

BAB I

PENDAHULUAN

A. Latar Belakang

Pada tahun 2022, Indonesia menghadapi ancaman signifikan di ruang digital. Badan Siber dan Sandi Negara (BSSN) mencatat sebanyak 370,02 juta serangan digital. Laporan tersebut mengalami peningkatan sebesar 38,72% kasus dibandingkan tahun sebelumnya yang berjumlah 266,74 juta serangan. Sektor administrasi pemerintahan menjadi target utama dengan 284,09 juta serangan. Kasus tersebut menunjukkan rentannya sektor publik terhadap serangan digital yang semakin canggih dan terorganisasi [1], [2]. Laporan Lanskap Keamanan Siber Indonesia 2023 melaporkan total lalu lintas anomali sebanyak 403,99 juta pada tahun 2023, dengan ancaman seperti Ransomware, *phishing*, dan *Advanced Persistent Threat (APT)*. Ancaman-ancaman tersebut menjadi perhatian utama dalam perlindungan ruang digital nasional [2]. Kasus tersebut mencerminkan urgensi dalam meningkatkan kapabilitas keamanan digital, termasuk deteksi dini terhadap serangan yang memanfaatkan berkas dokumen, seperti *file* PDF, yang sering menjadi medium penyebaran *malware*.

Dalam beberapa tahun terakhir, *file* PDF semakin sering digunakan sebagai media penyebaran *malware*, memanfaatkan sifatnya yang dianggap aman oleh banyak pengguna. Menurut McAfee Labs, pelaku ancaman memanfaatkan *file* PDF untuk menyisipkan tautan berbahaya atau JavaScript yang mengunduh *payload malware* ke perangkat korban [3]. Laporan Lanskap Keamanan Siber Indonesia 2023 mencatat bahwa serangan yang menggunakan elemen *phishing* dan *malware* terus meningkat, menunjukkan bahwa pengguna masih rentan terhadap ancaman digital berbasis dokumen [2]. Penelitian sebelumnya menunjukkan bahwa dokumen PDF dapat dimanipulasi untuk menyisipkan kode berbahaya, memungkinkan eksekusi tanpa sepengetahuan pengguna melalui fitur internal tertentu [4]. Sebagai contoh, *malware* dapat mengeksekusi perintah PowerShell untuk mengunduh *malware* tambahan, seperti Trojan atau

Ransomware, yang dapat mencuri data sensitif [3]. Selain itu, deteksi *malware* berbasis tanda tangan (*signature-based detection*) menghadapi tantangan signifikan, termasuk ketidakmampuan mendeteksi serangan *zero-day* dan *malware polymorphic*, sehingga solusi yang lebih maju berbasis *machine learning*, seperti Support Vector Machine (SVM), menjadi sangat relevan untuk menganalisis pola anomali dan mendeteksi ancaman dalam *file* PDF [5].

Deteksi *malware* dalam *file* PDF memerlukan pendekatan yang lebih adaptif, terutama untuk mengidentifikasi ancaman *evasive* yang sulit terdeteksi oleh algoritma *machine learning* konvensional. CIC-Evasive-PDFMal2022, sebuah *dataset* yang terdiri dari 18.126 sampel (9.063 *file malicious* dan 9.063 *file benign*), dirancang untuk menangani tantangan ini. *Dataset* ini dibuat dengan menggabungkan data dari Contagio dan VirusTotal, sehingga menghasilkan *dataset* yang lebih representatif terhadap distribusi *file* PDF di dunia nyata [6].

Pendekatan berbasis *machine learning*, khususnya Support Vector Machine (SVM) dan Random Forest, memiliki potensi besar dalam menganalisis fitur statis dari *file* PDF, seperti elemen *metadata*, pola *encoding*, dan keberadaan JavaScript berbahaya, yang diekstraksi dari *dataset* ini [6]. SVM telah terbukti mampu menghasilkan *hyperplane* optimal untuk memisahkan data yang kompleks dan sangat sesuai untuk mendeteksi pola anomali dalam *file* PDF *evasive*. Di sisi lain, Random Forest sebagai metode Ensemble Learning yang menggabungkan banyak pohon keputusan, menunjukkan kinerja yang kompetitif dalam klasifikasi data kompleks dan telah digunakan dalam berbagai studi deteksi *malware* [5], [7]. Dengan fitur-fitur yang lebih sulit untuk diidentifikasi oleh algoritma tradisional, *dataset* ini memberikan tantangan sekaligus peluang untuk menguji kemampuan model *machine learning* dalam mendeteksi ancaman yang lebih canggih [6], [8].

