KLASIFIKASI MALWARE PADA ANDROID MENGGUNAKAN METODE RANDOM FOREST

Nadhya Hassni

Jurusan Sistem Komputer,Universitas Sriwijaya, Palembang

Jl. Masjid Al Ghazali, Bukit Lama. Kec. Ilir Barat 1. Kota Palembang, Sumatera Sekatan, 30128, Indonesia

E-mail : nadhyahassni@gmail.com

Abstrak

Kemajuan teknologi membuat hamper semua lapisan masyarakat sudah mulai menggunakan teknologi, teknologi yang paling banyak digunakan saat ini adalah smartphone / android. Semakin banyak penggunaan teknologi ini maka semakin banyak juga orang-orang yang tidak bertanggung jawab yang ingin mengambil keuntungan illegal dari dampak kemajuan teknologi ini. Untuk mengatasi hal ini peneliti mencoba untuk mengklasifikasikan malware yang berbasis android agar dapat mencegah tindakan illegal yang digunakan oleh orang-orang yang tidak bertanggung jawab.

Kata kunci : Malware, Android, Security Management, Machine Learning

Abstrak

Advances in technology have made almost all walks of life have started using technology, the technology most widely used today is smartphone / android. More people use of this technology, more irresponsible people who want to take illegal profits from the impact of this technological progress. To overcome this, researchers try to classify malware based on Android in order to prevent illegal actions that are used by people who are not responsible.

Keywords: Malware, Android, Security Management, Machine Learning

1. Pendahuluan

Sistem android adalah salah satu system perangkat lunak seluler yang paling populer. Android memiliki system yang terbuka memunkinkan programmer untuk melakukan modifikasi bahkan pada tingkat system. Dengan ini ancaman keamanan system android lebih serius dibandingkan system lain. Dalam laporan keamanan *Mcafee* 97% dari 36.699 sampel malware pada 2012 didasarkan pada system android. Dilansir Alibaba pada tahun 2015, 18% peralatan android telah terjangkit virus dan 95% aplikasi populer telah dipalsukan. Terdapat 300 juta sample malware yang terdeteksi di platform android. Kebanyakan dari mereka adalah hooliganisme dengan 52.4% termasuk Popup anonym, iklan yang mendorong untk melakukan hal mengerikan. Varietas lainya adalah pembajakan sms, perusakan system dan penipuan[1]

sehingga perlu untuk menemukan cara yang efektif untuk mengkasifikasikan aplikasi yang dipakai pada android mengandung malware atau tidak. Saat ini berbagai metode digunakan dalam pendeteksian malware didasarkan pada metode pendeteksian virus tradisional gerdasarkan *perilaku* dari aplikasi tersebut.

Pada penelitian sebelumnya dimana peneliti menggunakan metode contrasting permission patterns. Dalam penelitian ini peneliti menggunakan beberapa metode untuk di bandingan yaitu Random Tree dan RBFNetwork. Random Tree menghasilkan akurasi sebesar 87.66% sedangkan RBFNetwork menghasilkan akurasi sebesar 85.78%. pada penelitian ini hasil akurasi cukup besar tetapi masih berada di bawah 90%[3]

Dari dua penelitian yang sudah di lakukan sebelumnya dengan metode – metode yang berbeda dan tada set yang berbeda sudah menunjukan hasil akurasi yang cukup baik. Tetapi dibutuhkan metode lain untuk mengklasifikasikan malwre pada android yang diharapkan dapat menghasilkan akurasi yang besar dengan waktu pengeksekusian yang singkat.

2. Tinjauan Pustaka

a. Android

Android merupakan sistem operasi yang digunakan untuk perangkat mobile berbasis Linux. Pada awalnya sistem operasi ini dikembangkan oleh Android.Inc, yang kemudian dibeli oleh Google pada tahun 2005. Android mengembangka usaha pada tahun 2007 dibentuklah Open Handset Alliance (OHA), sebuah konsorsium dari beberapa perusahaan, yaitu Texas Instrument, Broadcom Corporation, Google, HTC, Intel, LG, Marvell Technology Group, Motorola, Nvidia, Qualcom, Samsung Electronics, Sprint Nextel, dan T-Mobile dengan tujuan untuk mengembangkan standar terbuka untuk perangkat mobile Smartphone[3][2]

