KLASIFIKASI PDF MALWARE PADA GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE LOGISTIC REGRESSION

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S1)

OLEH:

VIRGINITA PUTRI LESTARI
09011382025136

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2024

LEMBAR PENGESAHAN

KLASIFIKASI PDF MALWARE PADA GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE LOGISTIC REGRESSION

SKRIPSI

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer (S1)

Oleh
VIRGINITA PUTRI LESTARI
09011382025136

Palembang, Juni 2024
Mengetahui,

Pembimbing Tugas Akhir I	Pembimbing Tugas Akhir II
Prof. Deris Stiawan, M.T., Ph.D.	Nurul Afifah, M.Kom.
NIP. 197806172006041002	NIP. 199211102023212049

Ketua Jurusan Sistem Komputer 24/6/24

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

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada

Hari : **Rabu**

Tanggal : **22 Mei 2024**

Tim Penguji

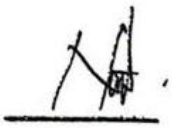
1. **Ketua** : **Huda Ubaya, M.T**

2. **Sekretaris** : **Abdurahman, M. Han**

3. **Penguji** : **Dr. Ahmad Zarkasi, M.T**

4. **Pembimbing I** : **Prof. Deris Stiawan, M.T., Ph.D**

5. **Pembimbing II** : **Nurul Afifah, M.Kom**



Mengetahui, 24/6/24
Ketua Jurusan Sistem Komputer

Dr. Ir. Sukemi, M.T.
NIP. 19661203200604100l

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Virginita Putri Lestari

NIM : 09011382025136

Judul : Klasifikasi PDF Malware Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti Dengan Metode Logistic Regression

Hasil Pengecekan Plagiat/Turnitin: 6%

Menyatakan bahwa laporan tugas akhir ini adalah hasil karya saya sendiri dan tidak mengandung unsur penjiplakan atau plagiat. Saya sepenuhnya menyadari bahwa jika terbukti adanya penjiplakan atau plagiat dalam laporan tugas akhir ini, saya siap menerima sanksi akademik dari Universitas Sriwijaya. Pernyataan ini saya buat dengan kesadaran penuh dan tanpa adanya paksaan dari pihak manapun.

Palembang, Juni 2024

Virginita Putri Lestari

NIM. 09011382025136

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh,

Puji dan syukur penulis ucapkan atas kehadirat Allah SWT. Yang telah melimpahkan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan penyusunan tugas akhir dengan judul **"Klasifikasi Pdf Malware Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti Dengan Metode Logistic Regression"**.

Pada penyusunan tugas akhir ini tidak terlepas dari peran berbagai pihak yang telah memberikan dukungan doa, semngat, motivasi dan bimbingan pada penulis. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Allah SWT. yang telah memberikan nikmat Kesehatan dan Kesempatan kepada penulis dalam penyusunan tugas akhir ini.
2. Kedua orang tua tercinta (Hengki Gunawan & Zuryati) yang selalu memberikan dukungan moral maupun finansial, serta do'a yang tiada hentinya.
3. Ayuk (Yunika Pratiwi) dan Adik (Yusuf Pratama) yang selalu memberikan dukungan semangat dan do'a.
4. Bapak Prof. Dr. Erwin, S.Si., M.Si selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Dr. Ir. H. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Universitas Sriwijaya.
6. Bapak Rahmat Fadli Isnanto, S.SI., M.SC. selaku Dosen Pembimbing Akademik.
7. Bapak Prof. Deris Stiawan, M.T., Ph.D., IPU., ASEAN-Eng., CPENT. selaku Pembimbing I Tugas Akhir Penulis yang telah meluangkan waktu untuk membimbing dan memberikan motivasi selama pengerjaan Tugas Akhir.
8. Mbak Nurul Afifah, M.Kom. selaku Pembimbing II Tugas Akhir yang telah meluangkan waktu untuk membimbing penulis dalam pengerjaan Tugas Akhir dari awal penelitian.
9. Mbak Sari Anhar selaku admin yang telah membantu dalam proses administrasi Tugas Akhir Penulis.

