p-ISSN:2527-5321
e-ISSN:2527-5941

KOMBINASI ALGORITMA DECISION TREE DAN NAIVE BAYES UNTUK MENINGKATKAN AKURASI DAN KECEPATAN WAKTU DETEKSI MALWARE

Fitri Puspitasari Budiana[1], Hanasa Shelviani[2]

[1]UIN Sunan Gunung Djati Bandung, [2]Universitas Indraprasta PGRI
[1]Fitritasari@uinsgd.ac.id, [2]goalsnasa64@gmail.com

ABSTRAK

Secara umum malware mencakup virus, keylogger, worm, trojan, rootkit, spyware, ransomware, dan software berbahaya lain yang perlu diwaspadai. Malware menyerang berkas seseorang lalu menyalin diri sehingga bisa merusak sistem kerja hardisk, mengambil data smartphone, maupun merusak sistem operasi pada komputer target. Data malware memiliki variabel bebas dan variabel terikat. Pengembangan menggunakan teknik data mining untuk mencari variabel terbaik untuk data training dan data testing sedangkan secara teknik dengan profiling. Data mining membantu menemukan pola, pengetahuan baru, formula baru, aturan, maupun insight dari suatu data. Penelitian ini penulis mengusulkan pendekatan baru melalui eksperimen menggabungkan dua algoritma mesin pembelajaran yang berbeda yaitu algoritma memiliki label ataupun tidak berlabel. Berlandaskan dari dua pendekatan yang masing-masing mempunyai kelebihan. Eksperimen sebanyak empat kali penggabungan dengan lima dataset berbeda untuk menguji akurasi dan kecepatan waktu. Algoritma yang diuji yaitu antara algoritma random forest dengan decision tree, algoritma decision tree dengan random forest, algoritma naive bayes dengan algoritma decision tree, serta algoritma decision tree dengan algoritma naive bayes. Hasil eksperimen tertinggi ada pada penggabungan algoritma decision tree dan algoritma naive bayes dengan rata-rata 96,12 % dan kecepatan waktu 5,36 milidetik.

Kata Kunci: *kejahatan siber, malware, algoritma*

ABSTRACT

In general, malware includes viruses, keyloggers, worms, trojans, rootkits, spyware, ransomware, and other malicious software that needs to be aware of. Malware attacks a person's files and then replicates itself so that it can damage the hard disk's working system, take smartphone data, or damage the operating system on the target computer. Data malware has independent variables and dependent variables. Development uses data mining techniques to find the best variables for training data and testing data while technically with profiling. Data mining helps find patterns, new knowledge, new formulas, rules, or insights from data. In this study, the author suggests a new approach through experiments that combine two different machine learning algorithms, namely labeled or unlabeled algorithms. Based on the two approaches, each of which has advantages. The experiments were conducted four times combined with five different datasets to test accuracy and speed of time. The algorithms tested were between the random forest algorithm and the Decision Tree algorithm, the Decision Tree algorithm with the Random Forest algorithm, the Naive Bayes algorithm with the Decision Tree algorithm, and the Decision Tree algorithm with the Naive Bayes algorithm. The highest experimental results were found in the combination of the Decision Tree algorithm and the Naive Bayes algorithm with an average of 96.12% and a time speed of 5.36 milliseconds.
Key Word: *Cyber crime, malware, algorithm*

PENDAHULUAN

Perkembangan teknologi yang kita rasakan terjadi setiap waktu, tanpa kita sadari memilki dampak negatif. Dampak negatif pun beragam mulai dari tindakan kejahatan perorangan hingga kelompok oknum tidak bertanggungtawab. Kejahatandunia maya disebutsebagai kejahatan siber (*cyber crime*). Kejahatan siber meningkat karena kemudahan yang lebih besar dan resiko yang lebih rendah untuk tertangkap pihak berwajib. Jenis kejahatan siber umum terjadi diantaranya malware, phishing, serangan DDoS, serangan *man-in-the-middle*, dan masih banyaklagi (Hull dkk., 2018).

Secara umum malware mencakup virus, *keylogger, worm, trojan, rootkit, spyware, ransomware*, dan software berbahaya lain yang perlu diwaspadai. *Malware* menyerang

berkas seseorang lalu menyalin diri sehingga bisa merusak sistem kerja hardisk, mengambil data *smartphone*, maupun merusak sistem operasi pada komputer target (Ye dkk., 2018, hlm. a survey).

