Visualisasi PDF *Malware* Menggunakan *Clustering K-Means* pada Layanan GARUDA Kemdikbud Dikti sebagai Agregator Nasional

TUGAS AKHIR

OLEH :

INDAH CAHYA RESTI

09011281823046

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2023

Visualisasi PDF *Malware* Menggunakan *Clustering K-Means* pada Layanan GARUDA Kemdikbud Dikti sebagai Agregator Nasional

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

OLEH :

INDAH CAHYA RESTI

09011281823046

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2023

HALAMAN PENGESAHAN

VISUALISASI PDF *MALWARE* MENGGUNAKAN *CLUSTERING K-MEANS* PADA LAYANAN GARUDA KEMDIKBUD DIKTI SEBAGAI AGREGATOR NASIONAL

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

Oleh

INDAH CAHYA RESTI

09011281823046

Indralaya, 12 Januari 2023

Pembimbing I Tugas Akhir	Pembimbing II Tugas Akhir
Deris Stiawan, M.T., Ph.D.	**Tri Wanda Septian, M.Sc.**
NIP. 197806172006041002	**NIK. 1901062809890001**

Mengetahui,

Ketua Jurusan Sistem Komputer

Dr. Ir. H. Sukemi, M.T.

NIP. 196612032006041001

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada :

Hari : Jum'at

Tanggal : 16 Desember 2022

Tim Penguji :

1. **Ketua** : **Ahmad Heryanto, M.T.**
2. **Sekretaris** : **Adi Hermansyah, M.T.**
3. **Penguji** : **Huda Ubaya, M.T.**
4. **Pembimbing I** : **Deris Stiawan, M.T., Ph.D.**
5. **Pembimbing II** : **Tri Wanda Septian, M.Sc.**

Mengetahui, 13/1/23
Ketua Jurusan Sistem Komputer

Dr.Ir.H. Sukemi, M.T.
NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Indah Cahya Resti

NIM : 09011281823046

Judul : Visualisasi PDF *Malware* Menggunakan *Clustering* K-Means pada Layanan GARUDA Kemdikbud Dikti sebagai Agregator Nasional

Hasil Pengecekan Software *iThenticate/Turnitin* : 10%

Menyatakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsru penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari universitas sriwijaya.

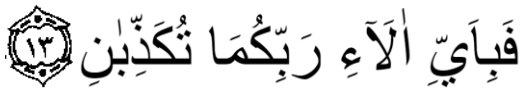
Demikian, pernyatan ini saya buat dalam keadaan sadar dan tanpa paksaan dari siapapun.

Indralaya, Januari 2023

Indah Cahya Resti

NIM.09011281823046

HALAMAN PERSEMBAHAN



“Maka nikmat Tuhanmu manakah yang kamu dustakan (wahai jin dan manusia)?”

(Q.S. Ar-Rahman : 13)

Aku berdoa pada Tuhan, mengeluh sebab seharian hawa di kotaku panas tak kepalang, aku meminta hujan turun untuk mendinginkan.

Aku bersungut-sungut sebab doaku tak kunjung dikabulkan.

Tanpa aku ketahui, ada doa seorang penjual minuman pinggir jalan yang ingin dagangannya laku sebab malam nanti harus menebus obat, ada doa seorang orang tua tunggal yang belum punya biaya membetulkan genteng dan atap rumah, ada seorang supir yang baru saja kehilangan kacamatanya tak mampu melihat jelas jalanan yang tertutup rintik hujan padahal sedang kejar setoran.

Doa manusia bermacam-macam.

Tunggu giliran.

Semuanya tak bisa semaumu.

Yang ditunda belum tentu tak indah.

-9996

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Puji syukur atas kehadirat Allah SWT serta segala karunia dan rahmat-Nya sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul **"Visualisasi PDF *Malware* Menggunakan *Clustering K-Means* pada Layanan GARUDA Kemdikbud Dikti sebagai Agregator Nasional"**.

