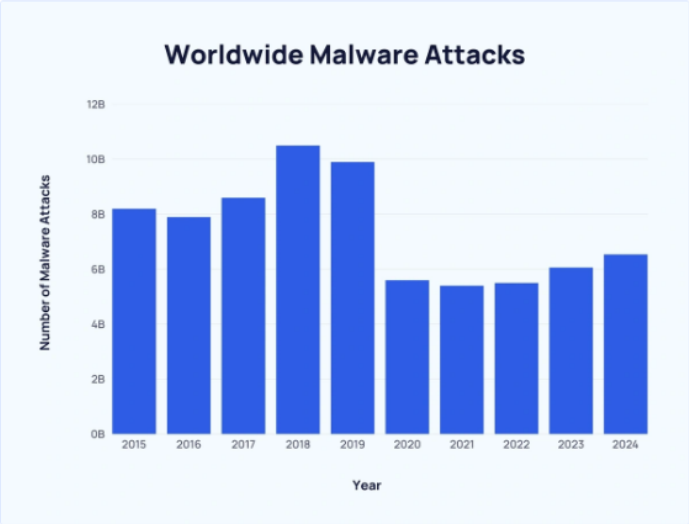
BAB I
PENDAHULUAN

1.1 Latar Belakang

Peradaban manusia secara bertahap mengalami modernisasi, yang berdampak pada berbagai aspek kehidupan. Era saat ini ditandai oleh kemajuan dalam teknologi dan informasi yang dimanfaatkan oleh manusia untuk menyederhanakan berbagai tugas. Internet adalah salah satu contoh implementasi kemajuan ini [1]. Seiring dengan perkembangan internet, keamanan siber menjadi semakin rentan terhadap berbagai jenis serangan. Salah satu jenis serangan yang dapat terjadi adalah *malware* [2].

Istilah umum "*Malware"* merujuk pada program atau perangkat lunak yang secara khusus diciptakan untuk menyusup maupun merusak sistem komputer maupun sistem operasi. Malware terdiri dari berbagai jenis, antara lain *virus, worm, trojan horse,* Sebagian besar *rookit, spyware, adware,* serta program berbahaya lainnya yang dapat mengancam komputer [3]. Ancaman ini bukan hanya sekedar ancaman biasa. Berdasarkan data dari explodingtopics.com/blog/cybersecurity-stats pada tiga tahun terakhir adanya peningkatan jumlah serangan *malware,* pada tahun 2024 tercatat lebih dari 6,5 miliar serangan *malware* diseluruh dunia, serangan *malware* ini meningkat sekitar 8% dibanding kan tahun sebelumnya. Dalam rentang 2020 hingga 2024, jumlah serangan *malware* terus berkisar miliaran, dengan puncak tertinggi serangan terjadi 2018 yaitu sampai mencapai 10,5 miliar serangan. Fakta ini menunjukkan bahwa serangan *malware* tetap menjadi ancaman serius dalam dunia digital.



Gambar 1. 1 Grafik jumlah serangan *malware* pada periode 2015-2034

Berdasarkan permasalahan tersebut, diperlukan metode deteksi *malware* efektif. Salah satunya menerapkan algoritma *machine learning* karena mampu mendeteksi pola baru dan mengenali karakteristik apakah ada serangan *malware* atau tidak. Proses yang akan dilakukan untuk deteksi *malware* ini adalah pengumpulan dataset, tahap *pre-processing*, pelatihan *Machine Learning*, dan pengujian kinerja model. Data yang diproses dengan *machine learning* menggunakan model *Random Forest, Support Vector Machine (SVM),* serta *Naïve Bayes* yang berasal dari dataset *malware* [4]. Dan agar dapat mengetahui apakah itu merupakan serangan *malware* atau tidak.

Machine Learning merupakan penerapan Kecerdasan Buatan *(Artificial Intelligence, AI*) yang bertujuan untuk menciptakan sistem yang dapat belajar sendiri tanpa perlu diprogram ulang berkali-kali. *Machine Learning* menggunakan data (data pelatihan) selama proses pembelajarannya sebelum memberikan hasil *output* yang diinginkan [5]. Pada penelitian ini akan menguji seberapa baik kinerja *Machine Learning* yang akan melakukan deteksi dengan menggunakan model seperti *Random Forest, Support Vector Machine (SVM),* serta *Naïve Bayes*. Alasan Penggunaan model-model tersebut dalam penelitian ini didasarkan pada

karakteristik serta keunggulan masing-masing model. Model *Random Forest* dipilih karena kemampuannya dalam mengurangi *overfitting* melalui kombinasi beberapa pohon Keputusan yang stabil dan akurat. Model *Support Vector Machine (SVM)* dipilih karena efektivitasnya dalam membedakan kelas secara optimal pada data berdimensi tinggi dengan membentuk *hyperplane* maksimum. Sementara, untuk model Naïve Bayes digunakan karena efisiensi komputasionalnya serta kecocokannya dalam menangani data klasifikasi berbasis probabilitas. Ketiga model ini dipertimbangkan untuk memberikan variasi pendekatan dalam meng*evaluasi* performa klasifikasi deteksi *malware*.

