Klasifikasi *PDF Malware* Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*

TUGAS AKHIR

OLEH :

ALIFAH FIDELA

09011281823039

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2023

Klasifikasi *PDF Malware* Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH :

ALIFAH FIDELA
09011281823039

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2023

HALAMAN PENGESAHAN

Klasifikasi *PDF Malware* Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh

ALIFAH FIDELA
09011281823039

Indralaya, 13 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 : Alifah Fidela

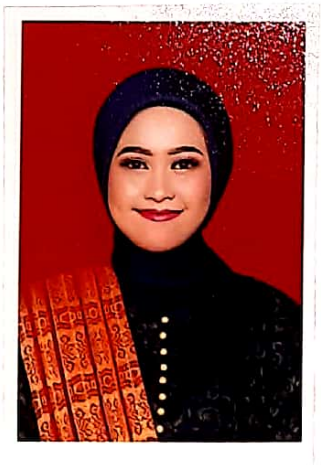
NIM : 09011281823039

Judul : Klasifikasi PDF *Malware* pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*

Hasil Pengecekan Software *iThenticate/Turnitin* : 15%

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



Indralaya, Januari 2023

Alifah Fidela

NIM.09011281823039

HALAMAN PERSEMBAHAN

“Tugas Akhir ini kupersembahkan untuk Papa, Mama, Kak Nia, Uni Dina, Adik Naya, Uwo dan keluarga besar yang selalu memberikan doa, dukungan dan semangat selama ini hingga dapat menyelesaikan masa studi serta teman-teman yang senantiasa selalu mendukungku”

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya”

(QS. Al-Baqarah 2 : 286)

“Maka sesungguhnya bersama kesulitan ada kemudahan,
sesungguhnya bersama kesulitan ada kemudahan”

(QS. Al-Insyirah : 5-6)

“It does not matter how slowly you go, as long as you do not stop”

(*confucius*)

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Puji syukur atas kehadirat Allah SWT, atas segala karunia dan rahmat-Nya sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul **"Klasifikasi PDF *Malware* Pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*"**.

Pada penyusunan tugas akhir ini, tidak terlepas dari bantuan, bimbingan, ajaran serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan rasa syukur dan terima kasih kepada :

1. Allah Subhanahu Wa ta'ala yang telah memberikan berkah dan hidayah-Nya serta nikmat yang tak terhitung.
2. Papa dan Mama serta adik-adik yang telah memberikan doa dan dukungannya serta memberikan motivasi.
3. Uwo, Om dan Tante serta Adik - Adik Sepupu yang telah memberikan dukungan dan semangat selama perkuliahan.
4. Bapak Jaidan Jauhari, M.T., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Dr. Ir. H. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.
6. Bapak Deris Stiawan, M.T., Ph.D. selaku Dosen Pembimbing I Tugas Akhir dan Pembimbing Akademik di Universitas Sriwijaya yang telah meluangkan waktunya untuk membimbing penulis serta memberikan saran dalam menyelesaikan Tugas Akhir ini.
7. Kak Tri Wanda Septian, S.Kom., M.Sc selaku Dosen Pembimbing II Tugas Akhir yang telah meluangkan waktunya untuk membimbing penulis serta memberikan saran dalam menyelesaikan Tugas Akhir ini.
8. Mbak Nurul Afifah M.Kom yang telah memberikan masukan dan saran dalam penulisan Tugas Akhir.

9. Mbak Renny Virgasari selaku admin Jurusan Sistem Komputer yang telah membantu mengurus seluruh berkas.
10. Teman-teman seperjuangan riset *Malware* Geeks Infosec Unsri Rani, Indah, Novi, Shena, dan Nata yang telah membersamai dan membantu dalam mengerjakan tugas akhir ini.
11. Rizki Valen Mafaza dan Indah Cahya Resti yang telah berbagai canda tawa dan membersamai selama ini.
12. Teman-teman lab elsidi dan robot, Furqon, Arif, Farhan, Imam, Dimas, Taufik, Realdi, Tedy, Hana, Ades, dan Alif.
13. Teman-teman riset Comnets dan Kakak-kakak tingkat yang telah membantu dalam menyelesaikan tugas akhir.
14. Teman-teman seperjuangan angkatan 2018
15. Serta semua pihak yang telah membantu yang tidak dapat disebutkan satu persatu dalam penyelesai tugas akhir ini.
16. Almamater.