Pelaksanaan dan penyusunan Tugas Akhir ini tidak mungkin berhasil tanpa adanya bantuan dari berbagai pihak terutama do'a dari kedua orangtua serta dukungan moral seperti bimbingan, nasihat, ide, saran serta semangat agar Tugas Akhir ini dapat terselesaikan. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada :

1. Allah Subhanahu Wa Ta'ala yang telah memberikan berkah dan hidayah-Nya serta nikmat yang tak terhitung.
2. Mama, Papa, Adik, Ibu, Ayah dan keluarga besar yang telah memberikan doa dan restu serta dukungan yang sangat besar selama penulis berjuang dalam perkuliahan.
3. Bapak Jaidan Jauhari, M.T., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Dr. Ir. H. Sukemi, M.T., selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer.
5. Bapak Deris Stiawan, M.T., PH.D., selaku Dosen Pembimbing I Tugas Akhir dan Dosen Pembimbing Akademik yang telah berkenan meluangkan waktunya agar dapat membimbing, memberikan saran dan motivasi terbaik untuk penulis dalam menyelesaikan Tugas Akhir ini.
6. Kak Tri Wanda Septian, S.Kom., M.Sc., selaku Dosen Pembimbing II Tugas Akhir yang bersedia membimbing, memberikan saran dan masukan untuk penulis dalam menyelesaikan Tugas Akhir ini.
7. Mbak Nurul Afifah M.Kom yang selalu memberikan arahan, masukan dan saran kepada Tim Riset PDF *Malware* (GARUDA).

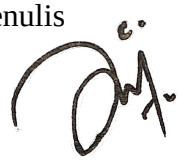
8. Mbak Renny Virgasari selaku Admin Jurusan Sistem Komputer yang telah membantu penulis dalam hal-hal administrasi.
9. Retjech caps (Alifah Fidela dan Rizki Valen Mafaza), rekan kacamata (Dimas, Furqon, Arif, Imam, Farhan, Taufik, Realdi) dan SKB 2018 yang juga telah banyak membantu dan menghibur penulis selama perkuliahan, glad to know all of you.
10. Teman-teman dalam Tim Riset PDF *Malware* (GARUDA), yaitu Alifah, Rani, Novi, Shena, dan Nata.
11. Penghuni Eldas x Lab Robotika x Bultang x Pingpong, yaitu Arif, Furqon, Alifah, Valen, Rani, Ades, Shena, Taufik, Imam, Farhan, Alif, Tedi, dan HanaNur.
12. Grup Riset COMNETS.
13. Seluruh teman seperjuangan dari SK 2018 serta kakak-kakak tingkat.
14. Oyen selaku hewan peliharaan yang menemani dan menghibur penulis selama perkuliahan ini.
15. Seluruh orang baik yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan semangat serta do'a yang baik.
16. Almamater.

Dalam penulisan Tugas Akhir ini, penulis menyadari bahwa masih banyak kekurangan. Maka dari itu, penulis memohon maaf serta menerima kritik dan saran sebagai bahan evaluasi penulis untuk di masa mendatang. Harapan penulis agar Tugas Akhir ini bermanfaat dan berguna bagi pembacanya.

Wassalamu'alaikum Wr. Wb.

Indralaya, Januari 2023

Penulis



Indah Cahya Resti

NIM. 09011281823046

VISUALIZATION OF MALWARE PDF USING CLUSTERING K-MEANS ON GARUDA SERVICE KEMDIKBUD DIKTI AS NATIONAL AGGREGATOR

INDAH CAHYA RESTI (09011281823046)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email : indahcahya666@gmail.com

ABSTRACT

K-Means clustering is a method to grouping data based on the similarity of features and detect the hidden patterns in dataset. The dataset is from GARUDA Repository which contains raw data of PDF files. GARUDA dataset extraction process used static analysis method. The data extraction process produced twenty-one features using PDFiD. GARUDA dataset has a multi-class and imbalanced data, therefore a SMOTE process is required. K-Means succeed to grouping 3 clusters with silhouette score is 0,71311. A best validation result is using K-Means label and support with Logistic Regression model at 5-Fold. The accuracy of K-Means label is 94,66%, hence K-Means labeling is better than GARUDA labeling that only obtained the accuracy of 87,16%.

