BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan pesat teknologi informasi dan komunikasi telah mendorong transformasi digital secara luas di berbagai sektor, yang pada satu sisi meningkatkan efisiensi dan konektivitas, namun di sisi lain juga memperluas permukaan serangan terhadap ancaman keamanan siber [1]. Peningkatan jumlah dan kompleksitas serangan tercermin dari analisis puluhan ribu insiden keamanan global yang menunjukkan eskalasi signifikan dalam lanskap ancaman modern [2]. Seiring dengan meningkatnya ketergantungan terhadap sistem digital, risiko terhadap gangguan operasional, pencurian data, serta eksploitasi kerentanan menjadi semakin signifikan, terutama pada organisasi dengan tingkat kematangan keamanan siber yang rendah [3].

Salah satu ancaman yang mengalami peningkatan paling signifikan adalah *ransomware*, yang tidak lagi dipandang sebagai *malware* konvensional, melainkan sebagai bagian dari ekosistem kejahatan siber yang terorganisir dan berorientasi pada keuntungan ekonomi [4]. Laporan menunjukkan bahwa eksploitasi kerentanan menjadi salah satu penyebab utama serangan *ransomware*, yang menegaskan bahwa kelemahan sistem masih menjadi titik masuk utama bagi pelaku ancaman [5]. Transformasi ini juga ditandai dengan meningkatnya profesionalisasi aktor ancaman melalui model layanan terstruktur yang memungkinkan serangan dilakukan secara lebih luas dan sistematis [6].

Lebih lanjut, perkembangan teknologi seperti otomatisasi dan kecerdasan buatan turut mempercepat siklus serangan siber, mulai dari tahap pengintaian hingga eksploitasi, sehingga meningkatkan kecepatan serta efektivitas serangan [7]. Dalam praktiknya, pelaku ancaman mampu memanfaatkan teknik dan alat otomatis untuk mengidentifikasi kerentanan secara masif dalam waktu singkat, yang memperbesar peluang keberhasilan intrusi [8]. Di sisi lain, kemajuan teknologi juga memperkuat kemampuan adversarial, termasuk dalam pengembangan *malware* dan teknik manipulasi seperti *phishing* berbasis kecerdasan buatan [9].

Selain itu, evolusi *ransomware* ditandai dengan perubahan pola serangan yang semakin kompleks, di mana pelaku tidak hanya mengenkripsi data tetapi juga melakukan pencurian informasi sensitif sebagai bentuk tekanan tambahan terhadap

korban [10]. Pendekatan ini dikenal sebagai *double extortion*, yang menggabungkan enkripsi data dengan ancaman publikasi informasi untuk meningkatkan efektivitas pemerasan [11]. Hal ini menunjukkan bahwa *ransomware* telah berkembang menjadi ancaman multidimensional yang menggabungkan aspek teknis dan psikologis dalam satu rangkaian serangan [12].

Di sisi lain, lanskap ancaman global menunjukkan peningkatan signifikan baik dari segi skala maupun intensitas serangan, dengan *ransomware* dan serangan terhadap ketersediaan sistem seperti *Distributed Denial of Service* (DDoS) menjadi ancaman yang paling dominan [13]. Sektor-sektor kritis seperti administrasi publik, kesehatan, dan infrastruktur digital menjadi target utama karena dampak operasional yang tinggi ketika terjadi gangguan layanan [14]. Dengan demikian, *ransomware* telah berkembang menjadi salah satu ancaman keamanan siber paling dominan di era digital saat ini dengan karakteristik yang semakin adaptif, terorganisir, dan sulit dideteksi [15]. Kondisi ini menunjukkan bahwa pendekatan keamanan yang bersifat reaktif tidak lagi memadai, sehingga diperlukan strategi yang lebih proaktif, adaptif, dan komprehensif dalam menghadapi dinamika ancaman siber yang terus berkembang [16].

Metode deteksi konvensional berbasis *signature* terbukti tidak efektif dalam mengenali varian *ransomware* baru maupun polimorfik, karena pendekatan tersebut hanya mampu mengenali pola yang sudah diketahui sebelumnya [17]. Sebagai respons terhadap keterbatasan ini, pendekatan *machine learning* (ML) muncul sebagai solusi yang menjanjikan karena mampu mempelajari pola perilaku berbahaya dari data historis dan menggeneralisasinya terhadap ancaman yang belum pernah ditemui sebelumnya [18]. Tinjauan sistematis menunjukkan bahwa lebih dari 72,8% studi deteksi *ransomware* kini mengadopsi pendekatan berbasis kecerdasan buatan, mencerminkan pergeseran paradigma dari metode heuristik statis menuju model yang mampu beradaptasi secara dinamis [17, 19]. Salah satu pendekatan yang banyak diadopsi adalah analisis berbasis fitur lalu lintas jaringan, yang mampu menangkap pola perilaku berbahaya secara *real-time* tanpa memerlukan akses mendalam ke sistem operasi korban [20].