Penulis menyadari bahwa masih banyak kekurangan dalam penulisan Tugas Akhir. Oleh karena itu kritik dan saran yang membangun sangat penulis harapkan. Akhir kata, semoga tugas akhir ini dapat bermanfaat dan berguna bagi khalayak.

Wassalamu'alaikum Wr. Wb.

Indralaya, Januari 2023

Penulis,

Alifah Fidela
NIM. 09011281823039

KLASIFIKASI *PDF MALWARE* PADA GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI DENGAN METODE *RANDOM FOREST*

ALIFAH FIDELA (09011281823039)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : alifahfidela16@gmail.com

ABSTRAK

Portable Document Format (PDF) merupakan salah satu format pembaca dokumen yang paling sering digunakan, Struktur objek pada PDF yang fleksibel dan mudah dalam penggunaanya. Sehingga, peretas menggunakan PDF untuk melakukan serangan. Dataset yang digunakan berasal dari dari Garba Rujukan Digital (GARUDA) yang terdiri dari kumpulan file PDF. File PDF akan diekstraksi menggunakan *tools* pdfid untuk mendapatkan fitur yang akan digunakan pada proses klasifikasi multiclass. Dataset penelitian ini memiliki kondisi data yang tidak seimbang. Mengatasi data yang tidak seimbang dengan melakukan proses *resampling* menggunakan *oversampling* dengan SMOTE dan *undersampling* dengan *NearMiss*. Hasil klasifikasi dengan metode *Random Forest* menghasilkan tingkat akurasi sebesar 99,94%, presisi sebesar 99,95%, recall sebesar 99,94%, F1-Score sebesar 99,94% dan OOB-Error sebesar 0,06%. Kemudian dilakukan validasi untuk keakuratan model menggunakan *Stratified Kfold Cross Validation* dan hasil rata-rata akurasi tertinggi diperoleh dengan menggunakan 7-fold sebesar 99,74%.

Kata Kunci *:* PDF Malware, *Random Forest, Stratified Kfold Cross Validation, Synthetic Minority Over-sampling Technique* (*SMOTE*) , *NearMiss*, pdfid.

MALWARE PDF CLASSIFICATION ON GARBA RUJUKAN DIGITAL (GARUDA) KEMDIKBUD DIKTI USING RANDOM FOREST METHOD

ALIFAH FIDELA (09011281823039)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email : alifahfidela16@gmail.com

ABSTRACT

The Portable Document Format (PDF) is one of the most commonly used document reader formats, the object structure in PDF is flexible and easy to use. Therefore, that hackers use PDFs to carry out the attacks. The dataset comes from the Garba Rujukan Digital (GARUDA), which consists of a collection of PDF files. PDF files will extract using the pdfid tools to get features used in the multiclass classification process. This research dataset has imbalanced data conditions. Overcoming imbalanced data by resampling using oversampling with SMOTE and undersampling with NearMiss. The classification results using the Random Forest method produce an accuracy rate of 99.94%, a precision of 99,95%, a recall of 99,94%, an F1-Score of 99.94%, and an OOB-Error of 0.06%. Then validation was carried out for the accuracy rate of the model using Stratified K-fold Cross Validation, and the highest average accuracy obtained using 7-fold was 99.74%.

Keywords *: Malware* PDF, *Random Forest, Stratified Kfold Cross Validation, Synthetic Minority Over-sampling Technique* (*SMOTE*) , *NearMiss, pdfid*.

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang

Portable Document Format atau umumnya disebut sebagai PDF merupakan format file yang banyak digunakan untuk berbagi dokumen, membaca serta memvisualisasikan. Struktur objek pada PDF yang fleksibel dan mudah dalam penggunaanya sehingga menjadikan tujuan peretas dalam mengeksploitasi dan melakukan serangan melewati file PDF [1][2]. File PDF yang telah terinfeksi *malware* akan mengganggu kinerja suatu sistem jika file tersebut dijalankan. Untuk menganalisis pendeteksi *malware* dengan menggunakan *machine learning*.

Garba Rujukan Digital (GARUDA) adalah sebuah portal yang digunakan sebagai sumber informasi publikasi karya ilmiah di Indonesia mencakup beberapa aspek ilmu pengetahuan seperti seni, ilmu perilaku, ilmu sosial, fisika, teknik, matematika dan komputer. GARUDA bertujuan untuk mempermudah dalam mengakses karya ilmiah serta mengelola informasi berupa ilmu pengetahuan.