10. Teman-teman seperjuangan sejak awal perkuliahan yang selalu memberikan semangat dan solusi dalam pengerjaan Tugas Akhir, yaitu Indah Ria Andina, Risky Wahyuni, dan Muhammad Ramadhanil. Sukses untuk kita semua guys!
11. Teman-teman satu kelompok riset yang selalu memberikan semangat dan solusi kepada penulis yaitu Krisna Agustini, Siti Khaeronisyah, dan Cynthia Anggraeni.
12. Teman-teman seperjuangan Jurusan Sistem Komputer Unggulan 2020.
13. Seluruh pihak yang tidak dapat penulis sebutkan satu per satu, yang telah memberikan semangat serta doa.
14. Almamater

Penulis menyadari bahwa laporan ini masih jauh dari kesempurnaan, oleh karena itu penulis dengan senang hati menerima kritik dan saran serta masukkan dari pembaca yang bersifat membangun agar lebih baik lagi dikemudian hari. Penulis berharap semoga laporan ini dapat bermanfaat bagi kita semua khususnya bagi mahasiswa Fakultas Ilmu Komputer Universitas Sriwijaya. Demikian yang dapat penulis sampaikan.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Palembang, Juni 2024
Penulis,

Virginita Putri Lestari
NIM. 09011382025136

KLASIFIKASI PDF MALWARE PADA GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE LOGISTIC REGRESSION

VIRGINITA PUTRI LESTARI (09011382025136)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer
Universitas Sriwijaya
Email: virginita08@gmail.com

ABSTRAK

Malware dapat masuk melalui file PDF yang tampak tidak mencurigakan merupakan salah satu faktor utama dalam serangan cyber security. Dataset GARUDA dianalisis secara statis menggunakan VirusTotal dan PDFiD untuk mengidentifikasi apakah sebuah file PDF berbahaya atau tidak, kemudian dilakukan Klasifikasi untuk mengetahui karakteristik file PDF tersebut menggunakan metode Logistic Regression jenis Multinomial. Dataset yang digunakan terdiri dari 10.000 file format PDF dengan 21 variabel prediksi dan terdapat 3 kategori kelas. Dataset GARUDA memiliki data yang tidak seimbang, oleh karena itu dilakukan teknik Random Oversampling untuk mengatasinya. Hasil penelitian menunjukkan bahwa model Logistic Regression jenis Multinomial mampu mencapai akurasi sebesar 93%. Hasil ini mengidikasikan bahwa model memiliki kinerja yang dapat diandalkan dalam melakukan klasifikasi.

Kata Kunci ***:*** *Logistic Regression, PDF Malware, Imbalance Dataset.*

CLASSIFICATION OF PDF MALWARE ON THE DIGITAL REFERENCE GARBA (GARUDA) OF THE KEMDIKBUD DIKTI USING THE LOGISTIC REGRESSION METHOD

VIRGINITA PUTRI LESTARI (09011382025136)

Department of Computer Systems, Computer Science Faculty
Sriwijaya University
Email: virginita08@gmail.com

ABSTRACT

Malware that can enter through PDF files that appear unsuspicious is one of the main factors in cyber security attacks. The GARUDA dataset was analyzed statically using VirusTotal and PDFiD to identify whether a PDF file is dangerous or not, then classification was carried out to determine the characteristics of the PDF file using the Logistic Regression method of the Multinomial type. The dataset used consists of 10,000 PDF format files with 21 prediction variables and there are 3 class categories. The GARUDA dataset has unbalanced data, therefore a Random Oversampling technique is used to overcome it. The results show that the Multinomial Logistic Regression model is able to achieve an accuracy of 93%. These results indicate that the model has reliable performance in performing classification.

