JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

Deteksi Dini Ancaman Malware pada OS Windows di Indonesia Menggunakan Analisis Data PE Header Berbasis Algoritma Random Forest

Angga Ginanjar[1], Ilham[2], Lyra Kudunga[3], Mutiara Laela Sukartini[4], Rizkyanti Ajeng Trias Marani[5], Rahmawati[6*]

[1,2,3,4,5,6] Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]anggauser33@gmail.com, [2]ilham@gmail.com, [3]lyrakudunga@gmail.com, [4]mutiaralaela@gmail.com, [5]rizkyantiajang@gmail.com, [6*]dosen02394@unpam.ac.id

Abstrak−Perkembangan infrastruktur digital di Indonesia berbanding lurus dengan meningkatnya risiko serangan siber, khususnya ancaman malware yang menargetkan sistem operasi Windows sebagai platform paling dominan. Teknik deteksi tradisional yang berbasis signature-based terbukti kurang efektif menghadapi malware varian baru atau serangan zero-day. Oleh karena itu, penelitian ini bertujuan untuk membangun sebuah sistem deteksi dini malware pada OS Windows melalui analisis data Portable Executable (PE) Header menggunakan algoritma Machine Learning Random Forest. Dataset yang digunakan mencakup 19.611 sampel file PE (14.599 malware dan 5.012 file aman/benign) dengan 78 fitur teknis header. Tahapan metodologi meliputi pembersihan data, pengisian missing values secara median, ekstraksi fitur dominan, serta pembagian data training dan testing sebesar 80:20. Hasil pengujian menunjukkan bahwa model Random Forest mampu mendeteksi ancaman malware dengan tingkat akurasi yang sangat impresif mencapai 99,13%, dengan presisi 1.00% untuk file benign dan recall 1.00% untuk file malware. Fitur PE Header yang paling berpengaruh secara signifikan dalam proses klasifikasi adalah MajorLinkerVersion (8,34%), MinorOperatingSystemVersion (8,02%), dan MajorSubsystemVersion (7,45%). Penelitian ini membuktikan bahwa analisis PE Header yang dikombinasikan dengan algoritma Random Forest dapat diandalkan sebagai mekanisme pertahanan dini yang responsif dan akurat untuk memperkuat ketahanan siber nasional di Indonesia.

Kata Kunci: Deteksi Malware, Windows OS, PE Header, Random Forest, Keamanan Siber.

Abstract−*The development of digital infrastructure in Indonesia is directly proportional to the increasing risk of cyberattacks, especially malware threats targeting the Windows operating system as the most dominant platform. Traditional signature-based detection techniques have proven ineffective against new malware variants or zero-day attacks. Therefore, this study aims to build an early detection system for malware on the Windows OS through Portable Executable (PE) Header data analysis using the Random Forest Machine Learning algorithm. The dataset used includes 19,611 PE file samples (14,599 malware and 5,012 safe/benign files) with 78 technical header features. The methodological stages include data cleaning, filling in missing values by median, extracting dominant features, and dividing training and testing data by 80:20. The test results show that the Random Forest model is able to detect malware threats with a very impressive accuracy rate of 99.13%, with a precision of 1.00% for benign files and a recall of 1.00% for malware files. The PE Header features that significantly influenced the classification process were MajorLinkerVersion (8.34%), MinorOperatingSystemVersion (8.02%), and MajorSubsystemVersion (7.45%). This study demonstrates that PE Header analysis combined with the Random Forest algorithm can be relied upon as a responsive and accurate early defense mechanism to strengthen national cyber resilience in Indonesia.*

Keywords: Malware Detection, Windows OS, PE Header, Random Forest, Cybersecurity.

1. PENDAHULUAN

Seiring dengan akselerasi transformasi digital yang masif di Indonesia, ketergantungan masyarakat, sektor korporasi, hingga instansi pemerintahan terhadap infrastruktur teknologi informasi semakin mutlak. Di antara berbagai platform sistem operasi yang digunakan, Microsoft Windows menempati posisi puncak sebagai OS yang paling banyak diadopsi di Indonesia karena kemudahan integrasi dan fungsionalitasnya. Namun, dominasi pasar ini menjadikan ekosistem Windows sebagai target utama para aktor ancaman siber, terutama pengembang perangkat lunak berbahaya (malware).

