BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang pesat pada saat ini juga diiringi dengan meningkatnya ancaman kejahatan siber, di mana internet menjadi media utama pertukaran informasi global sekaligus celah bagi *cybercrime*[1]. Salah satu pendekatan teknologi yang kini menjadi andalan dalam memperkuat keamanan siber adalah *Machine Learning*. *Machine Learning* memungkinkan sistem untuk mempelajari pola dari data historis guna mendeteksi ancaman tanpa perlu diprogram secara eksplisit untuk setiap aturan serangan yang baru[2].

Ancaman keamanan siber global menunjukkan tren yang mengkhawatirkan dengan tercatatnya 892.494 serangan phishing unik pada kuartal ketiga tahun 2025. Pada periode ini, sektor SaaS/Webmail menjadi target yang paling dominan dengan menyumbang 21,2% dari keseluruhan serangan. Kondisi tersebut mempertegas bahwa isu utama yang menjadi fokus dalam keamanan siber saat ini adalah serangan *phishing* melalui URL atau situs web[3]. *Phishing* merupakan teknik penipuan yang melibatkan penyamaran sebagai entitas terpercaya untuk mencuri data sensitif seperti username, kata sandi, dan informasi finansial. Serangan ini sering kali memanfaatkan URL palsu yang dirancang semirip mungkin dengan situs aslinya untuk mengelabui korban[4]. Dampak dari serangan ini tidak hanya menyebabkan kerugian privasi, tetapi juga kerugian finansial yang signifikan akibat penyalahgunaan data[5]. Oleh karena itu, diperlukan sistem deteksi otomatis yang mampu membedakan antara URL yang sah *legitimate* dan URL *phishing* dengan akurasi tinggi.

Dalam upaya melakukan deteksi *phishing*, berbagai arsitektur model *Machine Learning* telah diterapkan, seperti Naive Bayes, Random Forest, dan Decision Tree. Penelitian terdahulu menunjukkan bahwa algoritma seperti Random Forest dan Decision Tree cenderung memberikan performa yang lebih unggul dibandingkan metode sederhana seperti Naive Bayes dalam mendeteksi pola URL

berbahaya. Algoritma Naive Bayes, meskipun cepat sering kali memiliki kelemahan dalam akurasi ketika dihadapkan pada fitur data yang kompleks atau tidak seimbang[2].

Di antara berbagai model tersebut, Decision Tree dipilih sebagai model dasar (*base learner*) dalam penelitian ini karena memiliki keunggulan pada karakteristiknya. Decision Tree memiliki kemampuan interpretasi yang baik serta mampu menangani data kategorikal dan numerik tanpa memerlukan normalisasi data yang rumit[2]. Struktur pohon keputusan memungkinkan penelusuran logika deteksi yang transparan, sehingga alasan mengapa sebuah URL diklasifikasikan sebagai *phishing* dapat dijelaskan dengan lebih mudah dibandingkan model lainnya. Selain itu, Decision Tree terbukti memiliki efisiensi komputasi yang baik dengan tingkat akurasi yang kompetitif[6].

Namun penggunaan Decision Tree sebagai model tunggal memiliki kelemahan mendasar, yaitu kecenderungan untuk mengalami *overfitting* yang menyebabkan performanya menurun pada data baru. Untuk mengatasi kelemahan tersebut, penelitian ini mengusulkan pendekatan yang mengintegrasikan teknik pemrosesan fitur lanjutan dengan optimasi algoritma. Pada tahap ekstraksi fitur, peneliti menggunakan ekstrasi fitur leksikal yang berfungsi untuk menganalisis struktur dan karakteristik dari URL, seperti panjang karakter dan keberadaan simbol khusus, mengingat metode ini terbukti efektif dalam menangkap anomali pada alamat website berbahaya. Selanjutnya, untuk menyempurnakan kemampuan model dan meningkatkan performa prediksi, pendekatan ini diperkuat melalui penerapan teknik *ensemble* dengan metode Boosting[2].

