

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan pesat teknologi informasi telah mendorong peningkatan penggunaan internet dalam berbagai aspek kehidupan, mulai dari transaksi digital, komunikasi, hingga layanan berbasis web. Namun, pertumbuhan ini juga diikuti oleh meningkatnya ancaman siber, seperti phishing yakni teknik penipuan yang dirancang untuk mencuri informasi sensitif pengguna melalui situs web palsu yang menyerupai situs resmi [1]. Phishing biasanya dilakukan dengan menyebarkan URL yang tampak sah, sehingga pengguna yang kurang berhati-hati sering terjebak membuka situs berbahaya tersebut dan memasukkan data pribadi mereka.

Metode konvensional seperti blacklist dan aturan heuristic umumnya hanya efektif terhadap URL phishing yang sudah dikenal, dan cenderung gagal menangani variasi baru dari URL yang terus berubah. Hal ini dikarenakan metode tersebut hanya memeriksa kecocokan dengan pola tertentu atau daftar URL berbahaya, tanpa memahami karakteristik inheren dari struktur URL itu sendiri [2]. Oleh karena itu, diperlukan pendekatan yang lebih adaptif yang mampu mengenali pola abstrak dalam struktur URL untuk mendeteksi phishing secara otomatis.

Pendekatan machine learning (ML) telah terbukti efektif dalam tugas klasifikasi data kompleks seperti deteksi phishing karena kemampuannya dalam belajar dari data historis dan memprediksi pola yang belum pernah terlihat sebelumnya [3]. Dalam konteks URL phishing, fitur-fitur struktural seperti panjang URL, jumlah subdomain, penggunaan karakter khusus, serta kedalaman jalur path merupakan informasi yang sangat berguna bagi model ML untuk membedakan antara URL phishing dan URL legitimate [1], [3]. Analisis terhadap fitur-fitur ini memungkinkan pembuatan detektor phishing yang lebih akurat dibandingkan metode berbasis aturan tradisional.

Berbagai penelitian telah mengevaluasi efektivitas algoritma machine learning dalam deteksi URL phishing. Penelitian oleh Foozy dkk. [1] menunjukkan bahwa model ensemble seperti Random Forest dan LightGBM menghasilkan

performa lebih baik dibandingkan model sederhana seperti Naive Bayes dan Logistic Regression ketika dievaluasi terhadap dataset URL phishing yang luas. Hasil tersebut menunjukkan bahwa teknik ensemble learning dan boosting memberikan keunggulan dalam menangani fitur kompleks pada URL.

Selain itu, hasil studi Albishri & Dessouky [4] juga memperkuat temuan tersebut dengan menunjukkan bahwa ensemble methods seperti Gradient Boosting dan XGBoost mampu mencapai akurasi tinggi dalam tugas deteksi phishing dibandingkan dengan model klasik lainnya. Model-model ini secara konsisten menunjukkan performa stabil dan kuat, terutama ketika dilatih dan diuji menggunakan validasi silang (cross validation) untuk mengurangi bias pembagian data.

Penelitian oleh Kautsar, Aida, & Yulistia [5] melaporkan bahwa Random Forest sangat efektif dalam mengidentifikasi URL phishing dengan mempertimbangkan fitur seperti rata-rata panjang URL, jumlah angka, serta karakter spesial. Hasil ini sejalan dengan temuan lain yang menekankan bahwa pendekatan feature-based learning dapat memberikan performa terbaik untuk deteksi phishing yang adaptif.

Studi oleh Ulum, Taryo, & Sudarno [6] menunjukkan bahwa algoritma boosting seperti XGBoost dan LightGBM unggul dalam hal metrik AUC-ROC serta F1-Score dibandingkan beberapa algoritma lain ketika diaplikasikan pada dataset phishing yang telah dilabeli. Penelitian tersebut juga menekankan bahwa metode gradient boosting memiliki kemampuan menangkap pola interaksi fitur yang tidak linier, yang sering kali terjadi pada URL phishing kompleks.

Studi lain oleh Sukant dkk. [7] menjelaskan deteksi URL berbahaya dalam kasus phishing yang termasking, yaitu ketika domain phishing dirancang sedemikian rupa agar terlihat legit dari segi visual. Penelitian ini menunjukkan bahwa algoritma seperti Random Forest dan XGBoost masih mampu mendeteksi pola yang membedakan URL phishing dari URL legitimate meskipun ada strategi penyamaran yang rumit (masked phishing).