Badan Siber dan Sandi Negara (BSSN) mencatat ratusan juta anomali tren serangan siber yang masuk ke Indonesia setiap tahunnya, di mana infeksi malware mendominasi jenis serangan tersebut. Ancaman ini tidak hanya berdampak pada kebocoran data personal, melainkan berpotensi

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

melumpuhkan sistem operasional publik, menyebabkan kerugian finansial yang signifikan, hingga mengancam kedaulatan informasi nasional.

Metode proteksi konvensional seperti antivirus berbasis signature-based detection mendeteksi malware dengan mencocokkan nilai hash file terhadap basis data virus yang sudah diketahui. Kelemahan fatal dari metode ini muncul saat menghadapi serangan malware modern yang bersifat polimorfik, metamorfik, atau serangan zero-day yang belum pernah terdaftar di database. Oleh sebab itu, diperlukan pendekatan proaktif yang mampu menganalisis struktur struktural file sebelum file tersebut dieksekusi.

Setiap program aplikasi yang dapat dijalankan pada Windows menggunakan format berkas standar bernama Portable Executable (PE). Di dalam struktur berkas PE terdapat komponen vital yang disebut PE Header. PE Header menyimpan metadata teknis mengenai karakteristik berkas, struktur memori, fungsi pustaka (DLL) yang diimpor, versi compiler yang digunakan, susunan bagian berkas (sections), hingga pengaturan subsistem operasional berkas. Pola penyusunan metadata pada file malware cenderung memiliki kejanggalan struktural yang mencolok dibandingkan dengan file aman (benign).

Melalui pemanfaatan teknologi Machine Learning, kejanggalan atau pola tersembunyi pada struktur PE Header tersebut dapat diekstrak dan dipelajari secara otomatis tanpa perlu melakukan eksekusi file secara dinamis yang berisiko merusak sistem sandboxing. Algoritma Random Forest dikenal memiliki keunggulan performa yang sangat tinggi dan stabil untuk klasifikasi data tabular berdimensi besar karena kemampuannya memproses banyak fitur tanpa mengalami overfitting yang parah.

Penelitian ini berfokus pada pengembangan model klasifikasi klasifikasi deteksi dini ancaman malware di Indonesia berbasis analisis data PE Header menggunakan algoritma Random Forest. Dengan mengoptimalkan 78 fitur struktural dari dataset riil yang berisi belasan ribu sampel file PE, diharapkan penelitian ini dapat melahirkan kontribusi akademis serta solusi praktis yang adaptif dalam memperkuat benteng pertahanan digital nasional.

2. METODE PENELITIAN

2.1 Tinjauan Pustaka

Penelitian terdahulu mengenai deteksi malware berbasis klasifikasi data tabular telah banyak dikembangkan. Pratama dkk. (2026) dalam penelitiannya menerapkan algoritma Extreme Gradient Boosting (XGBoost) berbasis TF-IDF untuk mendeteksi pelanggaran etika siber pada media sosial X dan membuktikan bahwa algoritma ensemble berbasis pohon memiliki akurasi superior dalam mengenali pola anomali teks digital. Kesimpulan serupa didapatkan oleh Murni dkk. (2026) yang membandingkan performa Support Vector Machine (SVM) menggunakan platform Orange Data Mining untuk klasifikasi komentar cyberbullying pada media sosial Instagram, menunjukkan bahwa pemilihan parameter yang tepat pada algoritma machine learning menentukan performa final deteksi.

Penelitian deteksi malware statis pada file Portable Executable (PE) mulai mendapatkan perhatian setelah Raff et al. mengembangkan platform untuk analisis berbasis urutan byte end-to-end dengan jaringan konvolusi dalam Malware Detection by Eating a Whole EXE. Makalah tersebut menunjukkan bahwa informasi dasar dari header dan konten file PE mudah diperoleh dari model pembelajaran mendalam tanpa memerlukan pengembangan fitur manual. Untuk mendukung replikasi dan perbandingan penelitian, Anderson dan Roth merilis dataset EMBER yang menyediakan lebih dari 2.000 atribut statis (struktur header, impor API, informasi versi) dari jutaan sampel Windows. Dataset ini dapat digunakan untuk melakukan eksperimen deteksi dini dengan pembelajaran mesin dalam skala besar. Kehadiran EMBER memungkinkan perbandingan kinerja yang mudah dan memungkinkan pendekatan ensemble ringan seperti Gradient Boosting dan arsitektur pembelajaran mendalam khusus untuk header PE.

