KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. 14, No. 1, April 2025, ISSN : 2089-9033, e- ISSN : 2715-7849
DOI: 10.34010/komputa.v14i1. 15595

Studi Pustaka: Optimalisasi Deteksi *Malware* melalui Integrasi Pembelajaran Mesin Heuristik dan *Big Data* untuk Keamanan Siber

Devi Tiana Octaviani Supriyadi[1], Bisyron Wahyudi[2], Danang Rimbawa[3]

[1,2,3] Program Program Studi Rekayasa Pertahanan Siber, Universitas Pertahanan
E-mail : devi.supriyadi@tp.idu.ac.id[1]

Abstrak

Ancaman *malware* yang semakin kompleks dan dinamis mendorong perlunya strategi deteksi yang lebih adaptif daripada metode konvensional berbasis tanda tangan. Penelitian ini bertujuan untuk mengevaluasi efektivitas pendekatan pembelajaran mesin, heuristik, dan *big data* dalam mendeteksi *malware* modern. Permasalahan utama yang diangkat adalah keterbatasan metode tradisional dalam mengidentifikasi *malware* varian baru, khususnya yang menggunakan teknik *obfuscation* seperti *polymorphism* dan *metamorphism*. Dengan menggunakan pendekatan studi pustaka sistematis terhadap literatur tahun 2016-2024 dari berbagai sumber bereputasi, penelitian ini membandingkan performa masing-masing pendekatan berdasarkan akurasi, efisiensi, dan ketahanan terhadap serangan manipulatif (*adversarial attacks*). Hasil analisis menunjukkan bahwa model *deep learning* seperti *Convolutional Neural Network* (CNN) memiliki akurasi deteksi tertinggi, sedangkan metode heuristik unggul dalam efisiensi deteksi awal, dan *big data* memberikan keunggulan dalam skalabilitas sistem deteksi secara *real-time*. Penelitian ini menyimpulkan bahwa integrasi ketiga pendekatan secara *hybrid* berpotensi menciptakan sistem deteksi *malware* yang lebih adaptif dan tangguh terhadap serangan siber, meskipun validasi empiris lanjutan masih diperlukan untuk implementasi di dunia nyata.
Kata kunci : Deteksi *Malware*, Pembelajaran Mesin, Heuristik, *Big Data*, Keamanan Siber.

Literature Study: Optimizing Malware Detection Through Integration of Heuristic Machine Learning and Big Data for Cybersecurity

Abstract

The increasingly complex and dynamic threat of malware drives the need for a more adaptive detection strategy than conventional signature-based methods. This study aims to evaluate the effectiveness of machine learning, heuristics, and big data approaches in detecting modern malware. The main problem raised is the limitation of traditional methods in identifying new malware variants, especially those that use obfuscation techniques such as polymorphism and metamorphism. Using a systematic literature study approach to the 2016-2024 literature from various reputable sources, this study compares the performance of each approach based on accuracy, efficiency, and resistance to adversarial attacks. The results of the analysis show that deep learning models such as the Convolutional Neural Network (CNN) have the highest detection accuracy, while heuristic methods excel in initial detection efficiency, and big data provides advantages in the scalability of real-time detection systems. This study concludes that the hybrid integration of these three approaches has the potential to create a malware detection system that is more adaptive and resilient to cyberattacks, although further empirical validation is still needed for real-world implementation.
Keywords : *Malware Detection, Machine Learning, Heuristics, Big Data, Cyber Security.*

1. Pendahuluan

Keamanan siber telah menjadi salah satu topik utama dalam teknologi informasi, terutama dengan meningkatnya ancaman *malware* yang semakin kompleks. *Malware* merupakan perangkat lunak berbahaya yang dirancang untuk menginfeksi sistem, mencuri data, atau mengganggu infrastruktur jaringan yang kritis

Naskah dikirim: 17-02-2025; direvisi 10-04-2025; diterima: 16-04-2025

[1], [2]. Berdasaran laporan Check Point dan Statista, pada tahun 2023 rata-rata organisasi menghadapi lebih dari 1.200 serangan *malware* per minggu, dengan total serangan global mencapai lebih dari 6 miliar kasus [3], [4].

Ancaman ini semakin diperparah oleh kemunculan teknik serangan yang canggih seperti *polymorphism* dan *metamorphism*, yang memungkinkan *malware* untuk mengubah struktur dan perilakunya secara dinamis, sehingga menyulitkan deteksi dengan metode tradisional berbasis tanda tangan digital [5], [6]. *Polymorphic malware* mengenkripsi dirinya sendiri secara acak setiap kali dijalankan, sementara *metamorphic malware* dapat memodifikasi seluruh kode internalnya sambil tetap mempertahankan fungsinya [7], [8].

Dalam menghadapi tantangan ini, pendekatan tradisional seperti *signature-based detection* terbuktu tidak memadai [9]. Oleh karena itu, teknologi berbasis kecerdasan buatan (AI) dan pembelajaran mesin (ML) mulai diadopsi secara luas karena kemampuannya dalam mengenali pola yang tidak terdeteksi oleh metode konvensional [10], [11]. Model seperti *Convolutional Neural Network* (CNN), *Support Vector Machine* (SVM), dan *Random Forest* digunakan dalam berbagai penelitian untuk mengklasifikasikan dan mengenali *malware* berdasarkan karakteristik *file* dan perilakunya.

