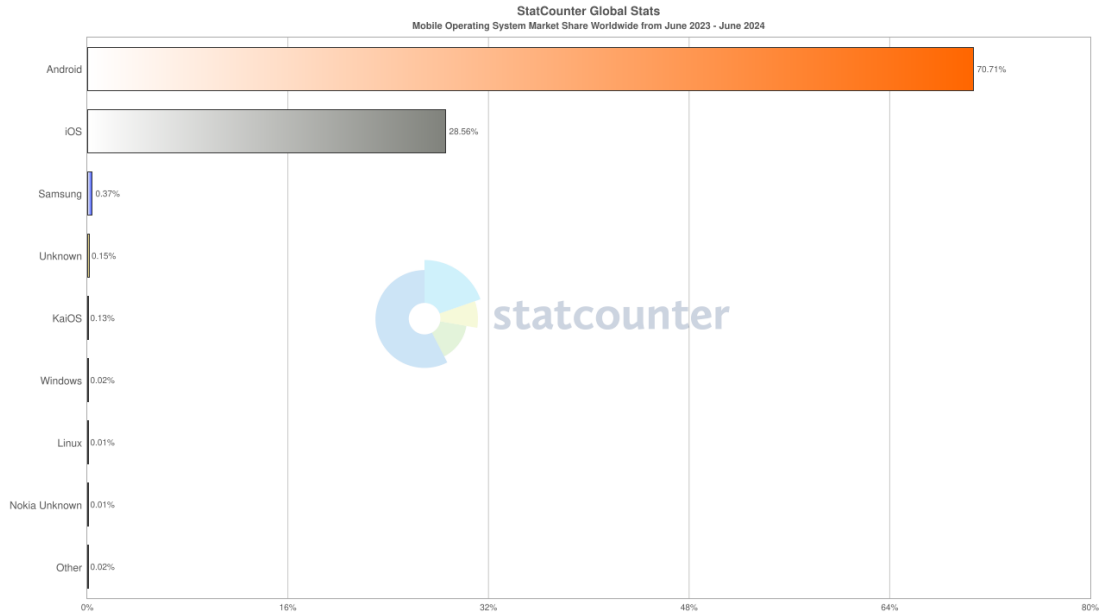
PENINGKATAN PERFORMA DAN PENERAPAN ALGORITMA RANDOM FOREST UNTUK DETEKSI MALWARE PADA ANDROID

BAB I

PENDAHULUAN

1.1Latar Belakang

Penggunaan perangkat mobile berbasis Android terus meningkat pesat, menjadikannya target utama bagi berbagai jenis serangan siber, khususnya malware. Menurut analisis Statcounter(2024) , Android menguasai lebih dari 70% pangsa pasar perangkat seluler global, sehingga rentan terhadap ancaman malware yang disebarkan melalui aplikasi berbahaya. Serangan malware pada perangkat Android berpotensi membahayakan privasi pengguna, memberikan akses tidak sah ke perangkat, dan mencuri data pribadi. Strategi baru yang lebih mudah beradaptasi diperlukan karena sistem deteksi malware tradisional yang mengandalkan teknik deteksi berbasis tanda tangan mengalami kesulitan dalam mengidentifikasi serangan baru atau zero-day.



Gambar 1.1 Market share held by mobile operating systems in 2023-2024

Machine learning dapat mengenali pola aktivitas malware dan membedakannya dari aplikasi yang aman, algoritma ini sangat menjanjikan untuk memecahkan masalah mengenai pola aktivitas malware(Thorat et al., 2024). Salah satu algoritma yang dapat digunakan yaitu Random Forest , yang dapat mendeteksi malware dengan akurasi tinggi dan dapat mengelola data yang sangat besar dan kompleks(Rafrastara et al., 2023). Dalam upaya untuk meningkatkan ketepatan dan efektivitas sistem pendeteksi malware, penelitian ini berfokus pada penerapan dan peningkatan kinerja algoritma Random Forest dalam mengidentifikasi *malware* pada Android. Informasi yang digunakan dalam penelitian ini dikumpulkan dari berbagai Android yang `aman dan bebas malware dan tersedia sebagai dataset terbuka pada Android Permision Dataset. Dengan optimalisasi algoritma ini, diharapkan sistem deteksi malware dapat memberikan perlindungan lebih baik bagi pengguna Android tanpa mengorbankan kinerja perangkat