Keywords *: **Malware PDF**, **Static Analysis**,* **SMOTE***, **Clustering**,* **K-Means, Silhouette Score, Stratified K-Fold***

VISUALISASI PDF *MALWARE* MENGGUNAKAN *CLUSTERING* K-MEANS PADA LAYANAN GARUDA KEMDIKBUD DIKTI SEBAGAI AGREGATOR NASIONAL

INDAH CAHYA RESTI (09011281823046)
Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
Email : indahcahya666@gmail.com

ABSTRAK

K-Means clustering adalah suatu metode untuk mengelompokkan data berdasarkan kesamaan fitur dan mendeteksi pola tersembunyi yang ada dalam kumpulan data. Dataset berasal dari GARUDA Repository yang berisi data mentah file PDF. Proses ekstraksi dataset PDF GARUDA menggunakan metode analisis statik. Proses ekstraksi data menghasilkan 21 fitur menggunakan PDFiD. Dataset GARUDA memiliki banyak kelas dan data tidak seimbang, oleh karena itu diperlukan proses SMOTE. K-Means berhasil mengelompokkan 3 clusters dengan silhouette score sebesar 0,71311. Hasil validasi terbaik menggunakan label K-Means dengan bantuan model Logistic Regression pada 5-Fold. Akurasi menggunakan label K-Means sebesar 94,66%, sehingga pelabelan K-Means lebih baik dibandingkan menggunakan label PDF GARUDA yang hanya mendapatkan akurasi sebesar 87,16%.

Kata Kunci : **PDF *Malware*, *Static Analysis*, SMOTE, *Clustering*, K-Means, *Silhouette Score*, *Stratified K-Fold***

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1. Latar Belakang

PDF (*Portable Document Format*) adalah format *file* yang ditemukan oleh *Adobe* untuk menyajikan, bertukar, dan mengarsipkan dokumen yang tidak bergantung pada perangkat keras, perangkat lunak, dan sistem operasi. Sebagai salah satu format *file* yang paling banyak digunakan, yaitu dokumen PDF telah menjadi salah satu vektor utama dari serangan *malware* [1]. Fleksibilitas struktur *file* PDF dalam menyematkan berbagai jenis konten seperti kode *JavaScript*, aliran yang disandikan dan objek gambar, dll. Fitur ini dapat dimanfaatkan oleh penyerang untuk menyematkan *malware* dalam *file* PDF menggunakan alat seperti *Metasploit* [2]. Misalnya, dilaporkan bahwa *Ransomware* populer saat ini dapat disembunyikan di dalam dokumen PDF untuk memasukkan serangannya.

Salah satu *platform* yang menyediakan *file* PDF yang beberapa artikelnya dapat diakses secara *open-access* adalah GARUDA *repository*. GARUDA (Garba Rujukan Digital) adalah *platform* sumber informasi publikasi ilmiah di Indonesia yang dikelola oleh Kemenristekdikti. Garuda memiliki seluruh aspek ilmu pengetahuan seperti ilmu sosial, pendidikan, komputer, hukum, biologi-fisika, ekonomi, dan masih banyak yang lainnya. *Repository* ini mempunyai lebih dari 1.789.340 artikel, sehingga kemungkinan juga terdapat *file* PDF yang disusupi oleh *malware*.