Namun demikian, efektivitas metode tersebut masih tergantung pada ketersediaan *dataset* yang representatif. *Dataset* populer seperti *EMBER* dan *Drebin* sering digunakan, namun memiliki keterbatasan karena tidak selalu mencerminkan *malware* terbaru atau lingkungan serangan dunia nyata [12], [13]. Selain itu, tantangan lainnya adalah tinggi *false positive* dan kerentanan terhadap *adversarial attacks* yang dapat mengecoh model deteksi [14].

Seiring berkembangnya teknologi, pendekatan berbasis *big data* dan komputasi terdistribusi (seperti *Spark* dan *edge computing*) mulai dieksplorasi untuk mendeteksi anomali secara *real-time* di lingkungan berskala besar [15]. Meskipun menjanjikan, pendekatan ini menghadapi tantangan dari sisi biaya infrastruktur, kebutuhan akan tenaga ahli, serta integrasi data multiformat.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengevaluasi efektivitas berbagai pendekatan *malware* modern–terutama pembelajaran mesin, heuristik, dan *big data*–serta mengidentifikasi celah dan tantangan yang dihadapi. Kajian ini diharapkan memberikan kontribusi dalam pengembangan sistem keamanan siber yang lebih adaptif, akurat, dan berkelanjutan.

2. Metodologi

2.1 Metode Penelitian

Penelitian ini menggunakan pendekatan studi pustaka sistematis yang disusun berdasarkan prosedur eksplisit dan terukur. Meskipun tidak secara formal menggunakan model PRISMA, proses pencarian, seleksi dan klasifikasi literatur dilakukan mengikuti prinsip-prinsip sistematis yang serupa, meliputi penyaringan awal berdasarkan abstrak dan kata kunci, penilaian kualitas metodologi, serta klasifikasi berdasarkan pendekatan deteksi *malware*. Validasi dilakukan oleh tiga peneliti independen, memastikan reliabilitas hasil tinjauan literatur. Semua tahapan terdokumentasi dengan rinci untuk memudahkan replikasi oleh peneliti lain. Sumber data yang diperoleh dari jurnal yang telah terindeks dan memiliki *Internasional Standard Serial Number* (ISSN), yang terpublikasi melalui berbagai *database* akademik terkemuka seperti *Google Scholar*, *IEEE Xplore*, *SpringerLink*, *ScienceDirect*, MDPI, dan *arXiv*. Proses pengumpulan data dilakukan melalui pencarian daring menggunakan kombinasi kata kunci tertentu.

2.2 Objek Penelitian

Objek dalam penelitian ini adalah pendekatan deteksi *malware* yang menggunakan tiga teknologi utama, yaitu pembelajaran mesin (*machine learning*), heuristik, dan analisis data besar (*big data*). Ketiga pendekatan ini dipilih karena merupakan metode yang paling dominan dan berkembang pesat dalam literatur terkini terkait deteksi *malware*. Selain itu, masing-masing pendekatan menawarkan keunggulan yang saling melengkapi: pembelajaran mesin unggul dalam analisis pola kompleks dan otomatisasi deteksi, heuristik efektif dalam deteksi cepat berbasis aturan, sedangkan *big data* mendukung pemrosesan data dalam skala besar dan waktu nyata.

Penelitian ini memfokuskan kajian pada efektivitas dan efisiensi dari masing-masing pendekatan dengan menggunakan parameter-parameter yang relevan. Beberapa parameter utama yang diamati dalam literatur meliputi:

1. **Akurasi deteksi:** Diukur dalam bentuk persentase prediksi benar terhadap total data, seperti *True Positive Rate* (TPR) dan *Precision*.
2. ***False Positive Rate* (FPR):** Frekuensi terdeteksinya aktivitas normal sebagai ancaman.
3. **Waktu komputasi (*processing time*):** Digunakan untuk mengukur efisiensi eksekusi model.
4. **Kebutuhan sumber daya (*resource usage*):** Seperti penggunaan CPU, RAM, dan skalabilitas sistem.
5. ***Robustness* terhadap *adversarial attacks:*** Kemampuan model bertahan terhadap *input* yang dimanipulasi untuk mengecoh sistem deteksi.

Setiap artikel dalam studi pustaka ini diklasifikasikan dan dianalisis berdasarkan parameter-parameter tersebut guna mendapatkan gambaran menyeluruh mengenai kelebihan dan keterbatasan masing-masing pendekatan dalam konteks deteksi *malware* modern.

2.3 Langkah-Langkah Pengumpulan Data

1. Kriteria Seleksi Artikel
   a. Artikel yang dipilih berasal dari jurnal bereputasi.
   b. Kata kunci pencarian meliputi "*malware analysis*".
   c. Artikel yang dipilih adalah yang diterbitkan dalam tujuh tahun terakhir (2016-2024) untuk memastikan relevansi dengan tren terkini.
2. Prosedur Pencarian dan Penyaringan
   a. Database yang digunakan meliputi *Google Scholar*, *IEEE Xplore*, *SpingerLink*, *ScienceDirect*, MDPI, dan *arXiv*.
   b. Penyaringan artikel dilakukan berdasarkan abstrak, relevansi topik, dan kualitas metodologi. Artikel dengan hasil yang kurang signifikan dieliminasi.
3. Proses Klasifikasi Literatur
   a. Artikel yang relevan diklasifikasikan berdasarkan metode yang digunakan dalam deteksi *malware*, seperti:
      (a) **Pembelajaran Mesin (*Machine Learning*):** Mencakup *supervised learnig*, *deep learning*, dan metode hybrid.
      (b) **Heuristik dan Statistik:** Pendekatan berbasis aturan (*rule-based*) atau analisis pola statistik.
      (c) **Big Data:** Teknologi seperti *Apache Hadoop* dan *Spark* untuk pengolahan data besar.
4. Validasi Data
   a. Semua artikel yang terpilih dievaluasi oleh tiga peneliti independen untuk memastikan validitas hasil tinjauan.
   b. Diskusi dilakukan untuk menyepakati analisis hasil, termasuk identifikasi kekuatan, kelemahan, dan kontribusi penelitian.