Bentuk *malware* pun bermacam-macam seperti berupa dokumen, *software*, kode atau user yang menjalankan system ilegal tanpa persetujuan pemilik. Teknik deteksi malware beragam mulai dari cara tradisional sampai berbasis komputasi awan. Tiga teknik analisis *malware* yaitu : *static analysis* menginfeksi komputer dari berbagai unduhan dokumen dan program, dynamic analysis menggunakan kode, grafik maupun sistem dan vision based analysis dengan proses Analisa *biner malware* dengan memvisualisasikannya sebagai gambar (Roseline dkk., 2020).

Proses analisis malware secara otomatis menggunakan dua tahapan yaitu dengan *feature extraction* yang berfungsi membedakan malware. Malware kemudian diekstrak secara statik maupun dinamik dengan pendekatan klasfikasi dan klastering. Klasifikasi membuat kelasfile yang belum dikenal menjadi malware benign. Klastering dengan cara mengelompokkan sampel yang memiliki kemiripan menjadi beberapa grup malware (Tang dkk., 2017).

Grup malware secara menyeluruh dikumpulkan menjadi satu datasetbesar yang jenisnya beragam baik secara terstruktur dan tidak tersruktur. Data malware memiliki variable bebas dan variable terikat. Variabel mewakili keadaan pendekatan dandapat berubah selama simulasi yang sifatnya fluktuasi (naik dan turun). Dalam pengembangan dataset mencari variable terbaik untuk data training dan data testing sedangkan secara teknik dengan profiling (Canbek dkk., 2018).

Dataset malware yang kompleks dipelajari menggunakan teknik data mining. Data mining membantu menemukan pola, pengetahuan baru, formula baru, aturan maupun inshight dari suatu data. Peran utama data mining dapat membuat estimasi, forecasting, klasifikasi, klastering, dan asosiasi. Pendekatan data mining dibuktikan pada efek pola data yang digunakan akan menghasilkan kualitasdata dari dimensi akurasi, data terkini, kelengkapan, konsisten dan interprestasi mudah (Bushuyeva dan Kutsenko, 2019).

Saat ini algoritma machine learning memiliki banyak pendekatan dari tiga tipe algoritma *supervised learning, unsupervised learning* dan *reinforcement. Machine learning* mengenali secara otomatis pola kompleks dan membuat keputusan cerdas berdasarkan data. Algoritma yang dikelompokkan berdasarkan *supervised* learning memetakan masukan yang dikehendaki dengan label sedangkan *unsupervised* learning mependekatankan himpunan masukan dengan cara mengelompokkan *atauclustering* (Ghazi dan Moulay Rachid, 2020).

Algoritma *clustering* dapat mengklasifikasikan seluruh kumpulan data dengan mudah. Standar pengelompokan clustering dengan mengelompokkan objek data mirip satu sama lain dalam clusteryang sama dan berbeda dengan objek di cluster lain (Dan dkk., 2011). Algoritma *unsupervised learning* tidak hanya mendeteksi anomali, juga dapat menghematwaktu dalam menggali sebuah data besar. Sisi positif penggunaan algoritma unsupervised learning adalah dapat mendeteksi apa yang tidak dapat dipahami oleh mata manusia. *Unsupervised learning* memiliki kompleksitas yang rendah sehingga pengguna tidak harus menafsirkan label (Shalev-Shwartz dan Ben-David, 2014).

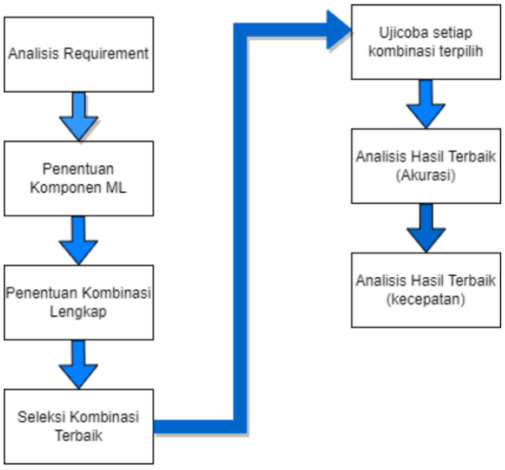
Penelitian sebelumnya sebanyak 50 penelitian bersifat komparatif membahas satu algoritma, 10 penelitian diantaranya membahas kombinasi algoritma *supervised.* Namun, belum ada penelitian yang melakukan penggabungan dua algoritma yang berbeda jenis antara algoritma *supervised* dan algoritma *unsupervised.* Melihat fenomena populasi malware dan kaitannya antara algoritma maka dua hal tersebut menjadi pendukung bahwa terdapat celah penelitian. Celah penelitian yang diusulkan yaitu menggabungkan dua jenis algoritma antara *unsupervised learning* dan *supervised learning* sebagai pendekatan alternatif yang dituju.

