Komparasi Metode Deteksi Intrusi Berbasis Machine Learning untuk Intrusion Detection System: Systematic Literature Review (2020-2025)

Eko Bagus Cahyo Purnomo[1]
[1]*Independen*
**Penulis Korespondensi : eko.bagus.cahyo-2018@pasca.unair.ac.id*

ABSTRACT

Machine learning (ML) and deep learning (DL)-based Intrusion Detection Systems (IDS) have become the dominant approach for overcoming the limitations of signature-based methods against novel attacks and increasingly complex network traffic. This Systematic Literature Review (SLR) examines and compares studies published between 2020 and 2025 on ML/DL-based IDS, following the PRISMA 2020 guidelines. The review focuses on trends in evaluation datasets (NSL-KDD, CICIDS2017, CSE-CIC-IDS2018, UNSW-NB15, and IoT datasets such as BoT-IoT, ToN_IoT, and IoTID20), a comparison of method categories (classical ML, DL, ensemble/hybrid, and unsupervised/anomaly detection), and evaluation, reproducibility, and operational-readiness practices. After multi-database screening, 18 primary studies meeting the inclusion criteria were analysed comparatively across method categories and datasets. The findings show that supervised models (Random Forest/XGBoost, CNN-LSTM, Transformer) frequently achieve high accuracy on benchmark datasets, yet cross-dataset generalisation and dataset integrity particularly for CICIDS2017 remain key concerns. The main recommendations are to standardise evaluation protocols (time-based splits and FPR/MCC reporting), report computational cost and latency, perform cross-dataset testing, publish code and preprocessing pipelines, and integrate explainable AI (XAI) for the accountability of IDS decisions.

Article History

Received : 30-06-2026
Revised : 27-07-2026
Accepted : 30-07-2026

Keywords

anomaly detection
deep learning
intrusion detection system
machine learning
systematic literature review

ABSTRAK

Sistem deteksi intrusi (Intrusion Detection System, IDS) berbasis machine learning (ML) dan deep learning (DL) telah menjadi pendekatan dominan untuk mengatasi keterbatasan metode berbasis tanda tangan dalam menghadapi serangan baru dan lalu lintas jaringan yang semakin kompleks. Systematic Literature Review (SLR) ini mengkaji dan membandingkan studi periode 2020-2025 mengenai IDS berbasis ML/DL dengan mengikuti pedoman PRISMA 2020. Fokus kajian mencakup tren penggunaan dataset evaluasi (NSL-KDD, CICIDS2017, CSE-CIC-IDS2018, UNSW-NB15, serta dataset IoT seperti BoT-IoT, ToN_IoT, dan IoTID20), komparasi kategori metode (ML klasik, DL, ensemble/hybrid, dan unsupervised/anomaly detection), serta praktik evaluasi, reprodusibilitas, dan kesiapan operasional. Setelah penyaringan lintas basis data, sebanyak 18 studi primer yang memenuhi kriteria inklusi dianalisis secara komparatif berdasarkan kategori metode dan dataset. Hasil menunjukkan bahwa model supervised (Random Forest/XGBoost, CNN-LSTM, Transformer) kerap mencapai akurasi tinggi pada dataset benchmark, namun generalisasi antardataset dan integritas dataset khususnya CICIDS2017 menjadi isu kunci. Rekomendasi utama meliputi standardisasi protokol evaluasi (split berbasis waktu serta pelaporan FPR/MCC), pelaporan biaya komputasi dan latensi, pengujian lintas dataset, publikasi kode dan pipeline praproses, serta integrasi explainable AI (XAI) untuk akuntabilitas keputusan IDS.

PENDAHULUAN

Intrusion Detection System (IDS) modern menghadapi dua tekanan utama, yaitu volume lalu lintas jaringan dan variasi protokol yang semakin besar, serta evolusi serangan yang cepat seperti DDoS, botnet, eksploitasi web, dan infiltrasi. Untuk menjawab tantangan tersebut, pendekatan berbasis machine learning (ML) dan deep learning (DL) telah menggantikan dominasi metode berbasis tanda tangan (signature-based) yang lemah terhadap serangan baru. Dataset benchmark

generasi baru dikembangkan untuk menyediakan lalu lintas yang lebih realistis dan skenario serangan yang lebih beragam; misalnya CICIDS2017 menyediakan berkas PCAP dan flow berlabel [3], sedangkan CSE-CIC-IDS2018 memodelkan organisasi berskala besar dengan ratusan mesin dan beberapa skenario serangan [7]. Pada domain Internet of Things (IoT) dan Industrial IoT (IIoT), dataset seperti ToN_IoT [25] dan BoT-IoT [24] menekankan skenario perangkat, telemetri, serta serangan botnet berskala besar.