2.4 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini menggunakan studi pustaka (*library research*), yaitu metode pengumpulan informasi dari berbagai sumber tertulis, berupa jurnal ilmiah. Semua prosedur pencarian, penyaringan, dan klasifikasi didokumentasikan dengan rinci, termasuk kata kunci, database, dan parameter pencarian, untuk memastikan metodologi ini dapat diterapkan kembali oleh peneliti lain dalam konteks serupa.

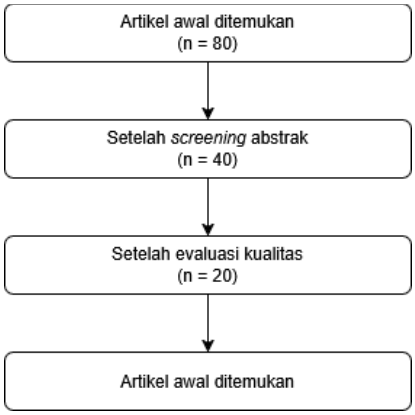
2.5 Analisis Data

Data dari artikel yang terpilih disusun dalam bentuk tabel dan grafik untuk mempermudah analisis tren dan perbandingan hasil penelitian. Visualisasi ini mencakup perbandingan tingkat akurasi, efisiensi waktu komputasi, serta kelebihan dan kekurangan dari masing-masing pendekatan (*machine learning*, heuristik, dan *big data*).

Untuk memastikan transparansi dan replikasi dalam proses peninjauan literatur, penelitian ini menyusun tabel ringkas dan *flowchart* yang menggambarkan jumlah artikel pada tiap tahap seleksi. Proses ini melibatkan identifikasi awal artikel dari *database* akademik, penyaringan berdasarkan abstrak dan kesesuaian topik, evaluasi kualitas metodologi, hingga penentuan artikel yang layak dianalisis lebih lanjut. Berikut adalah ringkasan tahapan seleksi artikel:

Tabel 1. Tabel Seleksi Artikel

Tahap Seleksi	Jumlah Artikel
Artikel awal ditemukan	80
Setelah *screening* abstrak	40
Setelah evaluasi kualitas	20
Artikel akhir dianalisis	9



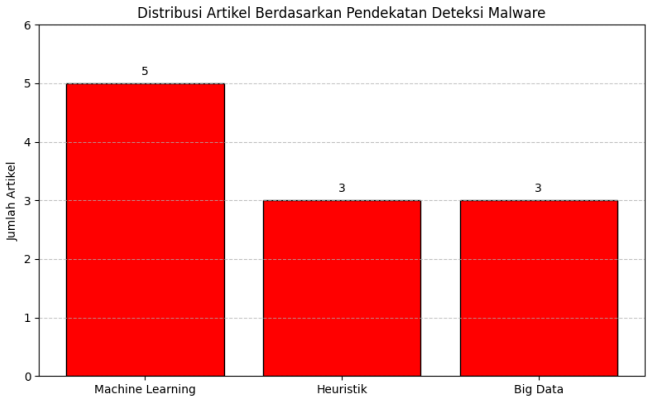
Gambar 1. *Flowchart* Seleksi Artikel

Flowchart seleksi artikel juga disusun secara visual untuk memperjelas alur penyaringan dan eliminasi artikel. Meskipun penelitian ini tidak secara eksplisit model PRISMA, alur seleksi disusun berdasarkan prinsip-prinsip sistematis yang serupa. Langkah-langkah ini dilakukan untuk memastikan bahwa analisis dilakukan secara objektif, transparan, dan dapat direplikasi oleh peneliti lain dalam kajian serupa.

3. Hasil Dan Pembahasan

Bab ini menyajikan hasil analisis dari 9 artikel terpilih yang dikelompokkan berdasarkan tiga pendekatan utama dalam deteksi *malware*, yaitu *machine learning*, heuristik, dan *big data*. Pembagian ini mencerminkan fokus literatur yang ditemukan, dimana:

a. ***Machine Learning*** menjadi pendekatan paling dominan, dianalisis pada 5 dari 9 artikel (56%),
b. **Heuristik dan statistik** dibahas pada 3 artikel (33%), dan
c. ***Big Data*** menjadi fokus pada 3 artikel (33%), dengan beberapa tumpang tindih antar pendekatan.



Gambar 2. Distribusi Artikel Berdasarkan Pendekatan Deteksi *Malware*

3.1 Pendekatan Berbasis Pembelajaran Mesin

Pendekatan berbasis pembelajaran mesin (*machine learning*) menjadi yang paling dominan dalam literatur yang dianalisis (56%), menunjukkan tren kuat ke arah otomatisasi dan deteksi berbasis pola.

Model-model seperti *Support Vector Machine* (SVM), *Random Forest* (RF), dan *Convolutional Neural Networks* (CNN) digunakan secara luas untuk menganalisis baik fitur statis (struktur *file*) maupun fitur dinamis (perilaku *runtime*) [16], [17].