Keywords ***:*** *Logistic Regression, PDF Malware, Imbalance Dataset.*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Portable Document Format (PDF) merupakan salah satu format file paling populer yang diperkenalkan oleh Adobe Systems pada tahun 1993, bertujuan untuk menyediakan format dokumen yang dapat dengan mudah dipertukarkan dengan sistem operasi dan perangkat keras yang berbeda. Dalam perkembangannya, Portable Document Format (PDF) telah menjadi standar pertukaran dokumen elektronik yang dirancang untuk menyimpan dan menyajikan informasi dalam bentuk teks, gambar, grafik, dengan cara terstruktur seperti pada *e-journal*, karya ilmiah, dan lain sebagainya [1]. Diperkirakan lebih dari 114 juta dokumen PDF ada di internet, dimana lebih dari 27 juta (24%) mudah diakses tanpa pembayaran dan registrasi [2]. Banyak perpustakaan digital yang menyajikan dokumen karya ilmiah dalam bentuk format PDF yang dapat diakses dengan mudah salah satunya Garba Rujukan Digital (GARUDA).

Garba Rujukan Digital (GARUDA) merupakan perpustakaan digital yang berisikan rujukan ilmiah, e-journal, hasil penelitian dan tesis dari peneliti akademisi Indonesia yang dikelola oleh KEMDIKBUD DIKTI [3]. Peneliti dari akademisi mengandalkan perpustakaan digital ilmiah untuk mengakses dokumen format PDF pada sebuah cloud penyimpanan, yang bisa saja terdapat celah sehingga dapat dimanfaatkan *hacker* untuk memasukkan berbagai jenis konten malware ke dalam file PDF. Celah dalam file PDF dapat dimanfaatkan oleh *hacker* untuk memasukkan kode berbahaya seperti *JavaScript* dan URL yang tertanam pada file PDF yang dapat mengakibatkan pengguna memasang malware tanpa sepengetahuan atau izin pengguna [4]. Terdapat beberapa tujuan hacker dalam melakukan aktifitas berbahaya seperti pencurian informasi pribadi, penyadapan, melakukan manipulasi data, dan meminta akses tanpa sepengetahuan dan izin pengguna sehingga menyebabkan banyak kerugian yang dialami pengguna.

Logistic regression merupakan salah satu teknik machine learning yang dapat digunakan untuk menyelesaikan masalah klasifikasi yang bertujuan untuk

memprediksi label kategori dari data berdasarkan pada fitur-fitur yang ada. Dengan menggunakan Logistic Regression, kita dapat mengetahui bagaimana hubungan antara variabel respon dengan variabel prediktor, yang dimana pada variabel respon merupakan kategorik dan variabel prediktor merupakan kategorik maupun numerik [5]. Logistic Regression terbagi menjadi dua yaitu *Dichotomus* yang memiliki skala nominal dua kategori dan *Polychotomus* yang memiliki skala nominal lebih dari dua kategori [6]. Penelitian ini menggunakan model Logistic Regression yang bersifat *Polychotomus* atau Multinomial karena variabel respon yang diamati berskala nominal tiga kategori.

Menurut penelitian [7] mengenai Klasifikasi Web Berbahaya menggunakan Metode Logistic Regression, analisis metode Logistic Regression memiliki ketepatan klasifikasi sebesar 94% dan tingkat kesalahan yang rendah sebesar 6%. Dengan menggunakan 1000 data web yang berasal dari kanggle yang berisikan Malicious dan Benign website dan menggunakan 10 variabel sebagai variabel prediksi (Prediktor) untuk mengidentifikasi sebuah website berbahaya atau tidak tanpa melihat langsung isi dari konten website tersebut.

