Jurnal Teknologi Informasi dan Komunikasi

Vol: 15 No 01 2024 **E-ISSN: 2477-3255**

Accepted by the Editor: 28-04-2024 | Revision: 21-04-2024 | Published: 31-05-2024

Comparative Study of the Effect of Datasets and Machine Learning Algorithms for PDF Malware Detection

Salman Abdul Jabbaar Wiharja[1], Deden Pradeka[2], Wirmanto Suteddy[3]

[1,2,3]Program Studi Teknik Komputer, Kampus Universitas Pendidikan Indonesia di Cibiru

[1,2,3]Jl. Pendidikan No. 15, Cibiru Wetan, Cileunyi, Bandung, telp. (022) 7801840

e-mail: [1]salmanfc207@upi.edu, [2]dedenpradeka@upi.edu, [3]wirmanto.suteddy@upi.edu

Abstract

This research presents an innovative approach to detecting malicious PDFs through machine learning algorithms, focusing on the expansion of the Evasive-PDFMal2022 dataset. The objective is to enhance the accuracy of detecting malicious PDFs by enriching the dataset, augmenting its representation and diversity, and developing a practical tool—a website—for extracting and detecting malicious PDFs. The methodology involves updating and enlarging the dataset with additional malicious PDFs sourced from CVE and Exploit-db, along with non-malicious PDFs from diverse origins. Features are then extracted using the PDFID tool, and these 20 features serve as the foundation for implementing K-Nearest Neighbor (KNN), Random Forest, and Random Committee algorithms. The outcomes demonstrate that the model trained with the expanded dataset achieves a remarkable 99% accuracy, surpassing the performance of models relying solely on the Evasive-PDFMal2022 dataset. Additionally, this research significantly enhances the representation and diversity of the dataset while delivering a practical solution in the form of a website tailored for the extraction and detection of malicious PDFs.

Keywords: *Machine learning, PDF, Malware, Random forest, Random committee.*

Studi Komparasi Pengaruh Dataset dan Algoritma *Machine Learning* untuk Pendeteksian *Malware* PDF

Abstrak

Penelitian ini menyajikan pendekatan inovatif untuk mendeteksi PDF berbahaya melalui algoritme pembelajaran mesin, dengan fokus pada perluasan dataset Evasive-PDFMal2022. Tujuannya adalah untuk meningkatkan akurasi pendeteksian PDF berbahaya dengan memperkaya dataset, menambah representasi dan keragamannya, dan mengembangkan alat praktis - sebuah situs web - untuk mengekstraksi dan mendeteksi PDF berbahaya. Metodologi ini melibatkan pembaruan dan perluasan dataset dengan PDF berbahaya tambahan yang bersumber dari CVE dan Exploit-db, bersama dengan PDF tidak berbahaya dari berbagai sumber. Fitur-fitur tersebut kemudian diekstraksi menggunakan alat PDFID, dan 20 fitur ini berfungsi sebagai fondasi untuk mengimplementasikan algoritme K-Nearest Neighbor (KNN), Random Forest, dan Random Committee. Hasilnya menunjukkan bahwa model yang dilatih dengan dataset yang diperluas mencapai akurasi 99% yang luar biasa, melampaui kinerja model yang hanya mengandalkan

https://doi.org/10.31849/digitalzone.v15i1.19744

dataset Evasive-PDFMal2022. Selain itu, penelitian ini secara signifikan meningkatkan representasi dan keragaman dataset sambil memberikan solusi praktis dalam bentuk situs web yang dirancang untuk mengekstraksi dan mendeteksi PDF berbahaya.

Kata kunci: Pembelajaran Mesin, PDF, Malware, Random forest, Random committee

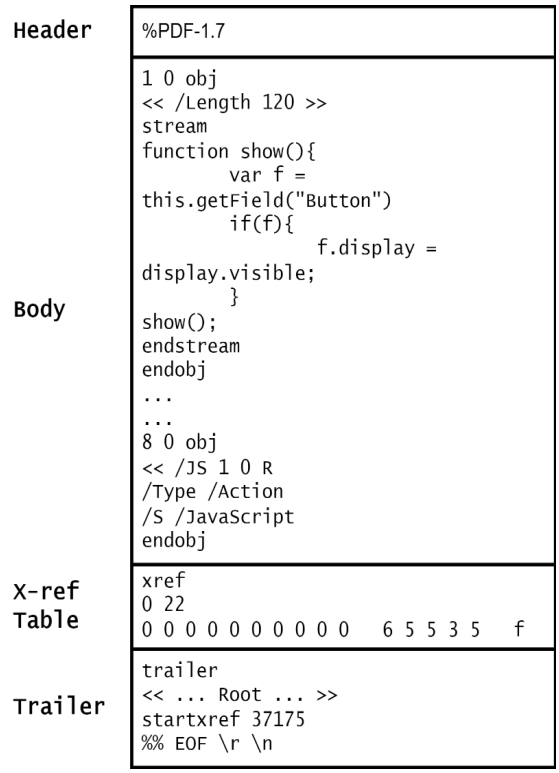
1. Pendahuluan

Portable Document Format (PDF) telah menjadi standar yang lebih disukai untuk pertukaran dan penyebaran dokumen karena kemampuannya untuk beradaptasi, fitur-fitur yang dapat disesuaikan, dan portabilitas yang mudah di berbagai *platform.* [1], [2]. Namun demikian, penggunaan PDF yang meluas telah menarik perhatian para penyerang siber yang bertujuan untuk mengeksploitasi kerentanan dan memanipulasi fitur file [3]. PDF muncul sebagai jenis file lampiran email berbahaya yang paling umum, mencakup 66% dari total lampiran. Kurangnya kesadaran tentang potensi jahat PDF dan kemampuannya untuk menghindari deteksi antivirus modern membuatnya menjadi vektor yang disukai untuk ancaman dunia maya [4].

Dalam artikel ini, dokumen berbahaya menonjol sebagai metode utama yang digunakan oleh penyerang untuk menyebarkan *malware*, melalui berbagai cara seperti alur enkripsi, file yang dapat dieksekusi (exe), *JavaScript*, perintah sistem, dan objek tersembunyi. Kompleksitas format file PDF menimbulkan tantangan besar dalam mendeteksi konten berbahaya, diperparah oleh teknik mengelak yang digunakan oleh penyerang [2], [3].

Dalam pengembangan sistem deteksi *malware* PDF menggunakan *machine learning*, faktor penting yang mempengaruhi akurasi model adalah dataset dan algoritma *machine learning* yang dipilih [5]. Penelitian ini bertujuan untuk mengembangkan dataset serta membandingkan berbagai algoritma *machine learning* guna mencapai pendeteksian *malware* PDF yang terbaik.

Portable Document Format (PDF) telah menjadi solusi serbaguna yang sangat diandalkan untuk berbagi konten yang beragam, mulai dari teks hingga media dan gambar. Dalam struktur dasarnya, dokumen PDF terdiri dari empat komponen utama [6].