Pendekatan *machine learning* menunjukkan performa yang unggul terutama dalam mendeteksi *malware* dengan pola kompleks. Model CNN yang diuji pada *dataset KaggleMalware* mencapai akurasi hingga 95%, lebih tinggi dibandingkan SVM pada *MalwareX* (93,5%) dan *Random Forest* pada *Ember* (88%) [16]. Tabel 2 berikut merangkum perbandingan hasil uji model pada berbagai *dataset*:

Tabel 2. Perbandingan Pendekatan *Machine Learning*

Jurnal	Metode	Dataset	Akurasi (%)	Kelebihan	Kekurangan
Analysis of Malware Detection Using Various Machine Learning Approach [18]	SVM	*MalwareX*	93.5	Interpretasi jelas, cocok untuk data kecil	Kurang efisien untuk data besar
Dynamic Malware Analysis Without Feature Engineering [16]	*Random Forest*	*Ember*	88	Andal pada fitur campuran	Memerlukan *tuning* parameter
Malware Analysis in IoT & Android Systems with Defensive Mechanism [19]	CNN	*KaggleMalware*	95	Deteksi otomatis pola kompleks	Rentan terhadap *adversarial attacks*

a. Karakteristik *dataset*

(a) ***MalwareX*:** *Dataset benchmark* berskala menengah yang mencakup 10.000+ sampel *malware* dan *non-malware*, cocok untuk *supervised learning*.

(b) ***Ember*:** *Dataset* terbuka dengan representasi fitur statis dari *executable Windows* (*.exe*), banyak digunakan dalam kompetisi ML di bidang keamanan siber.

(c) ***KaggleMalware*:** *Dataset* dari kompetisi *Kaggle*, terdiri dari lebih dari 20.000 *binary file* yang direpresentasikan sebagai urutan *byte*, cocok untuk CNN karena memungkinkan pemrosesan berbasis gambar.

Dari perbandingan di atas, CNN menunjukkan kinerja tertinggi pada *dataset* besar dengan struktur sekuensial, sedangkan SVM unggul dalam kejelasan interpretasi namun tidak optimal untuk data skala besar. *Random Forest* menempati posisi tengah – fleksibel dan stabil, tetapi perlu penyesuaian parameter secara manual [16].

Dalam sebagian literatur, terdapat studi yang membandingkan langsung performa CNN, SVM, dan *Random Forest* pada *dataset* yang sama (misalnya, *MalwareX*). Hasilnya menunjukkan:

(a) CNN unggul dalam akurasi, terutama saat fitur kompleks tersedia.

(b) SVM unggul dalam interpretabilitas dan kecepatan pelatihan pada data yang tidak terlalu besar.

(c) *Random Forest* lebih seimbang, tetapi performanya sensitif terhadap jumlah pohon dan kedalaman.

b. *Adversarial Attacks* dan Penanggulangan

Salah satu tantangan utama pendekatan ini adalah kerentanannya terhadap *adversarial attacks*, yaitu manipulasi *input* (misalnya *byte sequence*) untuk mengecoh model deteksi. Salah satu metode teknis yang banyak digunakan untuk menanggulangi ini adalah:

(a) ***Adversarial Training*****:** Melatih model dengan kombinasi data asli dan data yang telah dimodifikasi secara *adversarial*, agar model lebih tahan terhadap *input* manipulatif.
(b) ***Feature Squeezing*****:** Mengurangi dimensi *input* untuk membatasi kemungkinan variasi manipulatif.
(c) ***Ensemble Voting*****:** Menggunakan lebih dari satu model dan menggabungkan hasilnya, untuk mengurangi dampak manipulasi terhadap satu model tunggal.

Dengan mempertimbangkan konteks *dataset* dan teknik penguatan terhadap serangan, pendekatan *machine learning* terus berkembang sebagai tulang punggung deteksi *malware* modern yang lebih adaptif dan akurat.

3.2 Pendekatan Heuristik dan Statistik

Pendekatan heuristik dan statistik tetap menjadi pilar penting dalam sistem deteksi *malware*, terutama di lingkungan yang memerlukan respon cepat dan sumber daya terbatas. Heuristik umumnya bekerja dengan mengenali pola-pola perilaku mencurigakan berdasarkan aturan yang telah ditentukan sebelumnya (*rule-based*), seperti akses *file* sistem yang tidak biasa, perubahan *registry* mendadak, atau aktivitas jaringan anomali [20].

Sementara itu, pendekatan statistik mengandalkan distribusi numerik dan hubungan antar variabel untuk mendeteksi *outliner* atau anomali. Misalnya, analisis probabilistik terhadap frekuensi paket data atau penggunaan regresi linier untuk mendeteksi tren serangan secara pasif [21].

Namun, kedua pendekatan ini memiliki keterbatasan yang signifikan:

a. Heuristik sering gagal mengenali *malware polymorphic* dan *metamorphic*, yang terus-menerus berubah untuk menghindari deteksi berbasis pola.
b. Analisis statistik rentan terhadap *false positives*, terutama saat aktivitas legal menyerupai pola anomali.

Untuk menjawab keterbatasan tersebut, beberapa studi mengusulkan pendekatan *hybrid* yang menggabungkan heuristik dengan pembelajaran mesin (*supervised learning*). Dalam sistem ini, heuristik berperan sebagai penyaring awal, sementara model pembelajaran mesin seperti *Random Forest* digunakan untuk klasifikasi akhir berdasarkan *dataset* yang telah dilabeli.

A. Contoh Studi Hybrid

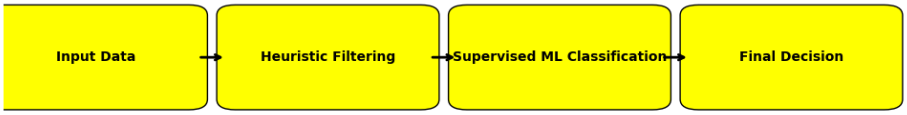
Yadav et al. (2022) menunjukkan bahwa integrasi *rule-based detection* dengan *Random Forest* pada sistem *Android* dan IoT meningkatkan akurasi dari sekitar 85% (heuristik saja) menjadi 94,5% serta menurunkan *false positive rate* hingga 18% [22].

