yang lama. Sebaliknya, algoritma yang lebih sederhana seperti *Logistic Regression [7]* dapat teteap kompetitif bila dikombinasikan dengan pemilihan fitur yang tepat, khususnya hanya berbasis tautan mencurigakan [8], [9]

Algoritma ini unggul dalam efisiensi, kecepatan, serta kemudahan interpretasi, sehingga cocok untuk implementasi di sisi klien [10], [11]. Berdasarkan latar belakang tersebut, penelitian ini mengusulkan implementasi model deteksi tautan palsu [12] berbasis *Logistic Regression* hanya dengan tautan laman web yang diintegrasikan dalam ekstensi *Google Chrome.* Sistem ini diharapkan mampu memberikan peringatan langsung tanpa bergantung pada pemrosesan server eksternal, sehingga pengguna mendapatkan perlindungan tambahan yang cepat, ringan, dan efisien [13], [14].

1.2 Rumusan Masalah

Penelitian ini dirancang untuk menjawab pertanyaan berikut :

1. Bagaimana merancang dan membangun model deteksi ancaman tautan palsu menggunakan algoritma *Logistic Regression* yang efisin secara komputasi?
2. Bagaimana kinerja model *Logistic Regression* dalam membedakan tautan palsu dan tautan sah berdasarkan metrik evaluasi seperti *Accuracy, Recall, F1 Score, Precision,* dan *ROC-AUC?*
3. Bagaimana mengimplementasikan model tersebut ke dalam ekstensi Chrome agar dapat memberikan peringatan langsung pada pengguna?

1.3 Batasan Masalah

Penelitian ini memiliki beberapa batasan masalah, diantara lain:

1. Penelitian hanya difokuskan pada deteksi penyerangan berbasis fitur tautan palsu tanpa analisis isi halaman web atau elemen visual.
2. Algoritma yang digunakan terbatas pada *Logistic Regression* tanpa perbandingan mendalam dengan algoritma lain.
3. Implementasi sistem hanya mencakup ekstensi *Google Chrome*.

1.4 Tujuan Penelitian

Penelitian ini bertujuan mengembangkan model deteksi *phising* berbasis *Logistic Regression* hanya menggunakan laman tautan yang efisien secara komputasi, sekaligus mengevaluasi kinerjanya menggunakan metrik standar klasifikasi. Selain itu, penelitian ini diarahkan untuk mengimplementasikan model ke dalam ekstensi *Google Chrome* sehingga mampu memberikan peringatan langsung kepada pengguna ketika mengakses laman web berpotensi penipuan.

1.5 Manfaat Penelitian

Penelitian ini memiliki beberapa manfaat, diantara lain:

1. Memberikan kontribusi akademis dalam literatur tentang penerapan *Logistic Regression* secara langsung hanya berdasarkan tautan web.
2. Menyediakan solusi praktis berupa ekstensi *Chrome* yang dapat mendeteksi serangan langsung.
3. Menjadi landasan bagi penelitian lanjutan yang mengintegrasikan algoritma lain atau dataset lokal.
4. Meningkatkan kesadaran masyarakat terhadap pentingnya keamanan siber dan perlindungan data pribadi.

BAB II
METODE PENELITIAN

2.1 Jenis Penelitian

Penelitian ini merupakan penelitian terapan dengan pendekatan eksperimen kuantitatif. Fokus utama penelitian adalah pengujian konsep, perancangan model deteksi serangan siber pada laman web berbasis *Logistic Regression*, serta implementasinya dalam lingkungan nyata. Hasil yang diharapkan berupa ekstensi *Google Chrome* yang mampu mendeteksi dan memperingatkan pengguna terhadap serangan *phising* secara langsung.

2.2 Sumber Data dan Dataset

Data penelitian diperoleh dari dua sumber dataset *PhiUSIIL* dari *UCI Machine Learning Repository [15]* dan koleksi laman berbahaya dari *URLhaus* (https://urlhaus.abuse.ch/api/). Kombinasi kedua sumber ini menghasilkan dataset yang representatif dan mutakhir. Proses pengolahan data meliputi normalisasi format, penghapusan duplikasi, pembersihan data tidak valid, serta penyelarasan fitur. Untuk mengatasi ketidakseimbangan kelas, digunakan teknik *sampling* terkontrol seperti *oversampling* dan *undersampling*.

