BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi saat ini telah membawa banyak kemajuan di berbagai bidang, termasuk pada perangkat bergerak yaitu *smartphone*. Ponsel pintar menyediakan berbagai aplikasi dan sistem operasi, seperti *iOS, Blackberry OS,* dan *Android,* untuk memudahkan penggunanya. Android merupakan salah satu sistem operasi utama yang mendominasi pasar dengan sekitar 1,4 juta aktivasi harian, seperti yang diungkapkan oleh CEO Google, Erich Schmidt (Hadiprakoso et al., 2022). Menurut data dari gsmaintelligence.com, pada September 2022 *Android* mendominasi pasar sistem operasi seluler dengan persentase sebesar 71,54% (Islam et al., 2023). Popularitas yang dimiliki *Android* tidak selalu berdampak positif bagi penggunanya, karena seiring dengan tingginya jumlah pengguna *Android,* hal ini bisa menjadikan target utama serangan *malware* (Muzaffar et al., 2022).

Malware merupakan perangkat lunak berbahaya yang dirancang untuk menyebabkan kerusakan pada sistem, mencuri data, dan mendapatkan akses tanpa izin pada suatu sistem. Penyebaran *malware* dapat terjadi melalui berbagai metode, termasuk serangan *phishing* melalui *email,* rekayasa sosial, dan melalui perangkat lunak tidak aman yang diunduh oleh pengguna. Serangan *malware* bertujuan untuk mengambil informasi rahasia seperti kata sandi dan *email* dari pemilik, serta menyebarkan *spam*. Diperlukan analisis lebih lanjut dalam mengatasi ancaman ini, karena risiko dan dampak negatif yang ditimbulkan oleh serangan *malware* sangat besar (Ramadhan et al., 2023).

Penerapan teknik *Machine Learning* (ML) untuk deteksi *malware* telah menjadi fokus penelitian yang signifikan pada beberapa dekade terakhir. ML merupakan konsep komputasi yang menggunakan algoritma matematika dan komputer untuk mempelajari pola-pola dalam data serta menghasilkan prediksi untuk kejadian di masa depan. Proses pembelajaran ini melibatkan dua tahapan kunci, yakni pelatihan (*training*) dan pengujian (*testing*). Ada tiga kategori utama

dalam *machine learning*, yaitu *Supervised Learning, Unsupervised Learning*, dan *Reinforcement Learning* (Roihan et al., 2020).

Penelitian mengenai deteksi *malware android* berbasis ML dapat dibagi menjadi dua kategori yaitu analisis statis dan analisis dinamis. Analisis statis dilakukan dengan mengekstrak file instalasi bernama *Android Package Kit* (APK) untuk mendapatkan informasi tentang *manifest, permission, API calls, intent,* dan lain-lain. Sementara analisis dinamis berfokus pada pelacakan aktivitas aplikasi seperti *logcat error, shared memory dirty, system calls, process,* dan lain-lain (Islam et al., 2023). Penelitian ini mengusulkan algoritma *machine learning* yaitu *Random Forest* (RF) dalam klasifikasi *malware android* berdasarkan analisis fitur dinamis pada dataset CCCS-CIC-AndMal2020. Analisis fitur dinamis dipilih dalam penelitian ini karena semakin banyak malware Android menghindari deteksi statis melalui teknik seperti *repackaging* dan *code obfuscation*, sehingga diperlukan metode analisis dinamis berdasarkan karakteristik perilaku yang dapat menyelesaikan masalah ini dengan baik (Lu et al., 2020).

Klasifikasi *malware android* akan lebih baik jika diterapkan pada kumpulan data yang besar dan berisi semua jenis *malware* yang terbaru (Rahali et al., 2020). Penelitian ini akan menggunakan dataset CCCS-CIC-AndMal2020, yang diterbitkan pada tahun 2020 dan mencakup 12 kategori utama malware. Algoritma *Random Forest* (RF) diusulkan untuk digunakan pada proses klasifikasi malware android, karena algoritma ini memiliki performa yang sangat baik ketika digunakan pada dataset yang besar dan kompleks. Kemampuan lain dari algoritma ini yaitu dapat mengatasi overfitting, menangani data yang tidak seimbang, mengatasi variabel yang tidak relevan, skalabilitas dan memiliki fleksibilitas dalam tipe data apapun. Algoritma RF termasuk kedalam metode ensemble machine learning yang memanfaatkan gabungan dari pohon keputusan (*decision tree*) dengan tujuan untuk meningkatkan kinerja dan ketangguhan model dengan memanfaatkan kekuatan kolektif dari model yang lebih sederhana (Parmar et al., 2019).