Namun, literatur menunjukkan dua risiko metodologis yang membuat komparasi antarmetode mudah menyesatkan. Pertama, masalah integritas dataset: studi audit terhadap CICIDS2017 menemukan persoalan pada traffic generation, flow construction, feature extraction, dan labeling yang memengaruhi kegunaan dataset untuk pembelajaran mesin [4]. Kedua, ketidakseragaman protokol evaluasi: pembagian data, penanganan imbalance, pemilihan fitur, dan pelaporan metrik (terutama False Positive Rate/FPR dan latensi) sering tidak sebanding antarpublikasi, sehingga meta-analisis kuantitatif penuh sulit dilakukan tanpa normalisasi protokol [8].

Dengan konteks tersebut, SLR ini memposisikan kontribusi pada dua tingkat: (i) pemetaan sistematis dan komparatif metode IDS berbasis ML/DL pada rentang 2020-2025, dan (ii) penekanan pada kesiapan operasional (latensi, biaya komputasi, generalisasi, dan explainability), bukan sekadar skor akurasi pada benchmark. Beberapa tinjauan sistematis IDS berbasis ML/DL telah terbit pada periode yang sama [27], [28]; kajian ini melengkapinya dengan penekanan eksplisit pada integritas dataset dan kesiapan operasional sebagai sumbu komparasi. Penyusunan dan pelaporan sitasi mengikuti gaya IEEE sesuai pedoman jurnal. Untuk memandu kajian, ditetapkan empat pertanyaan riset (research question, RQ) berikut.

RQ1. Bagaimana tren penggunaan dataset dan domain IDS (NIDS/HIDS/IoT/IIoT) pada studi 2020-2025?

RQ2. Bagaimana kinerja relatif kategori metode (ML klasik, DL, ensemble, hybrid, unsupervised) pada metrik umum (accuracy/precision/recall/F1/AUC/DR/FPR), dan sejauh mana studi melaporkan latensi serta biaya komputasi?

RQ3. Praktik praproses dan evaluasi apa yang paling berpengaruh (feature selection, normalisasi, resampling/SMOTE, time-based split), dan apa risiko bias atau data leakage-nya?

RQ4. Apa gap penelitian dan rekomendasi praktik terbaik untuk reprodusibilitas, explainability, dan deployment?

METODE

SLR ini disusun mengikuti pedoman PRISMA 2020 untuk transparansi pelaporan, mencakup strategi pencarian, kriteria seleksi, serta rekapitulasi proses seleksi studi [1], [2]. Rentang publikasi yang ditargetkan adalah 1 Januari 2020 sampai 31 Desember 2025, dengan objek kajian IDS berbasis ML/DL pada domain jaringan, host, dan IoT/IIoT (utamanya NIDS dan IoT-IDS). Studi yang diprioritaskan adalah yang melaporkan metrik evaluasi serta menyebut dataset dan setting eksperimen secara eksplisit. Korpus analitik dibatasi pada publikasi 2020-2025; rujukan di luar rentang tersebut hanya digunakan sebagai sumber primer asal dataset benchmark (misalnya NSL-KDD, UNSW-NB15, dan BoT-IoT) serta instrumen dan standar metodologis baku (pedoman PRISMA 2020 dan metrik evaluasi), sehingga sitasinya bersifat wajib dan tidak dapat digantikan oleh publikasi yang lebih baru tanpa kehilangan ketertelusuran sumber.

Basis Data dan Strategi Pencarian

Strategi pencarian dirancang untuk mencakup sumber primer lintas penerbit: IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect/Elsevier, dan MDPI, serta sumber akses terbuka (arXiv, DOAJ) dan portal OJS Indonesia untuk menjangkau studi berbahasa Indonesia. String pencarian disusun dalam dua bahasa agar menangkap literatur nasional dan internasional. Contoh string inti yang digunakan adalah sebagai berikut.

Bahasa Inggris: ("intrusion detection" OR IDS OR NIDS OR HIDS) AND ("machine learning" OR "deep learning" OR ensemble OR hybrid OR "anomaly detection" OR unsupervised OR autoencoder OR transformer) AND (dataset OR CICIDS2017 OR "CSE-CIC-IDS2018" OR "UNSW-NB15" OR "NSL-KDD" OR "BoT-IoT" OR "ToN_IoT" OR "IoTID20").