p-ISSN:2527-5321
e-ISSN:2527-5941

Penelitian ini dibuat komparatif dengan analisis dari masing-masing algoritma terpilih. Kombinasi dua algoritma *unsupervised learning* dan *supervised learning* diharapkan dapat meningkatkan keakuratan dan memberikan pertahanan tambahandalam deteksi malware dari pemilihan algoritma yang tepat didasarkan pada karakteristik data yang ada, kebutuhan spesifik dan evaluasi kinerja yang berkelanjutan. Kombinasi teknik clustering yang menghemat waktu dan classification yang memiliki efektivitas deteksi malware. Kombinasi jenis ini dapat memberikan peningkatan akurasi dan menghemat waktu.

METODE PENELITIAN

Metode penelitian ini, mendesain kombinasi algoritma sebagai pendekatan alternatif. Kombinasi bertujuan menganalisis akurasi dan kecepatan waktu seperti pada gambar Gambar 1



Gambar 1. Metode Penelitian

Penelitian ini terdiri dari 7 tahapan utama: (1) analisis requirement, (2) penentuan komponen machine learning (ML), (3) penentuan kombinasi lengkap, (4) seleksi kombinasi terbaik, (5) ujicoba setiap kombinasi terbaik, (6) analisis hasil akurasi terbaik, (7) analisis hasil kecepatan terbaik. Metodologi ini berfokus pada pengembangan pendekatan sebagai pendekatan alternatif baru. Setiap kotak menunjukkan langkah-langkah yang harus dilakukan sedangkan tanda panah menunjukkan alur pelaksanaan tahapan penelitian sesuai kebutuhan.

A. Analisis Requirement

Tahapan pertama termasuk adalah bagian vital dengan analisis requirement kebutuhan dengan studi literatur menggunakan Systematic Literature Review (SLR) dari berabgai sumber yang berkaitan dengan topik penelitian deteksi malware. Secara umum tahap ini diuraikan pada tabel 1. berisi identifikasi tunggal maupun komparatif algoritma ML. Hasilnya adalah algoritma ML yang digunakan sebagai pendekatan.

Tabel 1. Analisis Requirement

Masukan	Jurnal dan conference mengenai deteksi malware; Observasi kebutuhan.
Pendekatan	Studi literatur; Identifikasi masalah dan celah penelitian algoritma tunggal maupun komparatif; Konsultasi dengan dosen pembimbing;
Keluaran	Algoritma yang berpotensi untuk pendekatan kombinasi

B. Penentuan Komponen Machine Learning

Tahapan kedua dilakukan eksplorasi untuk menganalisis data. Dengan mengumpulkan dataset untuk pengetahuan awal, mengevaluasi kualitas data dengan memilih sebagian kelompok kecil dari data untuk melihat pola permasalahan. Dataset didapatkan secara publik untuk penelitian di UCL Machine learning dan *kaggle repository*. Data lainnya dari data privat berasal konsol cloud managemen antivirus suatu bank daerah di Indonesia customer dari perusahaan jasa antivirus PT Vaksincom. Dataset yang dikumpulkan adalah dataset malware tahun 2019 hingga Mei tahun 2022.

Dataset disimpan dalam format .CSV, data tersebut memiliki atribut beragam maka sebelum preprocessing data diurutkan sesuai tipe ini. Catatan untuk tools rapid miner menampilkan sampai 5000 baris data dan kolom:

1. Polynominal = untuk atribut yang memiliki lebih dari 2 kategori
2. Binominal = untuk atribut yang memiliki 2 kategori saja
3. Real = untuk tipe data yang memiliki nilai koma, atau decimal

4. Integer = untuk file bernilai integer atau bilangan buat tanpa koma.

C. Penentuan Kombinasi Lengkap

Tahap ketiga adalah menyiapkan data dengan memperbaiki data dari missing value. Terkhusus penelitian komparatif yang menjadi objek penelitian adalah algoritma dan dataset malware. Pengelompokkan data dengan memilih atribut dan variable yang ingin ditampilkan menggunakan tools rapid miner. Melakukan perubahan pada beberapa variable sesuai kebutuhan.

