DETEKSI *MALWARE* PADA *FILE PORTABLE DOCUMENT FORMAT* (PDF) DENGAN *BYTE FREQUENCY DISTRIBUTION* (BFD) DAN PENDEKATAN *SUPPORT VECTOR MACHINE* (SVM)

TESIS

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister

OLEH :

HERU SAPUTRA

09012682125008

PROGRAM STUDI MAGISTER ILMU KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2024

LEMBARAN PENGESAHAN

DETEKSI MALWARE PADA FILE PORTABLE DOCUMENT FORMAT (PDF) DENGAN BYTE FREQUENCY DISTRIBUTION (BFD) DAN PENDEKATAN SUPPORT VECTOR MACHINE (SVM)

TESIS

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister

OLEH :

HERU SAPUTRA
09012682125008

Palembang, Januari 2024

Pembimbing I	**Pembimbing II**
Prof. Deris Stiawan, M.T., Ph.D.	**Hadipurnawan Satria, Ph.D.**
NIP. 197806172006041002	**NIP. 198004182020121001**

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

HALAMAN PERSETUJUAN

Pada hari senin tanggal 8 Januari 2024 telah dilaksanakan ujian sidang tesis oleh Magister Ilmu Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Heru Saputra

NIM : 09012682125008

Judul : Deteksi *Malware* Pada *File Portable Document Format* (PDF) dengan *Byte Frequency Distribution* (BFD) dan Pendekatan *Support Vector Machine* (SVM)

1. Pembimbing I

   Prof. Deris Stiawan, M.T., Ph.D. .....
   NIP. 197806172006041002

2. Pembimbing II

   Hadipurnawan Satria, Ph.D. .....
   NIP. 198004182020121001

3. Penguji I

   Dian Palupi Rini, M. Kom., Ph.D. .....
   NIP. 197802232006042002

4. Penguji II

   Dr. Ahmad Zarkasi, M.T. .....
   NIP. 197908252023211007

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

LEMBAR PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	: Heru Saputra
NIM	: 09012682125008
Program Studi	: Magister Ilmu Komputer
Judul Tesis	: Deteksi *Malware* Pada *File Portable Document Format* (PDF) dengan *Byte Frequency Distribution* (BFD) dan Pendekatan *Support Vector Machine* (SVM)

Hasil Pengecekan Software iThenticate/Turnitin : **17 %**

Menyatakan bahwa laporan tesis saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan tesis ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, 29 Desember 2023

Heru Saputra

NIM. 09012682125008

KATA PENGANTAR

Puji dan syukur kehadirat Allah SWT karena atas rahmat-Nya penulis dapat menyelesaikan laporan tesis ini. Tesis yang berjudul "***Deteksi Malware* Pada *File Portable Document Format* (PDF) dengan *Byte Frequency Distribution* (BFD) dan Pendekatan *Support Vector Machine* (SVM)**" ini disusun untuk memperoleh gelar magister pada Program Studi Magister Ilmu Komputer Universitas Sriwijaya.

Pada kesempatan ini, penulis ingin menyampaikan ucapan terima kasih yang tak terhingga kepada pihak-pihak telah memberikan dukungan, bimbingan, motivasi dan kemauan kepada penulis untuk menyelesaikan tesis ini, yaitu kepada:

