KLASIFIKASI PDF MALWARE PADA GARUDA KEMDIKBUD SEBAGAI AGREGATOR NASIONAL DENGAN METODE K-NEAREST NEIGHBOR

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH:

TRI SHENA ORIVIA PASIN
09011181823029

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA

2022

LEMBAR PENGESAHAN

Klasifikasi *PDF Malware* pada GARUDA Kemdikbud sebagai Agregator Nasional dengan Metode *K-Nearest Neighbor*

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh

Tri Shena Orivia Pasin
09011181823029

Indralaya, 22 Desember 2022

Mengetahui,

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

Ketua Jurusan Sistem Komputer

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

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada:

Hari : Selasa

Tanggal : 01 November 2022

Tim Penguji :

1. Ketua : Ahmad Heryanto, M.T.

2. Sekretaris : Adi Hermansyah, M.T.

3. Penguji : Huda Ubaya, M.T.

4. Pendamping I : Deris Stiawan, M.T., Ph.D.

5. Pendamping II : Tri Wanda Septian, M.Sc.

Mengetahui,

Ketua Jurusan Sistem Komputer

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

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : **Tri Shena Orivia Pasin**
NIM : **09011181823029**
Judul : **Klasifikasi *PDF Malware* pada GARUDA Kemdikbud sebagai Agregator Nasional dengan Metode *K-Nearest Neighbor***

Hasil Pengecekan *Software iThenticate* / *Turnitin* : 4%

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

Demikian, pernyataan ini saya buat dalam keadaan sadar dan tidak dipaksakan.

Indralaya, November 2022

Tri Shena Orivia Pasin
NIM. 09011181823029

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh.

Segala puji dan syukur atas kehadirat Allah SWT, atas segala karunia dan rahmat-Nya sehingga penulis dapat menyelesaikan penyusunan laporan Tugas Akhir yang berjudul **Klasifikasi *PDF Malware* pada GARUDA Kemdikbud sebagai Agregator Nasional dengan Metode *K-Nearest Neighbor*** yang digunakan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer (Strata 1) pada Jurusan Sistem Komputer di Universitas Sriwijaya.

Dalam penyusunan laporan Tugas Akhir ini, penulis banyak mendapatkan bimbingan, bantuan, ide, serta saran baik moril maupun materil dari berbagai pihak secara langsung maupun tidak langsung. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya, yang terhormat:

1. Allah Subbhanahu Wata'ala yang telah memberikan berkah serta nikmat kesehatan dan kesempatan kepada penulis dalam pelaksanaan pembuatan Tugas Akhir ini.
2. Kedua orang tua, ayah Baharuddin Pasin dan mama Aslamiyah, S.E. serta abangku Ir. Harasa Ramdhany Pasin, S.T., M.Eng. yang tak hentinya selalu memberikan do'a dan motivasi serta dukungan selama ini.
3. Bapak Jaidan Jauhari, S. Pd, 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 Dr. Erwin, M.Si. selaku Dosen Pembimbing Akademik penulis di Jurusan Sistem Komputer.

6. Bapak Deris Stiawan, M.T., Ph.D. selaku Pembimbing I Tugas Akhir penulis di Jurusan Sistem Komputer yang telah meluangkan waktu untuk membimbing penulis dalam menyelesaikan TA, serta motivasi dan nasihat yang diberikan selama perkuliahan.
7. Kak Tri Wanda Septian, M.Sc. selaku Pembimbing II Tugas Akhir penulis di Jurusan Sistem Komputer yang telah meluangkan waktu untuk membimbing dari proses awal mengolah *dataset* hingga selesainya laporan TA ini, serta motivasi dan nasihat yang diberikan selama perkuliahan.
8. Mbak Nurul Afifah, M.Kom. yang membimbing penulis dalam pembuatan Tugas Akhir di Jurusan Sistem Komputer yang telah meluangkan waktu dari proses awal mengolah *dataset* hingga membenarkan penulisan yang dibuat dalam laporan TA sehingga dapat lebih baik dari sebelumnya.
9. Mbak Renny Virgasari selaku Admin Jurusan SK yang baik hati yang banyak membantu penulis dalam melakukan pemberkasan.
10. Teman seperjuangan TA yaitu Alifah Fidela, Indah Cahya Resti, Nata Arista, Novi Yuningsih, dan Rani Octaviani yang membantu dan memberikan semangat dalam menyelesaikan laporan TA.
11. Teman – teman Lab Elektronika Dasar dan Sistem Digital serta Lab Robotika dan Sistem Kendali yaitu Arif Tumpal Leonardo Sianturi dan Muhammad Furqon Rabbani, Ades Harafi Duri, Alif Almuqsit, Muhammad Imam Rafi, Dimas Aditya Kristianto, Muhammad Farhan Alharits, M. Taufik, Hana Nur Shofwa, dan M. Tedi Bustami.
12. Kakak tingkat dan teman – teman dari Grup Riset COMNETS yaitu mbak Febi Rusmiati, kak Deri Andany, Budiman Alfian, Chendy Maulana, Rifqi Abiyyu, Christoper Marlo, Arief Saifullah, Ageng Raharjo, Rizki Ridho, dll…