Selain itu, kombinasi ini memungkinkan interpretasi lebih baik terhadap hasil klasifikasi karena keputusan berbasis aturan masih digunakan dalam proses awal, sebelum dilakukan pembelajaran pola yang lebih kompleks melalui algoritma ML.

Tabel 3. Perbandingan Pendekatan Heuristik dan Statistik

Jurnal	Metode	Dataset	Akurasi (%)	Kelebihan	Kelemahan
A Survey on Malware Detection and Analysis [20]	Heuristik	Statistik jaringan	~82	Cepat, efisien, tidak memerlukan pelatihan	Tinggi *false positive*
Malware Detection and Analysis Using Reverse Engineering [21]	Statistik	*Log* keamanan	~85	Mudah diimplementasikan, cocok untuk *baseline*	Tidak efektif terhadap *malware* kompleks
Malware Analysis in IoT & Android Systems with Defensive Mechanism [23]	IoT & *Android*	IoT & *Android*	94,5	Akurasi tinggi, kombinasi deteksi + klasifikasi	Membutuhkan model dan data pelatihan memadai

Gambar 3 di bawah ini memperjelas tahapan dalam sistem deteksi *hybrid*, dimana heuristik bertindak sebagai *filter* awal dan *machine learning* berperan dalam klasifikasi lanjutan. Alur ini menggambarkan sinergi antara pengetahuan *domain* berbasis aturan dan kekuatan pembelajaran dari data historis.



Gambar 3. Alur Kerja Sistem Deteksi *Hybrid*

Berikut adalah visual alur kerja sistem deteksi *hybrid* yang menggabungkan pendekatan heuristik dan *machine learning*:

1. ***Input* Data:** Data aktivitas sistem, *log*, atau trafik jaringan masuk ke sistem.
2. ***Heuristic Filtering*:** Aturan berbasis pola digunakan untuk menyaring aktivitas mencurigakan secara awal.
3. ***Supervised ML Classification*:** Aktivitas yang lolos disaring lebih lanjut menggunakan model *supervised learning* seperti *Random Forest* atau SVM.
4. ***Final Decision*:** Sistem mengklasifikasi apakah aktivitas tersebut termasuk *malware* atau bukan.

Dengan demikian, pendekatan heuristik dan statistik tetap relevan, terutama bila dikombinasikan dengan pembelajaran mesin. Model *hybrid* dapat menjadi jembatan antara kebutuhan *real-time*, interpretabilitas tinggi, dan akurasi berbasis data historis, menjadikannya pilihan yang sangat potensial untuk sistem deteksi *malware* yang adaptif dan efisien.

3.3 Analisis Data Besar (*Big Data*)

Pendekatan berbasis *big data* menjadi salah satu solusi yang semakin relevan dalam menghadapi deteksi *malware* berskala besar. Teknologi seperti *Apache Hadoop*, *Apache Spark*, dan *cloud computing* digunakan untuk memproses *volume* data yang sangat besar secara paralel dan *real-time*, mencakup data dari *log* keamanan, trafik jaringan, hingga *file executable [24]*.

Keunggulan utama pendekatan ini adalah skalabilitas dan kemampuannya untuk mendeteksi anomali dalam sistem kompleks yang terdiri dari ribuan *node* dan perangkat. Misalnya, *Spark* memungkinkan pemrosesan *real-time* pada data *streaming* dari *firewall* atau *endpoint security system*, yang sangat krusial untuk mencegah penyebaran *malware* secara masif [24].

Namun, potensi besar tersebut tidak lepas dari tantangan yang cukup signifikasn. Salah satunya adalah biaya infrastruktur dan sumber daya manusia. Penerapan sistem *big data* yang efektif memerlukan:

a. *Hardware* khusus untuk penyimpanan terdistribusi (*cluster Hadoop*, HDFS).
b. Koneksi jaringan berkecepatan tinggi dan reliabel.
c. *Engineer* & *data scientist* yang paham teknologi seperti *MapReduce*, *Spark*, dan *pipeline data*.

A. Diskusi *Cost-Benefit*

Meskipun investasi awal cukup besar, beberapa studi menunjukkan bahwa *return on investment* (ROI) dari sistem *big data* dalam deteksi *malware* bisa signifikan dalam jangka menengah hingga panjang, khususnya pada organisasi yang:

a. Mengelola infrastruktur TI berskala besar (perbankan, telko, *cloud provider*).
b. Memerlukan deteksi anomali secara *real-time* tanpa toleransi *downtime*.
c. Memiliki tim keamanan siber yang mampu mengelola data terdistribusi.