b. Malware

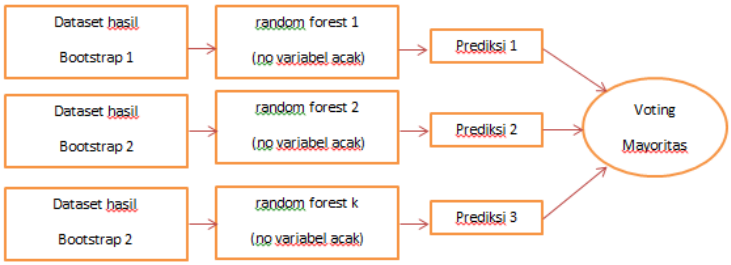
Malware adalah singkatan dari Malicious Ware yang berarti perangkat lunak yang dirancang untuk mengganggu kerja dari sebuah sistem komputer. Perangkat lunak ini diperintahkan untuk melakukan perubahan diluar kewajaran kerja dari system komputer. Malware biasanya menyusup pada sistem jaringan komputer tanpa diketahui oleh pemilik jaringan komputer, dari jaringan komputer ini malware tersebut akan memasuki sebuah sistem komputer. Pemilik komputer juga tidak mengetahui bahwa komputernya telah disusupi oleh malware. Tujuan seseorang untuk menyusupkan program jahat bisa bermacam-macam, mulai hanya sekedar iseng ingin mencoba kemampuan, merusak data, mencuri data, sampai menguasai computer orang lain dan mengendalikannya dari jarak jauh melalui jaringan computer[4][5]

c. Random Forest

Metode random forest mulai banyak diperbincangkan sejak tulisan Breiman (2001) muncul pada jurnal Machine Learning. Liaw dan Wiener (2002) menyatakan bahwa random forest dikembangkan oleh Breiman (2001). Secara jeli, Breiman (2001) berupaya untuk memperbaiki proses pendugaan yang dilakukan menggunakan metode bagging. Metode random forest adalah pengembangn dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection. Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Secara sederhana,

algoritma pembentukan random forest dapat disebutkan sebagai berikut. Andaikan data training yang kita miliki berukuran n dan terdiri atas p variabel penjelas (prediktor). Tahapan penyusunan dan pendugaan menggunakan random forest adalah[6]

- (Tahapan bootstrap) tarik sampel acak dengan pengembalian berukuran n dari data training
- Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pengembalian). Susun pohon berdasarkan data tersebut, namun pada setiap proses pemisahan pilih secara acak m < p peubah penjelas, dan dilakukan pemisahan terbaik (tahapan random sub-setting).
- Ulangi langkah 1-2 sebanyak l kali sehingga terbentuk sebuah hutan yang terdiri atas l pohon.
- Lakukan pendugaan gabungan berdasarkan l buah pohon tersebut (misal menggunakan majority vote untuk kasus klasifikasi atau rata-rata untuk kasus regresi)



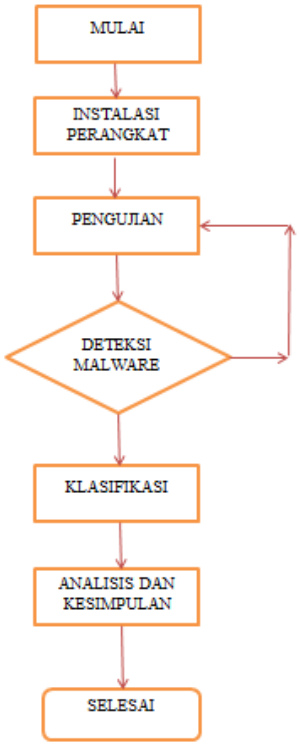
Gambar.1 [6]

3. Metodologi Penelitian
   a. Pengenalan dataset

      Data set yang digunakan didapatkan dari kaggle.com, ata diperoleh dengan proses yang terdiri dari membuat biner vektor izin yang digunakan untuk setiap aplikasi yang dianalisis {1 = digunakan, 0 = tidak digunakan}. Selain itu, sampel malware / jinak dibagi berdasarkan "Type"; 1 malware dan 0 non-malware.

   b. kerangka kerja penelitian.

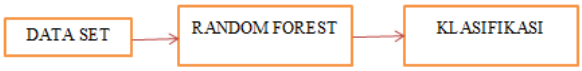
      Penelitian ini melalui beberapa tahapan,. Tahapan yang dibuat mengikuti kerangka kerja yang telah di rancang sehingga penelitian ini terstruktur dan mengikuti alur. Tahapan awal penelitian ini adalah melakukan pemilahan data – data yang ada pada data set, yang akan di jadikan focus klasifikasi. Kemudian melakukan pengajuan pada data yang sudah di seleksi untuk mendapatkan data dan besar persentasi klasifikasi yang di hasilkan.diagram alir dar kerangka kerja di tunjukan pada gambar di bawah ini:



Gambar.2.

c. Skenario pengujian

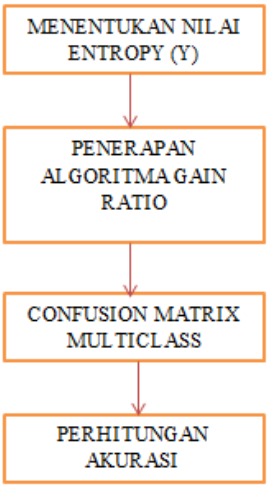
Pengujian penelitian ini daoat berjalan sesuai dengan tujuan dikarenkan adanya sebuah scenario yang baik agar tahap pengujian dapat berjalan baik dan terarah juga lebih teratur.