Dengan adanya berbagai bukti empiris tersebut, terlihat bahwa meskipun sejumlah algoritma machine learning sudah dievaluasi untuk deteksi phishing URL, kombinasi dan perbandingan tiga algoritma, yaitu *XGBoost*, *LightGBM*, dan *Random Forest*, dalam satu kerangka analisis yang sama masih jarang ditemukan pada literatur saat ini. Oleh karena itu, penelitian ini akan mengembangkan model deteksi phishing berbasis *Machine Learning* yang fokus pada analisis struktur URL sebagai fitur utama, kemudian melakukan perbandingan performa algoritma tersebut untuk menentukan algoritma yang paling optimal berdasarkan metrik klasifikasi seperti *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *AUC-ROC*.

1.2 Rumusan Masalah

Website phishing menjadi ancaman serius bagi keamanan informasi karena bentuk rekayasanya yang semakin sulit dikenali secara manual oleh pengguna awam. Di sisi lain, struktur URL sebenarnya memiliki pola dan karakteristik leksikal spesifik yang dapat digunakan untuk membedakan antara website phishing dan legitimate. Namun, proses identifikasi dan analisis pola tersebut secara manual sangat tidak efisien dan rentan terhadap kesalahan. Hal ini memunculkan tantangan komputasi mengenai bagaimana menganalisis karakteristik fitur struktural URL yang telah diformulasikan dalam dataset untuk membedakan website phishing dan legitimate? dan bagaimana perbandingan kinerja *XGBoost*, *LightGBM*, dan *Random Forest* pada dataset tersebut?

1.3 Batasan Masalah

Agar penelitian ini lebih terfokus dan terarah serta menghindari pembahasan yang terlalu luas, maka diberikan beberapa batasan masalah sebagai berikut:

1. Objek penelitian pada skripsi ini dibatasi pada deteksi website phishing berdasarkan analisis struktur URL, tanpa melakukan analisis terhadap konten halaman web, *source code HTML*, atau perilaku pengguna.
2. Dataset yang digunakan berupa kumpulan URL yang telah diberi label phishing dan legitimate, yang bersumber dari dataset publik dan telah digunakan dalam penelitian terdahulu.

3. Fitur yang digunakan dalam penelitian ini bersumber dari dataset publik Phishing Websites, di mana terminologi "Analisis Struktur URL" pada penelitian ini mencakup representasi 30 atribut fitur yang secara eksplisit dikategorikan ke dalam tiga kelompok: fitur leksikal/struktural URL murni (seperti panjang URL dan karakter khusus), fitur keamanan protokol (seperti validitas SSL), serta fitur reputasi domain eksternal (seperti PageRank dan Google Index). Penelitian ini tidak melakukan web scraping atau ekstraksi mandiri terhadap konten antarmuka maupun source code HTML halaman web.
4. Algoritma machine learning yang digunakan dalam penelitian ini dibatasi pada Algoritma *XGBoost*, *LightGBM* dan *Random Forest*.
5. Metode evaluasi performa model dibatasi pada metrik evaluasi klasifikasi, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *AUC-ROC*, serta menggunakan metode data split dan *K-Fold Cross Validation*.
6. Lingkungan pengembangan dan pengujian model dilakukan menggunakan bahasa pemrograman *Python* dengan pustaka machine learning yang tersedia, serta dijalankan pada platform *Google Colab*.
7. Penelitian ini tidak membahas integrasi model ke dalam infrastruktur keamanan jaringan skala produksi, meskipun prototipe antarmuka berbasis web telah dikembangkan sebagai bukti konsep (proof of concept).

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah dan batasan masalah yang telah dijelaskan sebelumnya, Tujuan yang akan dicapai dalam penelitian ini adalah sebagai berikut:

1. Menganalisis fitur-fitur struktural URL yang telah terencode di dalam dataset sekunder untuk mengklasifikasikan antara website phishing dan website legitimate.
2. Menerapkan tiga algoritma machine learning dalam proses deteksi website phishing berdasarkan analisis struktur URL, yaitu *XGBoost*, *LightGBM*, dan *Random Forest*.