D. Seleksi Kombinasi Terbaik

Tahap keempat membuktikan bahwa pendekatan yang telah dikombinasi dapat berfungsi semestinya. Hasil ini diidentifikasi bahwa masih ada celah penelitian. Maka termotivasi untuk memberikan usulan kombinasi antara dua algoritma. Algoritma machine learning dengan akurasi terbaik yang digunakan untuk penelitian yaitu algoritma *classification* pada *supervised learning*. Tahap pendekatan sebagai verifikasi yang bertujuan untuk melihat hasil dua kombinasi algoritma. Mulai eksperimen melakukan uji algoritma.

E. Ujicoba Setiap Kombinasi Terpilih

Tahap kelima dilakukan untuk menghitung akurasi dari pendeteksian malware. Ujicoba setiap kombinasi sebagai tahapan yang penting, agar data latih yang telah diuji dapat menghasilkan pendekatan alternatif. Impact analysis yang diharapkan dalam penelitian ini merupakan proses mengevaluasi dampak atau konsekuensi dari pendekatan kombinasi algoritma yang dilihat dari berbagai aspek. Melakukan Impact Analysis dalam penelitian desain membantu memastikan bahwa desain pendekatan yang dikembangkan memiliki efek yang diinginkan.

F. Analisi Hasil Terbaik Nilai Akurasi

Penelitian ini menganalisis hasil terbaik yang dihasilkan dari percobaan antara algoritma decision tree dengan naïve bayes. Jumlah dataset yang digunakan beragam. Jika Atribut data semakin kompleks maka kecendrungan untuk overfitting terjadi. Jika terjadi overfitting, pendekatan akan mempelajari semua detail tren pada atribut bukan atribut yang ingin dihasilkan sehingga akurasi tidak tercapai karena semakin kecil nilainya. Analisis akurasi dipengaruhi oleh parameter atribut dataset serta semakin besar dataset, algoritma akan menunjukkan tingkat akurasi yang sebenarnya. Algoritma decision tree dan naïve bayes sebagai algoritma klasifikasi yang memberikan hasil akurasi maksimal.

G. Teknik Pengumpulan Data

1. Preprocessing Data
2. Pembersihan Data
3. Data Integration and Data Transformation
4. Data Reduction

HASIL DAN PEMBAHASAN

Pada Pada bagian ini analisis dibahas alur analisis proses pendeteksian malware. Proses analisis dimulai dari memilih requirement yang dibutuhkan untuk peningkatan akurasi.

A. Analisis

1. Analisis Requirement

Langkah-langkah Analisis Kombinasi yaitu dari analisis requirement, penentuan komponen machine learning dengan komparatif dari kelebihan algorima klasifikasi dan seleksi algoritma untuk menjadi kombinasi terbaik.

2. Analisis Karakteristik Algoritma

Analisis karakteristik supervised learning menjelaskan dari kemampuan data latih hingga proses belajar jenis machine learning. Pada penelitian sebelumnya terdapat informasi hasil serta gap penelitian yang dapat menjadi celah penelitian selanjutnya. Adapun dasar pemikiran untuk penelitian ini yaitu mengusulkan pendekatan algoritma dengan dua jenis berbeda. Karakteristik algoritma menjawab bahwa terdapat perbedaan pola jenis machine learning. Perbedaan pola ini menjadi hal menarik untuk diteliti didukung dengan dataset. Dibuatlah penggabungan dua algorima berbeda yaitu algoritma klasifikasi dan klastering.

p-ISSN:2527-5321
e-ISSN:2527-5941

3. Penentuan Komponen Machine Learning

Pemilihan kombinasi yang akan dicoba disertai alasan mengapa kombinasi- tersebut yang dicoba, sesuai 4 langkah pertama dalam metodologi. Decision tree memiliki kesamaan dengan naïve bayes dalam lingkup algoritma klasifikasi. *Decision tree* untuk mengklasifikasi dataset sedangkan naïve bayes akan menampilkan seberapa besar probabilitas dari dataset. Karakteristik algoritma terdapat perbedaan pola jenis machine learning. Perbedaan pola ini menjadi hal menarik untuk diteliti didukung dengan dataset. Dibuatlah penggabungan dua algorima yaitu algoritma klasifikasi.

4. Penentuan Kombinasi Lengkap

Algoritma *supervised learning* yang diimplementasikan pada rapidminer adalah dua algoritma *decision tree* dan *naïve bayes*. Usulan pendekatan yang dibuat sebagai berikut : Algortima *Decision Tree* + Algoritma *Naïve Bayes.*

B. Uji Coba Setiap Kombinasi Terpilih

Menampilkan proses analisa data mining yang berisi operator dengan alur koneksi diantara satu operator ke operator lainnya. Operator memerlukan parameter untuk berfumgsi. Komponen dari proses data mining dasarnya adalah :

1) Deskripsi input,
2) Deskripsi output,
3) Aksi yang dilakukan;
4) Paramater dan atribut yang dipilih.

