

Deteksi Serangan Siber Berbasis *Machine Learning*: *Systematic Literature Review* terhadap Kinerja Algoritma, Tantangan Implementasi, dan Arah Penelitian

Wahyudi¹

¹Informatika, Universitas Horizon Indonesia, Karawang, Indonesia

Email: wahyudi008@gmail.com

Abstrak – Laju eskalasi ancaman siber yang melampaui kapasitas sistem deteksi berbasis aturan statis mendorong kebutuhan mendesak akan pendekatan deteksi yang adaptif dan berbasis data. Kajian ini bertujuan memetakan kinerja algoritma *machine learning* (ML) dalam mendeteksi serangan siber, mengidentifikasi tantangan implementasi yang berulang, serta mengevaluasi tingkat kematangan tren penelitian masa depan. Metode yang digunakan adalah *Systematic Literature Review* (SLR) mengacu pada panduan Carrera-Rivera et al. [6], dengan penelusuran sistematis pada enam basis data akademik: IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, Google Scholar, dan Portal Garuda. Dari 156 artikel yang teridentifikasi, 19 studi ditetapkan sebagai korpus final melalui prosedur PRISMA terdiri atas 9 studi primer (2020–2025) dan 10 studi landasan (sebelum 2020), setelah melewati penilaian kualitas dua dimensi: *rigor* metodologis dan relevansi industri. Hasil analisis menunjukkan bahwa CNN dan LSTM mencapai akurasi deteksi 97–99%, mengungguli *Random Forest* (95–98%) dan SVM (90–97%) dalam skenario klasifikasi kompleks. Empat tantangan implementasi yang paling konsisten dijumpai adalah ketidakseimbangan data latih, kendala latensi *real-time*, serangan *adversarial*, dan opasitas model *deep learning*. Dataset CICIDS2017 dan UNSW-NB15 semakin menggantikan posisi NSL-KDD sebagai standar evaluasi karena lebih merepresentasikan kondisi jaringan modern. Kajian ini menyimpulkan bahwa *federated learning* dan *explainable AI* (XAI) merupakan dua arah penelitian yang paling matang untuk mengatasi hambatan privasi data dan opasitas model. Kontribusi utama kajian ini terletak pada pemisahan eksplisit antara studi empiris terkini dan studi landasan historis dalam satu kerangka SLR terpadu, menghasilkan peta kinerja algoritma yang dapat langsung dijadikan acuan bagi praktisi keamanan dalam menentukan pendekatan ML yang sesuai dengan kapasitas infrastruktur organisasi.

Kata Kunci: Intrusion Detection System, Adversarial Machine Learning, Federated Learning, Explainable AI, PRISMA

Abstract – The escalating pace of cyber threats has outstripped the capacity of static rule-based detection systems, creating an urgent demand for adaptive, data-driven detection approaches. This study aims to map the performance of machine learning (ML) algorithms in detecting cyberattacks, identify recurring implementation challenges, and evaluate the maturity of emerging research directions. A Systematic Literature Review (SLR) was conducted in accordance with Carrera-Rivera et al. [6] guidelines for computer science research, with systematic searches across six academic databases: IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, Google Scholar, and Portal Garuda. From 156 identified articles, 19 studies were established as the final corpus through PRISMA procedures comprising nine primary studies (2020–2025) and ten foundational studies (pre-2020), following two-dimensional quality assessment covering methodological rigor and industry relevance. Results show that CNN and LSTM achieve detection accuracy of 97–99%, outperforming Random Forest (95–98%) and SVM (90–97%) in complex classification scenarios. Four implementation challenges recur most consistently: training data class imbalance, real-time latency constraints, adversarial attacks, and deep learning model opacity. CICIDS2017 and UNSW-NB15 are progressively replacing NSL-KDD as evaluation benchmarks due to superior representation of modern network conditions. This review concludes that federated learning and explainable AI (XAI) represent the most mature research directions for overcoming data privacy barriers and model opacity. The primary contribution of this study lies in the explicit temporal separation of recent empirical studies from foundational historical work within a unified SLR framework, yielding an algorithm performance map directly applicable by security practitioners in selecting ML approaches commensurate with organizational infrastructure capacity.

Keywords: Intrusion Detection System, Adversarial Machine Learning, Federated Learning, Explainable AI, PRISMA Review

1. PENDAHULUAN

Transformasi digital di berbagai sektor strategis telah memperluas permukaan serangan terhadap infrastruktur jaringan. Setiap perangkat terhubung, layanan cloud, dan proses digital membuka vektor eksploitasi baru. Laporan ENISA (2024) mencatat peningkatan signifikan volume dan kompleksitas ancaman

siber global, dengan proyeksi kerugian mencapai 10,5 triliun dolar AS pada tahun 2025 [1]. Angka ini melampaui PDB sebagian besar negara berkembang, menandakan ancaman siber telah menjadi risiko sistemik yang berdampak pada stabilitas ekonomi dan keamanan nasional.

Kompleksitas ancaman saat ini berbeda secara kualitatif dari serangan generasi sebelumnya. Advanced Persistent Threats (APT) beroperasi secara laten dalam waktu panjang dan memanfaatkan kerentanan yang belum dipublikasikan. Zero-day exploits menysasar kelemahan perangkat lunak sebelum patch tersedia. Polymorphic malware memodifikasi kodenya secara otomatis untuk mengelabui deteksi statis [2]. Ketiga ancaman ini mengekspos batasan fundamental sistem berbasis aturan yang hanya mengenali pola ancaman yang sudah diketahui sebelumnya.