Gambar 1. Struktur dari PDF

Pertama, header berisi informasi tentang versi standar PDF yang digunakan, memungkinkan aplikasi seperti Acrobat Reader untuk berfungsi secara tepat. Kedua, body mengandung konten yang

dapat dilihat pengguna seperti teks, gambar, dan kode skrip, serta menentukan operasi seperti dekompresi atau dekripsi data selama rendering file. Ketiga, tabel referensi silang (*X-ref Table*) berisi daftar offset dari setiap objek yang akan dirender dalam file, memfasilitasi akses acak ke setiap objek dan pembaruan tambahan pada dokumen. Terakhir, trailer terletak di bagian akhir file dan memberikan rincian tentang objek awal dokumen yang diidentifikasi oleh tag /Root, serta merangkum baris terakhir file dengan '%%EOF'. Proses membaca file PDF dimulai dari objek trailer dan menguraikan setiap objek yang dirujuk oleh tabel X-ref, sementara data secara progresif didekompresi untuk merender semua elemen, seperti teks, gambar, dan komponen lainnya, secara bertahap. Dokumen PDF diatur sebagai grafik objek yang berisi instruksi untuk menyajikan konten file kepada pengguna.

PDF, sebagai alat untuk berbagi informasi yang sah, juga membuka peluang bagi penyerang untuk mengeksploitasi potensi kerentanan keamanannya. Teknik canggih seperti histogram, pola piksel, dan steganografi digunakan oleh penyerang untuk menyembunyikan kode berbahaya dalam dokumen PDF [7], [8].

Tabel 1. Tag yang berpotensi menyebabkan kerusakan pada PDF

Tag	Keterangan
/JS dan /JavaScript	Skrip javascript yang biasa digunakan untuk membuka *backdoor*
/AA dan /OpenAction	Tidakan otomatis untuk memulai aksi tertentu
/GoTo	Tag yang memfasilitasi perpindahan ke halaman tertentu baik di dalam atau di luar file
/Launch	Tag untuk membuka dokumen atau menjalankan program
/URI	Tag yang memungkinkan untuk mengakses URL
/SubmitForm	Memfasilitasi pengiriman data ke URL yang ditentukan di dalam dokumen
/RichMedia	Tag untuk meyematkan Flash ke dalam PDF
/ObjStm	Tag untuk menyembunyikan Object Stream

Studi dan penelitian telah menyoroti kemungkinan penipuan melalui teknik ini, memperkuat kebutuhan akan perhatian yang lebih besar terhadap keamanan dokumen PDF. Selain itu, adanya tag-tag tertentu dalam struktur file PDF yang tidak terdeteksi dengan baik oleh sebagian besar alat pengamanan dapat menimbulkan risiko yang signifikan. Tag-tag seperti /JS dan /JavaScript, /AA dan /OpenAction, /GoTo, /Launch, /URI, /SubmitForm, /RichMedia, dan /ObjStm, dapat dimanfaatkan oleh penyerang untuk menyusupkan atau memanipulasi konten dengan potensi menyebabkan kerusakan atau ancaman keamanan [3]. Oleh karena itu, pemahaman mendalam tentang berbagai potensi ancaman ini menjadi kunci dalam menghadapi tantangan keamanan yang berkaitan dengan penggunaan dan distribusi dokumen PDF.

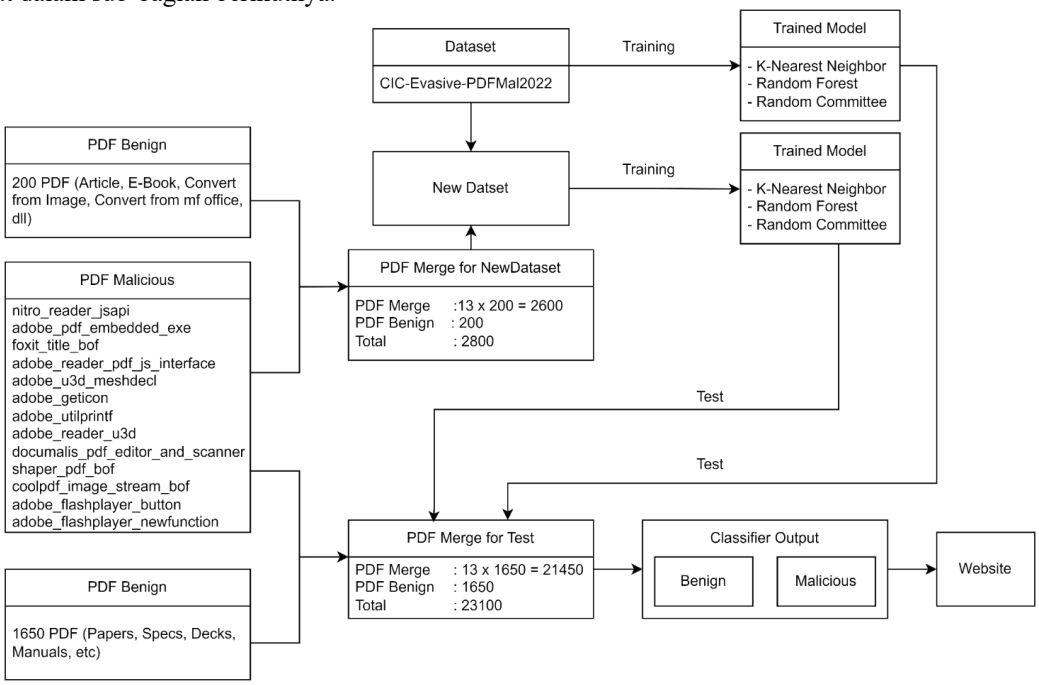
Machine learning (ML), merupakan kategori algoritme yang luas dan sering digunakan dalam beragam aplikasi, dari analisis sentimen komentar YouTube hingga pengenalan tulisan tangan [9]. Dalam dunia *Machine learning*, setiap algoritma memiliki keunggulannya masing-masing. Sebagai contoh, *Support Vector Machines* (SVM) terkenal karena kehandalannya dalam mengidentifikasi komponen analisis sentimen komentar YouTube [10]. Penelitian sebelumnya telah mengeksplorasi berbagai teknik pembelajaran kelompok, seperti Random Subspace, AdaBoost, Stacking, dan Random Committee, serta menerapkan Convolutional Neural Networks (CNN) untuk mengklasifikasikan PDF berbahaya [6], [11]. Namun, penelitian ini difokuskan pada memperluas pemahaman tentang kinerja relatif *Random Forest, Random Committee*, dan *K-Nearest Neighbor* (KNN) dalam mendeteksi *malware* PDF. Algoritma-algoritma ini dipilih karena kemampuannya dalam mengidentifikasi komponen berisiko dalam berkas PDF. Misalnya *Random Forest*, terkenal karena kemampuannya mengatasi overfitting dan toleransinya terhadap data yang tidak teratur atau tidak lengkap [12]. Di sisi lain, *Random Committee* menggabungkan beberapa model keputusan untuk meningkatkan keakuratan dan stabilitas prediksinya [6], sementara KNN [13] bekerja berdasarkan prinsip bahwa objek serupa berkumpul dalam ruang berdimensi tinggi, cocok untuk kasus deteksi PDF berbahaya dengan kumpulan fitur kompleks dan tidak linier. Dengan mempertimbangkan keunggulan masing-masing algoritma ini dalam menghadapi tantangan deteksi PDF berbahaya, penelitian ini bertujuan untuk menyediakan wawasan yang lebih dalam tentang kinerja mereka.

https://doi.org/10.31849/digitalzone.v15i1.19744

Salah satu kontribusi utama dari penelitian ini adalah pemanfaatan dataset Evasive-PDF Mal2022 yang telah diperbaharui dengan data baru, memperkaya representasi dataset tersebut. Dengan melakukan pembaruan ini, variasi dan kompleksitas sampel dalam dataset diperkaya, memberikan gambaran yang lebih akurat tentang ancaman *malware* PDF yang berkembang. Penggunaan dataset yang lebih representatif ini diharapkan dapat meningkatkan akurasi pemodelan *machine learning* dalam mendeteksi *malware* PDF, serta relevansinya dalam menghadapi ancaman yang terus berkembang. Selain itu, penelitian ini juga melibatkan pembuatan sebuah website untuk deteksi *malware* PDF yang menggabungkan teknologi *machine learning* [14]. Proses pembuatan website ini melibatkan serangkaian langkah kompleks, termasuk pengembangan antarmuka pengguna menggunakan HTML, CSS, dan JavaScript, serta penerapan kriptografi untuk melindungi data sensitif seperti model *machine learning* [15]. Dengan menggabungkan penggunaan dataset yang diperbaharui dan teknologi pembuatan website yang canggih, penelitian ini bertujuan untuk memberikan solusi yang lebih efektif dalam mendeteksi dan mengatasi ancaman *malware* PDF.