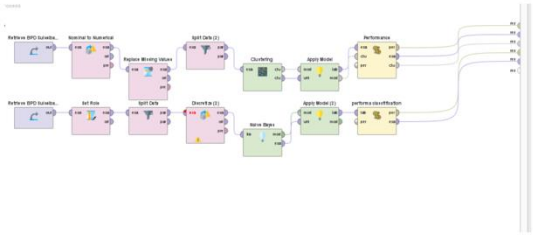
C. Proses Verifikasi

Secara umum, tanpa terikat pada istilah yang digunakan metodologi tertentu, hasil desain dapat diuji dan dibuktikan kualitasnya dengan proses verifikasi dan evaluasi. Verifikasi dan evaluasi adalah dua proses penting dalam penelitian desain.

Verifikasi dalam penelitian desain melibatkan langkah-langkah untuk memastikan bahwa desain memenuhi persyaratan dan spesifikasi yang telah ditetapkan. Hal ini dapat dilakukan dengan melakukan simulasi, pengujian, atau pengujian komponen individu dalam desain untuk memverifikasi bahwa mereka beroperasi sebagaimana yang diharapkan. Verifikasi juga melibatkan verifikasi analitis dari desain menggunakan pendekatan matematika atau perangkat lunak untuk memastikan bahwa desain memenuhi kriteria tertentu.

D. Eksperimen

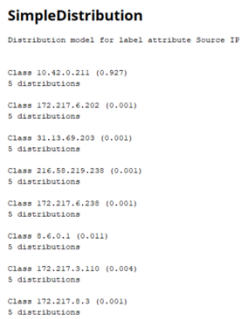
Pada Gambar 2 merupakan proses kombinasi *decision tree* + algoritma *naïve bayes*. Tahapan dimulai dari input data, setting parameter, split data menjadi data latih dan data uji, memasukkan algoritma decision tree dan *naïve bayes* , apply model dan performance algoritma untuk melihat akurasi.



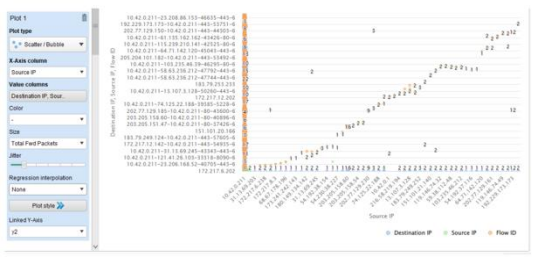
Gambar 2. Kombinasi Decision Tree +Naïve Bayes

E. Hasil Pengujian

Pada Gambar 3 Hasil Uji Algoritma *Decision Tree* + Algortima *Naïve Bayes* dengan dataset android malware



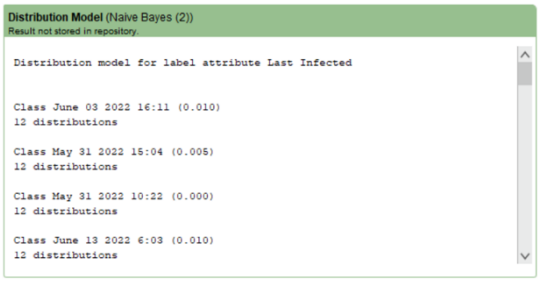
Gambar 3. Simple Distribution



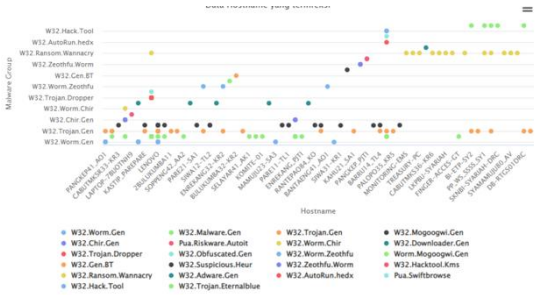
Gambar 4. Diagram Destination IP

p-ISSN:2527-5321
e-ISSN:2527-5941

Hasil Uji Algoritma *Decision Tree* + Algortima *Naïve Bayes* dengan dataset BPD Sulselbar



Gambar 5. Simple Distribution



Gambar 6. Grafik Jenis Hostname

Pada gambar 6 merupakan distribution model. Gambar 6 adalah Hasil uji berupa grafik hostname yang terkena serangan malware.