Machine Learning (ML) menawarkan paradigma deteksi yang berbeda secara konseptual. ML melatih model dari data historis sehingga pola anomali dikenali berdasarkan karakteristik statistiknya, bukan basis data tanda tangan [3]. Pendekatan ini secara teoretis memungkinkan deteksi serangan yang belum pernah terdokumentasi. Sejumlah penelitian melaporkan akurasi deteksi di atas 99% menggunakan deep learning pada dataset tertentu [4]. Namun, angka ini perlu diinterpretasi secara kritis karena mayoritas diperoleh dalam kondisi pengujian terkontrol, belum merepresentasikan heterogenitas jaringan produksi nyata.

Kesenjangan antara kinerja model dalam pengujian dan lingkungan operasional masih menjadi tantangan. False positive rate yang tinggi menghabiskan kapasitas analisis tim keamanan dan berpotensi memicu alert fatigue [5]. Di sisi lain, model deep learning berkinerja tinggi umumnya beroperasi sebagai black box, menyulitkan analis memahami dasar pengambilan keputusan model. Hal ini menjadi hambatan kritis ketika justifikasi teknis dibutuhkan untuk respons insiden maupun kepatuhan regulasi.

Beberapa tinjauan sistematis terdahulu telah memetakan lanskap ML untuk keamanan siber. Khraisat dkk. [2] mendokumentasikan teknik deteksi intrusi secara luas, tetapi tidak menyajikan perbandingan kinerja kuantitatif lintas algoritma. Ferrag dkk. [3] menyusun kajian komparatif deep learning yang komprehensif, namun cakupannya terbatas pada studi sebelum 2020. Ahmad dkk. [4] melakukan perbandingan sistematis ML dan DL, tetapi tidak membahas ketahanan terhadap serangan adversarial maupun kesiapan implementasi pada lingkungan IoT.

Dari analisis literatur, teridentifikasi tiga celah utama. Pertama, belum ada tinjauan yang memisahkan temuan studi kontemporer (2020–2025) dan studi historis dalam satu kerangka terpadu. Kedua, hubungan sebab-akibat antar tantangan implementasi, seperti dampak augmentasi data terhadap interpretabilitas model, belum dianalisis sistematis. Ketiga, penilaian tingkat kematangan tren penelitian masa depan berdasarkan bukti empiris masih sangat terbatas.

Bertolak dari celah tersebut, kajian ini memberikan tiga kontribusi orisinal. Pertama, pemisahan eksplisit antara studi primer (2020–2025) dan studi landasan (sebelum 2020) dalam satu kerangka SLR terpadu untuk atribusi temuan yang tepat secara temporal. Kedua, pemetaan sistematis rantai ketergantungan antar tantangan implementasi. Ketiga, penilaian kematangan tren penelitian masa depan berdasarkan bukti empiris, bukan proyeksi spekulatif.

Kajian ini merumuskan empat pertanyaan penelitian. RQ1 membahas algoritma ML dominan dan perbandingan kinerjanya secara kuantitatif. RQ2 mengkaji dataset acuan utama pengujian dan representativitasnya terhadap jaringan nyata. RQ3 menelaah hambatan implementasi paling konsisten dan keterkaitannya. RQ4 mengarahkan perkembangan penelitian berdasarkan tingkat kematangan bukti empiris.

Hasil kajian ditujukan bagi dua segmen pembaca. Bagi peneliti, peta algoritma dan identifikasi celah penelitian menyediakan titik tolak agenda riset lanjutan. Bagi praktisi keamanan, peta kinerja algoritma dapat menjadi referensi dalam menyelaraskan pendekatan ML dengan kapasitas infrastruktur dan prioritas operasional organisasi.

2. METODOLOGI PENELITIAN

2.1 Desain dan Strategi Pencarian

Penelitian ini menggunakan desain *Systematic Literature Review* (SLR) yang mengacu pada panduan Carrera-Rivera et al. [6] yang secara khusus dirancang untuk penelitian ilmu komputer. Kerangka ini dipilih karena menetapkan prosedur terstruktur yang memungkinkan setiap tahap proses seleksi direplikasi dan diaudit secara independen, sekaligus meminimalkan bias seleksi melalui dokumentasi kriteria yang ditetapkan sebelum pencarian dimulai. Penelusuran dilakukan secara sistematis pada enam basis data akademik: IEEE Xplore,

ScienceDirect, SpringerLink, ACM Digital Library, Google Scholar, dan Portal Garuda. Rentang tahun untuk studi primer dibatasi pada 2020–2025 guna memastikan relevansi terhadap konteks ancaman terkini, sementara studi landasan dipilih tanpa batasan tahun berdasarkan relevansinya terhadap RQ yang telah ditetapkan.

String pencarian dikonstruksi menggunakan operator *Boolean* dan diuji terlebih dahulu pada basis data utama sebelum diterapkan secara resmi. Untuk basis data IEEE Xplore, ScienceDirect, SpringerLink, dan ACM, string yang digunakan adalah: ("machine learning" OR "deep learning" OR "neural network" OR "random forest" OR "support vector machine") AND ("cyberattack" OR "intrusion detection" OR "malware detection" OR "network security" OR "phishing detection") AND ("detection" OR "classification" OR "prediction"). Untuk Google Scholar, pencarian dilakukan dengan query: machine learning cyberattack detection 2020–2025 | deep learning intrusion detection system | DDoS detection machine learning. Untuk Portal Garuda, string yang digunakan adalah: "machine learning" AND ("serangan siber" OR "deteksi intrusi" OR "keamanan jaringan" OR "deteksi malware").