2. Metode Penelitian

Peneliti mengusulkan pendekatan untuk merancang sistem deteksi PDF berbahaya menggunakan *machine learning*. Pendekatan ini memanfaatkan dataset yang telah disempurnakan, terdiri dari 20 fitur dan 12,625 baris data. Fitur-fitur ini diekstraksi dari file yang telah dilabeli dan dipisahkan sebagai set pelatihan. Dataset Evasive-PDFMal2022 terdiri dari 10.025 baris data, yang kemudian diperluas dengan hasil ekstraksi metadata dari 2.600 PDF yang merupakan hasil penggabungan antara 200 PDF *Benign* dengan 13 PDF *malware*. Dengan demikian, dataset baru yang dihasilkan terdiri dari total 12.625 baris data. Algoritma *K-Nearest Neighbor, Random forest*, dan *Random committee* digunakan untuk mempelajari karakteristik yang membedakan antara file PDF yang aman dan berbahaya. Sistem ini kemudian dapat memprediksi dan mengklasifikasikan file PDF tanpa label untuk menentukan apakah file tersebut aman atau berbahaya. Metode yang digunakan dalam pembangunan sistem akan dijelaskan lebih lanjut dalam sub-bagian berikutnya.



Gambar 2. Usulan pendekatan berbasis dataset yang disempurnakan.

2.1. Dataset

Evasive-PDFMal2022 yang diperkenalkan oleh [16] adalah perbaikan dari dataset PDF Contagio yang bertujuan untuk mengatasi kekurangan seperti proporsi sampel duplikat yang tinggi dan kurangnya keragaman sampel di setiap kelas. Dengan 10.025 sampel file PDF, terdiri dari 4.468 sampel jinak dan 5.557 sampel berbahaya, dataset ini dimaksudkan untuk memberikan representasi yang lebih realistis dari distribusi PDF. Perbaikan pada dataset ini diharapkan meningkatkan validitas dan keterwakilan dataset untuk penelitian deteksi PDF berbahaya.

2.2. PDF Collection

Selain Evasive-PDFMal2022, dua koleksi set data tambahan berkontribusi pada penelitian ini yaitu *PDF Private Collection* dan *Technically-oriented PDF Collection.*

2.2.1. *Private PDF Collection*

Terdiri dari 200 file PDF asli yang mencakup berbagai jenis konten seperti artikel, e-book, dan PDF yang telah dikonversi. Penggabungan hasil dengan Evasive-PDFMal2022 menghasilkan kumpulan data yang lebih kaya dan beragam.

2.2.2. *Technically-oriented PDF Collection*

Merupakan dataset yang berisi 1.650 file PDF yang dikumpulkan oleh TPN di repositori GitHub mereka. Koleksi ini mencakup dokumen teknis seperti makalah, spesifikasi, presentasi, dan manual. Digunakan sebagai kumpulan data uji, evaluasi ini menilai kapasitas model untuk menggeneralisasi seluruh dokumen teknis.

Penggabungan antara Evasive-PDFMal2022 dengan *Private PDF Collection* diharapkan dapat meningkatkan keterwakilan dan keragaman dataset, sementara penggunaan *Technically-oriented PDF Collection* sebagai data uji akan memberikan gambaran seberapa baik model dapat menggeneralisasi pada dokumen teknis. Dengan adanya dua koleksi tambahan ini, peneliti berharap dapat memberikan hasil yang Latihan dan test lebih baik dibandingkan penelitian sebelumnya.

2.3. Ekstraksi Fitur

Sistem deteksi yang diajukan oleh peneliti berdasarkan pada 20 fitur struktural yang diekstraksi menggunakan PDFMalyzer, sebuah perangkat lunak sumber terbuka yang memanfaatkan PDFID dan PyMuPDF. Fitur-fitur ini dipilih karena jarang ditemukan pada file PDF yang tidak bersifat berbahaya. Dalam eksperimen penelitian ini, fitur-fitur struktural tersebut diadopsi untuk analisis deteksi PDF berbahaya, sejalan dengan penelitian sebelumnya yang telah dilakukan [17].

Tabel 2. Kumpulan fitur awal yang terdiri dari 20 fitur

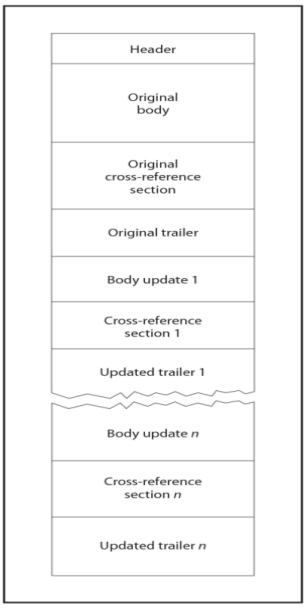
Nama Fitur	Deskripsi Fitur
Obj	Jumlah pembuka objek
Endobj	Jumlah penutup objek
Stream	Jumlah pembuka stream
Endstream	Jumlah penutup stream
Startxref	Jumlah awal x-ref
Xref	Jumlah tabel x-ref
Trailer	Jumlah trailer
/Objstm	Jumlah aliran objek
/JavaScript	Jumlah Javascript
/JS	Jumlah JS
/Encrypt	Indikasi PDF memiliki kata sandi
/OpenAction	Jumlah Tindakan otomatis saat PDF dibuka
/AA	Jumlah Tindakan otomatis dalam satu waktu
/Acroform	Jumlah formular acrobat
/launch	Jumlah Tindakan
/EmbeddedFile	Jumlah kata kunci embedded yang ditemukan
/JBIG2Decode	Apakah dokomen dikompresi dengan JBIG2Decode
/RichMedia	Jumlah media yang disematkan
/XFA	Jumlah kata kunci XML
/GoTo	Jumlah perpindahan halaman
/URI	Jumlah URL yang disematkan di dalam PDF

https://doi.org/10.31849/digitalzone.v15i1.19744

Tabel 2 memberikan detail lengkap tentang kumpulan awal 20 fitur struktural. Fitur-fitur ini merupakan karakteristik struktural yang tidak umum pada file PDF biasa. Keberadaan atau ketiadaannya dapat menjadi indikator penting untuk menentukan apakah suatu file PDF bersifat berbahaya atau tidak. Oleh karena itu, ekstraksi dan analisis fitur struktural ini menjadi langkah awal yang krusial dalam pengembangan sistem deteksi PDF.