Gambar.3

d. random forest

Pada penelitian ini metode random forest akan mendeteksi malware pada system android dengan klasfkasi label tertentu. Metode random forest akan mengkasifikasikan seberapa besar persentase malware pada system android. Diagram alir proses metode random forest dapat dilihat pada gambar di bawah ini:



Gambar.4

4. Hasil dan Pembahasan
   a. Visualisasi data set

um.chrome.shell.permission.SANDBOX	org.chromium.chromecast.shell.permission.SANDBOX	org.chromium.content_shell.permission.SANDBOX	test_permission	type
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

Gambar.5

Data yang digunakan sebanyak 5 baris dan 331 colom

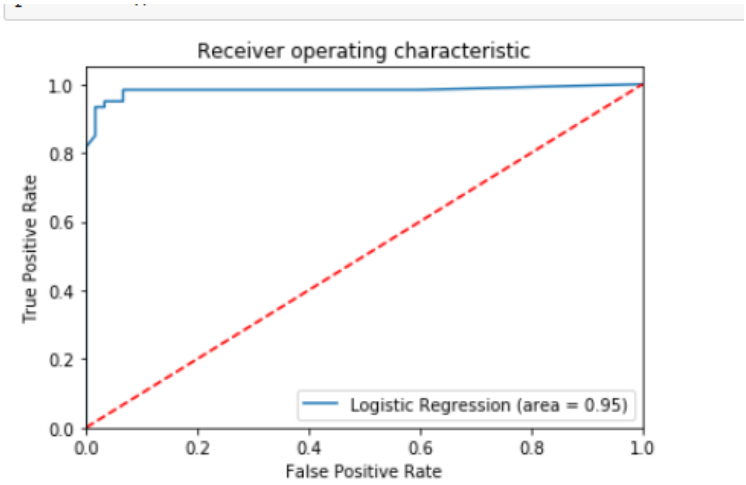
Dengan perbandingan 50% data malware dan 50% data normal.

   b. Hasil klasifikasi
   Dengan menggunakan metode random forest di dapatkan hasil akurasi sebesar 95% dengan sensitifitas sebesar 98% dan presisi sebesar 98%

```
Accuracy training 0.95
Specificity training 0.9833333333333333
Sensitivity training 0.9166666666666666
Precision training 0.9821428571428571
F1 Score training 0.9482758620689654
```

Gambar.6

Grafik ROC hasil traning akurasi



Gambar .7

5. Kesimpulan

Dapat dilihat dari hasil visualisasi dan hasil training sebesar 95% dapat sesuai tujuan awal dimana peneliti mengharapkan hasil training yang lebih besar dari 90%. Dengan demikian tergapai tujuan peneliti yang ingin membuktikan bahwa metode random forest dapat lebih baik di banding dengan beberapa metode lain tetapi bukan mungkin beberapa metode terbaru dapat mengasilkan akurasi yang lebih tinggi sesuai dengan metode, data set dan fitur yang digunakan.

REFRESNSI

[1] H. Jin *et al.*, “Analyzing and Recognizing Android Malware via Semantic-Based Malware Gene,” in *2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2017, pp. 17–20.

[2] P. Ravi Kiran Varma, K. P. Raj, and K. V. Subba Raju, “Android mobile security by detecting and classification of malware based on permissions using machine learning algorithms,” *Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017*, pp. 294–299, 2017, doi: 10.1109/I-SMAC.2017.8058358.

[3] P. Xiong, X. Wang, W. Niu, T. Zhu, and G. Li, “Android malware detection with contrasting permission patterns,” *China Commun.*, vol. 11, no. 8, pp. 1–14, 2014, doi: 10.1109/CC.2014.6911083.

[4] R. Grewal, “A hybrid approach of malware detection in Android,” 2017.

[5] A. Bettany and M. Halsey, “What Is Malware?,” in *Windows Virus and Malware Troubleshooting*, Springer, 2017, pp. 1–8.

[6] S. S. Pangastuti, “Perbandingan Metode Ensemble Random Forest Dengan Smote-Boosting Dan Smote-Bagging Pada Klasifikasi Data Mining Untuk Kelas Imbalance (Studi Kasus: Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur)-A Comparison Of The Ensemble Random Forest Methods With Smote-Boosting And Smote-Bagging On Data Mining Classification For Imbalance Class,” Institut Teknologi Sepuluh Nopember, 2018.