Dataset *PhiUSIIL* digunakan sebagai data pelatihan dasar, sementara *URLhaus* digunakan sebagai data pengujian untuk menilai kemampuan model dalam mengenali pola serangan terbaru.

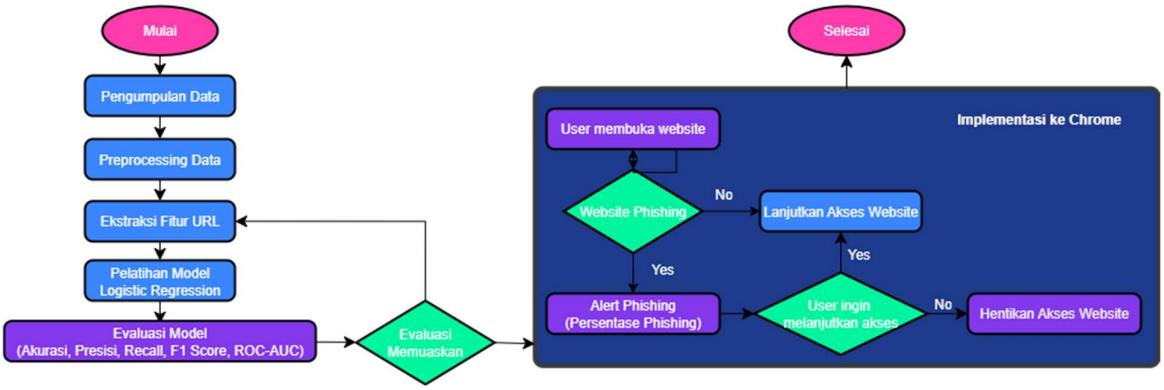
2.3 Instrumen Penelitian

Instrumen yang digunakan dalam penelitian ini meliputi:

1. Menggunakan perangkat lunak *Python* (versi 3.x) dengan *library Pandas, NumPy,* dan *scikit-learn* untuk pemrosesan data, pelatihan model dan pengembangan ekstensi.
2. Menggunakan perangkat keras laptop atau komputer dengan spesifikasi standar penelitian.
3. Alat pengumpulan data yang digunakan adalah Skrip *Python* untuk mengakses *API URLhaus* dan memuat dataset *PhiUSIIL*.
4. Manajemen kode *Github* untuk kontrol versi serta kolaborasi pengembangan.
5. Mengevaluasi model dengan *Confusion Matrix,* laporan klasifikasi,dan kurva *ROC* yang dihasilkan menggunakan pustaka *scikit-learn*.

2.4 Alur Penelitian

Diawali dengan tahap pengumpulan data dari dua sumber utama, yakni dataset *PhiUSIIL* dari *UCI Machine Learning Repository* dan *URLhaus API* sebagai penyedia data tautan berbahaya terkini. Data yang diperoleh kemudian melalui proses pembersihan dan penyeragaman agar layak digunakan pada tahap berikutnya. Tahap selanjutnya adalah ekstraksi fitur tautan yang berfokus pada karakteristik seperti panjang alamat, struktur domain, keberadaan parameter, serta indikasi penggunaan kata yang mencurigakan. Hasil ekstraksi ini dimanfaatkan dalam pelatihan model menggunakan algoritma *Logistic Regression* untuk mengklasifikasikan tautan phishing dan tautan sahih secara efisien. Model yang telah terbentuk dievaluasi dengan menggunakan ukuran performa berupa akurasi, presisi, *recall, F1-score,* dan *ROC-AUC* guna menilai tingkat keandalannya. Setelah memenuhi kriteria performa yang diharapkan, model diimplementasikan ke dalam ekstensi *Chrome* yang mampu bekerja secara langsung untuk mendeteksi dan memberikan peringatan kepada pengguna ketika mengakses situs yang terindikasi phishing, sekaligus memastikan keamanan saat pengguna mengunjungi situs yang valid.



Gambar 1.1 Alur Penelitian

2.5 Ekstraksi Fitur

Ekstraksi fitur dilakukan langsung dari struktur tautan tanpa memuat isi halaman. Fitur utama mencakup panjang tautan, jumlah *subdomain,* panjang *domain*, jumlah parameter *query*, penggunaan alamat IP, status HTTPS, kedalaman *path*, serta token mencurigakan seperti *"login"* atau *"secure"*. Fitur numerik dinormalisasi dan fitur kategorikal dikonversi ke bentuk numerik melalui proses *encoding* agar sesuai dengan model. Seleksi fitur dilakukan menggunakan *mutual information* dan regularisasi *L1* untuk mengurangi kolinearitas.