1. Kedua orang tua, Bapak Beny Rizal dan Ibu Yuliani serta Saudari-saudariku Heny Yuniarti, dan Hesti Melinda beserta seluruh keluarga atas semua bantuan yang tak dapat penulis hitung dan tuliskan satu persatu;
2. Istri, Dwi Yunita Sari dan Anak-anak tercinta, Alia Shanum Hediputri dan Arsyila Sabhira Hediputri yang selalu memberikan dukungan, motivasi, dan doanya.
3. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya;
4. Bapak Hadipurnawan Satria, Ph.D. selaku Ketua Program Studi Magister Ilmu Komputer.
5. Bapak Dr. Ali Ibrahim, S.Kom., M.T. selaku Ketua sidang.
6. Bapak Prof. Deris Stiawan, M.T., Ph.D. dan Bapak Hadipurnawan Satria, Ph.D. selaku pembimbing yang telah banyak memberikan bimbingan, masukan dan bantuan dalam proses penyelesaian tesis ini;
7. Ibu Dian Palupi Rini, M.Kom., Ph.D. dan Bapak Dr. Ahmad Zarkasi, M.T. selaku penguji yang telah banyak memberikan saran dan kata-kata yang membangun;
8. Bapak dan Ibu Dosen yang selama ini telah melimpahkan ilmunya kepada penulis selama proses belajar mengajar di Fakultas Ilmu Komputer Universitas Sriwijaya.
9. Seluruh teman-teman mahasiswa Magister Ilmu Komputer Angkatan 2021 dan seluruh teman-teman akademika Fakultas Ilmu Komputer.

10. Seluruh staf administrasi dan pegawai yang selalu membantu dan mendukung Penulis dalam hal administrasi perkuliahan.
11. Untuk semua pihak yang telah membantu dalam penyelesaian tugas akhir ini dan tidak dapat disebutkan satu-persatu.

Akhir kata, penulis menyadari bahwa tugas akhir ini jauh dari kata sempurna. Untuk itu penulis mengharapkan kritik dan saran yang membangun dari semua pihak untuk penyempurnaan laporan tesis ini dan semoga tesis ini dapat bermanfaat bagi pihak yang membutuhkan.

Palembang, Januari 2024

Penulis

MALWARE DETECTION IN PORTABLE DOCUMENT FORMAT (PDF) FILES WITH BYTE FREQUENCY DISTRIBUTION (BFD) AND SUPPORT VECTOR MACHINE (SVM)

Heru Saputra (09012682125008)
Dept. of Master Computer Science, Computer Science, Sriwijaya University
Email: saputra31.heru@gmail.com

ABSTRACT

Portable Document Format (PDF) files as well as files in several other formats such as (.docx, .hwp and .jpg) are often used to conduct cyber attacks. According to VirusTotal, PDF ranks fourth among document files that are frequently used to spread malware in 2020. Malware detection is challenging partly because of its ability to stay hidden and adapt its own code and thus requiring new smarter methods to detect. Therefore, outdated detection and classification methods become less effective. Nowadays, one of such methods that can be used to detect PDF files infected with malware is a machine learning approach. In this research, the Support Vector Machine (SVM) algorithm was used to detect PDF malware because of its ability to process non-linear data, and in some studies, SVM produces the best accuracy. In the process, the file was converted into byte format and then presented in Byte Frequency Distribution (BFD). To reduce the dimensions of the features, the Sequential Forward Selection (SFS) method was used. After the features are selected, the next stage is SVM to train the model. The performance obtained using the proposed method was quite good, as evidenced by the accuracy obtained in this study, which was 95.58% with an F1 score of 97.47%. The contributions of this research are new approaches to detect PDF malware which is using BFD and SVM algorithm, and using SFS to perform feature selection with the purpose of improving model performance. To this end, this proposed system can be an alternative to detect PDF malware.

Keywords: portable document format, malware, byte frequency distribution, sequential forward selection, support vector machine

DETEKSI MALWARE PADA *FILE PORTABLE DOCUMENT FORMAT* (PDF) DENGAN *BYTE FREQUENCY DISTRIBUTION* (BFD) DAN PENDEKATAN *SUPPORT VECTOR MACHINE* (SVM)

Heru Saputra (09012682125008)
Jurusan Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
Email: saputra31.heru@gmail.com