2.2 Kriteria Inklusi dan Eksklusi

Studi dibagi ke dalam dua kategori primer dan landasan dengan kriteria yang berbeda sebagaimana diuraikan pada Tabel 1. Studi primer mencakup artikel *peer-reviewed* (2020–2025) yang membahas penerapan ML atau *deep learning* untuk deteksi serangan siber dan menyajikan eksperimen empiris dengan metrik evaluasi kuantitatif yang dapat dibandingkan (akurasi, F1-score, FPR). Studi landasan mencakup karya yang memperkenalkan *dataset benchmark* yang masih aktif digunakan, meletakkan dasar arsitektur atau konsep yang dirujuk penelitian modern, atau menyediakan *baseline* metodologis yang digunakan dalam kajian ini. Artikel dieksklusikan apabila tidak menyajikan eksperimen kuantitatif, bersifat editorial atau opini, tidak melaporkan metrik yang dapat dibandingkan antarstudy, atau teks lengkapnya tidak dapat diakses [6].

Tabel 1. Kriteria Inklusi Berdasarkan Kategori Studi

Kategori Studi	Rentang Tahun	Kriteria Inklusi
Studi Primer	2020–2025	Artikel peer-review yang membahas ML/DL untuk deteksi serangan siber, menyajikan eksperimen empiris dengan metrik evaluasi kuantitatif (akurasi, F1-score, FPR)
Studi Landasan	Sebelum 2020	Karya yang: (a) memperkenalkan dataset benchmark yang masih aktif digunakan (NSL-KDD, UNSW-NB15, CICIDS); (b) meletakkan landasan arsitektur/konsep yang menjadi acuan penelitian modern; atau (c) memberikan baseline metodologi yang dirujuk kajian ini

2.3 Proses Seleksi PRISMA, Penilaian Kualitas, dan Ekstraksi Data

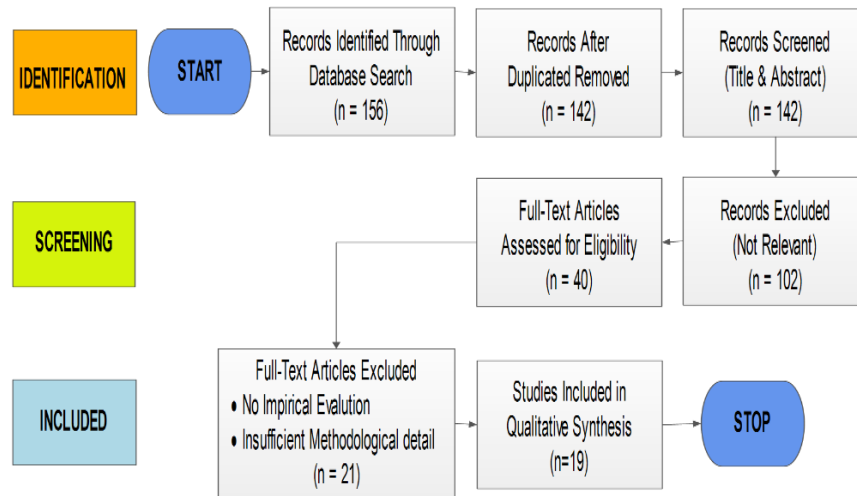
Proses seleksi mengikuti empat tahap PRISMA secara ketat (Gambar 1). Dari 156 artikel yang teridentifikasi melalui pencarian pada enam basis data, sebanyak 14 duplikat dieliminasi sehingga tersisa 142 artikel. Tahap skrining judul dan abstrak mengeliminasi 102 artikel yang tidak memenuhi kriteria tematik maupun inklusi. Dari 40 artikel yang dievaluasi secara penuh, sebanyak 21 artikel dieksklusikan dengan rincian: 11 artikel tidak menyajikan eksperimen kuantitatif, 7 artikel tidak melaporkan metrik yang dapat dibandingkan antarsatu studi dengan studi lainnya, dan 3 artikel tidak dapat diakses teks lengkapnya. Dengan demikian, 19 artikel ditetapkan sebagai korpus final. Komposisi 9 studi primer dan 10 studi landasan merupakan hasil seleksi berbasis kriteria, bukan kuota yang ditetapkan sejak awal penelitian.

Penilaian kualitas dilakukan dengan menggunakan kerangka Yang dkk. [7] yang mengevaluasi setiap studi pada dua dimensi, yaitu rigor (validitas desain eksperimen dan keandalan pelaporan) serta relevansi industri (tingkat keterterapan temuan dalam skenario produksi nyata). Kerangka ini dipilih karena dikembangkan melalui analisis komprehensif terhadap 241 SLR dalam bidang rekayasa perangkat lunak dan sistem berbasis komputer, sehingga instrumen penilaiannya mencerminkan praktik terkini komunitas riset. Studi yang dinilai lemah pada salah satu dimensi dieksklusikan pada tahap evaluasi teks penuh.

Ekstraksi data dilakukan secara terstruktur dari setiap studi yang lolos seleksi, mencakup delapan elemen, yaitu identitas publikasi, algoritma ML yang digunakan, jenis serangan yang ditargetkan, dataset yang dipakai, metrik evaluasi (akurasi, F1-score, dan FPR), kategori studi, keterbatasan yang dilaporkan penulis,

serta pembahasan mengenai interpretabilitas atau ketahanan terhadap serangan adversarial. Formulir ekstraksi seragam diterapkan pada seluruh 19 studi untuk menjamin konsistensi data.