Menurut penelitian [8] mengenai Logistic Regression For Polymorphic Malware Detection Using ANOVA F-Test, analisis metode Logistic Regression memiliki ketepatan deteksi 97% dalam melakukan deteksi polymorphic malware dan penggunaan ANOVA F-Test untuk pemilihan fitur. Penelitian ini melakukan pengujian menggunakan Kali Linux sebagai sistem penyerang dan Windows XP sebagai sistem yang rentan menunjukkan efektivitas sistem dalam mendeteksi malware tersebut.

Menurut penelitian [9] mengenai Aplikasi Klasifikasi SMS Berbasis Web menggunakan Logistic Regression, analisis metode Logistic Regression memiliki ketepatan klasifikasi 97% dalam mengklasifikasikan pesan spam dan pesan yang bukan spam. Dengan menganalisis pengembangan aplikasi website dengan kumpulan data 1.140 pesan SMS. Aplikasi yang dikembangkan dalam penelitian menggunakan sebuah web sederhana dengan bantuan *Flask Framework* dari python.

Menurut penelitian [10] mengenai Mutual Information based Logistic Regression for phishing URL Detection, analisis metode memiliki ketepatan akurasi sebesar 99% dalam mendeteksi URL Palsu dan URL Normal. Hasil yang didapatkan dari penerapan metode Logistic Regression pada penelitian dapat meningkatkan deteksi ancaman URL palsu dan meningkatkan pertahanan terhadap upaya phishing untuk para professional dibidang cyber security.

Salah satu hal yang harus diperhatikan ketika mengevaluasi model adalah tingkat akurasi sebuah model dalam memprediksi variabel respon dengan benar. Kualitas model dipengaruhi oleh adanya keseimbangan antara kelas mayor dengan kelas minor. Apabila dataset yang digunakan tidak seimbang atau imbalance dapat membuat kinerja algoritma model menurun dan menyebabkan peningkatan kesalahan klasifikasi pada kelas minor [11]. Oleh karena itu, penelitian ini menggunakan teknik Oversampling yaitu *Random Oversampling* untuk mengatasi masalah imbalance dataset. Proses rebalance data menggunakan teknik *random oversampling* dapat meningkatkan jumlah sampel kelas minoritas melalui replikasi acak atau random dari sampel kelas yang sudah ada [12]. Dengan menyeimbangkan distribusi kelas mayoritas dengan kelas minoritas, model dapat lebih akurat dalam mengidentifikasi dan memprediksi data.

Berdasarkan pembahasan diatas, penulis akan melakukan Klasifikasi PDF Malware menggunakan dataset yang berasal dari *repository* GARUDA. Adapun judul dari penelitian tugas akhir ini adalah **"KLASIFIKASI PDF MALWARE PADA GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE LOGISTIC REGRESSION".** Diharapkan penelitian tugas akhir ini dapat menjadi referensi bagi penelitian terkait dan diharapkan dapat menghasilkan nilai *accuracy, precision*, *recall* dan *f1-score* yang baik.

1.2 Rumusan Masalah

Berikut ini rumusan masalah penelitian Tugas Akhir yang akan dilakukan:

1. Bagaimana cara dalam menangani masalah data *imbalance* agar mendapatkan performa terbaik?

2. Bagaimana penggunaan metode Logistic Regression dalam proses klasifikasi PDF malware GARUDA?
3. Bagaimana hasil validasi dari klasifikasi menggunakan metode Logistic Regression pada dataset PDF malware GARUDA?

1.3 Batasan Masalah

Berikut ini batasan masalah penelitian Tugas Akhir yang akan dilakukan:

1. Dataset yang digunakan berasal dari Garba Rujukan Digital (GARUDA) Kemdikbud Dikti.
2. Menggunakan Random Oversampling untuk menangani masalah dataset *imbalance* pada PDF malware GARUDA.
3. Melakukan klasifikasi PDF malware menggunakan algoritma Logistic Regression pada program Python.
4. Nilai performasi yang diukur adalah *accuracy, precision, recall dan f1-score.*

1.4 Tujuan Penelitian

Berikut ini merupakan tujuan dari penelitian Tugas Akhir yang dilakukan:

1. Menerapkan *Random oversampling* untuk menangani masalah *imbalance* pada dataset PDF malware GARUDA.
2. Menggunakan Logistic Regression Multinomial untuk klasifikasi dataset PDF malware GARUDA.
3. Melakukan analisa hasil kinerja dari proses klasifikasi yang dihasilkan menggunakan Logistic Regression untuk memperoleh model terbaik.

