E-ISSN : 2623-1700
P-ISSN : 1693-3672

Analisis Static Malware Menggunakan Algoritma Random Forest Machine Learning.

Dimas Satrya Bhayangkara[1], Hasvid Dwi Putranto[2], Farhan Toriq[3] , Febri Wijayanto[4]

Prodi Teknik Komputer, Fakultas Ilmu Komputer

Universitas Amikom Yogyakarta

dimasatrya@students.amikom.ac.id

Abstrak

Malware adalah program software berbahaya yang bertujuan untuk merusak atau mencuri data dari sistem korban. Dibutuhkan malware analisis untuk mengetahui kinerja dari malware. Analisis static digunakan untuk ekstraksi fitur dari total 10 PEfile 5 file malware dan 5 benign file. Algoritma random forest pembelajaran mesin didapatkan akurasi 99.99 %.

Kata kunci: malware, analisis malware, analisis statis, random forest, pembelajaran mesin.

Abstract

Malware is a malicious software program that aims to damage or steal data from a victim's system. Malware analysis is needed to determine the performance of the malware. Static analysis was used for feature extraction from a total of 10 PEfiles, 5 malware files and 5 benign files. The random forest machine learning algorithm obtained an accuracy of 99.99%.

Keywords : *malware, malware analysis, static analysis, random forest, machine learning.*

E-ISSN : 2623-1700
P-ISSN : 1693-3672

PENDAHULUAN

Penggunaan dari internet saat ini tersebar cukup luas dan banyak orang menggunakannya setiap hari[1]. Seiring dengan perkembangan internet ini internet juga terkena dampak negatif dari cyber attack atau serangan siber[2]. Salah satu serangan tersebut adalah malware. Malware adalah program perangkat lunak berbahaya yang digunakan untuk menjalankan perangkat lunak berbasis kode biner dan niat malware didefinisikan sebagai tindakan merusak atau mencuri dari sistem komputer dengan memanfaatkan infrastruktur yang rentan[3].

Berdasarkan perilaku dan cara kerjanya malware dibagi ke dalam berbagai kategori, antara lain worm, trojan, virus, backdoor, spyware, etc[4]. Dengan adanya serangan malware ini dibutuhkan analisis untuk mengetahui kinerja dari malware tersebut. Analisis malware adalah sebuah metode untuk memeriksa struktur dan perilaku dari malware dengan mengidentifikasi karakteristik yang menunjukan tujuan jahatnya[5]. Umumnya terdapat dua teknik atau metode analisis yaitu analisis static dan dynamic[6]

Analisis static atau biasa juga dikenal dengan analisis kode malware, adalah sebuah proses atau metode untuk mempelajari kinerja dari malware dengan mengamati dan analisis kode program dari malware[7]. Dengan menggunakan analisis static fitur PORTABLE EXECUTABLE (PE) bagian header file diektraks menggunakan disassembly tool[8]. Tool yang digunakan adalah PEfile. Pefile adalah sebuah library dari python yang digunakan untuk mengekstraksi fitur dari Portable Executable (PE) files[3][6].

Dalam jurnal ini bertujuan untuk deteksi dan akurasi dari analisis static dengan menggunakan algoritma machine learning random forest. Algoritma random forest ini bagian dari algoritma decision tree. Decision tree adalah algoritma dari machine learning dengan terus membagi data berdasarkan parameter tertentu decision tree ini digunakan untuk klasifikasi dan random forest adalah algoritma versi ensemble dari decision tree[9]. Dengan data total 10 executable files 5 malware files dan 5 benign files atau file jinak diekstraksi menggunakan PEfile python. Terdapat 2 struktur PE files yaitu PE header dan section[7].

METODE

Dalam penelitian ini metode yang digunakan adalah kualitatif dan kuantitatif. Kualitatif digunakan analisis static dengan menggunakan PEfile library python untuk ekstraksi fitur dari data malware files dan benign files. Kuantitatif digunakan pada algoritma machine learning dengan uji persentase dari algoritma random forest.

A. Data collection

Dataset yang diambil adalah 2 data yaitu malware files dan benign files atau file jinak. Data pada penelitian ini diambil dari github dengan total 10 PE files, 5 malware files dan 5 benign files dan format executables files yaitu .exe

B. Analisis static

Dalam analisis static ini menggunakan environment yaitu dengan menggunakan virtual box dengan OS REMNUX. Dengan menggunakan analisis static ini melakukan ekstraksi fitur dari malware files dan benign files dengan menggunakan pefile library python untuk mendapatkan informasi fitur dari executable PE files. Mengambil data dari Pe headers yang berbeda, termasuk file header, optional

E-ISSN : 2623-1700
P-ISSN : 1693-3672

header, dan section. Pada bagian file header diambil informasi fitur seperti Machine, number_of_section, compile_date. Pada bagian optional header diambil fitur seperti *DebugSize, DebugRVA, DebugRVA, OSVersion, ExportRVA, ExportSize, IATRVA, ResSize, LinkerVersion, StackReserveSize, dan DLLCharacteristic*. Section name, address, size, dan informasi terkait lainnya semuanya terdapat pada bagian section atau section header, dalam bagian section header ini biasanya atau standarnya terdapat berbagai fitur seperti *.text, .bss, .rdata, .data, .rsrc, .edata, .idata, .stabstr, .crt, .debug, .reloc, .stab, .pdata, .tls, .exptbl*, *.rodata*. Namun dalam dataset yang diambil jarang ditemukan untuk bagian section ini, maka dalam penelitian ini diambil pada bagian file header dan optional header

HASIL DAN PEMBAHASAN

Dengan fitur diekstraksi dari data malware menggunakan pe-file library python dan pada bagian ini algoritma random forest digunakan.