ABSTRAK

Portable Document Format (PDF) serta beberapa *file* lain seperti (.docx, .hwp dan .jpg) sangat sering digunakan untuk melakukan serangan siber. Berdasarkan VirusTotal PDF menempati urutan ke empat dalam file dokumen yang sering digunakan untuk menyebarkan malware pada tahun 2020. Deteksi *Malware* menjadi topik yang kompleks karena kemampuannya untuk tetap tersembunyi dan semakin canggih sehingga membutuhkan metode baru yang lebih baik untuk mendeteksinya. Oleh karena itu, metode deteksi dan klasifikasi yang telah using menjadi kurang efektif. Saat ini, salah satu metode yang digunakan untuk mendeteksi file PDF yang terinfeksi malware yaitu menggunakan pendekatan machine learning. Pada penelitian ini akan digunakan algoritma *Support Vector Machine* (SVM) untuk mendeteksi PDF *malware* karena kemampuannya dalam memproses data non-linear, dan pada beberapa penelitian, SVM menghasilkan akurasi yang paling baik. Pada prosesnya *file* akan dikonversi menjadi format *byte*, dan selanjutnya akan sajikan ke dalam *Byte Frequency Distribution* (BFD), untuk mengurangi dimensi fitur akan digunakan metode *Sequential Forward Selection* (SFS). Setelah dilakukan seleksi fitur, langkah berikutnya menggunakan SVM untuk melatih model. Performa yang didapat menggunakan metode yang diusulkan ini cukup baik, terbukti dengan akurasi yang didapat pada penelitian ini cukup tinggi, yaitu 95,58% dengan F1 score 97,47%. Kontribusi dari penelitian ini adalah memperkenalkan pendekatan baru untuk mendeteksi PDF *malware* yaitu menggunakan BFD dan SVM, serta menggunakan SFS untuk melakukan seleksi fitur dengan tujuan untuk meningkatkan performa model. Pendekatan baru ini diharapakan dapat menjadi suatu alternatif untuk mendeteksi PDF *malware*.

Kata Kunci: *portable document format, malware, byte frequency distribution, sequential forward selection, support vector machine*

DAFTAR ISI

		Halaman
Lembaran Pengesahan		ii
Halaman Persetujuan		iii
Lembar Pernyataan		iv
Kata Pengantar		v
Abstract		vii
Abstrak		viii
Daftar Isi		ix
Daftar Gambar		xi
Daftar Tabel		xii
BAB I. PENDAHULUAN		1
1.1.	Latar Belakang Masalah	1
1.2.	Rumusan Masalah	3
1.3.	Batasan Masalah	3
1.4.	Tujuan	3
1.5.	Manfaat	4
1.6.	Sistematika Penulisan	4
BAB II. TINJAUAN PUSTAKA		6
2.1.	Penelitian Terkait	6
2.2.	*Malicious Software*	9
2.3.	*Malicious* PDF	10
2.4.	*Byte Frequency Distribution* (BFD)	10
2.5.	*Sequential Forward Selection* (SFS)	11
2.6.	*Support Vector Mechine* (SVM)	12
BAB III. METODOLOGI PENELITIAN		15
3.1.	Kerangka Kerja Penelitian	15
3.2.	Alur Deteksi PDF *Malware*	16
3.3.	Dataset	18
3.4.	*Preprocessing*	19
3.5.	Proses Deteksi *malware* menggunakan SVM	21
3.6.	Analisa Hasil	22
3.7.	Penarikan Kesimpulan	23
BAB IV. HASIL DAN ANALISIS		24
4.1.	Parameter Pengujian	24
4.2.	Hasil Pengujian Model	24

4.2.1. Hasil Model Pertama 25
4.2.2. Hasil Model Kedua 26
4.2.3. Hasil Model Ketiga 27
4.2.4. Hasil Model Keempat 29
4.2.5. Hasil Model Kelima 30
4.2.6. Hasil model Keenam 31
4.3. Hasil dan Analisis Keseluruhan Model 33
4.4. Hasil Pengujian dengan Data *Unseen* 34
BAB V. KESIMPULAN DAN SARAN 37
5.1. Kesimpulan 37
5.2. Saran 38
DAFTAR PUSTAKA 39