1.5 Manfaat Penelitian

Berikut ini merupakan manfaat dari penelitian Tugas Akhir yang dilakukan:

1. Menemukan solusi dari permasalahan dataset *imbalance* pada PDF malware GARUDA dengan menggunakan *Random oversampling*.
2. Dapat menerapkan Logistic Regression sebagai metode klasifikasi pada dataset PDF GARUDA.

3. Mampu menganalisa hasil validasi yang didapatkan menggunakan metode Logistic Regression sehingga diperoleh model terbaik.

1.6 Metodologi Penelitian

Berikut ini merupakan metodologi penelitian yang digunakan dalam penulisan Tugas Akhir ini:

1. Metode Studi Pustaka (Literature)

   Pada tahapan ini mencari dan mengumpulkan referensi berupa *literature* yang membahas mengenai PDF Malware, Logistic Regression, Random Oversampling dan lainnya yang diperlukan dalam proses penelitian.

2. Metode Pengumpulan Data

   Metode ini mengumpulkan raw data dari repository GARUDA yang dikelola oleh COMNETS Research Labs Universitas Sriwijaya untuk dilakukan *scanning* dan ekstraksi fitur pada file PDF menjadi dataset format .csv.

3. Metode Pengolahan Data

   Metode ini dilakukan dengan menganalisis file PDF menjadi dataset yang siap diproses. Penelitian ini menggunakan dataset *imbalance* sehingga diperlukan tahap *resampling* dengan menggunakan Teknik Oversampling yaitu Random Oversampling. Menerapkan model Logistic Regression untuk klasifikasi PDF *benign*.

4. Metode Analisa

   Metode analisa ini dilakukan dengan menganalisa hasil yang didapat dari pengolahan data Tugas Akhir yang kemudian di validasi untuk membuat kesimpulan.

5. Metode Kesimpulan dan Saran

   Metode ini dilakukan setelah menganalisa penelitian secara keseluruhan untuk membuat kesimpulan Tugas Akhir serta memberikan saran yang dapat dijadikan referensi bagi peneliti selanjutnya.

1.7 Sistematika Penelitian

Berikut ini merupakan sistematika penelitian yang digunakan dalam penulisan Tugas Akhir:

BAB I PENDAHULUAN

Bab ini terdapat Latar Belakang penelitian yang dilakukan, Rumusan Masalah, Batasan Masalah, Tujuan Penelitan, Manfaat Penelitian, Metodologi Penelitian dan Sistematika Penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini terdapat penelitian terkait, penjelasan mengenai file PDF, Fitur PDF, PDF Malware, Analisa yang dilakukan, Dataset yang digunakan, Metode Logistic Regression, Label Encoder, Imbalance dataset, dan Confussion matrix.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai Langkah-langkah (Metodologi) penelitian yang terdapat Spesifikasi perangkat yang digunakan, Kerangka kerja penelitian, Perancangan Sistem yang dilakukan, Persiapan dataset, flowchart pre-processing dan processing data, dan parameter pengujian yang digunakan.

BAB IV HASIL DAN ANALISA

Bab ini menjelaskan hasil dari proses pengolahan data yang sudah dilakukan, dan dari hasil tersebut akan dilakukan analisa supaya mendapatkan data yang akurat.