Random Forest merupakan *ensamble classifier* berdasarkan pada supervised learning [3]. Konsep dari *Random Forest* ini akan menghasilkan hutan dengan sejumlah pohon keputusan. *Random Forest* dalam proses pengklasifikasi akan dibagi menjadi kelas prediksi dan kelas dengan suara yang terbanyak (*majority vote*) akan menjadi prediksi modelnya [4]. Metode *Random Forest* akan menghasilkan tingkat akurasi yang lebih baik dalam klasifikasi, dapat mengatasi data *training* dalam jumlah yang besar secara efisien dan menghasilkan error yang lebih rendah.

Penelitian yang dilakukan ini memiliki dataset yang tidak seimbang (*imbalanced*) sehingga akan mempengaruhi hasil kinerjanya. Untuk mengatasi dataset yang tidak seimbang dilakukan teknik resampling dengan *oversampling* menggunakan SMOTE dan *undersampling* menggunakan *NearMiss*. Dilakukan dengan cara menambah kelas minoritas berdasarkan *k-nearest neighbor* menjadi kelas yang hampir sama.

Pada penelitian [5] untuk mengidentifikasi *malware* pada file windows PE, data yang digunakan yaitu 489 *malware* dan 700 *benign* serta 54 fitur. Dilakukan klasifikasi dengan enam *ensamble machine learning* untuk mengetahui hasil akurasinya. Hasil akurasi yaitu Bagging Decision Tree Classifier (BDT) 93,27%, *Random Forest* Classifier (RFC) 85,29%, Extra Trees Classifier (ETC) 92,01%, AdaBoost Classifier (ABC) 92,85%, Gradient Boosting Classifier (GBC) 92,85%, dan Voting Ensamble Classifier (VEC) 93,69%.

Pada penelitian [6] dilakukan deteksi *malware* dengan jumlah data 41.265 file *malware* dan 10.920 file *benign*. Pendekatan yang digunakan dengan N-gram dan API Call features dan juga beberapa metode seperti Naïve Bayes, *Random Forest*, SVM. Hasil akurasi yang didapatkan yaitu Naïve Bayes sebesar 88.5%, *Random Forest* sebesar 94,8%, SVM sebesar 95.7%, dan pendekatan N-gram dan API Call yaitu 98.6%.

Selanjutnya, pada penelitian [7] dilakukan klasifikasi dengan metode *Random Forest* dan Supoort Vector Machine pada file *malware* dan *benign*. Hasil akurasi yang didapatkan yaitu pada SVM sebesar 58% dan *Random Forest* sebesar 84%, penelitian lainnya [8], [9] menunjukkan bahwa *Random Forest* menghasilkan tingkat akurasi yang lebih baik daripada metode yang digunakan lainnya.

Penulis akan membahas mengenai pengklasifikasian terhadap data file PDF *Malware* ke dalam tiga kelas yaitu *benign, mal-pdf,* dan *non-pdfmal*. Pengerjaan tugas akhir ini akan diterapkan menggunakan algoritma *Random Forest* yang akan melakukan klasifikasi terhadapat data pdf *malware* berdasarkan fitur-fitur pada dataset yang telah di ektraksi sebelumnya. Proses klasifikasi yang dilakukan akan bermanfaat pada keamanan siber untuk menganalisa file PDF yang terindikasi *malware* atau tidak, sehingga dapat membantu mengatasi ancaman yang datang dari luar melalui file PDF. Berdasarkan latar belakang tersebut judul tugas akhir ini adalah "Klasifikasi *PDF Malware* pada Garba Rujukan Digital (GARUDA) Kemdikbud Dikti dengan Metode *Random Forest*".

1.2 Perumusan dan Batasan Masalah

Adapun perumusan masalah dan batasan masalah yang akan dilakukan dalam penelitian Tugas Akhir ini adalah sebagai berikut.

1.2.1 Perumusan Masalah

Rumusan masalah dari penulisan Tugas Akhir adalah sebagai berikut:

1. Bagaimana proses ekstraksi data yang akan digunakan dalam proses klasifikasi PDF *Malware*?
2. Bagaimana mengklasifikasi PDF *Malware* menggunakan algoritma *Random Forest*.
3. Bagaimana cara mengevaluasi hasil kinerja dari proses klasifikasi PDF *Malware* menggunakan *Random Forest*.

