Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

Deteksi Intrusi Jaringan Berbasis Machine Learning Menggunakan Model Boosting dengan Session-Level Feature Representation

Mochamad Sanwasih[1,*], Fajar Septian[2], Ristasari Dwi Septiana[3]

[1]Program Studi Teknologi Rekayasa Multimedia, Politeknik Digital Boash Indonesia, Bogor
Jl. Atang Senjaya No.KM 02, Bantarsari, Kec. Ranca Bungur, Kabupaten Bogor, Jawa Barat, Indonesia
[2]Program Studi Teknik Informatika, Jurusan Teknik Informatika dan Komputer, Politeknik Negeri Jakarta, Depok
Jl. Prof. DR. G.A. Siwabessy, Kukusan, Kecamatan Beji, Kota Depok, Jawa Barat, Indonesia
[3]Program Studi Teknik Informatika, Fakultas Teknologi, Institut Teknologi dan Bisnis Swadharma, Jakarta
Jl. Malaka No.3, RT.7/RW.3, Roa Malaka, Kec. Tambora, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta, Indonesia
Email: [1*]mochamadsanwasih11@gmail.com, [2]fajar.septian@lecturer.pnj.ac.id, [3]ristasari@swadharma.ac.id,
Email Penulis Korespondensi: mochamadsanwasih11@gmail.com
Submitted: **22/11/2025**; Accepted: **31/01/2026**; Published: **31/01/2026**

Abstrak–Meningkatnya kompleksitas ancaman keamanan jaringan menuntut sistem deteksi intrusi yang kontekstual dan adaptif. Pendekatan Intrusion Detection System (IDS) konvensional berbasis tanda tangan memiliki keterbatasan dalam mendeteksi pola serangan baru, sehingga pendekatan berbasis machine learning menjadi alternatif yang lebih fleksibel. Namun, representasi fitur pada level paket yang terfragmentasi masih membatasi kemampuan model dalam menangkap perilaku jaringan secara menyeluruh. Penelitian ini bertujuan untuk mengevaluasi kinerja model boosting XGBoost dan LightGBM menggunakan Cybersecurity Intrusion Detection Dataset yang tersedia secara publik di Kaggle, yang merepresentasikan aktivitas jaringan dalam bentuk sesi komunikasi. Pendekatan yang diusulkan mengembangkan session-level feature representation berbasis fitur agregat dan rasio untuk menangkap karakteristik perilaku jaringan secara lebih komprehensif. Hasil eksperimen menunjukkan bahwa penerapan session-level feature representation menghasilkan peningkatan konsisten pada berbagai metrik evaluasi. Nilai accuracy meningkat dari 0.8779 menjadi 0.8847, F1-score meningkat dari 0.8452 menjadi 0.8525 pada XGBoost dan dari 0.8455 menjadi 0.8523 pada LightGBM, serta ROC-AUC meningkat dari 0.8789 menjadi 0.8844 pada XGBoost dan dari 0.8793 menjadi 0.8859 pada LightGBM. Meskipun peningkatan akurasi relatif moderat, perbaikan pada F1-score dan ROC-AUC menunjukkan peningkatan kemampuan diskriminatif model serta keseimbangan yang lebih baik antara precision dan recall. Kontribusi utama penelitian ini terletak pada integrasi session-level feature engineering dengan model boosting dalam kerangka evaluasi yang sistematis, yang menekankan pentingnya kualitas representasi perilaku jaringan dalam meningkatkan performa deteksi intrusi.

Kata Kunci: Intrusion Detection System; Session-Level Feature Representation; Boosting; XGBoost; LightGBM

