BAB I
PENDAHULUAN

1.1 Latar Belakang

Keamanan jaringan merupakan aspek krusial dalam menjaga integritas dan kerahasiaan data di era transformasi digital saat ini. Peningkatan intensitas aktivitas digital pada sektor pemerintahan dan bisnis berbanding lurus dengan pertumbuhan ancaman siber yang memiliki pola serangan semakin kompleks. Sistem keamanan tradisional berbasis tanda tangan (*signature-based*) sering kali gagal mengenali ancaman jenis baru (*zero-day attack*) karena ketergantungannya pada basis data pola serangan yang sudah diketahui sebelumnya [1] Oleh karena itu, diperlukan pendekatan yang lebih adaptif dan cerdas untuk mendeteksi anomali pada lalu lintas jaringan secara real-time.

Penerapan teknologi *Machine Learning* dalam sistem deteksi intrusi (*Intrusion Detection System*) menjadi solusi yang efektif karena kemampuannya mempelajari karakteristik lalu lintas data secara otomatis tanpa memerlukan definisi aturan manual yang kaku. Studi terbaru menunjukkan bahwa pendekatan berbasis pembelajaran cerdas mampu mengekstraksi fitur spasial dan temporal yang kompleks dari data lalu lintas jaringan, sehingga dapat membedakan antara aktivitas pengguna yang sah dan upaya peretasan dengan akurasi yang lebih tinggi dibandingkan metode konvensional [2] Tantangan utama dalam implementasi ini adalah memilih algoritma yang tidak hanya memiliki performa tinggi, tetapi juga efisien dalam penggunaan sumber daya komputasi serta mampu meminimalkan tingkat kesalahan deteksi (*False Positive*).

Salah satu metode yang terbukti tangguh dalam menangani data serangan siber yang kompleks adalah *Ensemble Learning*, khususnya algoritma *Random Forest*. Sebagaimana didefinisikan oleh Breiman ,[3] algoritma ini bekerja dengan membangun banyak pohon keputusan (*decision trees*) dan menggabungkan hasil prediksinya melalui mekanisme voting. Pendekatan ini secara teoritis mampu mengurangi risiko *overfitting* yang sering terjadi pada penggunaan model tunggal, serta memberikan stabilitas generalisasi yang lebih baik saat menangani dataset

dengan fitur yang beragam [4] Dalam konteks keamanan siber modern, *Random Forest* terbukti memiliki ketahanan yang kuat terhadap noise dan efektif dalam menangani data berdimensi tinggi [5].

Penelitian ini berfokus pada analisis deteksi intrusi menggunakan algoritma *Random Forest* dengan memanfaatkan dataset keamanan siber terbaru. Berbeda dengan beberapa penelitian sebelumnya yang cenderung menggabungkan metode hybrid yang rumit, penelitian ini bertujuan untuk memvalidasi efektivitas model *Random Forest* standar (baseline) yang dioptimalkan melalui tahapan preprocessing data yang ketat. Fokus utamanya adalah pada penanganan data yang tidak seimbang dan penghapusan fitur yang bias. Selain evaluasi performa akurasi, penelitian ini juga menekankan pada analisis *Feature Importance* untuk mengidentifikasi faktor dominan penyebab serangan [6], seperti pola kegagalan login (*failed logins*) yang mengindikasikan adanya upaya peretasan paksa (*Brute Force*).

Melalui pendekatan ini, diharapkan dapat dihasilkan sebuah model deteksi intrusi yang tidak hanya akurat dan efisien secara komputasi, tetapi juga dapat diinterpretasikan dengan jelas. Analisis mendalam mengenai karakteristik fitur serangan diharapkan dapat memberikan kontribusi ilmiah dalam pengembangan sistem keamanan jaringan yang lebih responsif terhadap pola ancaman siber yang terus berkembang.

1.2 Rumusan Masalah

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

1. Bagaimana performa algoritma Ensemble *Random Forest* dalam mengklasifikasikan aktivitas jaringan ke dalam kategori normal dan intrusi, serta berapa tingkat akurasi yang dihasilkan?
2. Fitur apa saja yang memiliki kontribusi paling signifikan dalam membentuk pola perilaku anomali/intrusi pada dataset keamanan siber yang digunakan?