Sebagai solusi, penelitian ini menggunakan SVM untuk menganalisis CIC-Evasive-PDFMal2022, yang diharapkan dapat meningkatkan akurasi deteksi terhadap ancaman *zero-day* dan *polymorphic*. Model ini juga relevan untuk menutup kelemahan metode berbasis tanda tangan yang sering gagal dalam mendeteksi serangan baru, menjadikannya langkah signifikan dalam pengembangan sistem keamanan digital berbasis dokumen.

B. Rumusan Masalah

Berdasarkan latar belakang di atas rumusan masalah penelitian ini yaitu:

1. Bagaimana mendeteksi *malware* pada *file* PDF menggunakan algoritma Support Vector Machine (SVM) dan Random Forest?
2. Bagaimana hasil deteksi *malware* pada *file* PDF menggunakan algoritma SVM jika dibandingkan dengan algoritma Random Forest dalam hal *Accuracy*, *Precision*, *Recall*, dan *F1-Score*?

C. Batasan Masalah

Penelitian ini dibatasi pada beberapa aspek berikut untuk menjaga fokus dan ruang lingkup deteksi *malware* pada pdf:

1. Penelitian ini hanya berfokus pada deteksi *malware* dalam *file* PDF dan tidak mencakup jenis dokumen lain seperti Word, Excel, atau format dokumen lainnya.
2. Algoritma utama yang digunakan adalah Support Vector Machine (SVM) dan Random Forest, namun penelitian ini juga akan melakukan perbandingan kinerja dengan algoritma Random Forest untuk mengevaluasi efektivitas model.
3. *Dataset* yang digunakan adalah CIC-Evasive-PDFMal2022, yang terdiri dari 18.126 sampel (9.063 *file malicious* dan 9.063 *file benign*), dengan fokus pada fitur statis seperti *metadata*, pola *encoding*, dan keberadaan elemen JavaScript berbahaya.
4. Penelitian ini hanya menggunakan analisis statis untuk mendeteksi *malware*, sehingga tidak mencakup analisis dinamis yang melibatkan eksekusi *file* PDF untuk mempelajari perilaku *malware* secara langsung.
5. *Tools* yang digunakan dalam pengembangan model dan pengolahan data adalah RapidMiner.

D. Tujuan Penelitian

Berdasarkan rumusan masalah di atas maka didapatkan tujuan penelitian

yaitu:

1. Mengembangkan metode untuk mendeteksi *malware* pada *file* PDF menggunakan algoritma Support Vector Machine (SVM) dan Random Forest.
2. Melakukan perbandingan kinerja antara algoritma SVM dan Random Forest dalam mendeteksi *malware* pada *file* PDF berdasarkan metrik evaluasi seperti *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

E. Metode Penelitian

Adapun beberapa metode yang akan digunakan adalah sebagai berikut:

1. Studi Literatur

Tahapan awal penelitian ini adalah melakukan studi literatur yang komprehensif untuk memahami konsep-konsep dasar dan penelitian sebelumnya terkait deteksi *malware* pada *file* PDF. Literatur yang dikaji mencakup teknik *machine learning*, terutama Support Vector Machine (SVM) dan Random Forest, analisis statis *file* PDF, serta karakteristik ancaman digital yang relevan. Selain itu, kelemahan metode berbasis tanda tangan dalam mendeteksi *malware* serta keunggulan metode berbasis *machine learning* akan dianalisis secara mendalam untuk membangun dasar teoritis penelitian ini.

2. Pengumpulan *Dataset*

Dataset yang digunakan dalam penelitian ini adalah CIC-Evasive-PDFMal2022, yang diperoleh dari Kaggle. *Dataset* ini berisi 18.126 sampel *file* PDF, terdiri dari 9.063 *file malicious* dan 9.063 *file benign*, yang dikumpulkan dari sampel Contagio dan VirusTotal.

3. *Preprocessing* Data

Data yang telah dikumpulkan melalui proses ekstraksi fitur kemudian diproses lebih lanjut untuk memastikan kualitasnya. Tahap ini mencakup pembersihan data untuk menghilangkan duplikasi data dan data yang tidak relevan, normalisasi fitur agar sesuai dengan algoritma SVM, serta pembagian *dataset* menjadi data latih (*training data*) dan data uji (*testing data*).