F. Evaluasi

Evaluasi hasil dari pengujian menjelaskan tahapan evaluasi metodologi. Bab ini terdiri dari dua bagian evaluasi algoritma, pengujian dataset malware untuk mengetahui efektifitas dan efisiensi pendekatan pendekatan alteranitf algoritma usulan. Evaluasi Algoritma *Decision Tree* + Algoritma *Naive Bayes*. Atribut data berpengaruh akurasi. Meningkatnya kompleksitas atribut menghasilkan kecenderungan prediks iyang buruk pada data dengan fitur yang sangat berkorelasi akurasinya makin keci

Tabel 2 Evaluasi Decision Tree dan Naïve Bayes

Algoritma Tunggal	Akurasi	Recall	Precison	Cross Entrophy	Classification Erorr
Decision Tree	96,58 %	31,40 %	20,70 %	0.120	3,42 %
Naïve Bayes	92,41 %	6,90%	5,61 %	0.120	7,59 %
Hasil Rata-Rata	94,49 %	19,5 %	13,15 %	0.120	5,5 %

SIMPULAN DAN SARAN

Sebagai upaya untuk mencari alternatif yang lebih baik (akurasi untuk proses deteksi malware. Pada penelitian ini telah uji coba kombinasi teknik *Machine Learning*. Pertama dengan analisis untuk memberikan justifikasi bahwa kombinasi- tersebut berpotensi memberikan kinerja yang lebih baik. Penelitian ini mencapai akurasi terbaik dicapai dengan penggunaan kombinasi *Decision Tree* dan *Naïve Bayes* dengan akurasi 95,00 %. Berdasarkan hasil uji coba tersebut dapat disimpulkan bahwa kombinasi *decision tree* dan *naïve bayes* dapat menghasilkan akurasi yang lebih baik dan menghasilkan nilai probabilitas.

DAFTAR PUSTAKA

Bushuyeva, N., dan Kutsenko, M. (2019): Data Mining technics in projects with multinational teams, 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), IEEE, Lviv, Ukraine, 278–281. https://doi.org/10.1109/STC-CSIT.2019.8929801

Canbek, G., Sagiroglu, S., dan Taskaya Temizel, T. (2018): New Techniques in Profiling Big Datasets for Machine Learning with a Concise Review of Android Mobile Malware Datasets, 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), IEEE, ANKARA, Turkey, 117–121

Dan, J., Jianlin, Q., Yanyun, C., dan Li, C. (2011): Clustering method and its formalization, 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference, IEEE, Chongqing, China, 57–61. https://doi.org/10.1109/ITAIC.2011.6030150

Ghazi, A. E., dan Moulay Rachid, A. (2020): Machine learning and datamining methods for hybrid IoT intrusion detection, 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), IEEE, Marrakesh, Morocco, 1–6. https://doi.org/10.1109/CloudTech49835.2020.9365895

Hull, M., Eze, T., dan Speakman, L. (2018): Policing The Cyber Threat: Exploring the Threat from Cyber Crime and the Ability of Local LawEnforcement to Respond, 2018 European Intelligence and Security Informatics Conference (EISIC), IEEE, Karlskrona, Sweden, 15–22. https://doi.org/10.1109/EISIC.2018.00011

Roseline, S. A., Geetha, S., Kadry, S., dan Nam, Y. (2020): Intelligent VisionBased Malware Detection and Classification Using Deep Random Forest Paradigm, IEEE Access, 8, 206303–206324. https://doi.org/10.1109/ACCESS.2020.3036491

Shalev-Shwartz, S., dan Ben-David, S. (2014): Understanding Machine Learning: From Theory to Algorithms (1 ed.), Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Tang, X., Dong, M., Bi, S., Pei, M., Cao, D., Xie, C., dan Chi, S. (2017): Feature Selection Algorithm Based on K-means Clustering, 2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE, Honolulu, HI, 1522–1527. https://doi.org/10.1109/CYBER.2017.8446108

Yang, L., Ciptadi, A., Laziuk, I., Ahmadzadeh, A., dan Wang, G. (2021): BODMAS: An Open Dataset for Learning based Temporal Analysis of PE Malware, 2021 IEEE Security and Privacy Workshops (SPW), IEEE, San Francisco, CA, USA, 78–84. https://doi.org/10.1109/SPW53761.2021.00020