1.2.2 Batasan Masalah

Batasan masalah pada penulisan Tugas Akhir ini adalah sebagai berikut:

1. Dataset yang digunakan pada penelitian ini merupakan dataset yang berasal dari GARUDA *Repository*.
2. Klasifikasi PDF *Malware* yang dilakukan secara *multiclass* (*benign*, *mal-pdf*, dan *non-pdfmal*).
3. Metode yang digunakan untuk mengklasifikasikan PDF *Malware* dengan menggunakan *Random Forest*.

1.3 Tujuan

Adapun tujuan dari penulisan Tugas Akhir ini adalah sebagai berikut:

1. Ekstraksi dataset file PDF *Malware* GARUDA *Repository* dilakukan secara statis dengan menggunakan *tools* pdfid.
2. Melakukan klasifikasi *PDF Malware* menggunakan algoritma *Random Forest*.
3. Mengevaluasi hasil kinerja dari proses klasifikasi PDF *Malware* dengan menggunakan *confusion matrix*.

1.4 Manfaat

Adapun manfaat dari penulisan Tugas Akhir ini adalah sebagai berikut:

1. Data hasil ekstraksi akan digunakan sebagai dataset dalam proses klasifikasi.
2. Dapat melakukan klasifikasi PDF *Malware* dan mengetahui tingkat akurasi dari dataset Garuda dengan *Random Forest*

3. Dapat mengevaluasi hasil kinerja dari proses klasifikasi PDF *Malware* menggunakan *Confusion matrix*.

1.5 Metodelogi Penulisan

Metodelogi penulisan yang digunakan dalam Tugas Akhir ini adalah sebagai berikut :

1. Metode Studi Pustaka dan Literature

   Metode ini dilakukan dengan cara mencari dan mengumpulkan referensi yang berupa literature yang terdapat pada buku dan internet mengenai "PDF *Malware*".

2. Metode Konsultasi

   Metode ini melakukan konsultasi kepada pihak-pihak yang memiliki pengetahuan serta wawasan yang baik dalam mengatasi permasalahan yang ditemui pada penulisan tugas akhir "Klasifikasi *PDF Malware* dengan metode *Random Forest*".

3. Metode Pembuatan Model

   Metode ini akan mengolah dta yang akan digunakan dan membuat suatu perancangan pemodelan dengan menggunakan algoritma *Random Forest*.

4. Metode Pengujian

   Metode ini melakukan pengujian terhadap model yang telah dibuat, apakah model tersebut dapat menghasilkan nilai akurasi yang baik atau tidak.

5. Metode Analisa dan Kesimpulan

   Hasil pada pengujian tugas akhir ini akan dilakukan analisa untuk mengetahui kekurangannya, sehingga dapat digunakan sebagai saran untuk penelitian selanjutnya.

1.6 Sistematika Penulisan

Adapun sistematika penulisan dalam Tugas Akhir ini adalah sebagai berikut:

BAB I PENDAHULUAN

Pada Bab Pendahuluan berisi hal terkait latar belakang, tujuan penelitan, manfaat penelitian, ruang lingkup dan perumusan masalah, metodologi penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada Bab Tinjauan Pustaka berisi tentang dasar teori mengenai PDF *Malware, Data extraction*, dan *Random Forest*.

BAB III METODOLOGI PENELITIAN

Pada Bab Metodologi menjelaskan kerangka kerja penelitan, bagaimana langkah-langkah penelitian yang akan dilakukan. Penjelasan pada bab ini meliputi tahapan perancangan sistem dan penerapan metode yang digunakan penelitian.

BAB IV HASIL DAN ANALISA

Pada Bab Hasil dan Analisa membahas hasil pengolahan data, hasil *oversampling* menggunakan smote dan *undersampling* menggunakan *Nearmiss*, hasil *cross validation* dan analisa pengklasifikasian pada PDF *Malware*.

BAB V KESIMPULAN DAN SARAN

Pada Bab Kesimpulan dan Saran akan berisi kesimpulan dari hasil yang diperoleh pada penelitian yang telah dilakukan dan saran yang diharapkan dapat digunakan untuk penelitian berikutnya.

DAFTAR PUSTAKA

[1] X. Zhou and J. Pang, "Expdf: Exploits detection system based on machine-learning," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 2, pp. 1019–1028, 2019, doi: 10.2991/ijcis.d.190905.001.

[2] S. Dabral, A. Agarwal, M. Mahajan, and S. Kumar, "Malicious PDF files detection using structural and javascript based features," *Commun. Comput. Inf. Sci.*, vol. 750, no. May, pp. 137–147, 2017, doi: 10.1007/978-981-10-6544-6_14.