DAFTAR GAMBAR

	Halaman
Gambar 2.2. Penelitian Terkait Malware	9
Gambar 2.3. Kerangka Penelitian (Masoumi, Keshavarz and Fotohi, 2021)	11
Gambar 2.4. Hyperplane yang membagi data (Kowalczyk, 2017)	12
Gambar 3.1. Kerangka Kerja Penelitian	16
Gambar 3.2. Alur Deteksi PDF Malware	17
Gambar 3.3. Alur preprocessing dataset	18
Gambar 3.4. Dataset Penelitian dalam format csv	19
Gambar 3.5. Proses konversi file .pdf ke format byte	19
Gambar 3.6. Contoh Byte Frequency Distribution pada suatu file .pdf	20
Gambar 3.7. Potongan kode seleksi fitur menggunakan SFS	21
Gambar 3.8. Contoh confusion matrix	22
Gambar 4.1. Perbandingan Data Traning dan Testing	24
Gambar 4.2. Confusion Matrix Model Pertama	25
Gambar 4.3. Confusion Matrix Model Kedua	26
Gambar 4.4. Confusion Matrix Model Ketiga	28
Gambar 4.5. Confusion Matrix Model Keempat	29
Gambar 4.6. Confusion Matrix Model Kelima	30
Gambar 4.7. Confusion Matrix Model Keenam	32
Gambar 4.8. Evaluasi Kelesuruhan Model	34
Gambar 4.9. Perbandingan data uji dan unseen	36

DAFTAR TABEL

	Halaman
Tabel 2.1. Tinjauan Terhadap Penelitian Terkait	7
Tabel 4.1 Hasil Pengujian Model Pertama	26
Tabel 4.2. Hasil Pengujian Model Kedua	27
Tabel 4.3. Hasil Pengujian Model Ketiga	28
Tabel 4.4 Hasil Pengujian Model Keempat	30
Tabel 4.5. Hasil Pengujian Model Kelima	31
Tabel 4.6. Hasil Pengujian Model Keenam	32
Tabel 4.7. Hasil Pengujian Seluruh Model	33
Tabel 4.8. Hasil Pengujian Data Unseen	35

BAB I. PENDAHULUAN

Pada Bab ini berisi latar belakang dilakukannya penelitian yang berjudul Deteksi Malware pada *file Portable Document Format* menggunakan *Byte Frequency Distribution* dan *Machine Learning.* Sub-bab berikutnya akan merumuskan permasalahan yang akan dibahas berlandaskan dari latar belakang. Terdapat juga batasan masalah untuk mencegah meluasnya permasalahan. Selanjutnya, dirumuskan tujuan dari penelitian yang dibuat, dan metodologi yang digunakan dalam penelitian tersebut.

1.1. Latar Belakang Masalah

Digitalisasi merupakan salah satu fokus transformasi yang ada di roadmap kementrian BUMN yang bertujuan untuk meningkatan kinerja dan efisiensi perusahan-perusahaan BUMN yang ada di Indonesia. PT Semen Baturaja merupakan bagian dari Semen Indonesia Group (SIG), sebuah Badan Usaha Milik Negara (BUMN) yang bergerak di bidang industri semen.

Sesuai fokus transformasi PT Semen Baturaja terus berupaya untuk mengadopsi teknologi digital pada proses bisnis nya. Sejak tahun 2019, PT Semen Baturaja telah mengimplementasikan sistem *Enterprise Resource Planning* (ERP) berbasis SAP S4/HANA serta aplikasi pendukung lainnya sebagai upaya digitalisasi proses bisnis. Seiring dengan digitalisasi yang dilakukan, PT Semen Baturaja juga meningkatkan kemananan dibidang siber untuk memastikan agar data-data perusahaan tetap aman dari ancaman serangan siber. Data ini disimpan secara elektronik menggunakan berbagai format, salah satunya PDF.