Abstract–The increasing complexity of network security threats demands intrusion detection systems that are both contextual and adaptive. Conventional signature-based Intrusion Detection Systems (IDS) suffer from limitations in detecting emerging and previously unseen attack patterns, making machine learning–based approaches a more flexible alternative. However, fragmented packet-level feature representations still limit the ability of models to capture network behavior comprehensively. This study aims to evaluate the performance of boosting models, namely XGBoost and LightGBM, using the publicly available Cybersecurity Intrusion Detection Dataset from Kaggle, which represents network activity at the session level. The proposed approach develops a session-level feature representation based on aggregated and ratio-based features to capture network behavior characteristics more comprehensively. Experimental results demonstrate that the implementation of session-level feature representation yields consistent improvements across multiple evaluation metrics. Accuracy increased from 0.8779 to 0.8847, while the F1-score improved from 0.8452 to 0.8525 for XGBoost and from 0.8455 to 0.8523 for LightGBM. Furthermore, ROC-AUC increased from 0.8789 to 0.8844 for XGBoost and from 0.8793 to 0.8859 for LightGBM. Although the improvement in accuracy is relatively moderate, the gains in F1-score and ROC-AUC indicate enhanced discriminative capability and a better balance between precision and recall. The main contribution of this study lies in the integration of session-level feature engineering with boosting models within a systematic evaluation framework, emphasizing the critical role of behavioral feature representation in improving intrusion detection performance.

Keywords: Intrusion Detection System; Session-Level Feature Representation; Boosting; XGBoost; LightGBM

1. PENDAHULUAN

Perkembangan teknologi jaringan komputer yang pesat telah meningkatkan ketergantungan berbagai sektor terhadap sistem berbasis jaringan, seperti layanan publik, industri, pendidikan, dan sistem informasi daring. Seiring dengan peningkatan tersebut, ancaman keamanan jaringan juga semakin kompleks, mencakup serangan malware, brute force, distributed denial of service (DDoS), serta berbagai aktivitas tidak sah lainnya [1]. Ancaman-ancaman ini berpotensi mengganggu ketersediaan layanan, merusak integritas data, dan menurunkan keandalan sistem jaringan, sehingga mekanisme Intrusion Detection System (IDS) menjadi komponen penting dalam menjaga keamanan jaringan komputer [2].

Pendekatan IDS konvensional umumnya berbasis tanda tangan (signature-based) atau aturan statis (rule-based). Meskipun efektif dalam mendeteksi serangan yang telah dikenal, pendekatan tersebut memiliki keterbatasan dalam menghadapi pola serangan baru, variasi perilaku serangan, serta dinamika lalu lintas jaringan yang terus berubah. Oleh karena itu, pendekatan berbasis machine learning semakin banyak digunakan karena kemampuannya dalam mempelajari pola kompleks dari data dan melakukan deteksi intrusi secara adaptif.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

Berbagai penelitian sebelumnya telah membahas penerapan machine learning dalam deteksi intrusi dan anomali jaringan komputer. Terdapat studi yang menerapkan algoritma Naïve Bayes untuk mendeteksi serangan Low Rate DDoS dan memperoleh akurasi tertinggi sebesar 83,45%, menunjukkan efektivitas model sederhana dengan latensi dan waktu komputasi yang rendah [3].

Penelitian lain menggunakan pendekatan machine learning berbasis data lalu lintas jaringan yang dikumpulkan secara langsung, dengan hasil evaluasi yang menunjukkan nilai recall yang tinggi namun akurasi yang masih bervariasi antar kelas trafik [4]. Studi komparatif terhadap Naïve Bayes, Support Vector Machine (SVM), dan Decision Tree menunjukkan bahwa Decision Tree mampu menghasilkan presisi hingga 0,99, diikuti oleh SVM dengan presisi 0,98 [5].

Penelitian selanjutnya yang berfokus pada deteksi serangan botnet menunjukkan bahwa SVM mencapai performa terbaik dengan akurasi 84% dan F1-score 80%, sementara Naïve Bayes cenderung kurang sensitif terhadap serangan [6]. Selain itu, penerapan Decision Tree, Random Forest, dan SVM pada sistem deteksi intrusi secara real-time menunjukkan bahwa Random Forest mampu mencapai akurasi di atas 95% dengan keseimbangan yang baik antara akurasi dan waktu pemrosesan [7].

Meskipun beberapa penelitian telah memanfaatkan model ensemble, kajian yang secara khusus mengevaluasi model boosting pada data jaringan masih relatif terbatas, terutama pada skenario yang mengintegrasikan session-level feature representation secara sistematis. Model boosting memiliki keunggulan dalam menggabungkan sejumlah weak learners untuk membentuk model prediksi yang lebih kuat, sehingga efektif dalam menangani hubungan non-linear antar fitur serta permasalahan ketidakseimbangan kelas [8].