Pada salah satu riset yang pernah dilakukan sebelumnya, Penelitian ini menggunakan analisis statis terhadap file Windows Portable Executable (PE) untuk menilai seberapa baik algoritma Random Forest dan Support Vector Machine (SVM) dalam mendeteksi malware. Algoritma Random Forest menunjukkan akurasi yang mengesankan sebesar 98,53%, sementara SVM mencatat akurasi sedikit lebih rendah yaitu 97,14%, menurut studi yang menggunakan dataset file PE terkait malware dan file aman. Berdasarkan temuan ini, Random Forest merupakan pilihan yang lebih baik untuk deteksi malware dalam konteks penelitian ini karena lebih berhasil dalam mengidentifikasi file PE sebagai aman atau jahat. (Ismail et al., 2024)

Pada penelitian sebelumnya menunjukkan bahwa setelah menggunakan pendekatan feature selection, algoritma K-Nearest Neighbor (KNN) mampu mengklasifikasikan malware dan program jinak pada perangkat Android dengan akurasi 77%. Dengan tidak adanya seleksi fitur, akurasi yang dicapai hanya 44%. Studi ini juga menemukan bahwa 80% data harus digunakan untuk pelatihan dan 20% untuk pengujian dalam hal pelatihan dan pengujian. Selain itu, penelitian ini menyoroti pentingnya preprocessing dataset untuk meningkatkan kinerja model dan menyarankan investigasi tambahan untuk mengkategorikan malware selain aplikasi Android. (Chitayae et al., 2023)

Terdapat sejumlah permasalahan yang signifikan dalam penerapan machine learning untuk mengidentifikasi Malware pada Android, pada penelitian “Peningkatan Performa dan Penerapan Algoritma Random Forest untuk Deteksi Malware di Android.” Salah satu tantangan

terbesar adalah kinerja algoritma di lingkungan perangkat mobile yang terbatas, seperti CPU, memori, dan daya baterai. Meskipun algoritma Random Forest memiliki kemampuan klasifikasi yang kuat, penerapannya dalam mendeteksi malware secara real-time di Android dapat menyebabkan penurunan performa akibat keterbatasan sumber daya tersebut.(Lakshmanarao & Shashi, 2022)

keberagaman dan kualitas data malware juga merupakan masalah yang signifikan. Algoritma mengalami kesulitan untuk mengidentifikasi keberagaman malware baru karena sejumlah besar dataset yang digunakan untuk pelatihan model tidak memiliki keterwakilan yang memadai. Overfitting, yaitu ketika model cocok dengan data pelatihan dengan sangat baik sehingga kinerjanya menurun Ketika dihadapkan dengan data baru. Teknik Random Forest sering menghadapi masalah ini karena mereka memiliki sejumlah besar parameter yang perlu diatur secara optimal. Waktu pendeteksian yang lambat adalah salah satu masalah tambahan. Deteksi malware yang cepat terhambat oleh waktu komputasi Random Forest yang relatif lebih lama karena banyaknya pohon keputusan yang harus diproses. Selain itu, model ini mengalami kesulitan dalam mengidentifikasi malware yang menyembunyikan aktivitas berbahaya karena taktik penyamaran yang digunakan oleh pembuatnya.

Sebagai solusi, penelitian ini akan mengurangi jumlah fitur yang tidak berguna dan memangkas pohon keputusan untuk mempercepat proses prediksi, membuat algoritma Random Forest lebih efektif di lingkungan Android. Selain itu, malware yang menggunakan taktik penyamaran dapat ditemukan dengan menggunakan pendekatan ekstraksi fitur berbasis perilaku. Metode ini berkonsentrasi pada pemeriksaan perilaku aplikasi setelah instalasi, termasuk interaksi dengan API dan pola izin. Validasi K-Fold akan digunakan untuk mengurangi overfitting dan memastikan model berkinerja baik pada dataset malicious website yang lebih besar dan beragam. Disarankan juga agar batch processing atau deteksi berbasis cloud digunakan untuk mempersingkat waktu deteksi ketika analisis ekstensif dilakukan di server dengan sumber daya yang lebih besar. Diharapkan metode ini akan sangat meningkatkan kecepatan, akurasi, dan efisiensi pemanfaatan sumber daya untuk deteksi malware di aplikasi Android.