- True Positive (TP): jumlah file berbahaya yang diklasifikasikan dengan benar
- True Negative (TN): jumlah file jinak yang diklasifikasikan dengan benar.
- False Positive (FP): jumlah file jinak salah diklasifikasikan sebagai file berbahaya.
- False Negative (FN): jumlah file berbahaya salah diklasifikasikan sebagai file jinak.
- Accuracy: = TP+TN / TP+TN+FP+FN

Dengan menggunakan perhitungan tersebut menggunakan machine learning dan dengan algoritma random forest maka mendapat laporan seperti dibawah ini :

TN	FP
FN	TP

5	386
1	296

Normalisasi confusing matrix

0.01	0.99
0	1.0

Dan dibawah ini report dari algoritma random forest Confusion matrix

391	0
1	297

Normalisasi confusing matrix

1.0	0
0	1.0

Dengan menggunakan perhitungan machine learning dan dengan algoritma random forest maka Dalam penelitian ini mendapat AUC algoritma random forest 99.99 % karena mendekati 1.0000

KESIMPULAN DAN SARAN

Pada penelitian ini difokuskan pada mengukur deteksi dan akurasi pada analisis static dengan melakukan ekstraksi fitur berdasarkan informasi PE files atau executable files dan menggunakan algoritma random forest machine

E-ISSN : 2623-1700
P-ISSN : 1693-3672

learning untuk mendeteksi malware dan benign files. Hasil dari algoritma random forest pada machine learning mendapatkan akurasi 1.000 dan AUC (Area Under Curve) 99.99 %. Terbukti dari percobaan bahwa menggunakan analisis statis berdasarkan informasi PE dan memilih fitur data terkait juga dapat memberikan akurasi deteksi terbaik dan menggambarkan malware secara akurat. Dan untuk penelitian selanjutnya diperbanyak untuk dataset dan menggunakan analisis dinamis dan analisis hybrid, untuk machine learning ditambah untuk algoritma.

DAFTAR PUSTAKA

[1] S. Agarkar and S. Ghosh, "Malware detection & classification using machine learning," in *Proceedings - 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security, iSSSC 2020*, Dec. 2020. doi: 10.1109/iSSSC50941.2020.9358835.

[2] A. R. Mohammed, G. Sai Viswanath, K. Sai babu, and T. Anuradha, "Malware Detection in Executable Files Using Machine Learning," 2020, pp. 277–284. doi: 10.1007/978-3-030-24322-7_36.

[3] M. Zafar-uz-Zaman, National Centre for Physics, Centres of Excellence in Science & Applied Technologies, Institute of Electrical and Electronics Engineers. Islamabad Section, and Institute of Electrical and Electronics Engineers, *Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) : 8th-12th January, 2019*.

[4] J. Singh and J. Singh, "A survey on machine learning-based malware detection in executable files," *Journal of Systems Architecture*, vol. 112. Elsevier B.V., Jan. 01, 2021. doi: 10.1016/j.sysarc.2020.101861.

[5] N. Balram, G. Hsieh, and C. McFall, "Static malware analysis using machine learning algorithms on apt1 dataset with string and PE header features," in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, Dec. 2019, pp. 90–95. doi: 10.1109/CSCI49370.2019.00022.

[6] B. Wang, Z. Zhang, and F. Zhang, "Subliminal channels in the code-based ring signature scheme," in *Proceedings - 2019 14th Asia Joint Conference on Information Security, AsiaJCIS 2019*, Aug. 2019, pp. 146–150. doi: 10.1109/AsiaJCIS.2019.00011.

[7] N. S. Selamat and F. H. M. Ali, "Comparison of malware detection techniques using machine learning algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 1, pp. 435–440, Oct. 2019, doi: 10.11591/ijeecs.v16.i1.pp435-440.

[8] Rahul, P. Kedia, S. Sarangi, and Monika, "Analysis of machine learning models for malware detection," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 2, pp. 395–407, Feb. 2020, doi: 10.1080/09720529.2020.1721870.

E-ISSN : 2623-1700
P-ISSN : 1693-3672

[9] Institute of Electrical and Electronics Engineers and Manav Rachna International Institute of Research and Studies, *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing : trends, prespectives and prospects : COMITCON-2019 : 14th-16th February, 2019*.

[10] *2019 IEEE 9th International Conference on Advanced Computing (IACC)*. IEEE, 2019.

[11] I. Muhamad, M. Matin, and B. Rahardjo, "Malware Detection Using Honeypot and Machine Learning." [Online]. Available: https://public.gdatasoftware.com/Presse/Publikationen/Malware_Rep

[12] R. Sihwail, K. Omar, K. A. Z. Ariffin, and S. al Afghani, "Malware detection approach based on artifacts in memory image and dynamic analysis," *Applied Sciences (Switzerland)*, vol. 9, no. 18, Sep. 2019, doi: 10.3390/app9183680.