Di antara berbagai teknik boosting, XGBoost (Extreme Gradient Boosting) dan LightGBM (Light Gradient Boosting Machine) dipilih karena kinerjanya yang kompetitif pada data tabular dan efisiensinya dalam proses pelatihan [9]. XGBoost unggul dalam mekanisme regularisasi yang mampu mencegah overfitting dan memodelkan hubungan fitur yang kompleks secara stabil [10], sedangkan LightGBM menawarkan efisiensi komputasi tinggi melalui pendekatan leaf-wise tree growth yang efektif untuk data berskala besar [11]. Namun demikian, keunggulan model boosting tersebut sangat bergantung pada kualitas representasi fitur yang digunakan [12]. Pada sebagian besar penelitian IDS, fitur masih direpresentasikan pada level paket (packet-level) atau aliran (flow-level), yang cenderung berdimensi tinggi dan terfragmentasi sehingga menyulitkan model machine learning dalam menangkap pola perilaku serangan secara utuh. Oleh karena itu, session-level feature representation menjadi pendekatan yang relevan karena mampu merangkum aktivitas jaringan ke dalam satuan sesi yang merepresentasikan interaksi antar entitas jaringan dalam periode waktu tertentu.

Perbedaan utama penelitian ini dengan penelitian terdahulu terletak pada integrasi antara session-level feature representation, model boosting (XGBoost dan LightGBM), serta penerapan strategi optimasi berbasis feature engineering dan pembelajaran sensitif biaya dalam satu kerangka evaluasi yang sistematis. Penelitian ini tidak hanya membandingkan kinerja model pada kondisi baseline, tetapi juga menganalisis secara kuantitatif dampak optimasi terhadap performa deteksi, dengan penekanan khusus pada peningkatan kemampuan model dalam mendeteksi serangan (recall), yang merupakan aspek krusial dalam sistem IDS untuk meminimalkan false negative.

Dalam penelitian ini, representasi sesi didefinisikan berdasarkan unit komunikasi yang telah tersedia pada dataset, di mana setiap entri data merepresentasikan satu sesi interaksi jaringan yang teragregasi. Setiap sesi mencerminkan aktivitas komunikasi antara entitas jaringan dalam suatu rentang waktu tertentu, yang secara konseptual sejalan dengan agregasi aliran berbasis parameter koneksi seperti sumber dan tujuan komunikasi, port, serta jenis protokol dalam suatu rentang waktu tertentu. Dengan menggunakan struktur sesi yang telah tersedia pada dataset, penelitian ini berfokus pada pengembangan fitur turunan berbasis agregasi dan rasio untuk memperkaya representasi perilaku jaringan pada level sesi. Selain itu, dalam konteks sistem IDS, kesalahan klasifikasi pada kelas serangan (false negative) memiliki konsekuensi yang lebih serius dibandingkan kesalahan pada kelas normal. Meskipun distribusi kelas pada dataset relatif seimbang, dalam konteks IDS kesalahan deteksi serangan (false negative) tetap memiliki biaya operasional dan risiko keamanan yang lebih tinggi dibandingkan kesalahan pada kelas normal. Oleh karena itu, pendekatan pembelajaran sensitif biaya (cost-sensitive learning) diterapkan untuk memberikan bobot lebih besar pada kesalahan klasifikasi kelas serangan, sehingga model lebih terfokus pada peningkatan sensitivitas terhadap pola intrusi tanpa mengorbankan stabilitas klasifikasi secara keseluruhan.