Random Forest merupakan suatu algoritma dalam *Machine Learning* yang memiliki fungsi untuk melakukan klasifikasi pada data berjumlah besar. Metode ini mampu diterapkan dengan bermacam-macam dimensi dan tingkat skala yang berbeda serta memiliki performa yang sangat baik [6]. Selain itu, *Random Forest* menghasilkan berbagai pohon dibangun secara independent dari subset data pelatihan yang di ambil secara acak menggunakan metode bootstrap, serta pada tiap node dilakukan pemilihan variable input secara acak [7]. *Random Forest* juga bisa dianggap sebagai gabungan dari setiap pohon yang digunakan dalam satu model [8]. Algoritma ini, sebagai metode *ensemble,* mampu menangani kompleksitas serta pola *non-linear* pada dataset secara efisien [9]. Peneliti [4] mengatakan bahwa *Random Forest* memiliki nilai akurasi mencapai 99%. Hal ini menunjukan bahwa metode *Random Forest* efektif dalam mengidentifikasi *malware* dengan presisi tinggi.

Support Vector Machine (SVM) merupakan salah satu algoritma klasifikasi dalam *machine learning* yang terawasi, menggunakan dua garis vektor (*hyperplane*) dengan *margin* yang maksimal. Algoritma SVM telah mengalami perkembangan pada tahun 1960 dan kemudian dikembangkan lagi oleh *Boser, Vapnik,* dan *Guyon* pada tahun 1992. [10]. Ketika mengidentifikasi risiko keamanan dalam jaringan, SVM dimanfaatkan untuk mengembangkan model yang mampu membedakan antara data yang

normal dan data yang dianggap mencurigakan atau berbahaya. [11]. SVM memanfaatkan ruang kemungkinan yang terdiri dari fungsi linier dalam fitur dengan dimensi tinggi dan dilatih menggunakan algoritma pembelajaran yang didasarkan pada teori optimasi [12]. SVM unggul karena kemampuannya dalam menangani model *nonlinier* yang kompleks, serta ketahanannya yang lebih baik terhadap *overfitting* jika dibandingkan dengan metode lainnya [13]. Menurut penelitian [14] menunjukkan bahwa metode *Machine Learning* SVM mampu melakukan proses klasifikasi dengan baik dan mencapai tingkat akurasi yang memuaskan.

Naïve Bayes merupakan teknik klasifikasi probabilitas yang sederhana, yang mengestimasi sekumpulan probabilitas dengan cara melihat frekuensi serta kombinasi nilai dari data yang diberikan [15]. Menurut [16] *Naive Bayes* adalah algoritma klasifikasi yang memiliki rumus sederhana dan mudah diaplikasikan[17]. Metode ini melakukan klasifikasi kelas dengan membandingkan nilai posterior dari berbagai kelas yang ada. Kelas dengan nilai posterior tertinggi kemudian menjadi hasil klasifikasi. *Naive Bayes* bekerja berdasarkan pada prinsip fitur independen, yang berarti bahwa fitur-fitur dalam suatu dataset tidak bergantung oleh adanya fitur lain dalam dataset tersebut [18]. Menurut penelitian, metode ini hanya memerlukan sedikit data latih untuk menilai parameter yang diperlukan dalam proses pengklasifikasiannya, karena parameter ini dianggap sebagai variabel independen.

Penelitian ini memiliki urgensi dalam menghadapi ancaman serangan *malware* yang semakin kompleks dan merugikan, untuk memastikan perlindungan yang optimal terhadap suatu data sensitif. Atas dasar tersebut, penulis menggunakan beberapa model *Machine Learning* untuk mengklasifikasi deteksi serangan dari *Malware*, model *Machine Learning* yang di uji untuk penelitian ini yaitu *Random Forest*, *Support Vector Machine*, dan *Naïve bayes*. Serta penelitian ini akan diberi judul "KLASIFIKASI SERANGAN *MALWARE* MENGGUNAKAN *MACHINE LEARNING*".