[3] C. D. Morales-Molina and ..., "Methodology for malware classification using a random forest classifier," *... Autumn Meet. ...*, 2018, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8661441/

[4] V. Geetha, A. Punitha, M. Abarna, M. Akshaya, S. Illakiya, and A. P. Janani, "An Effective Crop Prediction Using Random Forest Algorithm," *2020 Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2020*, 2020, doi: 10.1109/ICSCAN49426.2020.9262311.

[5] V. Atluri, "Malware Classification of Portable Executables using Tree-Based Ensemble Machine Learning," *2019 SoutheastCon*, 2019, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9020524/

[6] M. Chowdhury and A. Rahman, "Malware Analysis and Detection Using Data Mining and Machine Learning Classi fi cation," 2018, doi: 10.1007/978-3-319-67071-3.

[7] 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.

[8] D. Maiorca, G. Giacinto, and I. Corona, "A pattern recognition system for malicious PDF files detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7376 LNAI, pp.

510–524, 2012, doi: 10.1007/978-3-642-31537-4_40.

[9] S. M. Hossain and M. A. Ayub, “Parameter Optimization of Classification Techniques for PDF based Malware Detection,” *2020 23rd Int. Conf. …*, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9392685/

[10] M. Alam, M. Swawibe, and U. Alam, “Detection of Malware in PDF files Using NiCAD4 Tool,” *Researchgate.Net*, no. December 2016, 2017, doi: 10.13140/RG.2.2.13232.76804.

[11] R. Vyas, X. Luo, N. McFarland, and C. Justice, “Investigation of malicious portable executable file detection on the network using supervised learning techniques,” *Proc. IM 2017 - 2017 IFIP/IEEE Int. Symp. Integr. Netw. Serv. Manag.*, pp. 941–946, 2017, doi: 10.23919/INM.2017.7987416.

[12] 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.

[13] M. Goyal and R. Kumar, “Machine Learning for Malware Detection on Balanced and Imbalanced Datasets,” *2020 Int. Conf. Decis. Aid Sci. Appl. DASA 2020*, pp. 867–871, 2020, doi: 10.1109/DASA51403.2020.9317206.

[14] 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.

[15] P. Singh, S. Tapaswi, and S. Gupta, “Malware Detection in PDF and Office Documents: A survey,” *Inf. Secur. J.*, vol. 29, no. 3, pp. 134–153, 2020, doi: 10.1080/19393555.2020.1723747.

[16] 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

[17] 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.

[18] S. R. Gopaldinne, H. Kaur, P. Kaur, G. Kaur, and Madhuri, "Overview of PDF Malware Classifiers," *Proc. 2021 2nd Int. Conf. Intell. Eng. Manag. ICIEM 2021*, pp. 337–341, 2021, doi: 10.1109/ICIEM51511.2021.9445341.

[19] K. R. Mahmudah, B. Purnama, F. Indriani, and K. Satou, "Machine learning algorithms for predicting chronic obstructive pulmonary disease from gene expression data with class imbalance," *Bioinforma. 2021 - 12th Int. Conf. Bioinforma. Model. Methods Algorithms; Part 14th Int. Jt. Conf. Biomed. Eng. Syst. Technol. BIOSTEC 2021*, no. January, pp. 148–153, 2021, doi: 10.5220/0010316501480153.

[20] B. Kovács, F. Tinya, C. Németh, and P. Ódor, "Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment," *Ecol. Appl.*, vol. 30, no. 2, pp. 321–357, 2020, doi: 10.1002/eap.2043.

[21] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[22] A. R. B. Alamsyah, S. Rahma, N. S. Belinda, and A. Setiawan, "A R B Alamsyah et al SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data Case Study: IFLS 5," pp. 305–314, 2021, [Online]. Available: https://proceedings.stis.ac.id/icdsos/article/download/240/29/2098

[23] J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions : A Case Study involving Information Extraction," vol. 4, no. 1, pp. 88–100.

[24] M. A. Alim, S. Habib, Y. Farooq, and A. Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model," *2020 3rd Int. Conf. Comput. Math. Eng. Technol. Idea to Innov. Build. Knowl. Econ. iCoMET 2020*, 2020, doi:

10.1109/iCoMET48670.2020.9074135.

[25] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection,” *IEEE Access*, vol. 6, no. c, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[26] A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[27] A. D. Kulkarni and B. Lowe, “Random Forest Algorithm for Land Cover Classification,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 4, no. 3, pp. 58–63, 2016.