3. Mengevaluasi performa masing-masing algoritma machine learning dalam mendeteksi website phishing menggunakan metrik evaluasi klasifikasi, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *AUC-ROC*.
4. Membandingkan kinerja algoritma machine learning berdasarkan hasil pengujian menggunakan metode data split dan K-Fold Cross Validation untuk memperoleh evaluasi yang lebih stabil dan objektif.
5. Menentukan algoritma machine learning terbaik dalam mendeteksi website phishing berdasarkan analisis struktur URL dengan mempertimbangkan hasil evaluasi dan perbandingan performa antar algoritma.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian “Deteksi Website Phishing Berdasarkan Analisis Struktur URL Menggunakan Algoritma Machine Learning” adalah sebagai berikut:

Penelitian ini diharapkan dapat menambah referensi ilmiah dan memperluas wawasan akademis di bidang keamanan siber, khususnya pada topik deteksi *website phishing* berbasis *machine learning*. Hasil penelitian ini juga diproyeksikan dapat menjadi bahan rujukan yang relevan bagi studi selanjutnya yang mengeksplorasi analisis struktur URL maupun pengembangan metode klasifikasi lainnya. Secara lebih spesifik, penelitian ini akan memberikan gambaran komparatif yang komprehensif mengenai performa berbagai algoritma klasifikasi, seperti *XGBoost*, *LightGBM*, dan *Random Forest*, dalam mendeteksi anomali pada *website*.

Secara praktis, penelitian ini menawarkan solusi alternatif yang efektif dalam mendeteksi *website phishing* secara otomatis dan akurat melalui analisis struktur URL. Implementasi dari temuan ini diharapkan mampu membantu pengguna internet dalam meminimalisir risiko penipuan digital melalui peningkatan sistem deteksi dini terhadap ancaman *website* berbahaya. Lebih jauh lagi, hasil penelitian ini memiliki potensi besar untuk dijadikan landasan dalam pengembangan sistem keamanan web tingkat lanjut atau aplikasi pendukung yang dapat diintegrasikan langsung ke dalam *browser* maupun infrastruktur keamanan

jaringan.

Bagi peneliti, penyusunan skripsi ini memberikan manfaat yang signifikan dalam meningkatkan pemahaman teoritis dan keterampilan teknis terkait penerapan algoritma *machine learning* pada studi kasus nyata di bidang keamanan siber. Selama proses pengerjaan, peneliti mendapatkan pengalaman langsung yang komprehensif, mulai dari tahap analisis data dan eksperimen pemodelan, hingga evaluasi performa dan interpretasi hasil secara ilmiah. Pada akhirnya, penelitian ini menjadi sarana esensial untuk mematangkan kompetensi akademik dan teknis guna mendukung penyelesaian studi, serta menjadi bekal yang kuat untuk eksplorasi lebih lanjut di *ranah data science dan machine learning*.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran yang jelas mengenai alur pembahasan pada setiap bab. Adapun sistematika penulisan dalam skripsi ini adalah sebagai berikut :

BAB I Pendahuluan, Bab ini berisi gambaran umum mengenai penelitian yang dilakukan, meliputi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan. Bab ini bertujuan untuk memberikan pemahaman awal mengenai alasan dan arah penelitian.

BAB II Tinjauan Pustaka, Bab ini membahas teori dan konsep yang mendukung penelitian, meliputi konsep phishing dan website phishing, struktur URL, machine learning, serta penjelasan mengenai algoritma yang digunakan, yaitu *XGBoost*, *LightGBM*, dan *Random Forest*. Selain itu, bab ini juga memuat penelitian terkait yang relevan sebagai dasar pengembangan penelitian.

BAB III Metodologi Penelitian, Bab ini menjelaskan tahapan penelitian yang dilakukan, mulai dari pengumpulan dan preprocessing dataset, ekstraksi fitur struktur URL, penerapan algoritma machine learning, serta metode evaluasi model yang digunakan, seperti data split dan K-Fold Cross Validation. Alur penelitian disajikan secara sistematis untuk menggambarkan proses eksperimen secara

keseluruhan.

BAB IV Hasil dan Pembahasan, Bab ini menyajikan hasil eksperimen dan pengujian model menggunakan masing-masing algoritma machine learning. Pembahasan meliputi hasil evaluasi performa berdasarkan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *AUC-ROC*, analisis *learning curve*, *confusion matrix*, serta perbandingan performa antar algoritma untuk menentukan model terbaik dalam mendeteksi website phishing.

BAB V Kesimpulan dan Saran, Bab ini berisi kesimpulan yang diperoleh dari hasil penelitian serta saran untuk pengembangan penelitian selanjutnya, baik dari sisi metode, fitur, maupun penerapan sistem deteksi phishing di masa mendatang.