4. Pengembangan Model SVM

Pada tahap ini, dua model klasifikasi akan dikembangkan menggunakan

algoritma Support Vector Machine (SVM) dan Random Forest. Kedua model dirancang untuk memisahkan *file* PDF menjadi dua kelas, yaitu *malicious* dan *benign*. Proses pelatihan dilakukan dengan menggunakan *training dataset* yang telah diproses pada tahap sebelumnya.

5. Pengujian Model

Setelah kedua model dikembangkan, dilakukan pengujian terhadap *testing dataset* untuk mengevaluasi performa masing-masing model. Proses ini bertujuan untuk mengukur kemampuan kedua model dalam mengklasifikasikan data baru yang belum pernah dilihat sebelumnya. Metrik evaluasi yang digunakan meliputi *Accuracy*, *Precision*, *Recall*, dan *F1-Score*.

6. Analisis Hasil

Tahap ini bertujuan untuk menganalisis dan membandingkan kinerja kedua model, yaitu SVM dan Random Forest, dalam mendeteksi *malware* pada *file* PDF. Analisis dilakukan berdasarkan metrik evaluasi yang telah disebutkan untuk menilai sejauh mana kedua model mampu mengklasifikasikan *file* PDF sebagai *malicious* atau *benign*. Hasil analisis ini diharapkan memberikan wawasan mengenai efektivitas masing-masing model, serta membantu dalam memilih algoritma yang lebih optimal untuk tugas deteksi *malware* berbasis *machine learning*.

F. Sistematika Penulisan

Secara sistematis isi dari laporan ini disusun sebagai berikut:

1. BAB 1 PENDAHULUAN

Bab ini menjelaskan latar belakang penelitian, perumusan masalah, batasan masalah, dan tujuan penelitian. Latar belakang menguraikan pentingnya penelitian dan isu yang diangkat. Rumusan masalah memuat pertanyaan yang akan dijawab melalui penelitian. Batasan masalah menentukan ruang lingkup penelitian untuk menjaga fokus, dan tujuan penelitian menggambarkan hasil yang ingin dicapai.

2. BAB 2 TINJAUAN PUSTAKA

Bab ini berisi penelitian terdahulu yang relevan dengan topik penelitian.

Setiap penelitian diuraikan secara singkat, mencakup metodologi, hasil, serta kelebihan dan kekurangannya. Perbandingan dengan penelitian yang sedang dilakukan disajikan dalam bentuk tabel untuk menonjolkan kontribusi penelitian ini terhadap pengembangan ilmu.

3. BAB 3 LANDASAN TEORI

Bab ini membahas teori-teori yang mendukung penelitian, seperti konsep dasar *machine learning*, Support Vector Machine (SVM), Random Forest, dan analisis statis pada *file* PDF. Penjelasan dalam bab ini mencakup definisi, prinsip kerja, dan relevansi teori terhadap penelitian. Semua teori tentang alat dan implementasi dikecualikan dari bagian ini untuk menjaga fokus pada fondasi keilmuan.

4. BAB 4 DATASET DAN PENGEMBANGAN MODEL

Bab ini membahas penggunaan *dataset* CIC-Evasive-PDFMal2022 untuk mengembangkan model SVM dan Random Forest dalam mendeteksi *malware* pada *file* PDF. Langkah-langkah meliputi analisis, *preprocessing data*, pelatihan model, optimasi parameter, dan pengujian dengan *dataset* baru untuk memastikan akurasi dan keandalan mode.

5. BAB 5 HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pengujian dan analisis kinerja model klasifikasi yang dibangun menggunakan algoritma Support Vector Machine (SVM) dan Random Forest dalam mendeteksi *file* PDF yang bersifat *malicious* atau *benign*. Evaluasi dilakukan menggunakan metrik umum seperti *Accuracy*, *Precision*, *Recall*, *F1-Score*, serta kurva ROC dan nilai AUC untuk menggambarkan performa model secara keseluruhan dalam berbagai skenario (rata-rata, optimis, dan pesimis).

6. BAB 6 PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan mengenai pengembangan dan evaluasi model klasifikasi untuk mendeteksi *malware* pada *file* PDF menggunakan algoritma Support Vector Machine (SVM) dan Random Forest.