Terdapat dua batasan metodologis yang perlu dicermati oleh pembaca. Dari sisi construct validity, variasi terminologi dalam literatur berpotensi menyebabkan sejumlah studi relevan tidak terjaring oleh strategi pencarian yang disusun. Dari sisi external validity, temuan yang diperoleh dari dataset benchmark tidak serta-merta berlaku untuk jaringan produksi yang memiliki karakteristik jauh lebih heterogen. Selain itu, seluruh proses dilakukan oleh satu peneliti, sehingga risiko bias dimitigasi melalui penerapan kriteria yang terdokumentasi secara ketat dengan cakupan bahasa yang dibatasi pada publikasi berbahasa Indonesia dan Inggris.



Gambar 1. PRISMA Flowchart Proses Seleksi Literatur

3. HASIL DAN PEMBAHASAN

3.1 Ringkasan Studi Terpilih

Sembilan belas studi yang memenuhi seluruh kriteria seleksi terbagi atas 9 studi primer (Tabel 2a) dan 10 studi landasan (Tabel 2b). Studi primer menyediakan bukti empiris langsung untuk menjawab keempat RQ; studi landasan digunakan untuk membangun konteks historis, menyediakan perbandingan baseline, dan menelusuri genealogi dataset benchmark yang masih aktif digunakan. Pemisahan ini bukan semata perbedaan kronologis, melainkan keputusan metodologis yang memungkinkan pembaca membedakan temuan yang berakar dari konteks ancaman terkini dengan fondasi yang dibangun pada era jaringan berbeda.

Tabel 2. Studi Primer (2020–2025) – Bukti Empiris Perkembangan Terkini

No	Penulis / Tahun	Algoritma ML	Jenis Serangan	Dataset	Akurasi
1	Ferrag et al. [3], 2020	CNN, LSTM, RNN, SVM	Intrusion (DoS, Probe)	NSL-KDD, CICIDS2017	98,2%
2	Ahmad et al. [4], 2021	DNN, RF, SVM, KNN	Intrusion (DoS, DDoS)	NSL-KDD, KDD99	99,1%
3	Khraisat et al. [2], 2020	C5, OC-SVM, Stacking	Intrusion (hybrid)	UNSW-NB15	97,6%
4	Rahman et al. [14], 2020	RF, SVM, MLP	Intrusi IoT	BoT-IoT, UNSW-NB15	97,8%
5	Gupta et al. [15], 2024	BERT + CNN	Phishing Detection	Dataset Email Kustom	97,5%
6	Mumuni & Mumuni [13], 2025	AutoML, Deep Learning	Feature Engineering	Multi-domain	N/A (Survei)
7	Khraisat et al. [21], 2024	Federated Learning IDS	Intrusion (federated)	CICIDS2017, UNSW-NB15	95,3%

8	Adadi & Berrada [20], 2023	XAI (SHAP, LIME)	Explainable IDS	Multi-domain	N/A (Tinjauan)
9	Ayub et al. [11], 2020	Adversarial ML (FGSM)	Evasion attack IDS	NSL-KDD, CICIDS2017	Evasion >85%

Tabel 3. Studi Landasan (Sebelum 2020) – Baseline Metodologi dan Perbandingan Historis

No	Penulis / Tahun	Algoritma	Fokus	Dataset	Akurasi	Peran dalam Kajian
1	Vinayakumar et al. [8], 2019	DL (MLP, RNN)	Intrusion	KDD99, UNSW-NB15	99,3%	Landasan komparasi DL vs ML tradisional
2	Vinayakumar et al. [9], 2017	CNN	Intrusion	NSL-KDD	97,4%	Landasan CNN untuk IDS
3	Kim et al. [10], 2017	DNN (MLP)	Intrusion	NSL-KDD	98,6%	Landasan DNN untuk IDS
4	Berman et al. [5], 2019	CNN, LSTM, Autoencoder	Multi-tipe	NSL-KDD, CICIDS	96,8%	Survei DL komprehensif; landasan autoencoder keamanan siber
5	Zhang et al. [12], 2019	GAN + DNN	Imbalanced data	NSL-KDD, CICIDS2018	98,0%	Landasan GAN untuk augmentasi data serangan
6	Yang et al. [20], 2019	Federated Learning	Deteksi terdistribusi	Simulasi	92,0%	Landasan konsep federated learning untuk IDS
7	Moustafa & Slay [16], 2016	Analisis Statistik	Evaluasi kompleksitas UNSW-NB15	UNSW-NB15 vs KDD99	85,0%	Evaluasi komparatif benchmark; validasi UNSW-NB15 vs KDD99
8	Sharafaldin et al. [17], 2018	RF, NB, KNN	Intrusion	CICIDS2017	97,0%	Pengembang dataset CICIDS2017
9	Moustafa & Slay [16], 2016	Naive Bayes, DT	Evaluasi NIDS	UNSW-NB15	85,0%	Pengembang dataset UNSW-NB15 (versi jurnal)
10	Lundberg & Lee[19], 2017	SHAP (XAI)	Interpretabilitas model	Multi-domain	N/A	Landasan teknik SHAP yang mendasari XAI modern

Catatan: Yang et al. [7] (Yang, L. et al., 2021 - kerangka penilaian kualitas SLR) tidak dimasukkan sebagai studi korpus karena bukan penelitian substantif ML untuk keamanan siber; berbeda dengan Yang et al. [19] (Yang, Q. et al., 2019 - federated learning) yang merupakan studi landasan dalam korpus. [16] dan [17] merujuk pada dua artikel dari penulis yang sama (Moustafa & Slay, 2016) untuk dataset UNSW-NB15.