1.3 Batasan Masalah

Untuk menjaga agar penelitian ini tetap fokus dan terarah, peneliti menetapkan batasan masalah sebagai berikut:

1. Objek Penelitian: Dataset yang digunakan adalah data sekunder *Cybersecurity Intrusion Data* yang diperoleh dari repositori publik Kaggle (diakses pada tahun 2024).
2. Volume Data: Dataset terdiri dari 9.537 baris data dengan 11 atribut, yang mencakup informasi lalu lintas jaringan dan aktivitas pengguna.
3. Algoritma: Algoritma utama yang digunakan adalah *Random Forest* sebagai metode *Ensemble Learning*, dengan membandingkan performa model baseline dan model hasil optimasi (Hyperparameter Tuning).
4. Ruang Lingkup *Preprocessing*: Penelitian ini mencakup tahapan pembersihan data (Data Cleaning), penanganan nilai kosong (Missing Values), penghapusan fitur yang tidak relevan (session_id), dan penskalaan fitur (Feature Scaling).
5. Target Klasifikasi: Penelitian hanya berfokus pada Klasifikasi Biner, yaitu mengklasifikasikan data ke dalam dua kategori: Normal (0) dan Intrusi (1).
6. Lingkungan Pengembangan: Implementasi model dilakukan menggunakan bahasa pemrograman Python melalui platform Google Colab/Jupyter Notebook.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai oleh peneliti dalam penelitian ini adalah:

1. Menganalisis performa algoritma *Random Forest* sebagai metode *ensemble learning* dalam mengklasifikasikan aktivitas jaringan menjadi kategori normal atau intrusi.
2. Mengevaluasi tingkat akurasi, presisi, *recall*, dan *f1-score* dari model baseline dibandingkan dengan model yang telah melalui proses hyperparameter tuning.
3. Mengidentifikasi fitur atau atribut yang paling berkontribusi signifikan (seperti fitur failed_logins) dalam proses deteksi serangan siber melalui analisis feature importance.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat, baik secara teoritis maupun praktis.

1. Manfaat Teoretis
    1. Menambah referensi ilmiah mengenai efektivitas algoritma *Random Forest* dalam menangani dataset keamanan siber yang memiliki dimensi fitur beragam.
    2. Memberikan wawasan mengenai pentingnya tahapan *preprocessing* data, seperti penanganan missing values dan penghapusan fitur identitas sesi, dalam membangun model yang objektif.
2. Manfaat Praktis
    1. Bagi Peneliti: Menjadi sarana untuk mengimplementasikan teknik *machine learning* dan *feature engineering* dalam kasus nyata keamanan jaringan.
    2. Bagi Pengembang Sistem Keamanan: Memberikan acuan teknis dalam merancang Intrusion Detection System (IDS) yang lebih adaptif dan ringan secara komputasi.
    3. Bagi Organisasi: Memberikan informasi mengenai parameter kunci serangan (seperti pola kegagalan login) sehingga dapat digunakan sebagai dasar penyusunan kebijakan keamanan dan penguatan sistem otentikasi.

1.6 Sistematika Penulisan

Penulisan skripsi ini disusun secara sistematis dengan alur yang terstruktur, yang terdiri dari lima bab, dengan uraian sebagai berikut:

1. BAB I PENDAHULUAN Bab ini berisi Latar Belakang Masalah yang menjelaskan urgensi penelitian, Batasan Masalah, perumusan Tujuan, Manfaat yang diharapkan, serta Sistematika Penulisan. Bab ini akan memberikan gambaran umum mengenai permasalahan dan tujuan yang hendak dicapai.
2. BAB II TINJAUAN PUSTAKA Bab ini berisi tinjauan pustaka yang relevan dengan topik penelitian, mencakup dasar-dasar teori tentang keamanan siber, konsep-konsep deteksi intrusi, dan teori-teori terkait *data*

mining seperti klasifikasi, *feature engineering*, dan algoritma *Ensemble Random Forest*.

3. BAB III METODE PENELITIAN Bab ini menjelaskan tahapan-tahapan yang dilakukan peneliti dalam menyelesaikan penelitian. Bab ini mencakup tinjauan umum objek penelitian, tahapan pra-pemrosesan data, metode *feature engineering*, rancangan model Ensemble Random Forest, serta skema evaluasi kinerja model yang digunakan.
4. BAB IV HASIL DAN PEMBAHASAN Bab ini menyajikan hasil dari seluruh tahapan yang telah dilakukan. Pembahasan meliputi hasil analisis karakteristik data (EDA), hasil pengujian dan evaluasi kinerja model Ensemble Random Forest, serta analisis fitur-fitur yang paling berpengaruh terhadap akurasi klasifikasi.
5. BAB V PENUTUP Bab ini berisi kesimpulan dari seluruh penelitian yang telah dilakukan, menjawab rumusan masalah yang diajukan. Selain itu, bab ini juga memuat saran dan rekomendasi untuk pengembangan atau penelitian lebih lanjut di masa depan.