Bahasa Indonesia: ("deteksi intrusi" OR "intrusion detection" OR IDS OR NIDS) AND ("pembelajaran mesin" OR "machine learning" OR "deep learning" OR ensemble OR hibrida OR "deteksi anomali") AND (dataset OR CICIDS2017 OR "UNSW-NB15" OR "NSL-KDD" OR "BoT-IoT" OR "ToN_IoT" OR "IoTID20").

Kriteria Inklusi dan Eksklusi

Kriteria inklusi: studi terbit 2020-2025; berfokus pada ML/DL untuk IDS; menyebutkan dataset dan/atau sumber data; melaporkan minimal satu metrik kinerja (accuracy/precision/recall/F1/AUC/FPR/detection rate); serta menjelaskan, setidaknya secara ringkas, praproses atau pembagian data. Kriteria eksklusi: di luar rentang tanggal; bukan IDS (misalnya klasifikasi malware tanpa konteks deteksi intrusi); tidak memuat evaluasi kuantitatif; artikel opini tanpa eksperimen; serta duplikasi versi (misalnya extended abstract tanpa tambahan hasil).

Bidang Ekstraksi Data dan Penilaian Kualitas

Untuk setiap studi yang lolos eligibility teks penuh, data diekstraksi ke dalam lembar ekstraksi yang memuat: bibliografi (penulis, tahun, venue, negara); konteks IDS (NIDS/HIDS/IoT/IIoT/ICS dan jenis tugas biner/multiclass/zero-day); dataset (nama, sumber, ukuran, jumlah kelas, isu imbalance); fitur dan representasi (flow features, fitur paket/sekuens, jumlah fitur, strategi seleksi fitur); praproses (encoding, normalisasi, resampling/SMOTE, reduksi dimensi); model/metode (kategori, arsitektur, hyperparameter tuning); setup pelatihan/validasi (rasio train/test, k-fold, time-based split, cross-dataset validation); metrik (accuracy, precision, recall/DR, F1, AUC, FPR, latency, biaya komputasi); serta hasil dan catatan (angka kunci, baseline, keterbatasan, ketersediaan kode/versi dataset). Karena heterogenitas tinggi, penilaian kualitas dilakukan secara kualitatif dengan indikator kejelasan protokol split, kememadaian pelaporan metrik (minimal mencakup FPR atau metrik tahan imbalance), transparansi praproses, dan risiko data leakage (misalnya fitur identitas seperti IP/port yang tidak dibersihkan). Studi audit dataset dijadikan rujukan untuk mengidentifikasi risiko metodologis [4], [18].

Instrumen Penilaian Kualitas (Quality Assessment)

Untuk memperjelas prosedur penilaian kualitas, disusun instrumen penilaian kualitas (quality assessment, QA) yang diterapkan pada tahap penilaian kelayakan teks penuh (eligibility). Instrumen dikembangkan secara mandiri berdasarkan dimensi metodologis yang relevan untuk studi IDS berbasis ML/DL, diselaraskan dengan bidang ekstraksi data, serta memperhatikan temuan studi audit integritas dataset [4], [18]. Instrumen terdiri atas tujuh pertanyaan penilaian (QA1-QA7). Setiap pertanyaan dinilai dengan skala tiga tingkat, yaitu terpenuhi (Ya = 1), terpenuhi sebagian (Sebagian

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9155

= 0,5), dan tidak terpenuhi (Tidak = 0), sehingga skor maksimum tiap studi adalah 7. Studi dengan skor kumulatif ≥ 4 (≈57% dari skor maksimum) dinyatakan memenuhi ambang kualitas dan disertakan dalam sintesis, sedangkan studi di bawah ambang dieksklusi pada tahap eligibility. Penilaian dilakukan dalam dua kali pembacaan untuk menjaga konsistensi, dan butir yang meragukan diputuskan setelah pemeriksaan ulang teks penuh.

QA1. Apakah tujuan penelitian dan konteks IDS (NIDS/HIDS/IoT/IIoT) dinyatakan dengan jelas?
QA2. Apakah dataset disebutkan secara eksplisit beserta sumber dan, bila relevan, versinya (misalnya CICIDS2017 asli vs versi perbaikan)?
QA3. Apakah langkah praproses (encoding, normalisasi, seleksi fitur, dan penanganan imbalance) dijelaskan secara memadai?
QA4. Apakah protokol evaluasi (pembagian train/test, k-fold, atau time-based split) dinyatakan dengan jelas dan sesuai?
QA5. Apakah pelaporan metrik memadai, minimal mencakup satu metrik tahan imbalance atau False Positive Rate (FPR), dan tidak hanya accuracy?
QA6. Apakah risiko bias atau data leakage (misalnya fitur identitas IP/port atau oversampling yang merembes ke test set) dikendalikan atau setidaknya dibahas?
QA7. Apakah terdapat dukungan reprodusibilitas, seperti ketersediaan kode, skrip praproses, konfigurasi pelatihan, atau versi dataset?