2.4. Membuat Dataset Baru

Evasive-PDFMal2022 merupakan langkah perbaikan yang signifikan dari dataset Contagio, dimaksudkan untuk mengatasi beberapa kekurangan yang ada, seperti tingginya proporsi sampel duplikat dan kurangnya keragaman sampel di setiap kelas [18]. Dengan mengumpulkan 10.025 sampel file PDF, terdiri dari 4.468 sampel jinak dan 5.557 sampel berbahaya, dataset ini bertujuan memberikan representasi yang lebih realistis dari distribusi PDF yang digunakan secara luas. Namun, meskipun menjadi langkah maju yang signifikan, penelitian terbaru menyoroti beberapa kekurangan dari Evasive-PDFMal2022, termasuk ketidakmampuannya dalam mengatasi berbagai jenis PDF *Malware* dari CVE dan Exploit-db secara komprehensif. Untuk mengatasi tantangan ini, dilakukan pembuatan dataset baru yang memperkuat metode gabungan berkas Benign dengan berkas PDF *Malware*, tanpa mengubah struktur utama berkas tersebut. Langkah ini diharapkan dapat meningkatkan keakuratan dan ketangguhan model deteksi *malware* PDF yang dikembangkan, serta memberikan fondasi yang lebih kokoh untuk penelitian dan pengembangan keamanan siber di masa depan.



Gambar 3. Struktur file PDF yang sudah diperbaharui.

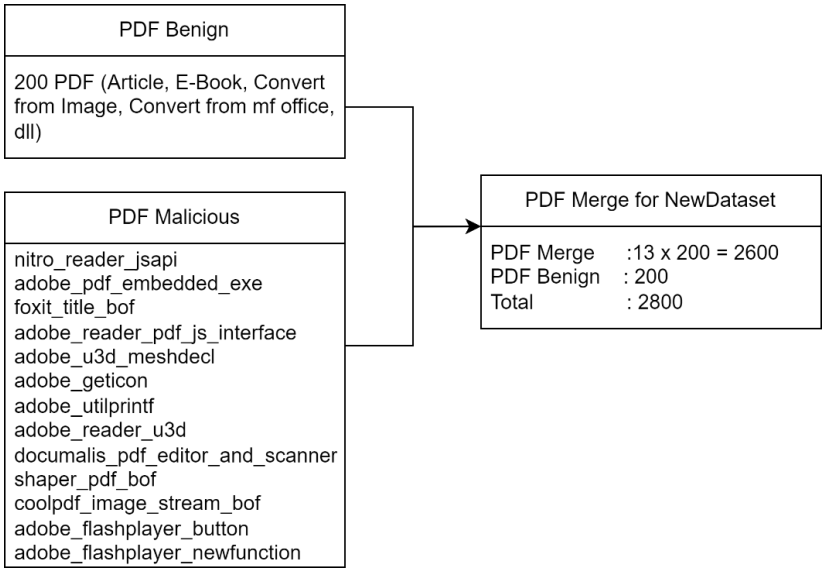
Metode yang dilakukan peneliti dalam membuat dataset baru terdiri dari dua Langkah yang dijelaskan sebagai berikut:

2.4.1. *Ekstraksi Metadata dari Benign PDF*

Prosesnya dimulai dengan ekstraksi *metadata* dari sekumpulan awal 200 PDF asli yang dipilih dengan cermat untuk mewakili beragam jenis konten, untuk memastikan fondasi yang kuat untuk dataset

2.4.2. *Integrasi Malicious PDF*

Selanjutnya, 13 PDF berbahaya, yang bersumber dari repositori Common Vulnerabilities and Exposures (CVE) dan Exploit-db, diintegrasikan secara strategis. Setiap PDF berbahaya mewakili eksploitasi yang unik, yang mencakup kerentanan dan teknik eksploitasi yang berbeda



Gambar 4. Proses penggabungan *Malicious* PDF ke dalam *Benign* PDF

Pendekatan metodologi kami mengikuti penelitian terkait [19] yang menyoroti pentingnya volume data pelatihan dalam meningkatkan akurasi model. Setiap jenis PDF berbahaya sengaja diduplikasi untuk meningkatkan data pelatihan, seperti yang diilustrasikan secara visual dalam Gambar 4. Komitmen terhadap set pelatihan yang komprehensif dan representatif tercermin dalam representasi ini. Dengan pendekatan ini, peningkatan kinerja model diharapkan melalui perluasan set data yang terarah, sesuai dengan praktik penelitian kontemporer.

PDF Berbahaya yang dipilih oleh peneliti untuk di integrasikan ke dalam PDF tidak berbahaya diambil dari CVE dan Exploit-DB. Deskripsi dari PDF berbahaya tersebut diantaranya berikut ini:

Tabel 3. Daftar *Malware* PDF yang diambil dari CVE dan Exploit-db

Nama PDF *Malware*	Deskripsi PDF *Malware*
nitro_reader_jsapi (CVE-2017-7442)	Modul ini memanfaatkan kerentanan API JavaScript pada Nitro PDF Reader versi 11. Fungsi saveAs() memungkinkan penulisan file sembarang ke sistem file, sementara launch() memungkinkan eksekusi file lokal tanpa dialog keamanan.
adobe_pdf_embedded _exe (CVE-2010-1240)	Modul ini menyisipkan muatan *Metasploit* ke dalam PDF yang ada, memungkinkan serangan rekayasa sosial dengan mengirimkan PDF yang dimodifikasi ke target.
foxit_title_bof (EDB-ID-16621)	Modul ini mengeksploitasi overflow buffer di Foxit PDF Reader sebelum versi 4.2 yag dipicu oleh membuka file PDF yang cacat, dengan string yang terlalu panjang di bidang Judul, menyebabkan penimpaan catatan penanganan pengecualian terstruktur.
adobe_reader_pdf_js _interface (CVE-2014-0514)	Modul ini menyisipkan eksploitasi peramban dari android/webview_addjavascriptinterface ke dalam PDF untuk memperoleh akses shell pada Adobe Reader versi < 11.2 karena mengekspos antarmuka asli yang tidak aman ke JavaScript.
adobe_u3d_meshdecl (CVE-2009-3953)	Modul ini mengeksploitasi overflow array di Adobe Reader dan Adobe Acrobat pada versi <7.1.4, <8.2, dan <9.3. Dengan memanfaatkan PDF berisi data U3D yang cacat, penyerang dapat menjalankan kode arbitrer.
adobe_geticon	Mengeksploitasi buffer overflow di Adobe Reader dan Acrobat pada versi