Menurut Laporan Tahunan Monitoring Keamanan Siber di tahun 2021 yang dikeluarkan oleh BSSN, PDF serta beberapa *file* lain seperti (.docx, .hwp dan .jpg) sangat sering digunakan untuk melakukan serangan siber (Maiorca et al., 2020), (BSSN, 2022). Berdasarkan VirusTotal PDF menempati urutan ke empat dalam *file* dokumen yang sering digunakan untuk menyebarkan *malware*, meningkat 97% dibandingkan tahun sebelumnya (VirusTotal, 2022).

Disisi lain, pada serangan siber salah satu metode yang digunakan untuk mendeteksi *file* PDF yang terinfeksi yaitu menggunakan pendekatan Machine Learning (Hossain Faruk et al., 2021), (Alshamrani, 2022). Cuan et al. pada penelitiannya menggunakan *Support Vector Machine* (SVM) untuk mendeteksi malware pada PDF. Pada peneliatian tersebut *file* PDF akan diekstrak menggunakan PDFiD, *tools* ini dikembangkan oleh Stevens pada 2006. Hanya saja dalam penelitian ini nilai sensitivitas dan presisi tidak diketahui (Cuan et al., 2018) (Stevens, 2006).

Teknik lainnya yang digunakan, yaitu teknik visualisasi dan teknik pengolahan citra untuk mendeteksi malware pada PDF. Pada penelitian tersebut *file* PDF dikonversi menjadi citra *greysacale* untuk selanjutnya diekstraksi, beberapa metode yang digunakan adalah *Random Forest*, *Decision Tree* dan *K-Nearest Neighbor*. Metode yang digunakan tergolong kompleks karena *file* PDF perlu dikonversi terlebih dahulu menjadi citra *greysacle*. Pada penelitian ini nilai akurasi, sensitivitas dan presisi tidak diketahui (Corum et al., 2019).

Kemudian pada tahun 2021, Seorang Peneliti menggunakan penggabungan metode statistik, yaitu *Byte Frequency Distribution* (BFD) dan *Sequential Forward Selection* algorithm (SFS). Metode ini digunakan untuk melakukan ekstraksi dan seleksi fitur dari *file fragment*. Pada penelitian ini model klasifikasi dibangun menggunakan beberapa algoritma, yaitu: *Multilayer Perceptron* (MLP), *Support Vector Machines* (SVM) dan *K-Nearest Neighbor* (KNN). Diantara ketiga algoritma tersebut, SVM menghasilkan akurasi terbaik (Masoumi et al., 2021).

Pada penelitian ini, akan dilakukan deteksi *malware* pada *file* PDF, diamana data yang digunakan berasal dari repositori PT Semen Baturaja (.pdf, .jpeg, dan .png). Untuk melakukan ekstraksi dan seleksi fitur akan digunakan metode BFD dan metode SFS, sedangkan penggunaan algoritma SVM digunakan untuk membuat model.

1.2. Rumusan Masalah

Dalam studi kasus repositori yang digunakan. dokumen PDF merupakan jenis *file* yang sering digunakan dalam pertukaran informasi, sehingga kemungkinan penyebaran *malware* semakin tinggi, oleh karena itu diperlukan suatu alat yang dapat digunakan untuk mendeteksi secara cepat dan akurat. Berdasarkan latar belakang permasalahannya, permasalahan tersebut dapat dijelaskan sebagai berikut:

1. Bagaimana proses ekstraksi dan seleksi fitur untuk digunakan dalam mendeteksi *malware* pada PDF?
2. Bagaimana mendeteksi *malware* di file PDF dari suatu kumpulan data?
3. Bagaimana mengukur kinerja model menggunakan SVM?

1.3. Batasan Masalah

Batasan masalah dalam melakukan proses deteksi yang dirancang pada tesis ini adalah:

1. Dataset yang digunakan berasal dari repositori PT. Semen Baturaja berupa kumpulan *file* PDF dan non-PDF (.jpeg dan .png).
2. Dataset PDF malware menggunakan dataset Contagio.
3. Metode *Machine Learning* yang digunakan untuk membuat model adalah Support Vector Machine.