13. Nur Riski Cahyati selaku teman penulis di SK yang tidak sengaja bertemu saat awal verifikasi pemberkasan. Semoga sukses selalu serta apapun impian dan tujuanmu selalu dipermudah dan dilancarkan. Aamiin aamiin allahumma aamiin.
14. Semua teman seperjuangan di Jurusan Sistem Komputer angkatan 2018 semoga jalan dan hambatan yang dilewati selama perkuliahan ini menjadi berkah dan kenangan kita kelak.
15. Karin Zikra Nisya sobat terbaik yang bisa diandalkan dalam susah maupun senang. Semoga apapun impian dan tujuanmu selalu dipermudah dan dilancarkan. Aamiin aamiin allahumma aamiin.
16. Tak lupa untuk semua orang baik yang datang di hidup penulis, tanpa disadari sangat banyak membantu di berbagai hal, yang tidak dapat disebutkan satu per satu.
17. Almamater Universitas Sriwijaya.

Penulis menyadari dalam penyusunan laporan Tugas Akhir ini masih terdapat banyak kekurangan, oleh karena itu penulis sangat menerima kritik dan saran terhadap isi dari laporan Skrispsi ini yang bersifat membangun. Semoga dengan laporan Tugas Akhir ini akan menjadi tambahan ilmu pengetahuan dan pengembangan wawasan kita serta bermanfaat bagi siapapun yang membacanya. Sekian dan terima kasih sebesar-besarnya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, 23 September 2022

Penulis

Tri Shena Orivia Pasin
NIM. 09011181823029

PDF MALWARE CLASSIFICATION ON GARUDA KEMDIKBUD AS NATIONAL AGGREGATOR USING K-NEAREST NEIGHBOR METHOD

TRI SHENA ORIVIA PASIN (09011181823029)
Computer Engineering Department, Computer Science Faculty, Sriwijaya University
Email : trishenaoriviapasin19@gmail.com

ABSTRACT

Garba Rujukan Digital (GARUDA) is a digital service used for collecting archives of national publication documents in PDF format. There are sections in the PDF that can be used by hackers to carry out attacks that the file becomes PDF Malware. The dataset obtained is unprocessed data, which was analyzed using PDFiD to obtain features. The feature extraction results were used to classify multiclass labels, namely PDF-Malware, PDF-HTML, and PDF-Benign using K-Nearest Neighbor. The best K-Nearest Neighbor results are K=1 obtained a precision value of 98,2%, a recall of 98,4%, an f1-score of 98.3%, an accuracy of 98.3%, and K=3 obtained a precision value of 98%, recall of 98.4%, f1-score of 98.2%, accuracy of 98.3%.

Keywords : *Classification*, *PDF Malware*, PDFiD, *Multiclass*, *K-Nearest Neighbor*.