(CVE-2009-0927)	<7.1.1, <8.1.3, dan <9.1. Dengan memanfaatkan PDF berisi Collab.getIcon() cacat, penyerang dapat menjalankan kode arbitrer.
adobe_utilprintf (CVE-2008-2992)	Modul ini mengeksploitasi buffer overflow pada Adobe Acrobat Professional sebelum 8.1.3 dan Adobe Reader. Dengan PDF yang berisi util.printf() cacat, penyerang dapat menjalankan kode arbitrer.
adobe_reader_u3d (CVE-2011-2462)	Kerentanan pada Adobe Reader disebabkan oleh penggunaan memori yang tidak diinisialisasi. Eksekusi kode arbitrer terjadi dengan menyematkan data U3D yang dibuat khusus ke dalam dokumen PDF. Metode ini menggunakan JavaScript untuk mengontrol memori yang digunakan oleh penunjuk yang tidak valid.
documalis_pdf_editor_and_scanner (CVE-2020-7374)	Documalis Free PDF Scanner v5.7.2.122 dan Documalis Free PDF Editor v5.7.2.26 rentan terhadap serangan buffer overflow karena tidak memvalidasi konten gambar JPEG dalam PDF. Penyerang dapat memanfaatkan ini untuk mendapatkan eksekusi kode jarak jauh saat perangkat lunak tersebut dijalankan.
shaper_pdf_bof (EDB-ID-37760)	PDF Shaper rentan terhadap kerentanan keamanan saat melakukan konversi PDF ke gambar dengan menggunakan file PDF yang dibuat secara khusus. Kerentanan ini telah diuji berhasil pada Windows XP, 7, 8, dan 10.
coolpdf_image_stream_bof (CVE-2012-4914)	Modul ini mengeksploitasi buffer overflow pada Cool PDF Reader versi < 3.0.2.256 ketika membuka file PDF yang cacat berisi aliran gambar khusus. Kerentanan ini telah diuji pada Cool PDF versi 3.0.2.256 di Win7 (SP-1) dan Windows XP (SP-3).
adobe_flashplayer_button (CVE-2010-3654)	Modul ini mengeksploitasi kerentanan dalam penanganan film SWF pada Adobe Flash Player versi 9.x.x dan 10.0.x, yang juga memengaruhi Adobe Reader, Acrobat, dan aplikasi lain yang menyematkan *Flash Player*. Eksekusi kode arbitrer terjadi dengan menyisipkan Flash khusus ke dalam dokumen PDF. Metode semprotan tumpukan AcroJS digunakan untuk mengontrol memori yang digunakan oleh penunjuk yang tidak valid. Modul ini menggunakan metode bypass DEP serupa dengan modul adobe_libtiff, namun tidak dapat digunakan secara universal pada berbagai versi Windows karena perbedaan nomor syscall.
adobe_flashplayer_newfunction (CVE-2010-1297)	Modul ini sama seperti adobe_flashplayer_button yang dijelaskan sebelumnya. Perbedaaannya terletak pada eksploitasi kerentanan dalam penanganan tag DoABC.

Setelah penggabungan dan ekstraksi *metadata*, dihasilkan 2800 PDF (2600 *malicous* dan 200 *benign*). Selanjutnya PDFID [18] digunakan untuk mengekstraksi *metadata*, yang kemudian digabungkan dengan dataset Evasive-PDFMal2022 menjadi NewDataset. Dengan demikian, total dataset yang digunakan mencapai 12.825 sampel file PDF.

2.5. Pelatihan Dataset

Pada tahap ini, proses pelatihan dataset yang digunakan dalam sistem yang diajukan. Algoritma *K-Nearest Neighbor*, *Random forest*, dan *Random committee* dipilih untuk pengklasifikasi, memanfaatkan dasar klasifikasi Random Tree [6]. Proses pelatihan terdiri dari dua tahap utama, menggunakan dataset Evasive-PDFMal2022 dan NewDataset yang diusulkan oleh peneliti.

2.6. Pembuatan *Website*

Setelah melalui tahapan pelatihan dataset dan evaluasi hasil pelatihan yang menghasilkan akurasi terbaik, peneliti melanjutkan dengan membuat sebuah website yang diharapkan dapat memberikan manfaat dalam ekstraksi *metadata* serta pendeteksian PDF berbahaya. Dengan demikian, diharapkan dapat menjadi alat yang berguna dalam mendukung pengguna dalam mengidentifikasi dan mengelola file PDF dengan potensi risiko keamanan.

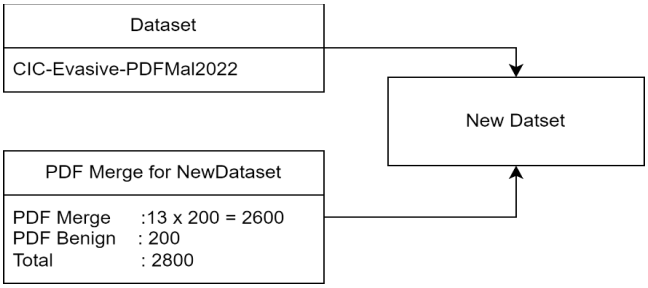
3. Hasil dan Pembahasan

Peneliti menyajikan hasil penelitian untuk mengembangkan sistem deteksi *Malware* dalam berkas PDF. Peneliti juga akan membandingkan pengaruh penggunaan dataset yang berbeda terhadap model *machine learning* yang digunakan untuk mendeteksi *Malware* dalam berkas PDF.

3.1. Pembuatan Dataset Baru

Pengembangan dataset yang canggih dan beragam dapat meningkatkan keandalan model *machine learning*, sebagaimana telah dibuktikan dalam penelitian sebelumnya [5]. Oleh karena itu, bagian ini menyajikan hasil eksperimen yang mengevaluasi dampak pembuatan dataset baru terhadap kinerja pengklasifikasi *K-Nearest Neighbor*, *Random forest*, dan *Random committee*. Sebelumnya, dijelaskan bagaimana peneliti berhasil menciptakan dataset baru untuk mengatasi kekurangan pada dataset Evasive-PDFMal2022.

Penelitian sebelumnya [20] mengungkapkan bahwa teknik steganografi dapat digunakan untuk menyembunyikan informasi dalam file PDF. Dalam penelitian ini, metode mergepdf digunakan untuk menggabungkan PDF, menciptakan dataset baru yang memadukan PDF berbahaya dengan PDF tidak berbahaya dapat dilihat dalam Gambar 5.



Gambar 5. Proses penggabungan PDF untuk menghasilkan dataset baru

Dimulai dengan menggabungkan 13 file PDF berbahaya dari dataset CVE dan Exploits-db dengan 200 Private PDF Collection, menghasilkan 2800 PDF. *Metadata* diekstraksi menggunakan PDFID dan diperkaya dengan skrip Python menjadi format CSV. Label kelas ditambahkan dan dataset diintegrasikan dengan Evasive-PDFMal2022, seperti yang terlihat dalam Gambar 5. Dengan pendekatan ini, diharapkan NewDataset dapat meningkatkan keandalan model tanpa mengorbankan akurasi klasifikasi. Analisis dampak NewDataset akan dibahas pada bagian selanjutnya.

3.2. Perbandingan Hasil Pelatihan

Dilakukan percobaan dasar untuk melatih model *K-Nearest Neighbor*, *Random forest*, dan *Random committee* menggunakan dataset Evasive-PDFMal2022 dan NewDataset. NewDataset merupakan gabungan dari Evasive-PDFMal2022 dengan dataset hasil ekstraksi *metadata* dari penggabungan PDF berbahaya dengan PDF tidak berbahaya. Analisis ini bertujuan untuk mengevaluasi dampak penambahan dataset baru terhadap kinerja model.

3.2.1 Konfigurasi *Hyperparameter*

Sebelum menganalisis hasil pelatihan, penting untuk memperinci konfigurasi *hyperparameter* yang digunakan untuk mengoptimalkan pengklasifikasi. Pengaturan *hyperparameter* memiliki dampak besar pada perilaku dan generalisasi model. Dalam penelitian ini, dilakukan pertimbangan teliti untuk menentukan satu set *hyperparameter* yang seimbang antara kompleksitas model dan ketangguhan untuk tugas deteksi file PDF berbahaya. Nilai-nilai *hyperparameter* dalam penelitian yang dilakukan dapat ditemukan pada Tabel 4.

