Journal of Computer Science and Computational Intelligence
April 2025 | Vol. 1 | No.1 | Page 0036-0040

Deteksi Serangan *Web Defacement* pada Infrastruktur Kritis Menggunakan *Machine learning*

Victor Eric Pattiradjawane[1*], Doms Upuy[1]

[1]Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Pattimura
Jl. Ir. M. Putuhena, Ambon, 97233, Indonesia
[*]Corresponding author's e-mail: *** victor.pattiradjawane@lecturer.unpatti.ac.id**

Abstract

The number of threats to the security of websites and web servers, which include things like Web Defacement, is increasing every year. This is a major concern in today's cybersecurity world. It includes websites that are part of critical infrastructure, like government, health, and energy systems. This research study looks at how machine learning (ML) models can automatically detect Web Defacement attacks. The main goal is to make sure these models are very accurate. We developed a supervised learning-based classification model using Web Defacement datasets from public archives and simulated mock sites. This research study looks at how well three types of classification models—Random Forest, support vector machine (SVM), and naive Bayes—perform at identifying defaced websites. The results of the experiment show that Random Forest is the best option, with up to 96% accuracy. This research shows that the Machine learning (ML) approach could be very important in developing a system that can detect cyberattacks early on. This system would protect important infrastructure in the country.

Keywords: *Anomaly Detection, Critical Infrastructure, Cybersecurity, Machine learning, Web Defacement.*

1. PENDAHULUAN

1.1. Latar Belakang

Dalam beberapa tahun terakhir, insiden keamanan siber mengalami peningkatan signifikan, baik dari segi frekuensi, kompleksitas, maupun target serangan. Salah satu bentuk serangan yang umum namun berdampak besar adalah *Web Defacement* [1]. Serangan ini mengubah tampilan atau isi halaman web dengan konten yang tidak sah, sering kali bersifat provokatif atau merusak citra organisasi. Ketika situs yang diserang termasuk dalam infrastruktur kritis seperti pemerintahan, kesehatan, dan energi, konsekuensinya mencakup gangguan layanan vital, kerugian ekonomi, serta penurunan kepercayaan publik.

Deteksi serangan *Web Defacement* secara manual atau berbasis signature sering kali tidak mampu mengimbangi kecepatan dan variasi serangan. Untuk itu, pendekatan *machine learning* (ML) muncul sebagai solusi yang adaptif dan cerdas. Model ML dapat dilatih untuk mengenali anomali pada konten halaman web berdasarkan pola-pola historis, sehingga mampu mendeteksi serangan yang sebelumnya belum dikenal *(zero-day attack)*.

Penelitian ini bertujuan menerapkan dan mengevaluasi performa beberapa model atau algoritma ML dalam mendeteksi halaman web yang telah di-deface. Hasil dari penelitian ini diharapkan dapat mendukung pengembangan sistem deteksi dini berbasis ML yang efisien dan akurat untuk perlindungan infrastruktur kritis di Indonesia.