3.2 Distribusi Temporal dan Evolusi Penelitian

Tabel 4 merekam pergeseran fokus penelitian secara kronologis. Pada periode 2009–2018, enam studi landasan memberikan kontribusi utama melalui pembangunan dataset acuan NSL-KDD [16], UNSW-NB15 [16], dan CICIDS2017 [17] - serta pengujian arsitektur *deep learning* awal yang masih bersifat eksplorasi. Pada 2019, empat studi landasan menandai ekspansi ke teknik yang lebih *sophisticated*: GAN untuk augmentasi data [12], *autoencoder* untuk deteksi anomali tanpa label [5], dan konsep awal *federated learning* [19] yang kala itu masih diperlakukan sebagai penelitian spekulatif. Memasuki 2020, studi primer mulai memperlihatkan pergeseran paradigma yang nyata dari optimasi akurasi semata menuju pertanyaan yang lebih substantif: ketahanan model

terhadap serangan *adversarial*, kecepatan inferensi pada perangkat IoT berdaya terbatas, dan transparansi keputusan model yang dibutuhkan untuk kepatuhan regulasi.

Tabel 4. Distribusi Temporal 19 Studi Terpilih

Periode	Jumlah Studi	Fokus Utama
2009–2018	6 (landasan)	Pembangunan dataset acuan benchmark (NSL-KDD, UNSW-NB15, CICIDS) dan pengujian metode klasik ML/DL awal
2019	4 (landasan)	Eksplorasi deep learning lanjutan: GAN untuk augmentasi data [12], autoencoder [5], federated learning awal [19]
2020	4 (primer)	Adopsi hybrid IDS [2][3], keamanan IoT [14], adversarial ML [11]
2021	1 (primer)	Studi komparatif ML/DL sistematis pada jaringan multi-tipe [4]
2022–2023	1 (primer)	XAI operasional: SHAP dan LIME untuk IDS [20]
2024–2025	3 (primer)	Phishing BERT+CNN [15], federated IDS skala besar [21], AutoML [13]

3.3 Algoritma ML yang Digunakan dan Evaluasi Kinerja (RQ1)

Perbandingan kinerja sepuluh kategori algoritma dirangkum pada Tabel 5. Referensi mencakup studi primer dan landasan karena penelusuran evolusi kinerja memerlukan perspektif lintas periode untuk memahami konteks perkembangan teknisnya secara utuh.

Tabel 5. Perbandingan Kinerja Algoritma *Machine Learning* dalam Deteksi Serangan Siber

Algoritma	Kasus Penggunaan Utama	Rentang Akurasi	Keb. Komputasi	Interpretabilitas
CNN [3][9]	Deteksi malware, analisis trafik berbasis citra	97–98%	Tinggi	Rendah
LSTM / RNN [4][10]	Deteksi DDoS, anomali temporal jaringan	97–99%	Tinggi	Rendah
Autoencoder [5][8]	Zero-day detection tanpa label serangan	94–97%	Sedang	Rendah
GAN (augmentasi) [12]	Pembangkit data sintesis kelas minoritas	96–98%	Sangat Tinggi	Rendah
DNN / MLP [4][8]	Klasifikasi intrusi multi-kelas	96–99%	Sedang	Sedang
Random Forest [2][14]	IDS umum, IoT, seleksi fitur	95–98%	Rendah	Tinggi
SVM [2][3]	Klasifikasi biner, dataset kecil	90–97%	Rendah	Tinggi
Hybrid DL+ML [3][14]	Keseimbangan akurasi dan efisiensi	97–99%	Sedang	Sedang
Federated Learning [19][21]	Deteksi terdistribusi, privasi terjaga	92–95%	Sedang	Sedang
BERT + CNN [15]	Phishing detection berbasis NLP	97,5%	Tinggi	Rendah

Deep Learning: CNN mengekstraksi pola spasial dari representasi visual trafik jaringan dan mencatat akurasi 97–98% pada studi yang diulas [3][9]. Untuk data berurutan seperti arus paket jaringan, LSTM dan RNN lebih unggul - keduanya mampu menangkap pola temporal anomali DDoS dengan akurasi 97–99% [4][10]. *Autoencoder* bekerja dengan cara yang berbeda: model dilatih hanya pada trafik normal, sehingga setiap pola yang menyimpang dari kondisi tersebut ditandai sebagai anomali. Pendekatan ini cocok untuk skenario *zero-day* [5][8]. GAN difungsikan bukan sebagai detektor, melainkan sebagai pembangkit data sintesis untuk menyeimbangkan distribusi kelas yang sangat timpang antara trafik normal dan serangan [12].

Machine Learning Tradisional: Random Forest mendominasi pilihan untuk *deployment* skala produksi karena efisien secara komputasi, memiliki mekanisme seleksi fitur bawaan melalui kalkulasi kepentingan fitur, dan menghasilkan keputusan yang dapat dijelaskan kepada pemangku kepentingan tanpa latar belakang ML [2][14]. Khraisat et al. [2] menegaskan bahwa interpretabilitas bukan sekadar akurasi menjadi faktor determinan

adopsi di lingkungan korporat. Temuan ini kontras dengan kecenderungan literatur pra-2020 yang memperlakukan akurasi sebagai satu-satunya kriteria keberhasilan, mengindikasikan pergeseran prioritas yang nyata dalam komunitas riset keamanan siber pascapandemi. SVM tetap relevan untuk klasifikasi biner pada dataset berukuran terbatas [3], namun kinerjanya terdegradasi secara signifikan pada dataset berskala besar dengan distribusi kelas yang kompleks.