Clustering adalah suatu cara mengelompokkan data mentah secara wajar dan mencari pola tersembunyi yang mungkin ada dalam kumpulan data. Proses pengelompokan objek data ke dalam *cluster* yang terputus-putus sehingga data-data yang berada di dalam kluster adalah serupa, namun data-data akan tergabung dalam *cluster* yang berbeda-beda. Dengan meningkatkan dan mempelajari informasi penting dari data, yang membuat teknik *clustering* banyak diterapkan di banyak bidang aplikasi seperti kecerdasan buatan, biologi, manajemen hubungan pelanggan, kompresi data, penambangan data, pencarian informasi, pemrosesan gambar, pembelajaran mesin, pemasaran, kedokteran, pengenalan pola, psikologi,

statistik, dan lain-lain [3]. *Clustering* juga dapat membantu mendeteksi suatu serangan ketika data *training unlabeled*, serta untuk mendeteksi serangan baru atau serangan yang tidak diketahui [4].

Pada penelitian [5], dalam melakukan klasifikasi hanya menggunakan algoritma C5.0 tanpa bantuan *SMOTE* diperoleh akurasi lebih dari 90% dan kelas prediksi yang benar kurang dari 82%. Sedangkan, model algoritma C5.0 menggunakan *SMOTE* dengan *k-fold cross validation* menunjukkan akurasi terbaik mencapai 99% dan mampu memperoleh kelas prediksi yang benar hingga 99%.

Penelitian [6] menyebutkan bahwa *K-Means* maupun FCM (*Fuzzy C-Means*) berhasil menemukan cekungan dan jenis kluster yang lain berbentuk arbitrer ketika sebuah kelompok tidak terpisah dengan baik. Di beberapa penelitian FCM akan memberikan hasil yang lebih baik dalam kumpulan data yang bising, akan tetapi KM akan menjadi pilihan yang baik untuk kumpulan data yang besar karena kecepatan eksekusinya. *K-Means* dengan beberapa permulaan direkomendasikan untuk analisis *cluster* karena akurasi dan kinerja waktu yang sebanding. Penelitian [7] juga menyebutkan algoritma *K-Means* menggunakan *dataset* Twonorm menunjukkan hasil *clustering* dengan baik mencapai 90% berdasarkan MI (*Mutual Information*) dan ARI (*Adjusted Rand Index*).

Lalu penelitian selanjutnya [8], algoritma *KMeans++ SMOTE* adalah gabungan dari algoritma *k-means++* dan *SMOTE* untuk mengatasi *random oversampling*. Algoritma ini bekerja dengan memilih titik pusat klaster yang lebih jauh dan lebih baik, lalu membagi tingkat regional titik pusat klaster dan mengontrol tingkat pengambilan sampel baru yang disintetis oleh kelas yang berbeda. Sehingga, hasil penelitian menunjukkan bahwa algoritma *KMeans++ SMOTE* memiliki kinerja yang lebih baik daripada algoritma lainnya (*SMOTE, Borderline-SMOTE, KM-SMOTE,* dan *C-SMOTE*) dalam mengatasi data yang tidak seimbang.

Berdasarkan latar belakang yang telah disebutkan sebelumnya, maka penulis akan melakukan *clustering* terhadap *file-file* PDF *malware* menggunakan *dataset* yang terdapat pada layanan GARUDA tersebut dengan memberi judul

penelitian, yaitu "Visualisasi *PDF Malware* Menggunakan *Clustering K-Means* pada Layanan GARUDA Kemdikbud Dikti sebagai Agregator Nasional".

1.2. Rumusan Masalah

Berikut adalah rumusan masalah yang ada pada penyusunan Tugas Akhir :

1. Bagaimana cara mengekstrak fitur *dataset* pada PDF GARUDA?
2. Bagaimana memvisualisasikan pola data PDF *benign,* PDF *malware,* serta non-PDF *malware* ke dalam bentuk grafik?
3. Bagaimana performa dari metode *K-Means* dalam mengelompokkan PDF *benign,* PDF *malware,* serta non-PDF *malware*?

1.3. Batasan Masalah

Berikut ini adalah batasan masalah yang ada pada penyusunan Tugas Akhir:

1. Penelitian dilakukan dalam permasalahan PDF *malware*.
2. Data yang digunakan adalah kumpulan *file* PDF yang berasal dari GARUDA (Garba Rujukan Digital) *repository*.
3. Metode yang diterapkan menggunakan algoritma *Clustering K-Means*.
4. Dalam penelitian ini tidak membahas tentang bagaimana pencegahan terhadap *file* PDF *malware*.