BAB V KESIMPULAN DAN SARAN

Bab ini menjelaskan kesimpulan yang yang didapatkan berdasarkan hasil dan analisa yang diperoleh setelah melakukan penelitian, kemudian memberikan saran untuk penelitian selanjutnya agar dapat dilakukan pengembangan.

DAFTAR PUSTAKA

[1] A. Castiglione, A. De Santis, and C. Soriente, "Security and privacy issues in the Portable Document Format," *J. Syst. Softw.*, vol. 83, no. 10, pp. 1813–1822, Oct. 2010, doi: 10.1016/j.jss.2010.04.062.

[2] N. Nissim *et al.*, "Sec-lib: Protecting scholarly digital libraries from infected papers using active machine learning framework," *IEEE Access*, vol. 7, pp. 110050–110073, 2019, doi: 10.1109/ACCESS.2019.2933197.

[3] "GARUDA - Gerba Rujukan Digital." https://garuda.kemdikbud.go.id/

[4] I. Corona, D. Maiorca, D. Ariu, and G. Giacinto, "Lux0R: Detection of malicious PDF-embedded javascript code through discriminant analysis of API references," *Proc. ACM Conf. Comput. Commun. Secur.*, vol. 2014-Novem, no. November, pp. 47–57, 2014, doi: 10.1145/2666652.2666657.

[5] B. Pavlyshenko, "Machine learning, linear and Bayesian models for logistic regression in failure detection problems," *Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 2046–2050, 2016, doi: 10.1109/BigData.2016.7840828.

[6] K. Pereira Teodoro da Silva, A. Kalbusch, and E. Henning, "Detection of unauthorized consumption in water supply systems: A case study using logistic regression," *Util. Policy*, vol. 84, no. August, p. 101647, 2023, doi: 10.1016/j.jup.2023.101647.

[7] A. Bimantara and T. A. Dina, "Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression," *Annu. Res. Semin.*, vol. 4, no. 1, pp. 173–177, 2019, [Online]. Available: https://seminar.ilkom.unsri.ac.id/index.php/ars/article/view/1932

[8] B. J. Kumar, H. Naveen, B. P. Kumar, S. S. Sharma, and J. Villegas, "Logistic regression for polymorphic malware detection using ANOVA F-Test," *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-January, pp. 1–5, 2017, doi: 10.1109/ICIIECS.2017.8275880.

[9] F. D. Pramakrisna, F. D. Adhinata, and N. A. F. Tanjung, "Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic

Regression," *Teknika*, vol. 11, no. 2, pp. 90–97, 2022, doi: 10.34148/teknika.v11i2.466.

[10] V. Vajrobol, B. B. Gupta, and A. Gaurav, "Mutual Information based Logistic Regression for phishing URL detection," *Cyber Secur. Appl.*, p. 100044, 2024, doi: 10.1016/j.csa.2024.100044.

[11] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 305–314, Jan. 2022, doi: 10.34123/icdsos.v2021i1.240.

[12] S. Diantika, "Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 19–25, 2023, doi: 10.36040/jati.v7i1.6006.

[13] C. Liu *et al.*, "A novel adversarial example detection method for malicious PDFs using multiple mutated classifiers," *Forensic Sci. Int. Digit. Investig.*, vol. 38, p. 301124, 2021, doi: 10.1016/j.fsidi.2021.301124.

[14] M. Issakhani, P. Victor, A. Tekeoglu, and A. Lashkari, "PDF Malware Detection based on Stacking Learning," no. Icissp, pp. 562–570, 2022, doi: 10.5220/0010908400003120.

[15] V. N and V. V, "Malicious-URL Detection using Logistic Regression Technique," *Int. J. Eng. Manag. Res.*, vol. 09, no. 06, pp. 108–113, 2019, doi: 10.31033/ijemr.9.6.18.

[16] M. Wadkar, F. Di Troia, and M. Stamp, "Detecting malware evolution using support vector machines," *Expert Syst. Appl.*, vol. 143, Apr. 2020, doi: 10.1016/j.eswa.2019.113022.