Pendekatan *hybrid* yang mengintegrasikan DL dan ML klasik semakin banyak dieksplorasi pada konteks IoT, di mana keterbatasan memori dan daya komputasi memaksa *trade-off* antara akurasi dan efisiensi [3][14]. Rahman et al. [14] mendemonstrasikan bahwa kombinasi RF dan MLP dapat mencapai akurasi 97,8% pada jaringan IoT dengan beban komputasi jauh di bawah arsitektur DL penuh - temuan yang relevan mengingat pertumbuhan pesat jumlah perangkat IoT sebagai vektor serangan baru. Secara keseluruhan, penelitian 2020–2025 menunjukkan transisi paradigmatis: akurasi tidak lagi menjadi satu-satunya tolok ukur, dan ketahanan terhadap serangan *adversarial*, kecepatan respons, serta kemampuan *self-explanation* model menjadi dimensi evaluasi yang sama pentingnya.

3.4 Jenis Serangan dan Dataset Benchmark (RQ2)

Intrusion detection mendominasi fokus studi, mencakup serangan DoS/DDoS, probe, R2L, dan U2R. Model DL (DNN: 96–99%) mengungguli RF (95–98%) rata-rata 1–4 poin persentase pada kategori ini [3][4], namun keunggulan tersebut datang dengan biaya komputasi yang substansial lebih tinggi. *Malware detection* berbasis analisis *hybrid* mencapai akurasi di atas 97% [3], sementara *phishing detection* menggunakan BERT+CNN memanfaatkan kapasitas pemahaman semantik NLP untuk mencapai 97,5% [15], membuka jalur bagi penerapan model bahasa besar dalam keamanan siber.

Dari sisi dataset, NSL-KDD yang dikembangkan dari KDD CUP 1999 sebagai penyempurnaan atas kelemahan statistik dataset aslinya masih mendominasi sebagai tolok ukur karena aksesibilitasnya yang tinggi, dokumentasi yang komprehensif, dan ketersediaan *baseline* perbandingan dari studi terdahulu. Namun, rentang usianya yang melampaui dua dekade menjadikannya kurang representatif terhadap pola trafik dan serangan modern: serangan berbasis enkripsi, dominasi trafik HTTPS, dan serangan terhadap infrastruktur *cloud* tidak terwakili secara memadai. CICIDS2017 [17] dan UNSW-NB15 [16], yang lahir sebagai studi landasan dalam korpus ini, semakin banyak diadopsi oleh studi primer 2020–2025 karena merepresentasikan kondisi jaringan yang lebih kontemporer. Moustafa & Slay [16] mendemonstrasikan secara empiris bahwa UNSW-NB15 jauh lebih kompleks dari KDD99 dan lebih tepat sebagai tolok ukur evaluasi NIDS modern. Adopsi data produksi nyata tetap terhambat oleh sensitivitas privasi yang membatasi berbagi data antarlembaga.

3.5 Tantangan Utama Implementasi (RQ3)

Tabel 6 merangkum tujuh tantangan yang muncul secara konsisten di berbagai studi. Aspek yang memerlukan perhatian khusus adalah relasi sebab-akibat antar tantangan: teknik augmentasi data menggunakan GAN untuk mengatasi ketidakseimbangan kelas berpotensi memperburuk interpretabilitas model; kompresi model untuk menekan latensi kerap mengorbankan akurasi secara tidak proporsional; dan mekanisme pertahanan terhadap serangan *adversarial* yang efektif saat ini dapat menjadi obsolet dalam waktu singkat seiring evolusi teknik serangan. Optimasi yang berfokus pada satu tantangan tanpa memperhitungkan rantai dampaknya terhadap dimensi lain kerap menghasilkan solusi yang fragil dalam lingkungan operasional.

Tabel 6. Tantangan Implementasi ML dalam Deteksi Serangan Siber

Tantangan	Deskripsi	Solusi yang Diusulkan	Keterbatasan yang Masih Ada
Ketidakseimbangan Data	Data serangan jauh lebih sedikit dari trafik normal; model bias ke kelas mayoritas	SMOTE, cost-sensitive learning, augmentasi GAN	Data sintetis GAN belum merepresentasikan serangan nyata [12]
Feature Engineering	Ekstraksi fitur manual lambat dan bergantung keahlian domain	AutoML, pipeline otomatis, pembelajaran fitur DL	Fitur DL sulit diinterpretasikan; skalabilitas AutoML terbatas [13]

Pemrosesan <i>Real-Time</i>	Model DL besar sulit berjalan latensi rendah pada trafik skala besar	Pruning, quantization, knowledge distillation, TinyML	Kompresi model sering menurunkan akurasi tidak proporsional [3][4]
Serangan Adversarial	Perubahan kecil input dapat mengelabui model klasifikasi secara signifikan	Adversarial training, defensive distillation	Teknik serangan dan pertahanan terus berkembang bergantian [11]
Interpretabilitas	Model DL bersifat black-box; sulit dijelaskan ke analis keamanan	XAI: LIME [18], SHAP, mekanisme attention	Overhead komputasi tinggi; belum ada standar industri [20]
Realisme Dataset	Dataset benchmark tidak merepresentasikan trafik produksi nyata	Adopsi CICIDS2017 [17], UNSW-NB15 [16], atau data produksi	Dataset produksi sensitif privasi; sulit dibagikan secara publik
Distribusi Data (FL)	Data tersebar di banyak organisasi dengan kebijakan privasi berbeda	Federated learning, differential privacy, secure aggregation	Data non-IID menurunkan kualitas agregasi model global [21]

3.6 Tren dan Arah Penelitian Masa Depan (RQ4)

Tabel 7 memetakan tujuh tren yang relevan berdasarkan tingkat kematangannya. Tanda (*) menandai tren yang belum terwakili dalam korpus kajian ini namun perkembangannya layak dipantau.