1.4. Tujuan

Adapun tujuan dari penyusunan Tugas Akhir ini adalah sebagai berikut :

1. Mengekstrak *file* PDF GARUDA untuk mendapatkan fitur-fitur yang digunakan sebagai *dataset* menggunakan metode analisis statik.
2. Melakukan visualisasi pola data PDF *benign,* PDF *malware,* non-PDF *malware* menggunakan *Parallel Coordinates*.
3. Melakukan analisis terhadap hasil kinerja *clustering* pada *file* PDF *benign,* PDF *malware,* non-PDF *malware*.

1.5. Manfaat

Adapun manfaat dari penyusunan Tugas Akhir ini adalah sebagai berikut :

1. *Dataset* dari PDF GARUDA *Repository* dapat digunakan untuk mengenali pola dan pengelompokkan PDF *malware*.

2. *Parallel Coordinates* dapat membuat visualisasi pola data PDF *benign*, PDF *malware*, non-PDF *malware* untuk mendapatkan informasi dalam bentuk grafik.
3. Hasil performa dari penelitian ini digunakan untuk mendapatkan nilai validasi seperti presisi, *recall*, *f1-score* dan akurasi dengan menerapkan algoritma *K-Means* sebagai metode *clustering* dalam mengelompokkan data sesuai kesamaan kelasnya.

1.6. Metodologi Penelitian

Metodologi yang digunakan dalam penyusunan Tugas Akhir ini sebagai berikut :

1. Studi Pustaka

   Tahap awal yang dilakukan dengan cara mencari dan mengumpulkan referensi yang berupa studi *literature* terhadap paper atau jurnal yang berhubungan dengan tugas akhir.
2. Pengolahan Data

   Tahap ini membahas mengenai proses pembuatan sebuah data mentah menjadi data siap olah, lalu dilakukan *oversampling* data menggunakan SMOTE.
3. Pengujian

   Tahap ini berupa pengujian terhadap rancangan pemodelan berdasarakan metodologi penelitian sehingga didapatkan data hasil uji yang sesuai dan tepat dengan algoritma *Clustering K-Means*.
4. Hasil dan Analisa

   Hasil dari pengolahan dan pengujian data akan dianalisis sesuai identifikasi permasalahan. Tahapan ini bertujuan untuk mendapatkan hasil yang objektif dari proses pengujian data yang telah diperoleh.
5. Kesimpulan dan Saran

   Tahap terakhir adalah membuat kesimpulan dari rumusan permasalahan, metodologi, dan analisa hasil pengujian. Tahapan ini juga terdapat saran untuk penelitian selanjutnya.

1.7. Sistematika Penulisan

Adapun sistematika penulisan dalam Tugas Akhir adalah sebagai berikut :

BAB I. PENDAHULUAN

Pada Bab ini berisi tentang landasan topik penelitian yang meliputi Latar Belakang, Rumusan Masalah, Batasan Masalah, Tujuan, dan Manfaat, serta termasuk Metodologi Penelitian dan Sistematika Penulisan.

BAB II. TINJAUAN PUSTAKA

Pada Bab ini menjelaskan dasar teori dari penelitian tugas akhir tentang PDF *Malware*, *Clustering K-Means*, serta teori yang berhubungan dengan penelitian.

BAB III. METODOLOGI PENELITIAN

Pada Bab III akan melakukan rancangan ataupun rincian sistematis terhadap penelitian yang akan dilakukan. Rincian mengenai kerangka kerja penelitian, tahapan pemrosesan data, hingga penerapan algoritma *Clustering K-Means*.

BAB IV. ANALISA DAN PEMBAHASAN

Pada Bab ini akan membahas dan menganalisa hasil yang telah diuji dan diperoleh dari tahap sebelumnya serta validasi hasil agar mendapatkan data yang akurat.