Semua perkembangan ini menunjukkan kebutuhan akan dataset yang lebih besar dan program yang diberi anotasi dengan baik. DeepImpact et al. mengembangkan MalNet [4], yang berisi puluhan juta representasi grafis dan metadata PE dan merupakan alat untuk memeriksa program dan hubungan antara keluarga malware. Kombinasi teknik yang diperkenalkan dan ketersediaan MalNet menempatkan penelitian deteksi ancaman dini di wilayah seperti Indonesia

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

dalam kerangka kerja di mana ekstraksi fitur header otomatis, transfer learning, dan kemampuan untuk menilai varian baru yang berkembang pesat berada di garis depan.

2.2 Metode Penelitian

Metodologi yang diterapkan dalam penelitian ini dirancang secara sistematis melalui beberapa tahapan utama:

2.2.1 Pengumpulan Dataset

Dataset yang dianalisis diperoleh dari repositori sekunder analisis malware publik yang memuat ekstraksi fitur PE Header dari berkas biner Windows riil. Dataset ini terdiri dari dua file utama, yaitu berkas training utama dataset_malwares.csv berjumlah 19.611 baris data, dan berkas pengujian mandiri dataset_test.csv yang berfungsi sebagai instrumen blind-test.

2.2.2 Pra-pemrosesan Data

Data mentah dievaluasi untuk mendeteksi keberadaan nilai kosong (missing values) atau noise. Langkah pembersihan dilakukan dengan membuang kolom identifikasi kontekstual seperti Name yang menyimpan nama file asli atau nilai hash identitas unik karena tidak memiliki korelasi logis terhadap sifat internal malware. Nilai kosong pada fitur numerik diisi menggunakan pendekatan nilai tengah (median imputation) guna mempertahankan integritas distribusi statistik data tanpa terpengaruh oleh nilai ekstrem (outliers).

2.2.3 Arsitektur Model Random Forest

Algoritma Random Forest bekerja dengan mengombinasikan hasil prediksi dari banyak pohon keputusan secara paralel. Rumus matematis utama pemilihan pemisahan node didasarkan pada tingkat ketidakmurnian Gini (Gini Impurity):

$$Gini = 1 - \Sigma (pi)2$$

Di mana pi mewakili probabilitas probabilitas kemunculan kelas i pada node data terkait.

Tabel 1. Distribusi Dataset Penelitian

Kategori Berkas	Label Target	Jumlah Sampel	Persentase
Aman (Benign)	0	5.012	25,56%
Malware	1	14.599	74,44%
Total Data	**-**	**19.611**	**100,00%**

3. HASIL DAN PEMBAHASAN

3.1 Distribusi dan Karakteristik Dataset

Uji ini akan menganalisis 19.611 sampel file PE yang diekstraksi dalam 78 fitur teknis struktural pada tahap desain. Sebelum memulai pelatihan model, dataset dikelompokkan untuk melihat keseimbangan data target.

Tabel 1. Distribusi Dataset Penelitian

Kategori Berkas	Label Target	Jumlah Sampel	Persentase
Aman (Benign)	0	5.012	25,56%

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

Kategori Berkas	Label Target	Jumlah Sampel	Persentase
Malware	1	14.599	74,44%
Total Data	-	19.611	100,00%

Distribusi Dataset Penelitian Kategori File Label Target Jumlah Sampel Persentase Aman (Benign) 0 5.012 25,56% Malware 1 14.599 74,44% Total Data - 19.611 100,00% Data dalam tabel 1 cukup miring dengan sampel malware muncul sebesar 74,44%. Untuk menghindari bias dari ketidakseimbangan ini dalam distribusi kelas pada tahap pembelajaran, kami akan membagi data menggunakan metode stratified splitting dengan 80% untuk data pelatihan (15.688 sampel) dan 20% untuk data uji (3.923 sampel). Jadi, persentase kelas benign dan malware dijaga dalam subset pelatihan dan pengujian. Nilai yang hilang dalam fitur numerik dihitung melalui imputasi median untuk membuat distribusi statistik se-stabil mungkin tanpa terdistorsi oleh outlier.

3.2 Evaluasi kinerja klasifikasi algoritma Random Forest.

Pohon keputusan dibagi dan dievaluasi berdasarkan tingkat Gini Impurity dari algoritma Random Forest dalam hal rumus berikut, di mana p_i adalah probabilitas kelas target i tiba di setiap node dalam node yang berbeda. Secara paralel, model random forest diuji pada dataset independen (blind test) sebanyak 3.923 sampel. Tabel 2. Matriks Kinerja Klasifikasi Model Kelas Target Presisi Recall F1-Score Dukungan 0 (Benign) 1,00 0,97 0,98 1.003 1 (Malware) 0,99 1,00 0,99 2.920 Akurasi Akhir 99,13% - 3.923 Berdasarkan hasil perhitungan kami dalam Tabel 2.