Tabel 7. Tren dan Arah Penelitian Masa Depan ML untuk Keamanan Siber

Tren	Deskripsi	Keunggulan	Tantangan	Kematangan
Federated Learning [21]	IDS kolaboratif lintas organisasi tanpa berbagi data mentah	Privasi terjaga; skalabel lintas institusi	Data non-IID menurunkan akurasi global	Tinggi
Explainable AI (XAI) [20]	Transparansi keputusan model untuk analis keamanan	Meningkatkan kepercayaan; mendukung forensik digital	Overhead komputasi; belum ada standar baku	Tinggi
Transfer Learning [15]	Model pra-latih untuk domain keamanan baru	Hemat data berlabel; pelatihan lebih cepat	Risiko negative transfer antar domain	Sedang
AutoML & NAS [13][14]	Otomatisasi arsitektur dan hyperparameter model	Adopsi ML tanpa keahlian teknis mendalam	Biaya komputasi NAS sangat tinggi	Sedang
Integrasi SOAR [4][21]	ML sebagai pengambil keputusan respons insiden otomatis	Respons cepat; beban tim keamanan berkurang	Risiko respons salah; regulasi masih ketat	Sedang
TinyML / Edge AI [14]	Inferensi langsung di perangkat IoT sumber daya terbatas	Latensi sangat rendah; privasi di perangkat	Kapasitas dan akurasi model terbatas	Berkembang
LLM for Security *	Analisis log, <i>threat intelligence</i> , deteksi phishing berbasis LLM	Pemahaman konteks tinggi; fleksibel	Biaya inferensi tinggi; risiko halusinasi	Eksplorasi Awal

* LLM for Security: belum ada studi primer dalam korpus kajian ini; dicantumkan sebagai emerging trend berdasarkan perkembangan terkini di luar periode kajian.

Federated learning [21] dan XAI [20] menonjol sebagai dua tren dengan kematangan bukti empiris tertinggi. FL menjawab hambatan privasi yang selama ini menghalangi kolaborasi antarlembaga dalam membangun model deteksi Bersama, sebuah kendala fundamental mengingat data jaringan mengandung informasi bisnis yang sensitif. XAI yang akar teknisnya dapat ditelusuri ke kerangka SHAP yang diperkenalkan oleh Lundberg & Lee [18] serta konsep dasar yang diletakkan Adadi & Berrada [20] menjawab tuntutan transparansi dari tim keamanan dan regulator yang semakin mendorong akuntabilitas sistem berbasis AI.

Transfer learning dan AutoML [13][14] relevan karena potensinya menurunkan biaya adopsi secara substansial: institusi dengan keterbatasan data berlabel atau kapabilitas ML internal dapat memanfaatkan model pra-latih atau *pipeline* otomatis tanpa membangun infrastruktur riset dari nol. Integrasi dengan platform SOAR membuka kemungkinan siklus respons insiden yang sepenuhnya otomatis, namun memerlukan tingkat kepercayaan pada keandalan model yang saat ini belum dapat dijamin secara konsisten. Pemanfaatan LLM untuk analisis log dan threat intelligence masih berada pada fase eksplorasi awal, namun kapasitasnya dalam memahami konteks bahasa alami membuka dimensi analitik yang belum dapat dijangkau oleh pendekatan berbasis fitur numerik.

4. KESIMPULAN

Kajian sistematis ini berhasil menjawab seluruh pertanyaan penelitian yang diajukan. Untuk RQ1, ditemukan bahwa CNN dan LSTM memberikan akurasi deteksi tertinggi (97–99%). Meskipun demikian, Random Forest dan SVM tetap menjadi pilihan rasional ketika efisiensi komputasi atau kebutuhan interpretabilitas menjadi prioritas. Temuan ini menunjukkan bahwa dikotomi antara deep learning dan machine learning klasik tidak lagi relevan. Pendekatan hibrida yang memadukan keduanya justru lebih prospektif untuk dikembangkan.

Untuk RQ2, CICIDS2017 dan UNSW-NB15 secara progresif menggeser dominasi NSL-KDD sebagai standar evaluasi karena lebih akurat merepresentasikan ancaman jaringan modern. Namun demikian, kajian ini menemukan kesenjangan kritis, yaitu belum ada dataset yang mampu sepenuhnya mereplikasi heterogenitas lalu lintas jaringan produksi nyata. Kesenjangan antara performa pada dataset benchmark dan realitas implementasi menjadi catatan penting bagi peneliti dan praktisi.

Untuk RQ3, teridentifikasi empat hambatan paling konsisten, yaitu ketidakseimbangan data latih, latensi real-time, serangan adversarial, dan opasitas model. Berbeda dengan kajian sebelumnya yang membahas hambatan secara terpisah, penelitian ini menemukan adanya relasi ketergantungan di antara keempatnya. Implikasinya, solusi parsial berisiko gagal sehingga pendekatan terpadu menjadi keharusan dalam perancangan sistem deteksi intrusi.

Untuk RQ4, federated learning dan XAI merupakan dua arah penelitian paling matang dengan potensi implementasi jangka pendek. Sementara itu, integrasi LLM untuk analisis keamanan masih berada pada fase eksplorasi awal dan memerlukan riset lebih lanjut.

Dari sisi metodologi, pemisahan eksplisit antara studi primer dan studi landasan dalam kerangka SLR terpadu terbukti menghasilkan atribusi temuan yang lebih akurat secara temporal. Pendekatan ini dapat diadopsi oleh bidang lain yang mengalami perubahan teknologi cepat untuk menghindari rekomendasi yang tidak relevan dengan kondisi terkini.