BAB V. KESIMPULAN DAN SARAN

Pada Bab terakhir ini akan menuliskan kesimpulan yang didapatkan selama proses penelitian sebagai jawaban dari target yang akan dicapai, serta saran yang diharapkan dapat dikembangkan lebih baik lagi.

DAFTAR PUSTAKA

[1] A. Corum, D. Jenkins, and J. Zheng, "Robust PDF Malware Detection with Image Visualization and Processing Techniques," *Proc. - 2019 2nd Int. Conf. Data Intell. Secur. ICDIS 2019*, pp. 108–114, 2019, doi: 10.1109/ICDIS.2019.00024.

[2] J. Park and H. Kim, "K-depth mimicry attack to secretly embed shellcode into PDF files," *Lect. Notes Electr. Eng.*, vol. 424, pp. 388–395, 2017, doi: 10.1007/978-981-10-4154-9_45.

[3] N. Shi, X. Liu, and Y. Guan, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," *3rd Int. Symp. Intell. Inf. Technol. Secur. Informatics, IITSI 2010*, pp. 63–67, 2010, doi: 10.1109/IITSI.2010.74.

[4] V. Kumar, H. Chauhan, and D. Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset," *Int. J. Soft Comput. Eng.*, vol. 3, no. 4, pp. 1–4, 2013.

[5] E. Kurniawan, F. Nhita, A. Aditsania, and D. Saepudin, "C5.0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in bandung regency," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, vol. 4, pp. 1–5, 2019, doi: 10.1109/ICoICT.2019.8835324.

[6] Z. Cebeci and F. Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures," *J. Agric. Informatics*, vol. 6, no. 3, pp. 13–23, 2015, doi: 10.17700/jai.2015.6.3.196.

[7] K. Sari, S. Efendi, and S. Nasution, "Combining the Active Learning Algorithm Based on the Silhouette Coefficient with PCKmeans Algorithm," *Mecn. 2020 - Int. Conf. Mech. Electron. Comput. Ind. Technol.*, pp. 232–237, 2020, doi: 10.1109/MECnIT48290.2020.9166596.

[8] C. Li, D. Ping, S. Wei, and Z. Yan, "Improving Classification of Imbalanced Datasets Based on KM++ SMOTE Algorithm," *Proc. - 2019 2nd Int. Conf.*

Saf. Prod. Informatiz. IICSPI 2019, pp. 300–306, 2019, doi: 10.1109/IICSPI48186.2019.9096022.

[9] H. Bae, Y. Lee, Y. Kim, U. Hwang, S. Yoon, and Y. Paek, "Learn2Evade: Learning-Based Generative Model for Evading PDF Malware Classifiers," *IEEE Trans. Artif. Intell.*, vol. 2, no. 4, pp. 299–313, 2021, doi: 10.1109/tai.2021.3103139.

[10] S. G. Sayed and M. Shawkey, "Data Mining Based Strategy for Detecting Malicious PDF Files," *Proc. - 17th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. 12th IEEE Int. Conf. Big Data Sci. Eng. Trust. 2018*, pp. 661–667, 2018, doi: 10.1109/TrustCom/BigDataSE.2018.00097.

[11] R. Fettaya and Y. Mansour, "Detecting malicious PDF using CNN," no. i, pp. 1–18, 2020, [Online]. Available: http://arxiv.org/abs/2007.12729.

[12] B. Cuan, A. Damien, C. Delaplace, and M. Valois, "Malware detection in PDF files using machine learning," *ICETE 2018 - Proc. 15th Int. Jt. Conf. E-bus. Telecommun.*, vol. 2, pp. 412–419, 2018, doi: 10.5220/0006884704120419.

[13] N. Srndic and P. Laskov, "Detection of Malicious PDF Files Based on Hierarchical Document Structure," *Proc. 20th Annu. Netw. Distrib. Syst. Symp.*, 2013, [Online]. Available: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2017/09/12_3_0.pdf%0Ahttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Detection+of+Malicious+PDF+Files+Based+on+Hierarchical+Document+Structure#0.