2.6 Evaluasi

Dilakukan melalui dua tahap, statistik dan implementatif. Evaluasi statistik menggunakan *k-fold cross-validation* dengan metrik *Accuracy, Precission, Recall, F1-Score,* dan *ROC-AUC*. Evaluasi implementatif menilai kecepatan inferensi, penggunaan memori, serta tingkat *false positive* pada ekstensi *Chrome*. Kriteria keberhasilan penelitian ditentukan berdasarkan nilai *AUC* yang tinggi, *recall* yang optimal, dan beban komputasi yang rendah.

DAFTAR PUSTAKA

[1] A. Almomani *et al.*, "Phishing Website Detection With Semantic Features Based on Machine Learning Classifiers," *Int J Semant Web Inf Syst*, vol. 18, no. 1, pp. 1–24, Feb. 2022, doi: 10.4018/IJSWIS.297032.

[2] F. Nourmohammadzadeh Motlagh, C. Meinel, M. Hajizadeh, M. Majd, P. Najafi, and F. Cheng, "Large Language Models in Cybersecurity: State-of-the-Art," *Proceedings of ACM Conference (Conference'17)*, vol. 1, doi: 10.48550/arXiv.2402.00891.

[3] R. MRM, N. F.A.M, S. A.M, and S. K.A.S, "A Comparative Analysis of Machine Learning Models for URL-Based Phishing Detection," Apr. 15, 2025. doi: 10.21203/rs.3.rs-6439154/v1.

[4] Q. E. ul Haq, M. H. Faheem, and I. Ahmad, "Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks," *Applied Sciences*, vol. 14, no. 22, p. 10086, Nov. 2024, doi: 10.3390/app142210086.

[5] A. John-Otumu, V. O. Aniugo, and V. C. Nwachukwu, "HyRANN-UPD: Enhancing Phishing URL Detection Using Ridge Regression-Based Feature Selection and Artificial Neural Networks," *Int J Comput Appl*, vol. 186, no. 78, pp. 56–62, Apr. 2025, doi: 10.5120/ijca2025924689.

[6] U. Daniel, E. Bartholomew, and F. Egbono, "Phishing URL Attack Detection using Logistic Regression and Convolutional Neural Network," 2025. [Online]. Available: https://www.kaggle.com/datasets/shashwatwork/phishing-

[7] H. Gonaygunta, "MACHINE LEARNING ALGORITHMS FOR DETECTION OF CYBER THREATS USING LOGISTIC REGRESSION," *International Journal of Smart Sensor and Adhoc Network.*, pp. 36–42, Jan. 2023, doi: 10.47893/IJSSAN.2023.1229.

[8] M. A. Tamal, M. K. Islam, T. Bhuiyan, and A. Sattar, "Dataset of suspicious phishing URL detection," *Front Comput Sci*, vol. 6, Mar. 2024, doi: 10.3389/fcomp.2024.1308634.

[9] S. Aghera and N. Y. Joshi, "Article in International Journal of Intelligent Systems and Applications in Engineering · May 2024 www.ijisae.org Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE," 2024. [Online]. Available: https://www.researchgate.net/publication/383984494

[10] M. Sanchez-Paniagua, E. F. Fernandez, E. Alegre, W. Al-Nabki, and V. Gonzalez-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," *IEEE Access*, vol. 10, pp. 42949–42960, 2022, doi: 10.1109/ACCESS.2022.3168681.

[11] D. Kalla and S. Kuraku, "Phishing Website URL's Detection Using NLP and Machine Learning Techniques," *Journal on Artificial Intelligence*, vol. 5, no. 0, pp. 145–162, 2023, doi: 10.32604/jai.2023.043366.

[12] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.004.

[13] M. Sannigrahi and R. Thandeeswaran, "Predictive Analysis of Network-Based Attacks by Hybrid Machine Learning Algorithms Utilizing Bayesian Optimization, Logistic Regression, and Random Forest Algorithm," *IEEE Access*, vol. 12, pp. 142721–142732, 2024, doi: 10.1109/ACCESS.2024.3464866.

[14] M. Maad M., S. Israa Ezzat, and I. Marwa M., "The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review," *Iraqi Journal for Computer Science and Mathematics*, pp. 87–101, Jan. 2023, doi: 10.52866/ijcsm.2023.01.01.008.

[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.