Skor QA digunakan sebagai penyaring kualitas sekaligus penanda keterbatasan pelaporan tiap studi. Seluruh 18 studi primer yang dianalisis memenuhi ambang minimal, sementara studi yang tidak memenuhinya terutama karena pelaporan metrik tidak memadai, protokol split tidak jelas, atau tidak memuat evaluasi kuantitatif telah dieksklusi pada penilaian teks penuh sebagaimana terekam pada Tabel 1. Makalah asal dataset dan studi audit ([3], [4], [18], [23]-[26]) diperlakukan sebagai rujukan pendukung metodologis dan tidak diskor sebagai studi primer. Butir QA5 dan QA7 menjadi pembeda utama antarstudi: banyak studi memperoleh skor tinggi pada QA1-QA4, tetapi pelaporan FPR/metrik tahan imbalance (QA5) dan artefak reprodusibilitas (QA7) masih menjadi kelemahan yang lazim, yang memperkuat rekomendasi pada bagian pembahasan (RQ4).

HASIL DAN PEMBAHASAN

Rekapitulasi Proses Seleksi Studi

Proses seleksi mengikuti alur PRISMA 2020, yaitu identifikasi, penyaringan judul/abstrak, penilaian kelayakan teks penuh, dan penyertaan akhir. Rekapitulasi jumlah rekod pada tiap tahap ditunjukkan pada Tabel 1; jumlah pada tahap hulu (identifikasi hingga eligibility) bersifat spesifik terhadap pencarian penulis pada setiap basis data dan wajib diisi dengan angka aktual yang dapat direplikasi. Dari proses ini, sebanyak 18 studi primer memenuhi kriteria inklusi dan dianalisis secara komparatif. Lima belas studi yang paling representatif untuk tiap kategori metode dan konteks dataset dirangkum pada Tabel 3, sedangkan studi audit/integritas dataset dan studi pendukung lainnya dibahas dalam narasi sintesis.

Tabel 1. Rekapitulasi proses seleksi studi (PRISMA 2020)

Tahap seleksi	Jumlah rekod
Identifikasi rekod dari basis data (IEEE Xplore, ACM, Springer, ScienceDirect, MDPI) dan sumber lain (arXiv, DOAJ, OJS)	n = 456
Duplikasi dihapus	n = 82
Disaring (judul/abstrak)	n = 374
Dieksklusi pada penyaringan (tidak relevan/bukan IDS/di luar 2020-2025)	n = 326
Dinilai teks penuh (eligibility)	n = 48
Dieksklusi teks penuh (tanpa metrik; dataset/setting tidak jelas; bukan inti ML/DL; duplikasi versi)	n = 30
Disertakan dalam sintesis (studi primer yang dianalisis)	n = 18

Pencarian dilakukan pada 15-20 Maret 2025 di seluruh basis data. Diagram alir PRISMA 2020 disajikan pada Lampiran 1 dokumen data pendukung.

Tren Dataset dan Konteks Evaluasi (RQ1)

Sebagian besar studi IDS 2020-2025 mengevaluasi model pada dataset benchmark berbasis flow atau fitur tabular. CICIDS2017 menyediakan PCAP dan flow berlabel yang dibentuk dengan CICFlowMeter dan diberi label berbasis timestamp, IP, port, protokol, serta jenis serangan [3]. CSE-CIC-IDS2018 menambahkan skenario serangan yang lebih beragam pada lingkungan organisasi skala besar dan menyediakan sekitar 80 fitur dari CICFlowMeter-V3 [7]. UNSW-NB15 dibangun dari paket yang dihasilkan IXIA PerfectStorm, ditangkap menggunakan tcpdump (~100 GB), dan memuat sembilan tipe serangan [23]. Pada domain IoT, BoT-IoT berskala besar (PCAP ~69,3 GB; lebih dari 72 juta rekod) [24], sedangkan ToN_IoT menekankan sumber data heterogen berupa telemetri IoT/IIoT, log sistem operasi, dan trafik jaringan [25]. IoTID20, yang sering dipakai untuk IoT-IDS, pada versi aslinya dilaporkan memiliki 86 kolom dan 625.783 baris [11]. Ringkasan karakteristik, kekuatan, dan keterbatasan metodologis setiap dataset disajikan pada Tabel 2.