Temuan kajian ini memberikan implikasi praktis. Bagi praktisi, hasil penelitian ini menjadi panduan dalam memilih algoritma: deep learning direkomendasikan ketika prioritas utama adalah akurasi, sedangkan Random Forest atau SVM lebih sesuai ketika sumber daya terbatas atau kebutuhan auditabilitas tinggi. Bagi pengembang dataset, kajian ini menegaskan urgensi membangun dataset berbasis lalu lintas produksi nyata yang memenuhi regulasi privasi, karena ketergantungan pada dataset benchmark lama berisiko menghasilkan model yang tidak siap menghadapi ancaman kontemporer.

Kajian ini memiliki tiga keterbatasan. Pertama, variasi terminologi dalam literatur internasional berpotensi menyebabkan studi relevan tidak terjaring. Kedua, proses seleksi dan ekstraksi data dilakukan oleh satu peneliti sehingga meningkatkan risiko bias subjektif, meskipun telah dimitigasi dengan dokumentasi kriteria yang ketat. Ketiga, generalisasi temuan ke lingkungan produksi perlu dilakukan dengan hati-hati.

Berdasarkan celah yang teridentifikasi, empat agenda riset mendatang direkomendasikan. Pertama, pengembangan teknik augmentasi data untuk menghasilkan sampel serangan sintesis yang realistis, khususnya untuk kelas serangan yang kurang terwakili, tanpa mengorbankan interpretabilitas model. Kedua, perancangan

arsitektur hibrida dengan latensi di bawah 10 milidetik pada perangkat IoT berdaya rendah dan akurasi di atas 95%. Ketiga, pengujian empiris metode pertahanan adversarial terhadap serangan yang belum dikenal menggunakan protokol evaluasi terstandarisasi. Keempat, pembangunan dataset publik berbasis lalu lintas jaringan produksi nyata dengan mekanisme anonimisasi yang mematuhi privasi, sebagai fondasi bagi riset selanjutnya.

REFERENCES

- [1] ENISA. (2024). ENISA Threat Landscape 2024. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024>
- [2] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2020). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, 2(1), 1–22. <https://doi.org/10.1186/s42400-019-0038-7>
- [3] Ferrag, M. A., Maglaras, L., Moschogiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419. <https://doi.org/10.1016/j.jisa.2019.102419>
- [4] Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1), e4150. <https://doi.org/10.1002/ett.4150>
- [5] Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A survey of deep learning methods for cyber security. *Information*, 10(4), 122. <https://doi.org/10.3390/info10040122>
- [6] Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Laso, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9, 101895. <https://doi.org/10.1016/j.mex.2022.101895>
- [7] Yang, L., Zhang, H., Shen, H., Huang, X., Zhou, X., Rong, G., & Shao, D. (2021). Quality assessment in systematic literature reviews: A software engineering perspective. *Information and Software Technology*, 130, 106397. <https://doi.org/10.1016/j.infsof.2020.106397>
- [8] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7, 41525–41550. <https://doi.org/10.1109/ACCESS.2019.2895334>
- [9] Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017). Applying convolutional neural network for network intrusion detection. *Proceedings of ICACCI 2017*, 1222–1228. <https://doi.org/10.1109/ICACCI.2017.8126009>
- [10] Kim, J., Shin, N., Jo, S. Y., & Kim, S. H. (2017). Method of intrusion detection using deep neural network. *Proceedings of IEEE BigComp 2017*, 313–316. <https://doi.org/10.1109/BIGCOMP.2017.7881684>
- [11] Ayub, M. A., Johnson, W. A., Talbert, D. A., & Siraj, A. (2020). Model evasion attack on intrusion detection systems using adversarial machine learning. *Proceedings of 54th CISS*, 1–6. <https://doi.org/10.1109/CISS48834.2020.1570617116>
- [12] Zhang, H., Yu, X., Ren, P., Luo, C., & Min, G. (2019). Deep adversarial learning in intrusion detection: A data augmentation enhanced framework. *arXiv:1901.07949*. <https://arxiv.org/abs/1901.07949>
- [13] Mumuni, A., & Mumuni, F. (2025). Automated data processing and feature engineering for deep learning and big data applications: A survey. *Journal of Information and Intelligence*, 3(2). <https://doi.org/10.1016/j.jiixd.2024.01.002>
- [14] Rahman, M. A., Asyhari, A. T., Leong, L. S., Satria, G. B., Tao, M. H., & Zolkipli, M. F. (2020). Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society*, 61, 102324. <https://doi.org/10.1016/j.scs.2020.102324>
- [15] Gupta, B. B., Gaurav, A., Arya, V., Attar, R. W., & Bansal, S. (2024). Advanced BERT and CNN-based computational model for phishing detection in enterprise systems. *Computer Modeling in Engineering & Sciences*, 141(3), 2165–2183. <https://doi.org/10.32604/cmescs.2024.056473>
- [16] Moustafa, N., & Slay, J. (2016). The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Information Security Journal: A Global Perspective*, 25(1–3), 18–31. <https://doi.org/10.1080/19393555.2015.1125974>

- [17] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of 4th ICISSP*, 108–116. <https://doi.org/10.5220/0006639801080116>
- [18] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 4765–4774. <https://arxiv.org/abs/1705.07874>
- [19] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1–19. <https://doi.org/10.1145/3298981>
- [20] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [21] Khraisat, A., Alazab, A., Singh, S., Jan, T., & Gomez, A. (2024). Survey on federated learning for intrusion detection system: Concept, architectures, aggregation strategies, challenges, and future directions. *ACM Computing Surveys*, 57(1), Article 7. <https://doi.org/10.1145/3687124>