2. TINJAUAN PUSTAKA

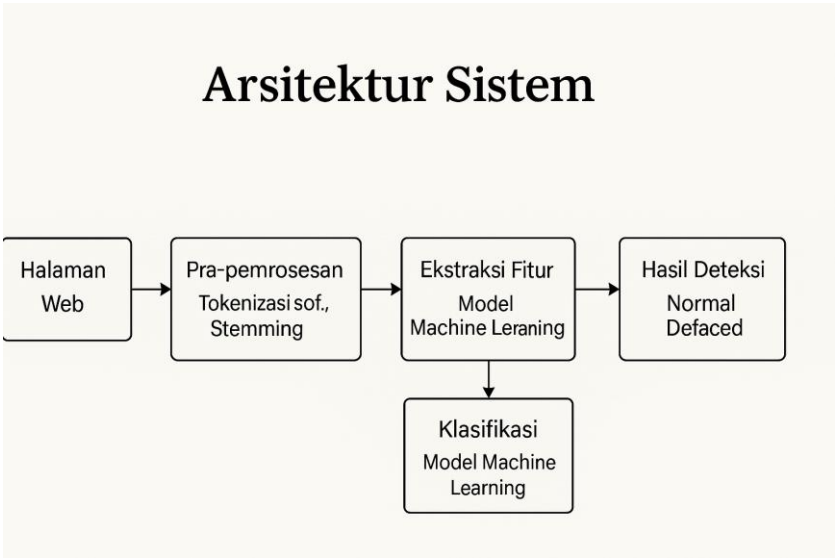
Berbagai penelitian telah dilakukan terkait deteksi serangan *Web Defacement*. Metode berbasis signature tidak mampu mendeteksi serangan baru secara efektif [2]. Pendekatan visual berbasis CNN telah digunakan untuk mendeteksi perubahan tampilan halaman web [3], sedangkan perbandingan algoritma klasifikasi menunjukkan efektivitas yang beragam dalam mendeteksi web berbahaya [4].

Metode deteksi anomali relevan untuk diterapkan dalam konteks *defacement* [5], [6]. Ekstraksi fitur HTML juga dianggap penting dalam membedakan konten sah dan konten yang telah diubah [7]. Pendekatan *hybrid* yang menggabungkan fitur log dan konten juga telah diusulkan untuk mendeteksi serangan web [8].

Pentingnya kualitas data dalam pelatihan model ML untuk keamanan siber telah banyak disorot [9]. Model *ensemble* seperti *Random Forest* dan *Gradient Boosting* menunjukkan efektivitas dalam mendeteksi intrusi [10]. Pendekatan berbasis AI dikembangkan untuk menghasilkan *threat intelligence* secara otomatis [11].

3. METODOLOGI PENELITIAN

Penelitian ini menerapkan pendekatan berbasis supervised *machine learning* untuk mendeteksi serangan *Web Defacement* pada infrastruktur kritis. Metodologi terdiri dari beberapa tahapan sistematis mulai dari akuisisi data, pra-pemrosesan, ekstraksi fitur, pelatihan model, hingga evaluasi kinerja. Diagram arsitektur sistem ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur sistem deteksi serangan *Web Defacement*

3.1. Akuisisi dan Klasifikasi Dataset

Dataset yang digunakan terdiri dari dua kategori utama yaitu halaman web normal dan halaman web yang telah mengalami *defacement*. Sumber data diperoleh dari arsip zone-h berupa situs-situs publik yang terdampak *defacement* (https://www.zone-h.org/archive), dan simulasi *defacement* lokal berdasarkan skenario umum seperti penggantian teks atau penyisipan pesan politik. Setiap halaman web disimpan dalam bentuk HTML atau plain text dan dilabeli sesuai kategorinya. Total data yang digunakan adalah sebanyak 100 sampel, dengan distribusi seimbang antara dua kelas.

3.2. Pra-Pemrosesan Data

Langkah ini bertujuan untuk membersihkan dan normalisasi data sebelum digunakan dalam model pembelajaran mesin. Tahapan pra-pemrosesan meliputi: 1) Tokenisasi: Memecah dokumen menjadi unit kata (tokens) ; 2) Stopword Removal: Menghapus kata-kata umum yang tidak memberikan nilai informatif (contoh: “yang”, “dan”, “dengan”) ; 3) Stemming: Mengubah kata ke bentuk dasarnya untuk menyatukan variasi morfologis. 4) Lowercasing: Menstandarkan semua kata menjadi huruf kecil. Setelah proses ini, setiap dokumen diwakili oleh kumpulan kata yang siap untuk diekstraksi fiturnya.

3.3. Ekstraksi Fitur

Fitur teks diekstraksi menggunakan teknik TF-IDF (Term Frequency–Inverse Document Frequency). TF-IDF mengukur pentingnya suatu kata dalam dokumen relatif terhadap kumpulan dokumen. TF mencerminkan frekuensi kata dalam dokumen, sedangkan IDF memberikan bobot lebih tinggi pada kata yang jarang muncul di seluruh dokumen. Output dari tahap ini adalah vektor fitur numerik berdimensi tinggi yang dapat digunakan oleh algoritma klasifikasi.