1.4. Tujuan

Adapun tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

1. Menggunakan metode *Byte Frequency Distribuion* (BFD) dan metode *Sequential Forward Selection* (SFS) untuk melakukan ekstraksi fitur dan seleksi fitur.
2. Mengembangkan model untuk mendeteksi *malware* pada PDF menggunakan SVM.

3. Mengukur hasil dari pengembangan model yang telah dibangun menggunakan *confusion matrix*.

1.5. Manfaat

Sedangkan, manfaat yang dapat diambil dari penelitian ini adalah:

1. Memberikan kontribusi untuk penelitian deteksi *malware* menggunakan metode BFD.
2. Memberikan kontribusi untuk penelitian deteksi *malware* khususnya pada *file* PDF menggunakan *machine learning.*
3. Hasil dari penelitian dapat menjadi referensi untuk penelitian dibidang deteksi *malware*.

1.6. Sistematika Penulisan

Untuk lebih memudahkan dalam menyusun proposal tesis ini dan untuk menjelaskan isi dari setiap bab pada laporan ini, maka dibuatlah sistematika penulisan sebagai berikut:

1. **BAB I Pendahuluan**

   Bab ini berisi latar belakang yang menjelaskan mengapa topik penelitian ini penting dan relevan untuk diteliti. Pada bab ini juga dijelaskan rumusan masalah, tujuan, manfaat serta batasan masalah untuk memberikan gambaran tentang bagaimana penelitian akan dilakukan.

2. **BAB II Tinjauan Pustaka**

   Bab ini berisi seluruh penjelasan terhadap tinjauan pustaka yang berkaitan dengan permasalahan yang dibahas dalam penulisan tesis ini. Hal ini dapat membantu penulis memperoleh pemahaman yang lebih baik tentang topik penelitian.

3. **BAB III Metodologi Penelitian**

   Bab ini berisi alasan dan metode penelitian, studi literatur metode pengambilan sampel, data yang digunakan, metode analisis data, dan metode penyajian data. Hal ini akan

digunakan untuk membuat kerangka kerja penelitian dalam penyelesaian tesis.

4. **BAB IV** **Hasil dan Analisis**

   Bab ini berisi hasil dan analisis terhadap hasil pengerjaan tesis yang telah dilakukan. Hasil dari analisis penelitian akan disajikan secara sistematis menggunakan berbagai macam teknik.

5. **BAB V** **Kesimpulan dan Saran**

   Bab ini berisi kesimpulan berdasarkan hasil yang telah diperoleh dari pengerjaan tesis. Pada bab ini juga menyajikan saran serta kekurangan yang mungkin dapat dikembangkan dari penelitian ini.

DAFTAR PUSTAKA

Adenansi, R., & Novarina, L. A. (2017). Malware dynamic. *Jurnal of Education and Information Communication Technology*, *1*(1), 37–43.

Al-Haija, Q., Odeh, A., & Qattous, H. (2022). PDF Malware Detection Based on Optimizable Decision Trees. *Electronics 2022*, 562–570. https://doi.org/10.5220/0010908400003120

Alshamrani, S. S. (2022). Design and Analysis of Machine Learning Based Technique for Malware Identification and Classification of Portable Document Format Files. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/7611741

BSSN. (2022). *Laporan Tahunan Monitoring Keamanan Siber 2021*.

Charim, A., Basuki, S., & Akbi, D. R. (2019). Detect Malware in Portable Document Format Files (PDF) Using Support Vector Machine and Random Decision Forest. *Jurnal Online Informatika*, *3*(2), 99. https://doi.org/10.15575/join.v3i2.196

Corum, A., Jenkins, D., & Zheng, J. (2019). Robust PDF Malware Detection with Image Visualization and Processing Techniques. *Proceedings - 2019 2nd International Conference on Data Intelligence and Security, ICDIS 2019*, 108–114. https://doi.org/10.1109/ICDIS.2019.00024