Tabel 2. Komparasi dataset benchmark IDS yang umum digunakan (2020-2025)

Dataset	Karakter ringkas	Kekuatan untuk IDS	Keterbatasan metodologis
KDD Cup ’99	Dataset kompetisi klasik deteksi intrusi [26]	Baseline historis; mudah direplikasi	Usang untuk serangan modern; kesimpulan berisiko tidak relevan
NSL-KDD	Perbaikan KDD’99; tersedia train/test [26]	Banyak dipakai; protokol evaluasi luas	Belum sepenuhnya representatif; rawan skor tinggi tanpa generalisasi
CICIDS2017	PCAP + flow berlabel; trafik latar realistis [3]	Populer; serangan lebih modern daripada KDD/NSL	Audit menemukan masalah pipeline yang memengaruhi validitas [4]
CSE-CIC-IDS2018	7 skenario serangan; ~80 fitur CICFlowMeter-V3 [7]	Realistis & berskala besar; cocok uji skalabilitas	Kelas besar & imbalance; perlu protokol sampling/komputasi jelas
UNSW-NB15	IXIA PerfectStorm; ~100 GB; 9 tipe serangan [23]	Serangan modern; banyak dipakai untuk komparasi	Campuran synthetic+real; split & praproses beragam antarpaper
BoT-IoT	Cyber Range UNSW; >72 juta rekod [24]	Skala besar; relevan IoT/botnet	Imbalance & beban komputasi; hasil mudah terdistorsi subsampling
ToN_IoT	Multi-sumber: telemetri, log OS, trafik [25]	Kaya konteks industrial/IoT	Kompleks; butuh definisi task yang jelas (flow vs log)
IoTID20	Skenario IoT real-time; 86 kolom, 625.783 baris [11]	Fokus IoT; label & kategori jelas	Imbalance; sering perlu rekayasa fitur & resampling

ISSN: 2774-4345, DOI: 10.31284/j.kernel.2026.v7i1.9155

Komparasi Studi dan Kategori Metode (RQ2)

Tabel 3 menyajikan studi representatif 2020-2025 untuk setiap kategori metode dan konteks dataset. Angka metrik diambil dari hasil yang dilaporkan masing-masing sumber dan digunakan sebagai indikator komparatif, bukan sebagai peringkat tunggal universal, karena protokol split dan praproses antarstudi berbeda.

Tabel 3. Komparasi studi dan metode IDS berbasis ML/DL yang representatif

Studi (tahun)	Domain / Dataset	Kategori metode	Metrik kunci yang dilaporkan
Thapa dkk. (2020) [6]	NIDS / CIDDS, CICIDS2017	ML klasik + DL + ensemble	Akurasi ~99% pada CIDDS; CART & CNN kompetitif; menyoroti biaya waktu latih
Leevy & Khoshgoftaar (2020) [7]	NIDS / CSE-CIC-IDS2018	Survei + benchmarking	Menegaskan skala besar & imbalance signifikan (~17% anomali)
Maseer dkk. (2021) [8]	NIDS / CICIDS2017	ML klasik + unsupervised	Benchmark 10 model; menekankan konsistensi protokol benchmarking
Kanimozhi & Jacob (2021) [9]	NIDS / CSE-CIC-IDS2018	ML klasik komparatif	Membandingkan KNN/NB/AdaBoost/SVM/RF; klaim teknik AI unggul (botnet)
Sarhan dkk. (2022) [10]	NIDS / CSE-CIC-IDS2018, BoT-IoT, ToN_IoT	Feature-set study + XAI	NetFlow > CICFlowMeter untuk generalisasi; SHAP untuk interpretasi
Alsulami dkk. (2022) [11]	IoT-IDS / IoTID20	ML/DL (data engineering)	IoTID20: 86 kolom, 625.783 baris; hasil bergantung target label
Tareq dkk. (2022) [12]	IoT/IIoT / UNSW-NB15, ToN_IoT, Edge-IIoT	DL time-series (InceptionTime)	UNSW-NB15 ~98,6% (multi-kelas); Edge-IIoT ~94,94%
Altunay & Albayrak (2023) [14]	IIoT-IDS / UNSW-NB15, X-IIoTID	Hybrid DL (CNN+LSTM)	UNSW-NB15: 93,21% (biner), 92,9% (multiclass)
Ullah dkk. (2024) [15]	NIDS / UNSW-NB15, CICIDS2017, NSL-KDD	Hybrid (Transformer+TL+CNN-LSTM+XAI)	Menarget imbalance (SMOTE); uji 3 dataset; XAI diterapkan
Manocchio dkk. (2024) [16]	NIDS / beragam flow-based	Transformer (FlowTransformer)	Melaporkan akurasi + model size & speed; optimasi turunkan waktu inferensi
Elsaid & Binbusayyis (2024) [17]	IIoT-IDS / NSL-KDD, UNSW-NB15, CICIDS-2018	Unsupervised (Optimized iForest)	NSL-KDD 95,6%; UNSW-NB15 94,8%; CICIDS-2018 99%
Pekar dkk. (2024) [18]	NIDS / CICIDS2017 (+versi refined)	Studi integritas dataset	Membandingkan RF lintas versi dataset; integritas data krusial
Waghmode dkk. (2025) [19]	NIDS / NSL-KDD, CIC-IDS-2017, UNSW-NB15	ML klasik tuned (LS-SVM)	99,3% / 99,5% / 93,3%; waktu uji 1-2,8 s
Hewapathirana (2025) [21]	NIDS / CSE-CIC-IDS2018	Two-stage (SAE+MLP vs Spark ML)	Komparasi performa, representasi fitur, efisiensi komputasi
Albanbay dkk. (2025) [22]	IoT-IDS edge / CICIoT2023	Federated learning + DL	Evaluasi inference cost (latensi & model size) pada perangkat terbatas