Penelitian mengenai algoritma RF sebagai model klasifikasi *malware* pada dataset *pe-files-malwares* pernah dilakukan, dengan tujuan mengetahui performa

dari algoritma RF dalam klasifikasi *malware*. Hasilnya model memberikan hasil akurasi yang tinggi sebesar 99% dalam proses klasifikasi meskipun tanpa dilakukan tahapan *preprocessing* data (Tjahjadi & Santoso, 2023). Penelitian lain menerapkan algoritma RF untuk klasifikasi *malware android* pada dataset *virusshare* dengan total data 13.076 APK. Teknik SMOTE (*Synthetic Minority Over-Sampling Techinque*) diterapkan untuk mengatasi masalah ketidakseimbangan kelas pada dataset. Hasil akhir dengan akurasi terbaik didapatkan dengan penerapan teknik SMOTE sebesar 92.96% pada proses klasifikasi (Turnip et al., 2023).

Perbandingan performa klasifikasi antara algortima *Support Vector Machine* (SVM) dengan *Random Forest* (RF) pada dataset *malware/benign permission android* pernah dilakukan, hasil akhir menunjukkan nilai akurasi sebesar 98,99% untuk model RF dan 96,23% untuk model SVM. Hasil tersebut menyatakan bahwa algoritma RF lebih unggul dari SVM (Sitorus et al., 2021). Metode *ensemble learning* dengan penggabungan algoritma *Decision Tree* (DT), *K-Nearest Neighbor* (KNN) dan *Random Forest* (RF) diterapkan pada klasifikasi *malware* dengan dataset CIC-AndMal2017. Hasil penelitian menunjukan bahwa penerapan *ensemble learning* memiliki performa model yang lebih baik dengan hasil akurasi sebesar 95,2% (Zakariya & Ramli, 2023).

Metode *ensemble learning* diterapkan pada penelitian lain dalam proses klasifikasi dengan menggunakan algoritma RF, KNN, *Multi-Level Perceptrons*, DT, SVM dan *Logistic Regression* pada dataset CCCS-CIC-AndMal2020 berdasarkan analisis fitur dinamis. Proses klasifikasi data melalui tahap *preprocessing* yang panjang meliputi, *missing data imputation random oversampling, outlier handling, feature scaling, feature transformation*. Hasil akhir didapatkan dengan nilai akurasi sebesar 95% setelah 60,2% fitur diseleksi. Hasil tersebut menyatakan bahwa teknik *ensemble learning* lebih disarankan dibandingkan teknik tunggal (Islam et al., 2023). Studi mendalam dilakukan pada penlitian lain dengan membandingkan beberapa algoritma yaitu SVM, DT, RF, *Naïve Bayes* (NB) dan KNN pada dataset CCCS-CIC-AndMal2020 berdasarkan

analisis fitur statis. Hasil akhir menunjukkan bahwa algoritma RF mendapat hasil akurasi yang terbaik yaitu sebesar 92,96% (Batouche & Jahankhani, 2021).

Fenomena serangan *malware* akan terus tumbuh beriringan dengan perkembangan teknologi, sehingga perlu dilakukan penelitian lebih lanjut mengenai *malware Android* dengan memanfaatkan algoritma *machine learning* dengan dataset yang terbaru (Diana et al., 2022). CCCS-CIC-AndMal2020 merupakan salah satu dataset yang baru-baru ini diterbitkan oleh *University Of New Brunswick, Canada,* yang berisi 12 kategori malware dengan 143 fitur yang menjadi karakteristik dari setiap kategori malware. Dataset yang terbaru sangat disarankan untuk digunakan pada klasifikasi *malware android,* karena serangan ataupun fitur dari *malware Android* akan selalu berkembang (Sitorus et al., 2021). Algoritma RF dapat dikembangkan dengan menerapkan beberapa *features* untuk pembangunan sistem klasifikasi yang lebih baik meskipun algoritma ini merupakan model partisirekursif yang bergantung pada partisi data karena ia bekerja pada pemisahan nilai fitur dan tidak melakukan perhitungan di dalamnya. Hal tersebut bisa dicoba pada kumpulan data yang lebih besar dan terbaru untuk melihat performa model (Tjahjadi & Santoso, 2023). Teknik *ensemble learning* sangat disarankan untuk digunakan pada klasifikasi dengan dataset yang besar dan kompleks, karena di nilai dari hasil yang didapat lebih baik jika dibandingkan algoritma klasifikasi tunggal (Zakariya & Ramli, 2023).