Cuan, B., Damien, A., Delaplace, C., & Valois, M. (2018). Malware detection in PDF files using machine learning. *ICETE 2018 - Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, *2*, 412–419. https://doi.org/10.5220/0006884704120419

Hossain Faruk, M. J., Shahriar, H., Valero, M., Barsha, F. L., Sobhan, S., Khan, M. A., Whitman, M., Cuzzocrea, A., Lo, D., Rahman, A., & Wu, F. (2021). Malware Detection and Prevention using Artificial Intelligence Techniques. *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, 5369–5377. https://doi.org/10.1109/BigData52589.2021.9671434

Issakhani, M., Victor, P., Tekeoglu, A., & Lashkari, A. (2022). PDF Malware Detection based on Stacking Learning. *Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP 2022)*,

Icissp, 562–570. https://doi.org/10.5220/0010908400003120

Jeong, Y. S., Woo, J., & Kang, A. R. (2019). Malware Detection on Byte Streams of PDF Files Using Convolutional Neural Networks. *Security and Communication Networks*, *2019*. https://doi.org/10.1155/2019/8485365

Kang He, Yuefei Zhu, Yubo He, Long Liu, Bin Lu, W. L. (2020). Chinese J of Electronics - 2020 - He - Detection of Malicious PDF Files Using a Two-Stage Machine Learning Algorithm.pdf. *Chinese Journal of Electronics*. https://doi.org/https://doi.org/10.1049/cje.2020.10.002

Khultsum, U., & Subekti, A. (2021). Penerapan Algoritma Random Forest dengan Kombinasi Ekstraksi Fitur Untuk Klasifikasi Penyakit Daun Tomat. *Jurnal Media Informatika Budidarma*, *5*(1), 186. https://doi.org/10.30865/mib.v5i1.2624

Kim, G. Y., Paik, J. Y., Kim, Y., & Cho, E. S. (2022). Byte Frequency Based Indicators for Crypto-Ransomware Detection from Empirical Analysis. *Journal of Computer Science and Technology*, *37*(2), 423–442. https://doi.org/10.1007/s11390-021-0263-x

Kowalczyk, A. (2017). *Support Vector Machines Succinctly*. 1–114.

Liu, C.-Y., Chiu, M.-Y., Huang, Q.-X., & Sun, H.-M. (2021). PDF Malware Detection Using Visualization and Machine Learning. In K. Barker & K. Ghazinour (Eds.), *Data and Applications Security and Privacy XXXV* (pp. 209–220). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-030-81242-3_12

Maiorca, D., Biggio, B., & Giacinto, G. (2020). Towards Adversarial Malware Detection. *ACM Computing Surveys*, *52*(4), 1–36. https://doi.org/10.1145/3332184

Masoumi, M., Keshavarz, A., & Fotohi, R. (2021). File fragment recognition based on content and statistical features. *Multimedia Tools and Applications*, *80*(12), 18859–18874. https://doi.org/10.1007/s11042-021-10681-x

Parkour, M. (2013). *contagio: 16,800 clean and 11,960 malicious files for signature testing and research.* https://contagiodump.blogspot.com/2013/03/16800-clean-and-11960-malicious-files.html

Polat, H., & Polat, O. (2020). Detecting DDoS Attacks in Software-Defined.pdf.

Mdpi.

PT Semen Baturaja (Persero) Tbk. (2019). *Laporan Tahunan PT Semen Baturaja (Persero) Tbk 2019*.

Stevens, D. (2006). *PDF Tools | Didier Stevens*. https://blog.didierstevens.com/programs/pdf-tools/

Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26. https://doi.org/10.2478/CAIT-2019-0001

VirusTotal. (2022). *Virustotal's 2021 Malware Trends Report* (Issue March). https://assets.virustotal.com/reports/2021trends.pdf

Zhang, J. (2018). *MLPdf: An Effective Machine Learning Based Approach for PDF Malware Detection*. 1–6. http://arxiv.org/abs/1808.06991