KLASIFIKASI PDF MALWARE PADA GARUDA KEMDIKBUD SEBAGAI AGREGATOR NASIONAL DENGAN METODE K-NEAREST NEIGHBOR

TRI SHENA ORIVIA PASIN (09011181823029)
Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
Email : trishenaoriviapasin19@gmail.com

ABSTRAK

Garba Rujukan Digital (GARUDA) merupakan layanan pengumpulan arsip dokumen publikasi nasional dalam bentuk *PDF*. Terdapat bagian dalam *PDF* yang dapat dimanfaatkan *hacker* untuk menjalankan serangan sehingga *file* tersebut menjadi *PDF Malware*. *Dataset* yang didapatkan berupa data mentah yang setiap *PDF* dianalisis menggunakan PDFiD untuk didapatkan fitur. Hasil ekstraksi fitur tersebut digunakan untuk klasifikasi label *multiclass* yaitu PDF-Malware, PDF-HTML, dan PDF-Benign menggunakan *K-Nearest Neighbor*. Didapatkan hasil klasifikasi *K-Nearest Neighbor* yang terbaik pada K=1 memperoleh nilai presisi sebesar 98,2%, *recall* sebesar 98,4%, f1-skor sebesar 98,3%, akurasi sebesar 98,3%, dan K=3 memperoleh nilai presisi sebesar 98%, *recall* sebesar 98,4%, f1-skor sebesar 98,2%, akurasi sebesar 98,3%.

Kata Kunci : Klasifikasi, *PDF Malware*, PDFiD, *Multiclass*, *K-Nearest Neighbor*.

DAFTAR ISI

Halaman

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penulisan penelitian ini menggunakan *dataset* dari layanan agregator Garba Rujukan Digital (GARUDA) [1] yaitu layanan untuk mengumpulkan arsip dokumen publikasi nasional dengan tujuan berbagi keilmuan dari publikasi ilmiah dalam bentuk jurnal penelitian. Dokumen publikasi ini biasanya dikumpulkan dalam format *file PDF*, yang banyak digunakan untuk mempermudah pengguna membaca dan melakukan pertukaran dokumen secara digital, *file PDF* mengandung beberapa bagian yang beresiko menyebabkan kerusakan sehingga adanya kemungkinan setiap *file PDF* dapat dimanfaatkan *hacker* untuk menyematkan *malware* dengan tujuan menjalankan serangan untuk mengeksploitasi pengguna. *Dataset* GARUDA berbentuk data mentah dalam *rar* yang berisikan *file PDF* yang akan diidentifikasi sehingga didapatkan kategori *multiclass* yang berupa *benign* yaitu data normal, *mal-html* yaitu *malware* berbentuk *html*, dan *mal-pdf* yaitu *malware* berbentuk *PDF*.

Penelitian ini mengacu dari beberapa penelitian sebelumnya [2] yang disarankan menggunakan metode lain untuk melihat seberapa akurat pembelajaran mesin dalam mendapatkan akurasi pada pengklasifikasian *malware* dalam *file PDF*. *K-Nearest Neighbor* dikenali sebagai algoritma pengklasifikasian dengan tingkat keakuratan yang baik dengan peningkatan keakuratan dipengaruhi dari nilai K [3] yang digunakan dan SMOTE yang menangani *data imbalanced* serta *Stratified K-Fold* sebagai pemisahan data.

Oleh karena itu, penelitian tugas akhir ini yaitu mengenai Klasifikasi *PDF Malware* pada GARUDA Kemdikbud sebagai Agregator Nasional dengan Metode *K-Nearest Neighbor*. Diharapkan dari penelitian ini dapat menghasilkan *accuracy*, *precision*, *recall*, dan *f1-Score* yang baik sehingga dapat menambah menjadi referensi ilmu pengetahuan terkait.

1.2 Rumusan Masalah

Dari topik yang diangkat di latar belakang maka fokus penelitian yang akan dibahas dari penelitian ini sebagai berikut:

1. Bagaimana pengolahan data mentah pada kumpulan *PDF* dari Layanan Agregator Nasional GARUDA dengan analisis *malware* secara statis sehingga dapat digunakan sebagai *dataset*.
2. Bagaimana cara yang dilakukan untuk menerapkan pengklasifikasian *PDF Malware* dalam label *multiclass* atau atribut yang memiliki banyak nilai yaitu *benign*, *mal-html*, dan *mal-pdf*.
3. Berapa baiknya hasil performa klasifikasi yang diterapakan pada *K-Nearest Neighbor* dalam *dataset imbalanced*.