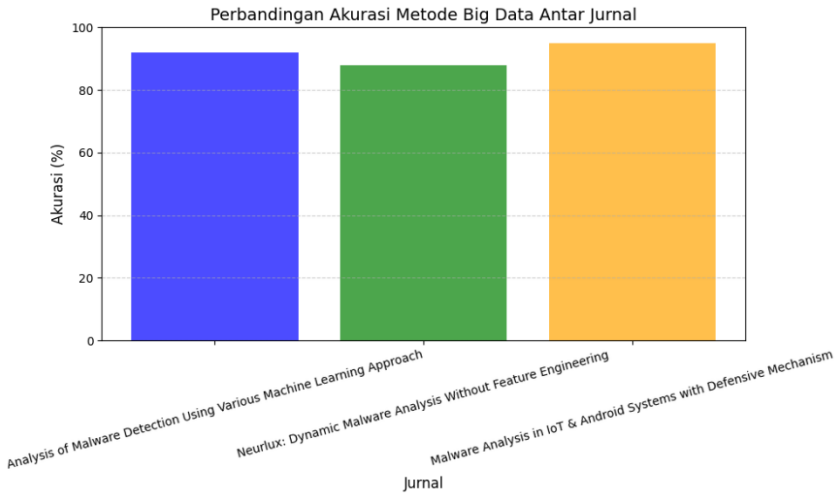
Sebaliknya, untuk organisasi kecil-menengah, penggunaan *big data* kadang tidak efisien, karena *volume data* yang terbatas dan kebutuhan respons yang tidak selalu *real-time*. Oleh karena itu, analisis *cost-benefit* harus disesuaikan dengan kebutuhan spesifik organisasi.

Tabel 4. Perbandingan Pendekatan *Big Data*

Aspek	Organisasi Skala Kecil (SME)	Organisasi Skala Besar (*Enterprise*)
Volume Data	Rendah hingga sedang	Sangat tinggi
Kebutuhan *Real-Time*	Tidak kritis	Sangat penting
Biaya Infrastruktur	Tinggi terhadap manfaat	Proporsional terhadap kebutuhan
Ketersediaan Tim Ahli	Terbatas	Tersedia (*dedicated team*)
Return on Investment	Rendah hingga sedang	Tinggi (jangka menengah-panjang)
Cocok untu	*Batch monitoring*, evaluasi berkala	*Monitoring real-time*, *auto-detection*

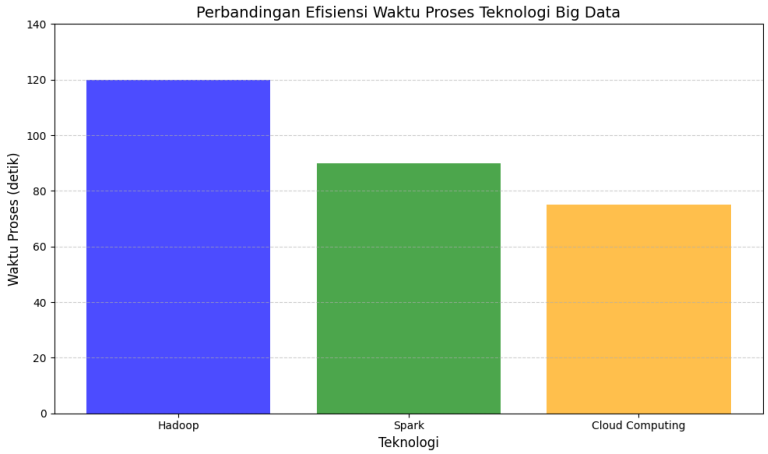
Tabel diatas menunjukkan bahwa meskipun *big data* memiliki potensi besar, penerapannya harus mempertimbangkan skala organisasi, kompleksitas sistem, dan kesiapan sumber daya. Untuk organisasi kecil, solusi yang lebih ringan atau *hybrid cloud-local* mungkin lebih efektif dibanding infrastruktur *big data* penuh.

Oleh karena itu, analisis *cost-benefit* harus disesuaikan dengan kebutuhan spesifik organisasi. *Big data* efektif jika digunakan untuk memantau infrastruktur besar secara berkelanjutan dan dinamis, tetapi bukan solusi terbaik untuk semua skenario.



Gambar 4. Perbandingan Akurasi Metode *Big Data* Antar Jurnal

Grafik diatas menunjukkan perbandingan akurasi metode *Bog Data* yang digunakan pada jurnal-jurnal yang dianalisis.



Gambar 5. Perbandingan Efisiensi Waktu Proses Teknologi *Big Data*

Grafik diatas menunjukkan perbandingan efisiensi waktu proses dari teknologi *Big Data* seperti *Hadoop*, *Spark*, dan *Cloud Computing*. *Spark* dan *Cloud* menunjukkan efisiensi waktu yang lebih baik dibandingkan *Hadoop*.

3.4 Tren dan Celah Penelitian

Penelitian deteksi *malware* terus berkembang seiring meningkatnya kompleksitas dan dinamika ancaman siber. Beberapa tren utama yang muncul dari analisis literatur antara lain:

a. Peningkatan penggunaan *deep learning*, khususnya CNN dan LSTM, dalam klasifikasi *file* eksekusi.
b. Eksplorasi teknologi graf, seperti *Graph Neural Network* (GNN), untuk mendeteksi hubungan tersembunyi antar entitas.
c. Pemanfaatan *cloud* dan *edge computing* untuk mendukung analisis data besar secara efisien.

Namun, di balik perkembangan tersebut, masih terdapat sejumlah celah yang menantang:

A. Kualitas dan Representasi *Dataset*

Ketersediaan *dataset* yang reprentatif masih menjadi salah satu hambatan utama. Beberapa *dataset* yang sering digunakan antara lain:

a. **EMBER:** *Dataset* statis untuk *executable Windows*. Sangat populer dan terbuka, namun sebagian besar berisi sampel dari lingkungan tertentu dan bisa kurang representatif terhadap *malware* muktakhir.
b. **Drebin:** *Dataset Android* yang digunakan untuk klasifikasi *malware* berbasis fitur dari APK. Meskipun lengkap, *dataset* ini kurang mencakup *malware Android* terbaru pasca-2020.
c. **Malimg, ClCMalDroid, dan *VirusShare*** juga sering digunakan namun memiliki keterbatasan akses atau dokumentasi.