Sebagai acuan, salah satu penelitian sebelumnya dilakukan oleh Evan Valdis Tjahjadi, Budy Santoso, dan Serwin pada tahun 2023, berjudul "Malware Classification Using Machine Learning Techinuques" Penelitian ini bertujuan untuk mengevaluasi kinerja metode *Random Forest* dalam *klasifikasi malware.* Hasil dari penelitian ini menunjukkan bahwa model *Random Forest* berhasil mencapai tingkat akurasi klasifikasi *malware* sebesar 99% [4].

1.2 Rumusan Masalah

Seberapa efektif model *Random Forest, Support Vector Machine* (SVM), dan *Naive Bayes* dalam mengklasifikasi *malware*, serta model mana yang memberikan hasil terbaik berdasarkan *accuracy, precision, recall*, dan *F1-score*?

1.3 Tujuan dan Manfaat

Mengacu pada masalah dan pertanyaan yang sudah diuraikan, tujuan dari penelitian ini adalah:

1. Mengetahui kinerja tiga model *machine learning Random Forest, Support Vector Machine* (SVM), dan *Naïve Bayes* dalam mengklasifikasi serangan *malware*.
2. Menentukan model *machine learning* yang paling efektif berdasarkan hasil evaluasi menggunakan metrik *accuracy, precision, recall,* dan *F1-Score.*

Adapun manfaat dari penelitian ini adalah:

1. Penelitian ini dapat menjadi referensi tambahan untuk pengembangan cara mendeteksi *malware* dengan menggunakan algoritma *machine learning* khususnya *Random Forest, Support Vector Machine* (SVM), dan *Naïve Bayes*.
2. Hasil dari penelitian ini dapat dijadikan pertimbangan dalam memilih model deteksi *malware* yang efektif berdasarkan *evaluasi metrix* seperti *accuracy, precision, recall,* dan *F1-score.*

1.4 Batasan Masalah

Berdasarkan penjelasan mengenai permasalahan dan tujuan dari penelitian ini, terdapat beberapa batasan penelitian yang tentukan untuk melakukan analisis yang relevan dengan isu yang dihadapi:

1. Data yang digunakan merupakan Dataset pe-files-*malware*s
2. Model *Machine Learning* yang diterapkan adalah *Random Forest, Support Vector Machine (SVM), Naïve Bayes*.
3. *Evaluasi* model dilakukan melalui *confussion matrix* (*Accuracy, Precision, Recall,* dan *F1-score)*

1.5 Metode Penelitian

Metode Penelitian ini menggambarkan langkah-langkah yang digunakan dalam menyelesaikan permasalahan pada penelitian ini. Proses penelitian dilakukan melalui tahapan-tahapan berikut:

1. ***Studi Literatur***

   Dilakukan untuk memahami teori, metode, dan penelitian sebelumnya yang berkaitan dengan deteksi *malware* serta penggunaan model *machine learning* seperti *Random Forest, Support Vector Machine* (SVM), dan *Naïve Bayes*

2. ***Data Selection***

   Data Selection dilakukan dengan penggunaan dataset berupa file Portable Executable (PE), yang mencakup proses pelabelan untuk membedakan file *malware* dan *benign*.

3. ***Exploratory Data Analysis* (EDA)**

   Exploratory Data Analysis (EDA) dilakukan untuk memperoleh pemahaman yang mendalam tentang karakteristik data, seperti proporsi kelas *malware* dan *benign.* Tahapan ini membantu dalam mengidentifikasi pola tersembunyi, ketidakseimbangan data yang berpotensi memengaruhi hasil pemodelan secara signifikan.

4. ***Pre-Processing***

Tahap Preprocessing data mencakup proses normalisasi, pemilihan fitur, serta penanganan ketidakseimbangan kelas (imbalanced data) pada dataset.

5. ***Modeling***

   Pada data mining akan dilakukan eksperimen model dengan mengimplementasikan ketiga model (*Random Forest, Support Vector Machine* (SVM), dan *Naïve Bayes*). Kemudian dilakukan penyesuaian *hyperparameter* serta evaluasi kinerja model menggunakan metrik seperti *accuracy, precision, recall,* dan *F1-score*.

6. ***Evaluation***

   Evaluation adalah analisis hasil yang betujuan untuk membandingkan performa masing-masing model serta menyimpulkan efektivitas metode dalam deteksi *malware*.