1.2Rumusan Masalah

1. Bagaimana cara menangani permasalahan overfitting pada model Random Forest yang digunakan untuk deteksi malware Android?

2. Bagaimana meminimalkan waktu deteksi malware tanpa mengurangi akurasi prediksi ?
3. Bagaimana meningkatkan representasi data malware dalam dataset yang digunakan agar dapat mencakup keberagaman malware yang lebih luas?
4. Bagaimana meningkatkan kinerja algoritma Random Forest dalam mendeteksi malware pada aplikasi Android tanpa mengorbankan performa perangkat?

1.3 Tujuan Penelitian

Dengan mengatasi sejumlah masalah signifikan, termasuk overfitting dan waktu deteksi yang lambat, penelitian ini bertujuan untuk meningkatkan efektivitas algoritma Random Forest dalam mengidentifikasi malware pada aplikasi Android. Melalui penggunaan prosedur validasi dan penyesuaian parameter model yang lebih ideal, penelitian ini berusaha untuk mengidentifikasi strategi praktis untuk mengurangi overfitting. Penelitian ini juga akan menyelidiki cara untuk mempercepat waktu deteksi malware tanpa mengorbankan akurasi prediksi, seperti pengurangan fitur dan optimasi pohon keputusan. Selain itu, penelitian ini juga fokus pada meningkatkan representasi data malware dalam dataset yang digunakan, agar model dapat mengenali lebih banyak variasi malware yang baru ditemukan. Selain itu, diharapkan algoritma Random Forest dapat dioptimalkan untuk bekerja secara efisien di lingkungan perangkat Android dengan keterbatasan sumber daya seperti CPU, memori, dan daya baterai.

1.4 Manfaat Penelitian

Manfaat penelitian ini ialah bahwa peneliti secara signifikan memajukan pembuatan sistem deteksi malware Android yang lebih akurat dan efisien. Penelitian ini akan meningkatkan keamanan aplikasi Android dengan mengatasi masalah overfitting dan menciptakan model yang lebih umum yang lebih mampu mengidentifikasi ancaman baru. Selain itu, sistem deteksi malware akan lebih responsif dan cocok untuk penggunaan real-time di Android jika waktu deteksi dioptimalkan tanpa mengorbankan akurasi prediksi. Peningkatan representasi data malware juga akan membantu model untuk lebih efektif dalam menghadapi beragam jenis malware yang terus berkembang. Penelitian ini juga memberikan solusi untuk meningkatkan efisiensi algoritma Random Forest dalam lingkungan perangkat mobile yang terbatas, sehingga pengguna dapat menikmati perlindungan yang lebih baik tanpa mengorbankan kinerja perangkat.

1.5 Batasan Masalah

1. Penelitian ini hanya menggunakan dataset malware yang tersedia secara terbuka, yaitu *Android Permission Dataset*.
2. Penelitian ini hanya berfokus pada optimasi dan peningkatan kinerja algoritma Random Forest.
3. Penelitian ini terbatas pada perangkat berbasis Android, dan tidak menguji algoritma di platform mobile lain seperti iOS, yang mungkin memiliki karakteristik sistem dan keamanan berbeda.
4. Penelitian ini difokuskan pada perangkat Android dengan sumber daya terbatas (CPU, memori, daya baterai), sehingga hasilnya mungkin tidak dapat diimplementasikan secara langsung pada perangkat Android dengan performa tinggi atau perangkat non-mobile seperti server.
5. Penelitian ini dilakukan dalam lingkungan simulasi dan tidak menguji penerapan deteksi malware dalam skala besar atau di lingkungan produksi yang sebenarnya, yang mungkin menghadapi tantangan tambahan seperti lalu lintas data besar dan berbagai kondisi jaringan.