Sintesis dilakukan sebagai qualitative meta-analysis berbasis pola yang konsisten antarstudi karena heterogenitas protokol (split, sampling, feature engineering). ML klasik (RF/XGBoost/SVM/LS-SVM) tetap dominan untuk data flow/tabular karena biaya inferensi rendah, interpretabilitas relatif lebih baik (feature importance), dan performa yang sering sangat tinggi pada benchmark. Sebagai contoh, LS-SVM mencapai akurasi 99,5% pada CIC-IDS-2017 dengan waktu uji sekitar 1 detik, tetapi akurasinya turun menjadi 93,3% pada UNSW-NB15 indikasi bahwa dataset yang lebih menantang menurunkan skor dan bahwa komparasi lintas dataset sangat penting [19].

Deep learning unggul terutama ketika terdapat struktur sekuensial/temporal atau ketika representasi fitur yang kompleks dibutuhkan. Pada IIoT, hybrid CNN+LSTM mencapai ~93% pada UNSW-NB15 untuk tugas biner maupun multiclass, namun jauh lebih tinggi pada dataset lain dalam studi yang sama menegaskan bahwa arsitektur yang sama dapat tampak unggul atau biasa bergantung dataset [14]. Pendekatan ensemble/hybrid kerap menggabungkan balancing (SMOTE), representasi (CNN/Transformer), dan classifier (LSTM/ML klasik); IDS-INT, misalnya, merancang pipeline imbalance-aware berbasis transfer learning Transformer yang dilengkapi XAI [15]; pendekatan berbasis Apache Spark juga dieksplorasi untuk klasifikasi serangan IoT pada data berskala besar [13]. Sementara itu, metode unsupervised/anomaly detection (Isolation Forest, Autoencoder) penting saat label terbatas dan untuk mendeteksi pola baru; OIFIDS melaporkan akurasi 94,8% pada UNSW-NB15 dan 99% pada CICIDS-2018, namun secara umum literatur menekankan trade-off berupa kenaikan false positive ketika sensitivitas terhadap serangan baru ditingkatkan [17].

Praktik Praproses, Integritas Dataset, dan Risiko Bias (RQ3)

Praktik praproses yang paling berpengaruh mencakup seleksi fitur, normalisasi/standarisasi, dan penanganan imbalance melalui resampling/SMOTE. Risiko utama adalah data leakage akibat fitur identitas (IP/port) yang tidak dibersihkan, serta oversampling yang merembes ke test set sehingga peningkatan metrik menjadi artefak. Audit CICIDS2017 mengungkap masalah pipeline yang dapat menginflasi atau mendistorsi evaluasi, dan menunjukkan bahwa perbaikan pemrosesan dapat mengubah proporsi data secara signifikan [4]. Studi lintas versi dataset juga menunjukkan bahwa model tertentu (misalnya Random Forest) dapat tampak robust terhadap variasi kualitas data yang memunculkan pertanyaan apakah metrik tinggi benar-benar mencerminkan kemampuan deteksi atau sekadar memanfaatkan artefak representasi [18]. Temuan ini menegaskan bahwa skor benchmark tidak boleh ditafsirkan tanpa uji yang lebih ketat seperti time-based split dan cross-dataset evaluation.