Menurut studi oleh Vinayakumar et al. (2019), sebagian besar model deteksi yang dilatih pada *dataset* tunggal memiliki tingkat generalisasi yang buruk saat diuji pada *dataset* lain, menunjukkan perlunya pendekatan lintas *dataset* atau augmentasi data.

B. Tabel Ringkas: Tren vs. Celah dan Potensi Aksi

Berikut adalah pembaruan tabel sebelumnya dengan penambahan kolom "Rekomendasi Aksi":

Tabel 5. Tabel Tren vs. Celah dan Potensi Aksi

Aspek Penelitian	Tren	Celah Penelitian	Rekomendasi Aksi
Teknologi	Adopsi *Graph Neural Networks*, *deep learning* (CNN, LSTM)	Kurangnya penerapan model *explainable* AI pada deteksi *malware*	Kembangkan model yang akurat dan dapat dijelaskan (*interpretable* DL)
Dataset	Penggunaan EMBER, Drebin, *KaggleMalware*, *MalwareX*	*Dataset* tidak selalu mencerminkan variasi *malware* terbaru dari realitis	Buat *benchmark dataset* terbuka yang diperbarui dan *multiplatform*
Infrastruktur	*Cloud* dan *edge computing* mendukung analisis data besar	Biaya dan kompleksitas integrasi sistem tinggi	Eksplorasi arsitektur *hybrid* (*cloud-edge* lokal) yang hemat biaya
Evaluasi	Penelitian lebih sering fokus pada akurasi model	Kurangnya standar evaluasi menyeluruh (mis. Waktu proses, konsumsi *resource*)	Bangun kerangka evaluasi standar lintas metode dan *platform*
Kolaborasi	Meningkatnya kerjasama akademik-industri	Minimnya komunikasi lintas disiplin dan sektor	Dorong ekosistem kolaboratif terbuka berbasis *platform* komunitas

4. Kesimpulan

Penelitian ini mengevaluasi efektivitas pendekatan pembelajaran mesin, heuristik, dan *big data* dalam mendeteksi *malware* modern yang semakin kompleks. Berdasarkan hasil studi pustaka, pendekatan berbasis pembelajaran mesin–terutama *deep learning* seperti *Convolutional Neural Network* (CNN)–menunjukkan akurasi deteksi yang tinggi terhadap berbagai pola *malware*. Sementara itu, metode heuristik tetap relevan dalam konteks deteksi awal yang cepat, dan *big data* memberikan keunggulan dalam skalabilitas serta kemampuan pemantauan *real-time* di infrastruktur berskala besar. Dengan demikian, integrasi ketiga pendekatan ini menawarkan solusi yang lebih komprehensif dan adaptif dalam menghadapi ancaman siber yang terus berkembang.

Temuan ini secara langsung menjawab tujuan utama penelitian, yaitu menilai keunggulan pendekatan pembelajaran mesin dibandingkan metode heuristik dalam konteks deteksi *malware*. Hasil analisis menunjukkan bahwa pembelajaran mesin memang unggul dalam hal akurasi dan ketahanan terhadap kompleksitas *malware* baru, meskipun masih menghadapi tantangan seperti tingginya tingkat *false positive* dan kerentanan terhadap serangan manipulatif (*adversarial attacks*). Oleh karena itu, kombinasi pendekatan berbasis aturan dan pembelajaran mesin dalam model *hybrid* menjadi strategi yang menjanjikan untuk menyeimbangkan antara akurasi, efisiensi, dan interpretabilitas.

Dari sisi praktis, hasil penelitian ini memberikan implikasi langsung bagi berbagai pemangku penting. Penyedia solusi keamanan siber dapat mengadopsi model deteksi *hybrid* untuk meningkatkan efektivitas sistem pertahanan, sementara tim keamanan TI di perusahaan dapat memilih pendekatan yang sesuai dengan kapasitas sumber daya dan kebutuhan *real-time*. Lembaga riset dan akademisi juga dapat menggunakan temuan ini sebagai dasar pengembangan metode baru yang lebih tahan terhadap manipulasi dan lebih efisien secara komputasi.

Dalam konteks kebijakan, hsi ini menggarisbawahi pentingnya dukungan terhadap intregrasi teknologi cerdas dalam sistem deteksi ancaman, termasuk melalui penyusunan regulasi yang mendorong kolaborasi akademik-industri dan pengembangan infrastruktur data terbuka. Pemerintah dan pembuat kebijakan perlu memperkuat kerangka kerja keamanan siber dengan memfasilitasi adopsi teknologi berbasis AI dan *big data* yang telah terbukti lebih adaptif terhadap pola ancaman baru.

Namun demikian, penelitian ini masih terbatas pada studi pustaka tanpa pengujian empiris langsung. Oleh karena itu, arah penelitian selanjutnya perlu mencakup validasi eksperimental di lingkungan riil, pengembangan *dataset benchmark* yang representatif dan mutakhir, serta penerapan pendekatan *Explainable* AI agar hasil deteksi lebih dapat ditafsirkan oleh pengguna manusia. Selain itu, eksplorasi arsitektur *hybrid* seperti *edge-cloud computing* dan studi *cost-benefit* jangka panjang juga menjadi agenda penting untuk memastikan bahwa adopsi teknologi deteksi canggih dapat disesuaikan dengan skala dan kebutuhan tiap organisasi.