Algoritma *Random Forest* (RF) merupakan salah satu algoritma machine learning yang sudah umum diterapkan pada proses klasifikasi dengan hasil performa yang sangat baik menurut penelitian terdahulu. Algoritma RF memiliki kelebihan yaitu dapat mengatasi *overfitting*, menangani data yang tidak seimbang, mengatasi variabel yang tidak relevan, skalabilitas dan memiliki fleksibilitas dalam tipe data apapun. Penelitian sebelumnya menyarankan untuk melakukan klasifikasi pada dataset *malware* yang terbaru, oleh karena itu pada penelitian ini digunakan dataset yang diterbitkan pada tahun 2020 yaitu CCCS-CIC-AndMal2020. Dataset ini berisi kumpulan data terbaru yang lebih besar dari dataset *malware* yang sudah ada sebelumnya. Minimnya penelitian mengenai klasifikasi malware android berdasarkan analisis fitur dinamis, menjadikan

peluang dan celah pada penlitian ini. Model *ensemble learning* sangat disarankan untuk digunakan pada klasifikasi dengan dataset yang besar dan kompleks menurut penelitian sebelumnya, karena di nilai dari hasil yang didapat lebih baik jika dibandingkan algoritma klasifikasi tunggal. Algoritma RF yang merupakan salah satu model *ensemble learning* akan diusulkan untuk melakukan proses klasifikasi karena dilihat dari penelitian sebelumnya, algoritma ini memiliki kemampuan yang sangat baik dalam pengklasifikasian dataset yang besar dan kompleks. Fokus pada penelitian ini terletak pada pengujian performa dari algoritma RF dalam proses klasifikasi *malware Android* pada dataset CCCS-CIC-AndMal2020. Melalui konsep yang lebih sederhana yaitu *preparation data, preprocessing data, classification report* dan *evaluasi,* penelitian ini akan menjadi pembanding dengan penelitian sebelumnya yang menerapkan metode *ensemble learning* dalam proses klasifikasi.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang , maka dapat dirumuskan masalah dalam penelitian ini, adalah sebagai berikut:

1. Bagaimana proses klasifikasi *malware android* menggunakan algoritma *Random Forest*?
2. Bagaimana performa algoritma *Random Forest* dalam klasifikasi *malware android*?

1.3 Tujuan Penelitian

Berdasarkan latar belakang masalah dan rumusan masalah, maka tujuan dari penelitian ini adalah:

1. Melakukan klasifikasi *malware android* menggunakan algoritma *Random Forest*
2. Mengukur performa algoritma *Random Forest* dalam klasifikasi *malware android*

1.4 Ruang Lingkup Penelitian

Terdapat beberapa batasan masalah yang digunakan agar penelitian dapat dilakukan secara spesifik. Adapun batasan masalah dalam penelitian yang dilakukan, adalah sebagai berikut :

1. Dataset yang digunakan dalam eksperimen adalah dataset dengan judul "CCCS-CIC-AndMal2020" yang diterbitkan pada tahun 2020 oleh *University Of New Brunswick, Canada,* bekerja sama dengan pusat keamanan siber dan institut keaman siber di Kanada (https: https://www.unb.ca/cic/datasets/andmal2020.html)
2. Klasifikasi *malware* yang dilakukan, berdasarkan hasil analisis fitur dinamis.
3. Implementasi model *Random Forest* menggunakan framework Scikit-learn
4. Pada tahap *preprocessing data* dilakukan tahapan seperti *missing data imputation, transformasi data, random oversampling (SMOTE), scaling* dan *split data*.
5. Pengujian performa dilakukan menggunakan *Confusion Matrix* dengan performa yang diukur yaitu *accuracy, precision, recal* dan *f1-score.*

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat meberikan manfaat bagi seluruh pihak yang terkait, diantaranya:

1. Kontibusi terhadap area penelitian mengenai bidang kajian *machine learning* untuk proses klasifikasi .
2. Kontribusi terhadap penelitian dalam penerapan metode klasifikasi menggunakan algoritma *Random Forest*.
3. Kontribusi terhadap klasifikasi *malware Android*.
4. Kontribusi terhadap pengembangan algoritma klasifikasi terhadap *malware android.*
5. Kontribusi terhadap pengetahuan dan pemahaman mengenai klasifikasi *malware Android* yang merupakan ancaman bagi individu maupun organisasi.