Tabel 2. Matriks Performa Klasifikasi Model

Kelas Target	Precision	Recall	F1-Score	Support
0 (Benign)	1.00	0.97	0.98	1.003
1 (Malware)	0.99	1.00	0.99	2.920
Akurasi Final	99,13%	-	-	3.923

Model Random Forest kami mencapai 99,13%. Secara lebih rinci, kami dapat menunjukkan kinerja prediktif dari setiap kelas sebagai berikut. Kelas Benign (0): Kami mencapai nilai Presisi sempurna sebesar 1,00. Ini berarti bahwa model kami memiliki kepastian 100% dalam mengklasifikasikan file normal dan tidak membuat kesalahan False Alarm (file baik yang ternyata malware). Dengan nilai Recall sebesar 0,97, kami melihat bahwa hanya 3% dari file benign yang diprediksi salah. Kelas Malware (1): Nilai Recall maksimum sebesar 1,00 berarti bahwa model kami sensitif terhadap semua ancaman malware dalam dataset uji tanpa mendeteksi sampel berbahaya (nol False Negative). Nilai Presisi sebesar 0,99 mengonfirmasi tingkat keandalan yang sangat tinggi dalam mendeteksi ancaman. Nilai F1-Score yang besar dari kedua kelas (0,98 untuk benign dan 0,99 untuk malware) menunjukkan bahwa algoritma ensemble ini sangat kuat, stabil dan memiliki kesalahan yang sangat tipis bahkan saat bekerja dengan dataset berdimensi besar.

3.3 Interpretasi fitur dominan (Feature Importance)

Kepentingan utama dari fitur dalam komponen PE Header adalah untuk mengidentifikasi pola tersembunyi dan anomali struktural yang disembunyikan oleh file malware. Oleh karena itu,

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

kami menganalisis kontribusi dari fitur yang paling populer. Kami mengekstraksi tiga parameter struktural utama dari fitur yang diekstraksi:

[MajorLinkerVersion: 8,34%] ⟶ [MinorOperatingSystemVersion: 8,02%] ⟶ [MajorSubsystemVersion: 7,45%] MajorLinkerVersion (8,34%)

Menjadi fitur paling penting dalam keputusan klasifikasi kami. Nilai versi linker compiler seringkali sangat tinggi dan fitur sangat dipenuhi anomali dalam file malware. Ini karena penyerang siber menggunakan compiler, packer, atau linker lama yang disesuaikan untuk merakit kode biner mereka untuk menghindari deteksi oleh mesin antivirus konvensional. MinorOperatingSystemVersion (8,02%): Ini adalah versi rendah dari sistem operasi dalam file eksekusi. Angka dalam fitur ini sering dimanipulasi (misalnya diatur ke nilai non-standar). MajorSubsystemVersion (7,45%): Ini adalah versi subsistem Windows minimum yang diperlukan untuk menjalankan file ini. Dari data kami, ketiga parameter di atas menunjukkan bahwa penjahat siber secara aktif memanipulasi informasi versi compiler, linker, dan subsistem operasional dalam struktur PE Header. Manipulasi ini dimaksudkan untuk menempatkan Windows dalam mode kompatibilitas yang berbeda (atau berjalan dalam subsistem tertentu di OS) dan memudahkan malware untuk memanfaatkan kerentanan keamanan lama (legacy vulnerabilities) dalam OS Windows.

3.4 Implikasi untuk keamanan siber di Indonesia

Analisis statis ini memiliki implikasi taktis penting untuk kejahatan siber di Indonesia. Metode perlindungan tradisional yang berbasis pencocokan tanda tangan tidak lagi efektif terhadap mutasi malware polimorfik atau serangan zero-day yang sering ditargetkan pada infrastruktur kritis nasional. Dengan memanfaatkan kinerja terbaik dari algoritma Random Forest dan analisis metadata PE Header, sistem pertahanan tidak perlu menghabiskan sumber daya komputasi dan memori yang besar untuk pembongkaran biner dinamis (analisis dinamis). Efisiensi komputasi model ini membuatnya cepat dan ringan. Implementasi model ini sangat andal jika diintegrasikan sebagai komponen deteksi tambahan di gerbang keamanan lembaga publik, sektor perbankan, pemerintah, dan korporasi di Indonesia untuk mencegat dan mengkarantina serangan malware secara real-time sebelum infeksi berhasil menyusup ke perangkat pengguna akhir.