Dari Skor Benchmark ke Kesiapan Operasional (RQ4)

Dua pola kuat muncul dari sintesis. Pertama, banyak studi mencapai akurasi/F1 yang sangat tinggi pada CICIDS2017/NSL-KDD, tetapi studi audit menegaskan bahwa pipeline dataset dapat membawa artefak; karena itu puncak skor tanpa evaluasi ketat berisiko menyesatkan [4]. Kedua, dataset seperti UNSW-NB15 cenderung menghasilkan skor lebih rendah (~93% pada studi LS-SVM maupun CNN-LSTM) sehingga justru lebih informatif untuk membedakan metode, khususnya bila disertai pelaporan FPR dan latensi [14], [19]. Untuk data yang imbalanced, metrik seperti Matthews Correlation Coefficient (MCC) memberikan gambaran yang lebih seimbang dibanding accuracy maupun F1 [5]. Karena IDS secara operasional sensitif terhadap false alarm (biaya SOC) dan miss detection (risiko insiden), pelaporan minimal yang direkomendasikan adalah precision, recall/DR, F1, FPR, AUC, serta satu metrik tahan imbalance (MCC).

Berdasarkan gap 2020-2025, empat agenda riset paling berdampak diidentifikasi. (1) Validasi lintas dataset dan lintas waktu: standardisasi eksperimen dengan time-based split dan cross-dataset evaluation (latih pada A, uji pada B) untuk mengukur generalisasi secara nyata [4], [29]. (2) Standardisasi representasi fitur: studi perbandingan NetFlow vs CICFlowMeter menunjukkan pemilihan feature set memengaruhi generalisasi dan explainability sehingga perlu dilaporkan secara jelas [10]. (3) Reprodusibilitas sebagai syarat utama: publikasi kode, skrip praproses, versi dataset, dan konfigurasi pelatihan agar perbandingan adil. (4) Pelaporan biaya komputasi/latensi dan keamanan model: studi edge/federated learning sudah mulai memasukkan inference cost dan hal ini perlu dibakukan pula untuk NIDS enterprise [16], [22]. Untuk praktisi, prinsip praktis yang direkomendasikan adalah memulai dari baseline kuat yang murah (RF/XGBoost/LS-SVM) lalu mengukur trade-off terhadap FPR dan latensi [19]; mempertimbangkan keterbatasan perangkat dan opsi federated learning pada domain IoT/edge [22]; serta mengintegrasikan XAI (misalnya SHAP) ketika dibutuhkan interpretasi dan audit keputusan [10], [20], [30].

Keterbatasan Kajian

Asumsi utama kajian ini adalah bahwa metadata yang tersedia pada halaman penerbit (judul, abstrak, dan sebagian hasil) cukup untuk mengisi tabel komparatif. Untuk beberapa studi berbayar, detail konfigurasi lengkap (parameter, seed, perangkat keras) tidak selalu dapat diverifikasi pada teks penuh. Selain itu, angka pada rekapitulasi PRISMA disajikan sebagai rekap proses seleksi berbasis pencarian web lintas penerbit dan OJS; angka dapat berubah bila dilakukan ekspor indeks terintegrasi (Scopus/WoS) dan deduplikasi otomatis pada skala besar. Sesuai semangat PRISMA 2020, keterbatasan ini dinyatakan secara transparan [1].

KESIMPULAN

SLR 2020-2025 ini menunjukkan bahwa komparasi metode IDS berbasis ML tidak dapat disimpulkan hanya dari siapa yang mencapai akurasi tertinggi, karena tiga alasan: (i) dataset berperan besar dan dapat mengandung isu integritas/pipeline, (ii) protokol evaluasi tidak seragam sehingga meta-analisis kuantitatif penuh sering tidak valid, dan (iii) kebutuhan deployment menuntut metrik tambahan (FPR, latensi, biaya komputasi) serta explainability. ML klasik yang dituning tetap kompetitif dan ringan untuk data flow/tabular, sedangkan DL/hybrid unggul pada skenario kompleks dengan konsekuensi biaya komputasi yang lebih tinggi, dan metode unsupervised menjanjikan untuk serangan baru dengan trade-off kenaikan false positive. Audit CICIDS2017 dan studi integritas dataset menegaskan perlunya pergeseran dari benchmark score chasing menuju evaluasi yang lebih tahan bias dan relevan secara operasional. Penelitian mendatang disarankan membakukan time-based split dan cross-dataset evaluation, melaporkan metrik tahan imbalance seperti MCC, mempublikasikan kode serta versi dataset, dan mengintegrasikan XAI untuk akuntabilitas keputusan IDS.

DAFTAR PUSTAKA

[1] M. J. Page et al., “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” BMJ, vol. 372, art. n71, 2021.

[2] M. J. Page et al., “PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews,” BMJ, vol. 372, art. n160, 2021.

[3] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108-116.