[17] R. Islam, M. I. Sayed, S. Saha, M. J. Hossain, and M. A. Masud, "Android malware classification using optimum feature selection and ensemble machine learning," *Internet Things Cyber-Physical Syst.*, vol. 3, no. March, pp. 100–111, 2023, doi: 10.1016/j.iotcps.2023.03.001.

[18] M. Ahmed, N. Afreen, M. Ahmed, M. Sameer, and J. Ahamed, "An inception V3 approach for malware classification using machine learning and transfer learning," *Int. J. Intell. Networks*, vol. 4, no. September 2022,

pp. 11–18, 2023, doi: 10.1016/j.ijin.2022.11.005.

[19] S. Yoo, S. Kim, S. Kim, and B. B. Kang, "AI-HydRa: Advanced hybrid approach using random forest and deep learning for malware classification," *Inf. Sci. (Ny).*, vol. 546, pp. 420–435, 2021, doi: 10.1016/j.ins.2020.08.082.

[20] Kiran Hassan Shivashankar, "Accurate Detection of Malicious Code in PDF Files using Machine Learning.," *Nas. Coll. Irel.*, 2020, [Online]. Available: https://norma.ncirl.ie/4494/1/kiranhassanshivashankar.pdf

[21] G. M. S. Hossain, K. Deb, H. Janicke, and I. H. Sarker, "PDF Malware Detection: Toward Machine Learning Modeling With Explainability Analysis," *IEEE Access*, vol. 12, no. December 2023, pp. 13833–13859, 2024, doi: 10.1109/ACCESS.2024.3357620.

[22] N. Fleury, T. Dubrunquez, and I. Alouani, "PDF-Malware: An Overview on Threats, Detection and Evasion Attacks," 2021, [Online]. Available: http://arxiv.org/abs/2107.12873

[23] "Virus Total." https://www.virustotal.com

[24] B. Deepa and K. Ramesh, "Epileptic seizure detection using deep learning through min max scaler normalization," *Int. J. Health Sci. (Qassim).*, vol. 6, no. April, pp. 10981–10996, 2022, doi: 10.53730/ijhs.v6ns1.7801.

[25] P. Subekti, "Model regresi logistik multinomial untuk menentukan pilihan sekolah lanjutan tingkat atas pada siswa SMP," *CAUCHY J. Mat. Murni dan Apl.*, vol. 3, no. 2, pp. 91–98, 2014, doi: 10.18860/ca.v3i2.2577.

[26] Z. Z. Zahroh and I. Zain, "Analisis Regresi Logistik Multinomial Pada Faktor-Faktor Yang Mempengaruhi Sumber Air Bersih Rumah Tangga Di Jawa Timur," *J. Sains dan Seni ITS*, vol. 7, no. 2, 2019, doi: 10.12962/j23373520.v7i2.34701.

[27] M. M. Rahman and D. N. Davis, "Cluster based under-sampling for unbalanced cardiovascular data," *Lect. Notes Eng. Comput. Sci.*, vol. 3 LNECS, no. July 2013, pp. 1480–1485, 2013.

[28] Y. Xiong, "Building text hierarchical structure by using confusion matrix," *2012 5th Int. Conf. Biomed. Eng. Informatics, BMEI 2012*, no. Bmei, pp. 1250–1254, 2012, doi: 10.1109/BMEI.2012.6513202.

[29] J. Dsouza and S. Senthil Velan, "Using Exploratory Data Analysis for

Generating Inferences on the Correlation of COVID-19 cases," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, pp. 2–7, 2020, doi: 10.1109/ICCCNT49239.2020.9225621.

[30] M. Ali *et al.*, "Analysis of Feature Selection Methods in Software Defect Prediction Models," *IEEE Access*, vol. 11, no. December, pp. 145954–145974, 2023, doi: 10.1109/ACCESS.2023.3343249.