3.4. Pelatihan Model *Machine learning*

Tiga algoritma *machine learning* diuji dan dibandingkan dalam penelitian ini, yaitu: *Random Forest* (RF): algoritma ensemble berbasis pohon keputusan yang membentuk beberapa pohon acak dan menggabungkan hasilnya untuk klasifikasi akhir; Support Vector Machine (SVM): model berbasis margin maksimal yang mencoba menemukan hyperplane terbaik untuk memisahkan dua kelas; dan Naive Bayes (NB): pendekatan probabilistik sederhana berdasarkan Teorema Bayes dengan asumsi independensi antar fitur. Setiap model dilatih dengan data TF-IDF dan dikonfigurasi menggunakan teknik validasi silang 10-fold untuk menghindari overfitting dan mendapatkan hasil generalisasi yang akurat.

3.5. Evaluasi Kinerja Model

Model yang telah dilatih dievaluasi menggunakan metrik evaluasi sebagai berikut: 1) Akurasi: Persentase klasifikasi yang benar terhadap total data. 2) Precision: Kemampuan model dalam mengklasifikasikan data positif secara benar. 3) Recall (Sensitivity): Kemampuan model untuk menangkap seluruh data positif. 4) F1-score: Harmonik rata-rata antara precision dan recall. Evaluasi dilakukan menggunakan confusion matrix untuk memahami distribusi prediksi model secara lebih mendalam.

4. Hasil dan Pembahasan

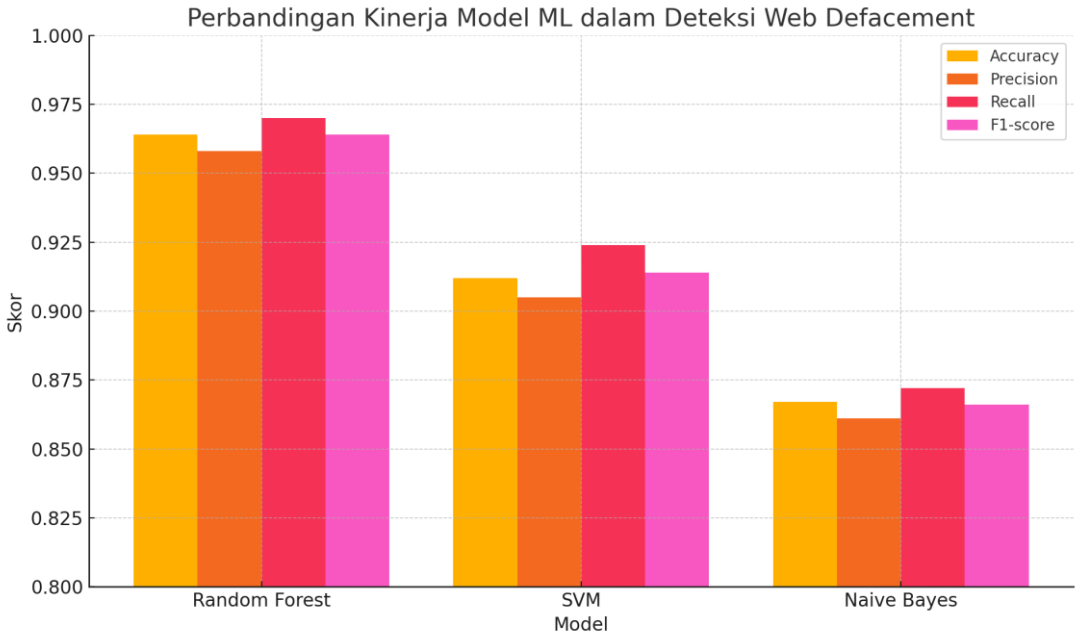
Hasil eksperimen menunjukkan bahwa model *Random Forest* unggul dengan akurasi sebesar 96%, precision 95%, recall 97%, dan nilai ROC-AUC sebesar 0.97. Model SVM memiliki

akurasi 91%, sedangkan Naive Bayes sebesar 87% (Tabel.1). *Random Forest* terbukti lebih stabil terhadap variasi struktur konten HTML. (Gambar.2)