4. KESIMPULAN

Penelitian ini sukses mengimplementasikan algoritma Random Forest untuk mendeteksi dini ancaman malware pada sistem operasi Windows melalui ekstraksi data PE Header. Akurasi pengujian yang mencapai 99,13% membuktikan bahwa pendekatan analisis statis berbasis metadata struktural berkas biner sangat solid, responsif, dan akurat. Implementasi model ini diharapkan mampu diintegrasikan menjadi mesin deteksi pendamping (add-on) pada gateway keamanan jaringan instansi di Indonesia untuk menangkal serangan siber sebelum masuk ke perangkat pengguna akhir.

REFERENCES

Applications, E. (2025). Implementation of TF-IDF and XGBoost Algorithms in Scientific Paper Classification. Journal of Cyber Security, 5(1), 1-5.

Felix Fernando. (2025). Klasifikasi Tweet Cyberbullying Dengan Menggunakan Algoritma Svm Dan Xgboost. Jurnal Ilmu Komputer Dan Sistem Informasi, 13(1). https://doi.org/10.24912/jiksi.v13i1.32857

Kairupan, I. Y., Angdresey, A., & Arif, H. (2023). An Extreme Gradient Boosting Approach for Classification and Sentiment Analysis. The Asian Journal of Technology Management (AJTM), 16(3), 211-225. https://doi.org/10.12695/ajtm.2023.16.3.5

Kipkosgei, D., & Mackenzie, S. (2026). Performance Evaluation of Hybrid SVM- RF and XGBoost-RF Architectures for Classifying Gender-Based Violence Tweets on X. International Journal of Data Science, 28(5), 61-72.

Kirana, A. S., Roeswidiah, R., & Pudoli, A. (2025). Analisis Sentimen Pada Media Sosial Terhadap Layanan Samsat Digital Nasional. Jurnal Teknologi Dan Sistem Informasi, 8, 53-63.

JRIIN : Jurnal Riset Informatika dan Inovasi
Volume 4, No. 5 Tahun 2026
ISSN 3025-0919 (media online)
Hal 1164-1169

Lukman, K., & Novianto, S. (2025). Komparasi Algoritma Naïve Bayes dan SVM untuk Identifikasi Cyberbullying Selebriti di Media Sosial Twitter. Jurnal Algoritma, 22(1), 970-981. https://doi.org/10.33364/algoritma/v.22-1.2196

Mahmudah, S. A., & Yudhistira, A. (2025). Analisis Sentimen Terhadap Cyberbullying pada Platform Media Sosial X Menggunakan Algoritma Naive Bayes. Jurnal Pendidikan Dan Teknologi Indonesia, 5(1), 189-200. https://doi.org/10.52436/1.jpti.628

Murni, Santoso, R. D., Salsabila, A. A., Farhan, A., Arfin, T., Marsiano, J., & Rahmawati. (2026). Klasifikasi Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Algoritma Support Vector Machine (SVM) Berbasis Orange Data Mining. JRIIN: Jurnal Riset Informatika dan Inovasi, 4(2), 441-449.

Munna, A., & Zuliarso, E. (2024). Interpretation of Stacking Ensemble model for sentiment analysis of online loan application reviews using LIME. Aiti, 21(2), 183-196.

Pratama, F. A., Silviana, F., Karifki, M., Muges, M. W., Septiani, S., & Rahmawati. (2026). Klasifikasi Pelanggaran Etika Siber pada Media Sosial X Menggunakan Algoritma Extreme Gradient Boosting Berbasis TF-IDF. JRIIN: Jurnal Riset Informatika dan Inovasi, 4(2), 459-464.

Putu, I., Purnama Widiarta, A., Dwiyansaputra, R., & Aranta, A. (2023). Analisis Sentimen Masyarakat Terhadap Kebijakan Penerapan PPKM Di Media Sosial Twitter Dengan Menggunakan Metode XGBoost. Jurnal Teknologi Informasi, Komputer Dan Aplikasinya (JTIKA), 5(2), 154-163. http://jtika.if.unram.ac.id/index.php/JTIKA/