Di antara berbagai algoritma ML, *Random Forest* (RF) dan *Extreme Gradient Boosting* (XGBoost) secara konsisten menunjukkan performa unggul dalam tugas deteksi *ransomware* dan *malware*. RF membangun banyak pohon keputusan melalui *bootstrap aggregation* sehingga stabil dan tahan terhadap *overfitting* [21], sementara XGBoost membangun model secara sekuensial berbasis gradien untuk memaksimalkan kecepatan dan akurasi prediksi [22]. Studi Aljabri

et al., yang membandingkan lima algoritma ML pada dataset *memory dump*, mencatat XGBoost mencapai akurasi 97,85% dan RF 97%, menunjukkan *trade-off* antara akurasi, jumlah fitur, dan *false positive rate* di antara kedua algoritma [21]. Sementara itu, Azugo et al. menerapkan RF pada dataset UGRansome2024, dataset lalu lintas jaringan yang menyertakan sebagian atribut turunan dari transaksi *blockchain* Bitcoin serta mengandung komponen sintetis (*synthetic signature*) untuk keperluan pengujian sistem deteksi dan mencapai akurasi 96% dengan *precision*, *recall*, dan F1-*score* masing-masing 0,96, sekaligus menjadi tolok ukur langsung bagi penelitian ini karena menggunakan dataset yang identik [20].

Temuan-temuan ini secara keseluruhan menegaskan bahwa perbandingan langsung antara RF dan XGBoost dalam satu lingkungan eksperimental yang terkontrol merupakan pendekatan yang valid dan kaya informasi untuk memandu implementasi deteksi *ransomware* yang optimal [23]. Di sisi lain, penelitian yang membangun *framework* deteksi *ransomware* berbasis *zero-day* menggunakan metode pembelajaran tidak terawasi menemukan bahwa meskipun pendekatan *deep learning* memberikan representasi fitur yang kaya, model-model *ensemble* seperti RF tetap kompetitif dalam hal akurasi dan interpretabilitas [24]. Hal ini menggarisbawahi bahwa kompleksitas arsitektur model tidak selalu berkorelasi dengan performa yang lebih baik, dan bahwa algoritma berbasis pohon keputusan seperti RF dan XGBoost menawarkan keseimbangan optimal antara akurasi, efisiensi komputasional, dan kemampuan interpretasi yang diperlukan dalam konteks keamanan siber operasional.

Meskipun demikian, terdapat kesenjangan penelitian yang belum diatasi. Azugo et al. hanya menerapkan satu algoritma (RF) tanpa pembanding seperti XGBoost, serta tidak menerapkan optimasi *hyperparameter*, validasi silang, maupun penanganan ketidakseimbangan kelas secara eksplisit [20]. Di sisi lain, Aljabri et al. membandingkan RF dan XGBoost secara langsung, namun menggunakan dataset berbasis *memory dump* yang karakteristiknya berbeda secara fundamental dari data lalu lintas jaringan, dan juga tidak menyertakan validasi silang maupun uji signifikansi statistik [21]. Dengan demikian, perbandingan RF dan XGBoost secara spesifik pada dataset UGRansome2024 berbasis fitur jaringan, lengkap dengan penanganan ketidakseimbangan kelas, optimasi *hyperparameter* terstruktur, dan uji signifikansi statistik, belum dikaji secara terfokus.

Berdasarkan uraian tersebut, penelitian ini bertujuan melakukan evaluasi komparatif yang sistematis terhadap kinerja RF dan XGBoost dalam mendeteksi *ransomware* menggunakan dataset UGRansome2024, dengan seluruh tahapan

penelitian disusun mengikuti kerangka kerja *Knowledge Discovery in Databases* (KDD). Penanganan ketidakseimbangan kelas dilakukan dengan SMOTE, dilengkapi dengan rekayasa fitur derivatif, analisis korelasi antar fitur, optimasi *hyperparameter* melalui *Randomized Search*, validasi menggunakan *Stratified K-Fold cross-validation*, serta uji signifikansi statistik menggunakan *Wilcoxon signed-rank test*. Evaluasi kinerja dilakukan secara menyeluruh melalui metrik akurasi, presisi, *recall*, F1-*score*, *ROC-AUC*, *false positive rate*, *confusion matrix*, serta analisis *feature importance*. Penelitian ini diharapkan dapat menutup kesenjangan metodologis dari studi Azugo et al. dan Aljabri et al., serta memberikan rekomendasi berbasis bukti bagi praktisi dan peneliti dalam memilih algoritma ML yang optimal untuk sistem deteksi *ransomware* berbasis jaringan [20, 21].