[14] Y. S. Jeong, J. Woo, and A. R. Kang, "Malware Detection on Byte Streams of PDF Files Using Convolutional Neural Networks," *Secur. Commun. Networks*, vol. 2019, 2019, doi: 10.1155/2019/8485365.

[15] P. O. Olukanmi, F. Nelwamondo, and T. Marwala, "K-Means-MIND: An Efficient Alternative to Repetitive k-Means Runs," *2020 7th Int. Conf. Soft Comput. Mach. Intell. ISCMI 2020*, pp. 172–176, 2020, doi: 10.1109/ISCMI51676.2020.9311598.

[16] M. Elingiusti, L. Aniello, L. Querzoni, and R. Baldoni, “PDF-Malware detection: A Survey and taxonomy of current techniques,” *Adv. Inf. Secur.*, vol. 70, pp. 169–191, 2018, doi: 10.1007/978-3-319-73951-9_9.

[17] M. Iwamoto, S. Oshima, and T. Nakashima, “A Study of Malicious PDF Detection Technique,” *Proc. - 2016 10th Int. Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2016*, pp. 197–203, 2016, doi: 10.1109/CISIS.2016.45.

[18] N. Fleury, T. Dubrunquez, and I. Alouani, “PDF-Malware: An Overview on Threats, Detection and Evasion Attacks,” no. July, 2021, [Online]. Available: http://arxiv.org/abs/2107.12873.

[19] C. Ulucenk, V. Varadharajan, V. Balakrishnan, and U. Tupakula, “Techniques for analysing PDF malware,” *Proc. - Asia-Pacific Softw. Eng. Conf. APSEC*, pp. 41–48, 2011, doi: 10.1109/APSEC.2011.41.

[20] H. Yan and J. Wang, “Visualization,” pp. 661–665, 2017, doi: 10.1109/DSC.2017.110.

[21] H. Choi, H. Lee, and H. Kim, “Fast detection and visualization of network attacks on parallel coordinates,” *Comput. Secur.*, vol. 28, no. 5, pp. 276–288, 2009, doi: 10.1016/j.cose.2008.12.003.

[22] H. I. Lin and M. C. Nguyen, “Boosting minority class prediction on imbalanced point cloud data,” *Appl. Sci.*, vol. 10, no. 3, 2020, doi: 10.3390/app10030973.

[23] W. Yanbo, L. Li, P. Xinfu, and F. Enpeng, “Load forecasting based on improved K-means clustering algorithm,” *China Int. Conf. Electr. Distrib. CICED*, no. 201804260000067, pp. 2751–2755, 2018, doi: 10.1109/CICED.2018.8592023.

[24] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[25] X. Wang, “An Improved K _ means Algorithm for Document Clustering

Based on Knowledge Graphs," *2018 11th Int. Congr. Image Signal Process. Biomed. Eng. Informatics*, pp. 1–5, 2018.

[26] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.

[27] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV : Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," no. August, 2022, doi: 10.3389/fnano.2022.972421.

[28] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis, "Unsupervised stratification of cross-validation for accuracy estimation," *Artif. Intell.*, vol. 116, no. 1–2, pp. 1–16, 2000, doi: 10.1016/S0004-3702(99)00094-6.

[29] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019*, pp. 135–139, 2019, doi: 10.1109/ICCSNT47585.2019.8962457.

[30] D. E. Zomahoun, "A semantic collaborative clustering approach based on confusion matrix," *Proc. - 15th Int. Conf. Signal Image Technol. Internet Based Syst. SISITS 2019*, pp. 688–692, 2019, doi: 10.1109/SITIS.2019.00112.

[31] F. Krüger, "Activity, Context, and Plan Recognition with Computational Causal Behaviour Models," *ResearchGate*, no. August, 2018, [Online]. Available: https://www.researchgate.net/figure/Confusion-matrix-for-multi-class-classification-The-confusion-matrix-of-a_fig7_314116591.