Daftar Pustaka

[1] “Malicious programs | Kaspersky IT Encyclopedia.” Accessed: Apr. 10, 2025. [Online]. Available: https://encyclopedia.kaspersky.com/knowledge/malicious-programs/
[2] “28th August – Threat Intelligence Report - Check Point Research.” Accessed: Apr. 10, 2025. [Online]. Available: https://research.checkpoint.com/2023/28th-august-threat-intelligence-report/
[3] “Surge in Cybercrime: Check Point 2023 Mid-Year Security Report Reveals 48 ransomware groups have breached over 2,200 victims - Check Point Software.” Accessed: Apr. 10, 2025. [Online]. Available: https://www.checkpoint.com/press-releases/surge-in-cybercrime-check-point-2023-mid-year-security-report-reveals-8-spike-in-global-cyberattacks/
[4] “Number of malware attacks per year 2023 | Statista.” Accessed: Apr. 10, 2025. [Online]. Available: https://www.statista.com/statistics/873097/malware-attacks-per-year-worldwide/
[5] “Polymorphic Malware Protection Best Practices - Identity Management Institute®.” Accessed: Apr. 10, 2025. [Online]. Available: https://identitymanagementinstitute.org/polymorphic-malware-protection-best-practices/
[6] “What is the Polymorphic Virus?” Accessed: Apr. 10, 2025. [Online]. Available: https://www.kaspersky.com/resource-center/definitions/what-is-a-polymorphic-virus
[7] “What is Metamorphic Virus? | Metamorphic Virus Definition.” Accessed: Apr. 10, 2025. [Online]. Available: https://www.kaspersky.com/resource-center/definitions/metamorphic-virus

[8] "Cloud Atlas APT upgrades its arsenal with polymorphic malware." Accessed: Apr. 10, 2025. [Online]. Available: https://www.kaspersky.com/about/press-releases/cloud-atlas-apt-upgrades-its-arsenal-with-polymorphic-malware
[9] "What Is Signature-based Malware Detection? | RiskXchange." Accessed: Apr. 10, 2025. [Online]. Available: https://riskxchange.co/1006984/what-is-signature-based-malware-detection/
[10] J. Ferdous, R. Islam, A. Mahboubi, and M. Z. Islam, "A Survey on ML Techniques for Multi-Platform Malware Detection: Securing PC, Mobile Devices, IoT, and Cloud Environments," *Sensors 2025, Vol. 25, Page 1153*, vol. 25, no. 4, p. 1153, Feb. 2025, doi: 10.3390/S25041153.
[11] T. Mane, P. Nimase, P. Parihar, and P. Chandankhede, "Review of Malware Detection Using Deep Learning," pp. 255–262, 2022, doi: 10.1007/978-981-16-5301-8_19.
[12] "Endgame Malware BEnchmark for Research (EMBER) Dataset – CyberCitadelLabs." Accessed: Apr. 10, 2025. [Online]. Available: https://www.cybercitadellabs.com/2022/03/10/endgame-malware-benchmark-for-research-ember-dataset/
[13] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and Explainable Detection of Android Malware in Your Pocket," May 2014, doi: 10.14722/NDSS.2014.23247.
[14] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," *ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, pp. 506–519, Feb. 2016, doi: 10.1145/3052973.3053009.
[15] "A Continuing Cyber-Storm with Increasing Ransomware Threats - Check Point Blog." Accessed: Apr. 10, 2025. [Online]. Available: https://blog.checkpoint.com/security/a-continuing-cyber-storm-with-increasing-ransomware-threats-and-a-surge-in-healthcare-and-apac-region/
[16] Vishal Borate, Dr. Alpana Adsul, Aditya Gaikwad, Akash Mhetre, and Siddhesh Dicholkar, "Analysis of Malware Detection Using Various Machine Learning Approach," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 314–321, Nov. 2024, doi: 10.48175/IJARSCT-22159.
[17] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," *IEEE Access*, vol. 7, pp. 46717–46738, 2019, doi: 10.1109/ACCESS.2019.2906934.
[18] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity," *ACM Comput Surv*, vol. 55, no. 8, Aug. 2022, doi: 10.1145/3547330/ASSET/A0E83E55-05EB-4519-BC6A-EB5E6A7BDBAA/ASSETS/GRAPHIC/CSUR-2021-0664-F04.JPG.
[19] C. Jindal, C. Salls, H. Aghakhani, K. Long, C. Kruegel, and G. Vigna, "Neurlux: Dynamic malware analysis without feature engineering," *ACM International Conference Proceeding Series*, pp. 444–455, 2019, doi: 10.1145/3359789.3359835.
[20] J. Smallman, "A Survey on Malware Detection and Analysis," *Journal of Science & Technology*, vol. 5, no. 4, pp. 1–14, 2024, doi: 10.55662/jst.2024.5401.
[21] Muhammad Taseer Suleman, "Malware Detection and Analysis Using Reverse Engineering," *International Journal for Electronic Crime Investigation*, vol. 8, no. 1, pp. 109–123, 2024, doi: 10.54692/ijeci.2024.0801191.
[22] C. S. Yadav *et al.*, "Malware Analysis in IoT & Android Systems with Defensive Mechanism," *Electronics (Switzerland)*, vol. 11, no. 15, 2022, doi: 10.3390/electronics11152354.
[23] A.-R. Belea, "Methods for Detecting Malware Using Static, Dynamic and Hybrid Analysis," vol. X, pp. 1–8, 2023.
[24] K. David *et al.*, "Real-Time Cybersecurity threat detection using machine learning and big data analytics: A comprehensive approach," *Computer Science & IT Research Journal*, vol. 4, no. 3, pp. 478–501, Dec. 2023, doi: 10.51594/CSITRJ.V4I3.1500.