1.3 Batasan Masalah

Untuk menghindari pembahasan yang terlalu jauh maka ditetapkan batasan agar sesuai dengan permasalahan yang diangkat yaitu:

1. Algoritma digunakan sebatas melakukan klasifikasi dengan *K-Nearest Neighbor* untuk dilihat performanya.
2. Nilai yang diukur yaitu *accuracy*, *precision*, *recall*, dan *f1-Score*.
3. Tidak membahas cara *malware* masuk kedalam *PDF* dan pencegahan.

1.4 Tujuan

Adapula hal yang diharapkan agar tercapainya penjabaran rumusan dari penelitian ini antara lain:

1. Melakukan pengolahan data mentah dari kumpulan *PDF* dalam Layanan Agregator Nasional GARUDA dengan menganalisis *malware* secara statis sehingga dapat digunakan menjadi *dataset*.
2. Melakukan klasifikasi *PDF malware* dalam kategori *benign*, *mal-html*, dan *mal-pdf*.
3. Mengetahui baiknya hasil performa klasifikasi yang diterapakan pada *K-Nearest Neighbor* dalam *dataset imbalanced*.

1.5 Manfaat

Adapula harapan yang didapatkan setelah melakukan dan menguji penelitian ini yaitu:

1. Mengetahui penggunaan analisis *malware* secara statis untuk mengidentifikasikan *file PDF* mengandung *malware* atau tidak.
2. Mengetahui penerapan algoritma *K-Nearest Neighbor* dan memberikan pemahaman terkait pengklasifikasian data *PDF* yang mengandung *benign, mal-html,* dan *mal-pdf*.
3. Dapat dijadikan rujukan untuk penelitian yang selaras membahas tentang *PDF Malware*.

1.6 Metodologi Penelitian

Proses yang dilakukan dalam pengumpulan data pada penelitian tugas akhir ini, yaitu:

1. Studi Pustaka (*Literature*)

   Kegiatan dilakukan dengan mencari literatur untuk menentukan kata kunci yang diangkat pada judul dengan tujuan untuk memperbanyak pengetahuan mengenai penelitian yang dilakukan.

2. Tukar Pikiran dan Perancangan Sistem

   Kegiatan dilakukan dengan diskusi bersama pembimbing untuk menemukan tahapan rancangan sistem yang sesuai untuk dilakukan. Sehingga dapat mempermudah untuk melakukan ke pengolahan data.

3. Pengumpulan Data

   Kegiatan dilakukan dengan memperoleh data dari Layanan Agregator Nasional GARUDA berbentuk data mentah yang dikelompokkan di dalam *rar* yang berisi 20000 *file* PDF dan dibagi menjadi 10000 *file PDF* yang di untuk mendapatkan informasi mengenai *PDF* mengandung *malicious* dan *non-malicious*. Selanjutnya

menghitung kemunculan setiap fitur pada *file PDF* sebagai *dataframe* kemudian dijadikan sebagai *dataset* untuk pengolahan data.

4. Pengolahan Data

   Kegiatan dilakukan dengan melakukan pemberian label, *Oversampling Data* dengan SMOTE, *Undersampling* dengan *Near Miss*, serta menerapkan pembagian data dengan *Stratified K-folds Cross Validation*. Setelah itu menerapkan pengklasifikasian algoritma *K-Nearest Neighbor* untuk dilihat baiknya hasil performa.

5. Hasil dan Analisa

   Kegiatan dilakukan dengan pengambilan hasil data yang diperoleh dan selanjutnya menganalisis perolehan hasil klasifikasi.

6. Kesimpulan dan Saran

   Kegiatan dilakukan dengan mengambil kesimpulan serta saran agar selanjutnya dapat dikembangkan oleh peneliti selanjutnya.

1.7 Sistematika Penulisan

Adapula rancangan penulisan yang digunakan sebagai arahan isi yang dibahas tiap bab – bab dari penelitian agar terstruktur yaitu:

BAB I PENDAHULUAN

Dalam isi bab satu merupakan awal gambaran seperti apa penelitian ini dilakukan yang isinya berupa Latar belakang, Rumusan, Batasan, Tujuan, Manfaat, Metodologi Penelitian, dan Sistematika Penulisan.