Tabel 4. Konfigurasi *hyperparameter* untuk setiap model

Algoritma	Konfigurasi
K-Nearest Neighbor	*n_neighbors* = 3
Random forest	*n_estimators* = 100
Random committee	*base_classifier* = *DecisionTree*, *n_estimators* = 100

https://doi.org/10.31849/digitalzone.v15i1.19744

3.2.2 Hasil Pelatihan model yang dilatih menggunakan Dataset Evasive-PDFMal2022

Tabel 5 menampilkan hasil validasi silang dari pengklasifikasi *K-Nearest Neighbor*, *Random forest*, dan *Random committee* pada dataset Evasive-PDFMal2022. Akurasi yang diperoleh adalah 98% untuk *K-Nearest Neighbor*, 99% untuk *Random forest*, dan 98% untuk *Random committee*. *Random forest* menunjukkan kinerja terbaik dalam pelatihan menggunakan dataset tersebut.

Tabel 5. Hasil Klasifikasi model menggunakan Dataset Evasive-PDFMal2022

	Precision			Recall			F1-Score		
	KNN	RF	RC	KNN	RF	RC	KNN	RF	RC
Benign	0.96	0.98	0.98	0.98	0.98	0.97	0.97	0.98	0.98
Malicious	0.98	0.99	0.98	0.97	0.99	0.99	0.98	0.99	0.98
Akurasi							0.98	0.99	0.98

3.2.3. Hasil Pelatihan model yang dilatih menggunakan NewDataset

Tabel 6 menampilkan hasil validasi silang dari pengklasifikasi *K-Nearest Neighbor*, *Random forest*, dan *Random committee* pada NewDataset. Akurasi yang diperoleh adalah 95% untuk *K-Nearest Neighbor*, 98% untuk *Random forest*, dan 98% untuk *Random committee*. Terjadi penurunan akurasi sebesar 3% untuk *K-Nearest Neighbor* dan 1% untuk *Random forest* dibandingkan dengan penggunaan dataset sebelumnya. Selain akurasi, terdapat penurunan sedikit pada beberapa metrik lain.

Tabel 6. Hasil Klasifikasi model menggunakan NewDataset

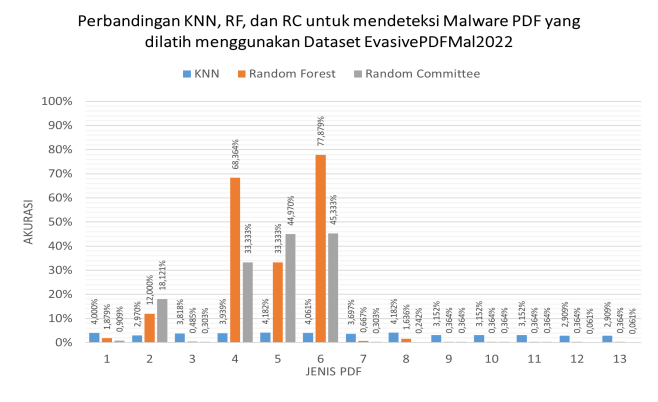
Metrik	Precision			Recall			F1-Score		
Algoritma	KNN	RF	RC	KNN	RF	RC	KNN	RF	RC
Benign	0.97	0.98	0.97	0.91	0.97	0.97	0.94	0.97	0.97
Malicious	0.95	0.98	0.98	0.98	0.99	0.98	0.96	0.99	0.98
Akurasi							0.95	0.98	0.98

3.3. Perbandingan Hasil Test

Bagian ini akan membandingkan hasil uji model *machine learning* dengan algoritma *K-Nearest Neighbor*, *Random forest*, dan *Random committee* yang dilatih menggunakan dataset Evasive-PDFMal2022 dan NewDataset.

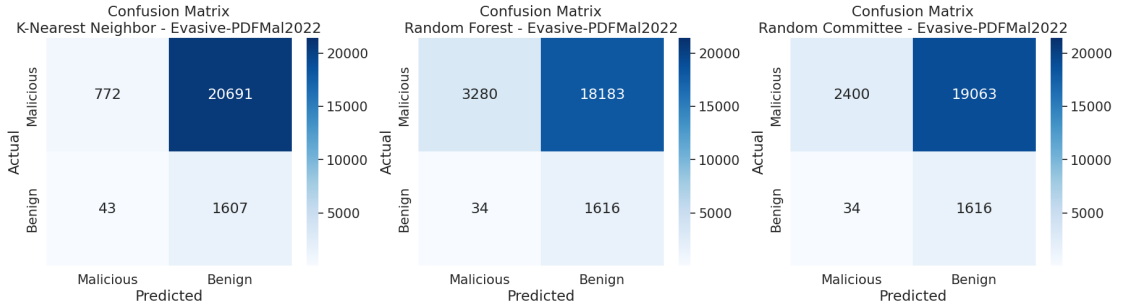
3.3.1. Hasil Test model yang dilatih menggunakan Dataset Evasive-PDFMal2022

Gambar 6 menyajikan analisis komparatif dari akurasi pelatihan model *K-Nearest Neighbor, Random forest*, dan *Random committee* yang dilatih menggunakan dataset Evasive-PDFMal2022 dalam mengidentifikasi 13 varian PDF berbahaya.



Gambar 6. Perbandingan model yang dilatih menggunakan Dataset Evasive-PDFMal2022

Model yang dilatih dengan dataset Evasive-PDFMal2022 menunjukkan kinerja deteksi yang kurang memuaskan. Hasil uji menunjukkan bahwa model tersebut hanya lebih efektif dalam mendeteksi tiga jenis malware PDF, yaitu jenis keempat (adobe_reader_pdf_js_interface), *malware* jenis kelima (adobe_u3d_meshdecl), dan *malware* jenis keenam (adobe_geticon) dengan akurasi di bawah 80%.

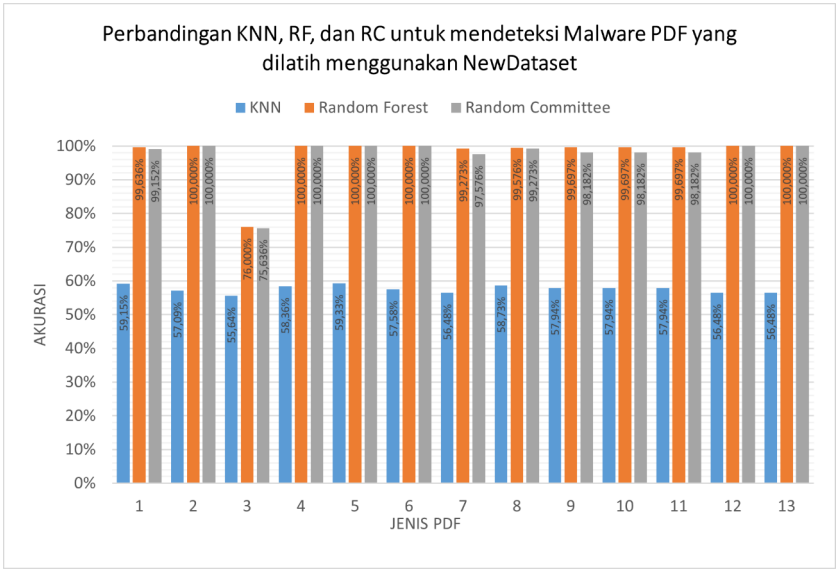


Gambar 7. *Confusion matrix* untuk Model yang dilatih dengan Evasive-PDFMal2022

Gambar 7 menampilkan *confusion matrix* dari pengujian tiga model (*K-Nearest Neighbor*, *Random forest*, dan *Random committee*) yang dilatih menggunakan dataset Evasive-PDFMal2022. Hasilnya menunjukkan kinerja yang kurang memuaskan dalam mendeteksi Malicious PDF, dengan nilai True Positive yang rendah dan dominasi False Positive. Ini mengindikasikan kelemahan dalam kemampuan deteksi malware dari model yang dilatih dengan dataset tersebut.