[4] G. Engelen, V. Rimmer, and W. Joosen, "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study," in Proc. IEEE Security and Privacy Workshops (SPW), 2021, pp. 7-12.

[5] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, art. 6, 2020.

[6] N. Thapa, Z. Liu, D. B. KC, B. Gokaraju, and K. Roy, "Comparison of machine learning and deep learning models for network intrusion detection systems," Future Internet, vol. 12, no. 10, art. 167, 2020.

[7] J. L. Leevy and T. M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 big data," Journal of Big Data, vol. 7, art. 104, 2020.

[8] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, and C. F. M. Foozy, "Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset," IEEE Access, vol. 9, pp. 22351-22370, 2021.

[9] V. Kanimozhi and T. Prem Jacob, "Artificial intelligence outflanks all other machine learning classifiers in network intrusion detection on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing," ICT Express, vol. 7, no. 3, pp. 366-370, 2021.

[10] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection," Big Data Research, vol. 30, art. 100359, 2022.

[11] A. A. Alsulami, Q. Abu Al-Haija, A. Alqahtani, and R. Alsini, "Symmetrical simulation scheme for anomaly detection in IoT traffic with improved data engineering," Applied Sciences, vol. 12, no. 23, art. 12336, 2022.

[12] I. Tareq, B. M. Elbagoury, S. El-Regaily, and E. M. El-Horbaty, "Analysis of ToN-IoT, UNSW-NB15, and Edge-IIoT datasets using deep learning in cybersecurity for IoT," Applied Sciences, vol. 12, no. 19, art. 9572, 2022.

[13] F. Anwar et al., "Comparison of artificial intelligence algorithms for IoT intrusion detection on Apache Spark," Procedia Computer Science, vol. 201, pp. 745-751, 2022.

[14] H. C. Altunay and Z. Albayrak, "A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks," Engineering Science and Technology, an International Journal, vol. 38, art. 101322, 2023.

[15] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, "IDS-INT: intrusion detection system using transformer-based transfer learning for imbalanced network traffic," Digital Communications and Networks, vol. 10, no. 1, pp. 190-204, 2024.

[16] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "FlowTransformer: a transformer framework for flow-based network intrusion detection systems," Expert Systems with Applications, vol. 241, art. 122564, 2024.

[17] S. A. Elsaid and A. Binbusayyis, "An optimized isolation forest based intrusion detection system for heterogeneous and streaming data in IIoT networks," Discover Applied Sciences, vol. 6, art. 165, 2024.

[18] A. Pekar, H. K. Jozsa, and M. Domb, "Evaluating ML-based anomaly detection across datasets of varied integrity: a case study," Computer Networks, vol. 251, art. 110617, 2024.

[19] P. Waghmode et al., "Intrusion detection system based on machine learning using least square support vector machine," Scientific Reports, vol. 15, art. 95621, 2025.

[20] V. Z. Mohale and I. O. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," Frontiers in Computer Science, vol. 7, art. 1520741, 2025.

[21] I. U. Hewapathirana, "A comparative study of two-stage intrusion detection using modern machine learning approaches on the CSE-CIC-IDS2018 dataset," Telecom, vol. 6, no. 1, art. 6, 2025.

[22] N. Albanbay et al., "Federated learning-based intrusion detection in IoT networks: performance evaluation and data scaling study," Journal of Sensor and Actuator Networks, vol. 14, no. 4, art. 78, 2025.

[23] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1-6.

[24] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Generation Computer Systems, vol. 100, pp. 779-796, 2019.

[25] N. Moustafa, “A new distributed architecture for evaluating AI-based security systems at the edge: network TON_IoT datasets,” Sustainable Cities and Society, vol. 72, art. 102994, 2021.

[26] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1-6.

[27] H. Zmaimita, A. Amine, and R. M. Mohammed, “Machine and deep learning for intrusion detection: a PRISMA-guided systematic review,” Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 11, no. 1, 2025.

[28] H. M. R. U. Rehman et al., “A systematic literature study of machine learning techniques based intrusion detection: datasets, models, challenges, and future directions,” Journal of Big Data, vol. 12, no. 1, art. 264, 2025.

[29] M. Iwanowski, D. Olszewski, W. Graniszewski, J. Krupski, and F. Pelc, “The choice of training data and the generalizability of machine learning models for network intrusion detection systems,” Applied Sciences, vol. 15, no. 15, art. 8466, 2025.

[30] B. Nugraha, A. V. Jnanashree, and T. Bauschert, “A versatile XAI-based framework for efficient and explainable intrusion detection systems,” Annals of Telecommunications, vol. 80, no. 11-12, pp. 1095-1120, 2025.