1.2 Rumusan Masalah

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

1. Bagaimana membangun model deteksi *ransomware* menggunakan algoritma *Random Forest* dan *XGBoost*?
2. Bagaimana performa model *Random Forest* dan *XGBoost* dalam mendeteksi *ransomware*?
3. Di antara *Random Forest* dan *XGBoost*, algoritma manakah yang lebih baik dalam mendeteksi *ransomware*?

1.3 Batasan Permasalahan

Agar penelitian ini tetap terfokus dan dapat dipertanggungjawabkan secara ilmiah, ditetapkan batasan-batasan permasalahan sebagai berikut:

1. Penelitian ini menggunakan dataset UGRansome2024 yang diperkenalkan oleh Azugo et al. [20], yang terdiri dari 149.043 sampel lalu lintas jaringan dari berbagai keluarga *ransomware*, dengan fitur yang mencakup protokol komunikasi, volume *netflow*, dan atribut transaksi *blockchain*. Dataset ini tidak mencakup fitur berbasis *memory dump* maupun artefak sistem operasi.
2. Algoritma yang dibandingkan dibatasi hanya pada *Random Forest* dan XGBoost, dengan *hyperparameter* kedua algoritma dioptimalkan menggunakan *RandomizedSearchCV*.

3. Rekayasa fitur dibatasi pada pembentukan lima fitur derivatif berbasis rasio, transformasi logaritmik, interaksi, dan biner, tanpa seleksi fitur berbasis statistik seperti Chi-Square sebagaimana digunakan pada studi acuan Aljabri et al. [21].

4. Penanganan ketidakseimbangan kelas dibatasi pada teknik SMOTE (*Synthetic Minority Oversampling Technique*), dan evaluasi model dibatasi pada kerangka *10-fold Stratified K-Fold cross-validation*.

5. Perbedaan kinerja antara kedua algoritma diuji secara statistik menggunakan *Wilcoxon signed-rank test* guna memastikan signifikansi perbedaan performa yang diperoleh. Penelitian ini tidak mencakup pengujian pada lingkungan produksi atau sistem deteksi secara *real-time*.

1.4 Tujuan Penelitian

Sesuai dengan rumusan masalah yang telah ditetapkan, tujuan penelitian ini adalah sebagai berikut:

1. Membangun model deteksi *ransomware* berbasis fitur lalu lintas jaringan menggunakan algoritma *Random Forest* dan *XGBoost*.

2. Melakukan analisis performa model *Random Forest* dan *XGBoost* dalam mendeteksi *ransomware*.

3. Membandingkan performa kedua algoritma untuk menentukan model yang lebih baik dalam mendeteksi *ransomware*.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini antara lain:

1. Memberikan referensi komparatif kinerja *Random Forest* dan XGBoost bagi penelitian deteksi *ransomware* selanjutnya.

2. Memberikan rekomendasi berbasis bukti empiris mengenai algoritma yang lebih tepat untuk diimplementasikan dalam sistem deteksi *ransomware*, khususnya bagi praktisi keamanan siber yang mempertimbangkan aspek akurasi sekaligus efisiensi komputasional.

3. Menyediakan metodologi eksperimental yang dapat direplikasi pada dataset UGRansome2024 atau dataset sejenis berbasis fitur jaringan.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran. Sistematika penulisan laporan adalah sebagai berikut:

1. Bab 1 PENDAHULUAN
   Bab ini menguraikan latar belakang penelitian, identifikasi kesenjangan penelitian, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.

2. Bab 2 LANDASAN TEORI
   Bab ini membahas landasan teori dan tinjauan pustaka yang relevan, mencakup konsep *ransomware*, algoritma *Random Forest* dan XGBoost, teknik seleksi fitur, serta penelitian-penelitian terdahulu yang menjadi acuan.

3. Bab 3 METODOLOGI PENELITIAN
   Bab ini menjelaskan metodologi penelitian yang digunakan, meliputi spesifikasi dataset, tahapan praproses data, teknik seleksi fitur, konfigurasi model, skema validasi, dan metrik evaluasi yang diterapkan.

4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil eksperimen secara lengkap beserta analisis dan pembahasan komparatif kinerja *Random Forest* dan XGBoost, termasuk analisis *trade-off* fitur, efisiensi komputasional, dan perbandingan dengan penelitian terdahulu.

5. Bab 5 SIMPULAN DAN SARAN
   Bab ini menyampaikan simpulan yang menjawab seluruh rumusan masalah penelitian serta saran untuk pengembangan penelitian selanjutnya.