3.3.2. Hasil Test model yang dilatih menggunakan NewDataset

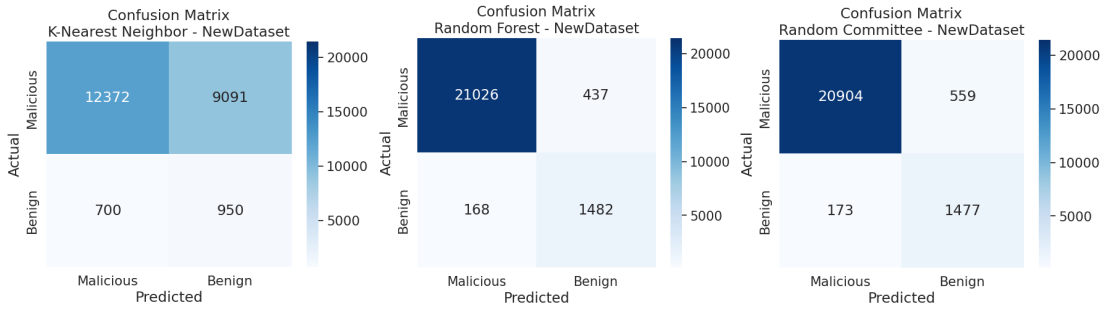
Gambar 8 menyajikan analisis komparatif dari akurasi pelatihan model *K-Nearest Neighbor, Random forest*, dan *Random committee* yang dilatih menggunakan NewDataset untuk mengidentifikasi 13 varian PDF berbahaya.



Gambar 8. Perbandingan model yang dilatih menggunakan NewDataset

Ketiga model yang dilatih menggunakan NewDataset telah menunjukkan peningkatan kinerja yang signifikan dibandingkan dengan model yang dilatih menggunakan dataset Evasive-PDFMal2022. Terutama, model dengan algoritma *Random Forest* dan *Random Committee*

mampu menghasilkan tingkat akurasi rata-rata sekitar 99%. Hal ini menunjukkan kemajuan yang signifikan dalam deteksi *malware* PDF dengan menggunakan dataset yang telah diperbarui.

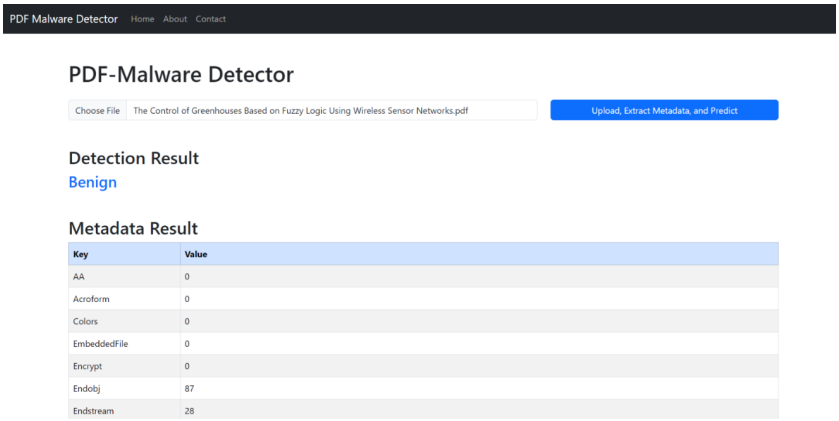


Gambar 9. *Confusion matrix* untuk Model yang dilatih dengan NewDataset

Gambar 9 menunjukkan hasil dari *confusion matrix* saat menguji tiga model (KNN, *Random forest*, dan *Random committee*) yang telah dilatih dengan NewDataset. Dapat dilihat bahwa ketiga model klasifikasi baik *Random forest, Random Committee,* dan *KNN* yang dilatih menggunakan NewDataset memberikan hasil akurasi yang jauh lebih baik apabila dibandingkan dengan model-model yang dilatih menggunakan dataset Evasive-PDFMal2022. Hal ini terlihat dari tingginya nilai *True Positive* dan *True Negative* pada setiap model.

3.4. Membuat Website

Pada tahap akhir penelitian, dibuat sebuah situs web dengan tujuan untuk memberikan manfaat di masa depan. Antarmuka web pendeteksi *malware* diilustrasikan pada Gambar 10.



Gambar 10. Antarmuka *Malware Detection* dan *Malware Extraction Webiste*

Seperti terlihat pada Gambar 10, situs web ini memungkinkan pengguna untuk mengunggah PDF dan memberikan hasil klasifikasi apakah file tersebut termasuk dalam kategori "*Malicious*" atau "*Benign*". Situs web ini merupakan alat penting untuk membantu pengguna mengidentifikasi dan mengelola file PDF dengan potensi risiko keamanan.

Pengujian menggunakan NewDataset menunjukkan bahwa ada ruang untuk pengembangan lebih lanjut pada dataset Evasive-PDFMal2022, meskipun dataset ini merupakan hasil pengembangan dari dataset Contagio, sebagaimana disoroti dalam penelitian sebelumnya [18]. Berbeda dengan penelitian sebelumnya, Penelitian ini menawarkan kontribusi baru dengan pembuatan NewDataset sebagai pengembangan dari Evasive-PDFMal2022, yang membuktikan bahwa peningkatan kualitas dataset sangat diperlukan untuk keakuratan pendeteksian malware PDF. Selain itu, penelitian ini juga memperkenalkan metode yang membandingkan hasil latihan model Random Committee, Random Forest, dan KNN yang dilatih dengan Evasive-PDFMal2022, dengan model yang sama yang dilatih menggunakan NewDataset. Hasil ini menegaskan pentingnya terus memperbaiki dan meningkatkan kualitas dataset serta pemilihan algoritma yang tepat. Seperti yang dijelaskan dalam penelitian sebelumnya [6], Random Committee dan Random Forest terbukti menjadi algoritma yang paling efektif dalam menyelesaikan masalah pendeteksian malware pada PDF menggunakan machine

learning. Oleh karena itu, pemilihan algoritma yang tepat juga memainkan peran kunci dalam meningkatkan kinerja sistem deteksi malware. Selain itu, penelitian ini juga memperkenalkan sebuah website yang mampu mengekstraksi metadata PDF serta mendeteksi malware pada PDF, yang merupakan langkah maju dalam mempermudah proses pendeteksian secara praktis dan cepat.