BAB II TINJAUAN PUSTAKA

Dalam isi bab dua mengenai bacaan dari buku maupun jurnal penelitian terkait *PDF Malware*, analisis *malware* yang dilakukan, pengetahuan tentang *Oversampling Data* dan

Stratified K-folds Cross Validation, serta pemahaman dari algoritma yang digunakan.

BAB III METODOLOGI

Dalam isi bab tiga mengenai penjelasan secara sistematis proses penelitan yang dilakukan meliputi tahapan perancangan sistem hingga penerapan metode yang digunakan dalam penelitian.

BAB IV HASIL DAN ANALISA

Dalam isi bab empat mengenai penjelasan proses dari pengambilan hasil dan analisis data perolehan hasil klasifikasi pada pengujian.

BAB V KESIMPULAN DAN SARAN

Dalam isi bab lima berupa penjelasan mengenai kesimpulan dan saran untuk acuan dikembangkannya penelitian ini oleh peneliti selanjutnya.

DAFTAR PUSTAKA

LAMPIRAN

DAFTAR PUSTAKA

[1] "Garuda - Garba Rujukan Digital." https://garuda.kemdikbud.go.id/ (accessed Mar. 05, 2022).

[2] A. Charim, S. Basuki, and D. R. Akbi, "Detect Malware in Portable Document Format Files (PDF) Using Support Vector Machine and Random Decision Forest," *J. Online Inform.*, vol. 3, no. 2, p. 99, 2019, doi: 10.15575/join.v3i2.196.

[3] E. M. F. El Houby, N. I. R. Yassin, and S. Omran, "A hybrid approach from ant colony optimization and K-nearest neighbor for classifying datasets using selected features," *Inform.*, vol. 41, no. 4, pp. 495–506, 2017.

[4] V. Atluri, "Malware Classification of Portable Executables using Tree-Based Ensemble Machine Learning," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2019-April, 2019, doi: 10.1109/SoutheastCon42311.2019.9020524.

[5] M. Chowdhury, A. Rahman, and R. Islam, "Malware analysis and detection using data mining and machine learning classification," *Adv. Intell. Syst. Comput.*, vol. 580, pp. 266–274, 2018, doi: 10.1007/978-3-319-67071-3_33.

[6] 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.

[7] 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://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Detection+of+Malicious+PDF+Files+Based+on+Hierarchical+Document+Structure#0.

[8] S. S. Pachpute, "Malware Analysis on PDF," 2019.

[9] "VirusTotal - Home." https://www.virustotal.com/gui/home/upload (accessed Mar. 05, 2022).

[10] "PDF Tools | Didier Stevens." https:// blog. didierstevens. com/ programs/ pdf-tools/ (accessed Mar. 05, 2022).

[11] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[12] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, “Classification of Imbalanced Data:Review of Methods and Applications,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012077, 2021, doi: 10.1088/1757-899x/1099/1/012077.

[13] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism,” *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, 2018, doi: 10.1109/JIOT.2018.2816007.

[14] L. Bao, C. Juan, J. Li, and Y. Zhang, “Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets,” *Neurocomputing*, vol. 172, pp. 198–206, 2016, doi: 10.1016/j.neucom.2014.05.096.

[15] M. Peng *et al.*, “Trainable undersampling for class-imbalance learning,” *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 4707–4714, 2019, doi: 10.1609/aaai.v33i01.33014707.

[16] R. Siringoringo, “Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor,” *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[17] H. Leidiyana, “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[18] N. S. Ramadhanti, W. A. Kusuma, and A. Annisa, “Optimasi Data Tidak Seimbang pada Interaksi Drug Target dengan Sampling dan Ensemble Support Vector Machine,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, p. 1221, 2020, doi: 10.25126/jtiik.2020762857.

[19] C. Jian, J. Gao, and Y. Ao, “A new sampling method for classifying imbalanced data based on support vector machine ensemble,” *Neurocomputing*, vol. 193, pp. 115–122, 2016, doi: 10.1016/j.neucom.2016.02.006.

[20] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016, doi: 10.1109/IACC.2016.25.