Tabel 1. Hasil evaluasi model berdasarkan empat metrik utama: akurasi, precision, recall, dan F1-score

Algoritma	Akurasi	Precision	Recall	F1-score
Random Forest	0.964	0.958	0.970	0.964
Support Vector Machine	0.912	0.905	0.924	0.914
Naive Bayes	0.867	0.861	0.872	0.866

Penelitian ini menunjukkan bahwa pendekatan *machine learning*, khususnya *Random Forest*, mampu mengidentifikasi serangan *Web Defacement* dengan tingkat akurasi yang sangat tinggi. Kombinasi dari fitur TF-IDF dan arsitektur memberikan keunggulan dalam mendeteksi perubahan anomali konten web. Selain itu, pendekatan ini bersifat generik dan dapat diadaptasi untuk berbagai bahasa atau struktur HTML, sehingga menjadikannya fleksibel untuk diintegrasikan ke dalam sistem keamanan *real-time*.



Gambar 2. Perbandingan visual dari hasil evaluasi

KESIMPULAN DAN SARAN

Penelitian ini membuktikan bahwa penerapan *machine learning*, khususnya algoritma *Random Forest*, efektif dalam mendeteksi serangan *Web Defacement*. Model ini memiliki potensi untuk diimplementasikan dalam sistem deteksi dini pada infrastruktur kritis nasional. Uji coba pada situs dummy yang meniru halaman pemerintahan menunjukkan model algoritma *Random Forest* (RF) mampu mendeteksi *defacement* dengan latensi rendah. Namun, tantangan muncul

pada halaman dengan konten dinamis atau berita yang mengandung kata-kata sensitif yang menyerupai konten *defacement*.

Saran untuk penelitian selanjutnya adalah menggabungkan analisis konten visual dan metadata halaman serta menerapkan pendekatan real-time berbasis *edge computing* dengan jumlah data yang lebih besar. Perlu juga dilakukan pengujian pada sistem nyata untuk mengevaluasi performa model dalam kondisi operasional sesungguhnya.

REFERENSI

[1] https://www.bssn.go.id/wp-content/uploads/2024/03/Lanskap-Keamanan-Siber-Indonesia-2023.pdf

[2] M. Hassen et al., '*Web Defacement* Detection Using Signature-Based Methods', International Journal of Cyber Security, 2019.

[3] J. Kim et al., 'Image-based Deep Learning for *Defacement* Detection', IEEE CyberSec Conference, 2020.

[4] R. Ahmad et al., 'Comparative Study of *Machine learning* Algorithms', Journal of Information Security, 2021.

[5] V. Chandola et al., 'Anomaly Detection: A Survey', ACM Computing Surveys, vol. 41, no. 3, 2009.

[6] A. Patcha and J. Park, 'An Overview of Anomaly Detection Techniques', Computer Networks, vol. 51, no. 12, pp. 3448–3470, 2007.

[7] G. Vrbančič et al., 'HTML Content Feature Extraction for Web Security', Information Sciences, vol. 484, 2019.

[8] Y. Zhang et al., 'Hybrid ML Approaches for Web Attack Detection', Computers & Security, vol. 74, 2018.

[9] R. Sommer and V. Paxson, 'Outside the Closed World: On Using ML for Security', IEEE S&P, 2010.

[10] R. Vinayakumar et al., 'Deep Learning Approaches for Cybersecurity', Future Generation Computer Systems, 2019.

[11] M. Alazab et al., 'AI-Driven Cyber Threat Intelligence', IEEE Access, 2020.