4. Kesimpulan

Penelitian ini mengusulkan pendekatan menggunakan machine learning untuk mendeteksi PDF berbahaya, dengan memanfaatkan dataset Evasive-PDFMal2022 yang terdiri dari 10.025 baris data yang diperluas menjadi NewDataset dengan 12.825 baris data. Model klasifikasi dilatih menggunakan fitur ekstraksi dari PDF dengan algoritma machine learning yaitu Random Forest (RF), K-Nearest Neighbor (KNN), dan Random Committee (RC). Hasil penelitian ini menunjukkan bahwa model yang dilatih dengan dataset awal, Evasive-PDFMal2022, hanya efektif dalam mendeteksi tiga jenis malware PDF dengan akurasi di bawah 80%. Namun, model yang dilatih dengan NewDataset jauh lebih efektif, terutama dengan algoritma Random Forest dan Random Committee yang mencapai akurasi di atas 99%.

Selain itu, penelitian ini menghasilkan sebuah situs web yang dapat digunakan untuk mengekstraksi metadata PDF dan mendeteksi apakah file tersebut termasuk dalam kategori berbahaya. Situs web ini diharapkan dapat menjadi alat yang berguna bagi pengguna yang ingin memeriksa keamanan file PDF yang diunggah.

Secara keseluruhan, penelitian ini berhasil menghadirkan langkah signifikan dalam pembaruan dan peningkatan representasi dataset, serta hasil pelatihan model yang lebih optimal. Pengembangan lebih lanjut menjadi penting untuk mengeksplorasi dan mengatasi tantangan yang mungkin timbul dalam mendeteksi malware PDF pada skenario yang lebih kompleks. Ini termasuk pengujian model deteksi dalam situasi yang lebih beragam dan realistis, serta penyesuaian terhadap teknik dan algoritma machine learning yang digunakan sesuai dengan karakteristik dan evolusi ancaman siber yang terus berkembang. Dengan demikian, penelitian lanjutan terkait deteksi malware PDF dengan machine learning diharapkan dapat memberikan kontribusi yang lebih besar serta lebih baik dalam memperkuat dan meningkatkan kemampuan deteksi malware PDF, sehingga dapat lebih efektif dalam menghadapi ancaman keamanan siber yang semakin kompleks dan canggih.

Daftar Pustaka

[1] H. Bae, Y. Lee, Y. Kim, U. Hwang, S. Yoon, dan Y. Paek, "Learn2Evade: Learning-Based Generative Model for Evading PDF Malware Classifiers," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 4, hlm. 299–313, Agu 2021, doi: 10.1109/tai.2021.3103139.

[2] International Organization for Standardization, *ISO 32000-2:2020 (PDF 2.0)*, 2 ed. Switzerland: PDF Association, Inc., 2020.

[3] P. Singh, S. Tapaswi, dan S. Gupta, "Malware Detection in PDF and Office Documents: A survey," *Information Security Journal*, vol. 29, no. 3, hlm. 134–153, Mei 2020, doi: 10.1080/19393555.2020.1723747.

[4] Paloalto Networks, "Network Threat Trends Research Report," 2023.

[5] F. Baharuddin dan A. Tjahyanto, "Peningkatan Performa Klasifikasi Machine Learning Melalui Perbandingan Metode Machine Learning dan Peningkatan Dataset," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 1, hlm. 25–31, Mar 2022, doi: 10.32736/sisfokom.v11i1.1337.

[6] S. Y. Yerima dan A. Bashar, "Explainable Ensemble Learning Based Detection of Evasive Malicious PDF Documents," *Electronics (Basel)*, vol. 12, no. 3148, Jul 2023, doi: 10.3390/electronics12143148.

[7] D. Pradeka, "Implementasi Aplikasi Kriptografi Berbasis Android menggunakan Metode Subtitusi dan Permutasi," *In Search – Informatic, Science, Entrepreneur, Applied Art, Research, Humanism*, vol. 18, no. 01, hlm. 161–168, Apr 2019.

[8] M. Elingiusti, L. Aniello, L. Querzoni, dan R. Baldoni, "PDF-Malware detection: A Survey and taxonomy of current techniques," *Advances in Information Security*, vol. 70, hlm. 169–191, 2018, doi: 10.1007/978-3-319-73951-9_9.

[9] W. Suteddy, D. Aprianti, R. Agustini, A. Adiwilaga, dan A. Atmanto, "End-To-End Evaluation of Deep Learning Architectures for Offline Handwriting Writer Identification: A Comparative Study," *JOIV : Int. J. Inform. Visualization*, vol. 7, no. 1, hlm. 178185, Mar 2023, doi: 10.30630/joiv.7.1.1293.

[10] A. N. Syafia, M. F. Hidayattullah, dan W. Suteddy, "Studi Komparasi Algoritma SVM dan Random Forest pada Analisis Sentimen Komentar Youtube BTS," *Jurnal Informatika: Jurnal pengembangan IT (JPIT)*, vol. 8, no. 3, hlm. 207–212, Sep 2023, doi: 10.30591/jpit.v8i3.5064.

[11] R. Fettaya dan Y. Mansour, "Detecting malicious PDF using CNN," Jul 2020, doi: 10.48550/arXiv.2007.12729.

[12] S. A. Roseline, S. Geetha, S. Kadry, dan Y. Nam, "Intelligent Vision-Based Malware Detection and Classification Using Deep Random Forest Paradigm," *IEEE Access*, vol. 8, hlm. 206303–206324, 2020, doi: 10.1109/ACCESS.2020.3036491.

[13] N. F. Munazhif, G. J. Yanris, dan M. N. S. Hasibuan, "Implementation of the K-Nearest Neighbor (kNN) Method to Determine Outstanding Student Classes," *SinkrOn*, vol. 8, no. 2, hlm. 719–732, Apr 2023, doi: 10.33395/sinkron.v8i2.12227.

[14] D. Pradeka, A. Adiwilaga, D. A. R. Agustini, M. B. Hidayatullah, dan A. Suheryadi, *Belajar Dasar Pemrograman Web serta Pengenalan Kriptografi dan Plugin Moodle*, vol. 1. Bandung: Widina Media Utama, 2023.

[15] D. Avelino, L. Cancerlon, M. K. Ryanta, Y. H. Christianto, dan W. Wangnardy, "Penggunaan Bahasa Pemrograman Python dalam Menganalisis Perbedaan Desain Website Tren di Negara Jepang dan Dunia," *Journal of Student Development Information System (JoSDIS)*, vol. 3, no. 2, hlm. 51–61, 2023, doi: 10.36987/josdis.v3i2.4525.

[16] M. Issakhani, P. Victor, A. Tekeoglu, dan A. H. Lashkari, "CIC-Evasive-PDFMal2022," Canadian Institute for Cybersecurity. Diakses: 26 Desember 2023. [Daring]. Tersedia pada: https://www.unb.ca/cic/datasets/pdfmal-2022.html

[17] R. Dubin, "Content Disarm and Reconstruction of PDF Files," *IEEE Access*, vol. 11, hlm. 38399–38416, 2023, doi: 10.1109/ACCESS.2023.3267717.

[18] M. Issakhani, P. Victor, A. Tekeoglu, dan A. H. Lashkari, "PDF Malware Detection based on Stacking Learning," dalam *International Conference on Information Systems Security and Privacy*, Science and Technology Publications, Lda, 2022, hlm. 562–570. doi: 10.5220/0010908400003120.

[19] A. Althnian *dkk.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Applied Sciences (Switzerland)*, vol. 11, no. 2, hlm. 1–18, Jan 2021, doi: 10.3390/app11020796.

[20] K. Koptyra dan M. R. Ogiela, "Distributed steganography in PDF files - Secrets hidden in modified pages," *Entropy*, vol. 22, no. 6, Jun 2020, doi: 10.3390/E22060600.