Metode Boosting bekerja dengan cara membangun model secara berurutan (*sequential*), di mana model baru difokuskan untuk memperbaiki kesalahan yang dihasilkan oleh model sebelumnya. Algoritma berbasis Boosting seperti Gradient Boosting dan Adaptive Boosting terbukti mampu meningkatkan akurasi secara signifikan dan lebih tangguh terhadap data yang kompleks dibandingkan model pohon tunggal[6][7]. Oleh karena itu, penelitian ini akan menerapkan algoritma Decision Tree yang dioptimalkan dengan teknik Boosting untuk mendeteksi URL

phishing, dengan harapan dapat menghasilkan model yang tidak hanya mudah diinterpretasikan tetapi juga memiliki akurasi yang tinggi.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah:

1. Apa Pengaruh metode Gradient Boosting dan Adaptive Boosting pada kinerja algoritma Decission Tree dalam mendeteksi URL Phising?
2. Bagaimana perbandingan hasil evaluasi antara model Decision Tree tunggal (tanpa boosting) dengan model yang menggunakan metode Boosting dalam mendeteksi URL *phishing*?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan fokus pada tujuan yang ingin dicapai, penulis membatasi ruang lingkup masalah sebagai berikut:

1. Algoritma yang digunakan adalah Decission Tree dengan pengoptimalan menggunakan metode boosting.
2. Pengoptimalan algoritma dilakukan pada tahap training
3. Metode boosting yang digunakan ditahap training Adalah Gradient Boosting dan Adaptive Boosting.
4. Dataset yang digunakan adalah dataset URL yang terdiri dari kelas *phishing* dan *legitimate* yang diperoleh dari repositori publik.
5. Proses ekstraksi fitur difokuskan pada metode Fitur Lexical
6. Evaluasi kinerja model diukur menggunakan parameter *Confusion Matrix* yang meliputi Akurasi, Precision, Recall, dan F1-Score.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai oleh peneliti dalam penelitian ini adalah menghasilkan:

1. Model klasifikasi Decision Tree yang mampu mendeteksi URL *phishing* berdasarkan fitur teks URL.
2. Peningkatan performa deteksi *phishing* melalui penerapan teknik Boosting pada algoritma Decision Tree.
3. Analisis perbandingan kinerja antara model Decision Tree standar dan model yang telah dioptimasi dengan Boosting untuk membuktikan efektivitas metode tersebut.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Manfaat Teoretis Memberikan kontribusi wawasan ilmiah mengenai penerapan teknik Boosting untuk mengatasi kelemahan *overfitting* pada algoritma Decision Tree dalam konteks keamanan siber.
2. Manfaat Praktis Menghasilkan model deteksi yang akurat yang dapat diimplementasikan sebagai sistem keamanan untuk membantu pengguna internet mengidentifikasi tautan berbahaya dan mencegah pencurian data..

1.6 Sistematika Penulisan

Untuk memudahkan pemahaman terhadap isi penelitian, skripsi ini disusun secara sistematis dalam beberapa bab yang saling berkaitan. Adapun sistematika penulisan skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, serta sistematika penulisan yang digunakan dalam penyusunan skripsi.

BAB II TINJAUAN PUSTAKA

Bab ini memuat tinjauan pustaka dari penelitian-penelitian sebelumnya serta dasar teori yang mendukung penelitian. Pembahasan meliputi konsep dasar URL

phishing, *Machine Learning*, algoritma Decision Tree, serta metode boosting yang digunakan yaitu Adaptive Boosting dan Gradient Boosting.

BAB III METODE PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, meliputi objek dan dataset penelitian, tahapan penelitian, proses preprocessing data, pembagian data menjadi data training, validasi, dan testing, pemodelan menggunakan algoritma Decision Tree dan metode boosting, serta skenario pengujian dan evaluasi model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pengujian model menggunakan metrik evaluasi seperti accuracy, precision, recall, F1-score, dan *confusion matrix*. Selain itu, dilakukan pembahasan terhadap perbandingan performa antara Decision Tree tanpa boosting dan Decision Tree dengan metode Adaptive Boosting serta Gradient Boosting.

BAB V PENUTUP

Bab ini berisi kesimpulan yang diperoleh dari hasil penelitian serta saran untuk pengembangan penelitian selanjutnya agar dapat meningkatkan performa sistem deteksi URL *phishing*.