6. Kontribusi terhadap peningkatan identifikasi *malware Android* serta diharapkan memberi pemahaman lebih kepada pengguna mengenai karakteristik *malware* yang bisa menjadi ancaman pada sistem operasi *Android*

1.6 Metodologi Penelitian

Metodologi penelitian berisi mengenai waktu dan tempat penelitian, tahapan penelitian, pendekatan penelitian, jenis penelitian, variabel penelitian, serta objek penelitian. Metode penelitian yang digunakan untuk menyelesaikan penelitian ini diantaranya:

1. Studi Literatur

   Studi Literatur merupakan tahapan dalam mengeksplorasi area penelitian, konsep, teori serta data-data dari berbagai sumber yang relevan dengan penelitian.

2. Perancangan Persoalan Penelitian

   Perancangan persoalan penelitian merupakan tahapan analisis untuk mengidentifikasi kesenjangan, kekurangan dan saran dari penelitian-penelitian terdahulu.

3. *Preparation Data*

   Preparation Data merupakan tahapan pertama sebelum dilakukan eksperimen, dataset yang disiapkan akan melalui tahapan *preprocessing* terlebih dahulu sebelum masuk ke proses pengembangan dan pengujian model

4. *Preprocessing Data*

   Preprocessing data merupakan tahapan persiapan sebelum melakukan eksperimen, beberapa tugas dilakukan pada tahapan *preprocessing* yaitu *missing data imputation, transformasi data, random oversampling (SMOTE), scaling* dan *split data*

5. Pengembangan Model

   Pengembangan model merupakan tahapan untuk membangun dan melatih algoritma RF yang akan digunakan dalam proses pengklasifikasian *malware*

Android. Data training akan digunakan untuk melatih model dalam proses klasifikasi

6. Evaluasi Model

   Evaluasi model merupakan tahapan pengukuran hasil yang dilakukan untuk mengetahui performa hasil klasifikasi dan evaluasi c*onfusion matrix* dari model yang sudah dikembangkan, kemudian akan di analisis dan dilakukan penarikan kesimpulan dari hasil evaluasi yang didapat

7. Penarikan Kesimpulan

   Tahapan ini merupakan tahapan akhir dari penelitian yang menyimpulkan performa hasil klasifikasi *malware Android* menggunakan algoritma RF.

1.7 Sistematika Penulisan

Sistematika penulisan digunakan dengan maksud agar penulisan laporan penelitian dapat terarah dan tersusun sesuai tahapan penelitian. Sistematika yang digunakan dalam penelitian ini yaitu sebagai berikut::

BAB I PENDAHULUAN

Bab ini berisi mengenai latar belakang atau dasar dilakukannya penelitian, rumusan permasalahan yang akan diteliti, tujuan penelitian, manfaat dari dilakukannya penelitian, metodologi penelitian dan bagaimana sistematika penulisan untuk melaporkan penelitian

BAB II TINJAUAN PUSTAKA

Bab ini berisi pembahasan teori - teori yang berhubungan dengan penelitian seperti konsep, metode dan algoritma yang digunakan di penelitian ini. Pada bab ini juga berisi penjelasan dari penelitian sebelumnya yang relevan dan penjelasan tentang keterbaruan penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Bab ini berisi uraian metode yang digunakan dalam melakukan penelitian, mulai dari waktu dan tempat penelitian, objek penelitian, variabel penelitian, matriks penelitian serta tahapan-tahapan yang dilakukan pada penelitian ini.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi pemaparan hasil serta pembahasan terhadap perancangan pada bab sebelumnya. Dalam pembahasan tersebut terdiri atas pembuatan rancangan model dan algoritma yang akan digunakan, serta eksperimen yang dilakukan bersamaan dengan model lainnya.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi mengenai kesimpulan dari hasil eksperimen yang dilakukan dan saran untuk penelitian selanjutnya berdasarkan batasan dan hasil penelitian yang membahas topik sejenis atau terkait.