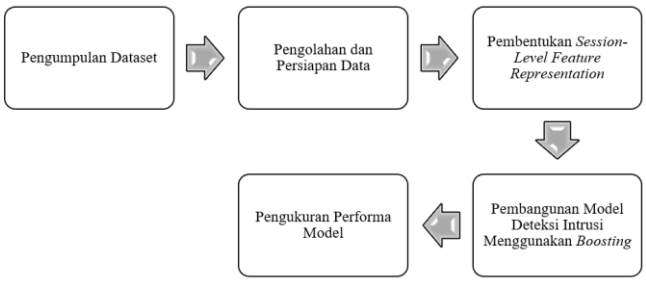
Tujuan dari penelitian ini adalah untuk mengevaluasi kinerja model boosting XGBoost dan LightGBM dalam mendeteksi intrusi jaringan berbasis session-level feature representation, serta menganalisis pengaruh penerapan feature engineering dan pembelajaran sensitif biaya terhadap performa model. Evaluasi dilakukan dengan membandingkan kinerja model pada skenario baseline dan skenario optimasi menggunakan metrik akurasi, presisi, recall, F1-score, dan ROC-AUC. Kontribusi utama dari penelitian ini adalah mengusulkan penerapan session-level feature representation untuk merepresentasikan perilaku jaringan secara lebih informatif dalam sistem deteksi intrusi, mengevaluasi kinerja model boosting XGBoost dan LightGBM pada data IDS berbasis representasi sesi, serta menganalisis dampak integrasi feature engineering dan pembelajaran sensitif biaya terhadap peningkatan kinerja deteksi, khususnya dalam meningkatkan kemampuan model dalam mendeteksi serangan.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pengembangan model deteksi intrusi jaringan berbasis machine learning dalam penelitian ini dilakukan melalui serangkaian tahapan yang saling terintegrasi dan disusun secara sistematis. Metode penelitian ini disusun agar pendekatan yang diusulkan dapat diterapkan kembali dan digunakan dalam konteks serupa [13]. Alur tahapan penelitian yang diterapkan dalam penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Prosedur Penelitian yang Diterapkan

Gambar 1 menyajikan tahapan penelitian yang diimplementasikan, dengan perincian sebagai berikut:

1) Pengumpulan Dataset
   Tahap awal penelitian dimulai dengan pengumpulan dataset lalu lintas jaringan yang digunakan untuk membangun model deteksi intrusi. Dataset yang digunakan merupakan dataset publik yang diperoleh dari platform Kaggle dengan nama Cybersecurity Intrusion Detection Dataset [14]. Dataset ini digunakan karena menyediakan data aktivitas jaringan yang telah dilabeli, sehingga sesuai untuk pengembangan model deteksi intrusi berbasis supervised learning. Dataset tersebut merepresentasikan aktivitas jaringan dalam suatu sesi komunikasi (session-based), di mana setiap baris data mencerminkan karakteristik perilaku jaringan selama satu sesi tertentu. Atribut yang tersedia dalam dataset meliputi session_id, login_attempts, failed_logins, session_duration, network_packet_size, ip_reputation_score, unusual_time_access, protocol_type, source_location, serta satu atribut target attack_detected. Atribut-atribut ini mencakup informasi terkait aktivitas autentikasi, durasi sesi, intensitas lalu lintas jaringan, karakteristik protokol, serta indikator perilaku anomali yang berpotensi mencerminkan aktivitas intrusi. Dalam dataset yang digunakan, sesi telah didefinisikan dan disediakan secara langsung oleh penyusun dataset sebagai unit komunikasi teragregasi. Oleh karena itu, penelitian ini tidak melakukan proses rekonstruksi sesi dari data mentah berbasis paket (misalnya menggunakan mekanisme timeout atau terminasi protokol TCP seperti FIN/RST). Setiap entri pada dataset telah merepresentasikan satu sesi komunikasi yang mencerminkan interaksi jaringan dalam suatu periode tertentu. Dengan demikian, kontribusi penelitian ini tidak terletak pada pembentukan sesi, melainkan pada pengayaan representasi fitur pada level sesi melalui feature engineering berbasis agregasi dan rasio.
2) Pengolahan dan Persiapan Data
   Tahap pengolahan dan persiapan data dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan dalam pelatihan model [15]. Atribut yang tidak memiliki nilai prediktif, seperti pengenal sesi, dihapus dari dataset. Selanjutnya, nilai hilang dan inkonsistensi data ditangani menggunakan strategi imputasi yang sesuai dengan karakteristik masing-masing atribut. Atribut kategorikal dikonversi ke bentuk numerik menggunakan one-hot encoding, sedangkan atribut numerik dipertahankan dalam skala aslinya karena algoritma boosting berbasis pohon keputusan tidak sensitif terhadap perbedaan skala fitur. Untuk menghindari data leakage, seluruh proses transformasi data dilakukan dengan pendekatan train-based fitting. Dataset terlebih dahulu dibagi menjadi data latih (80%) dan data uji (20%) menggunakan stratified split. Proporsi ini dipilih untuk memberikan data latih yang memadai bagi proses pembelajaran model sekaligus menjaga representativitas data uji [16]. Selanjutnya, perhitungan statistik yang digunakan dalam proses imputasi nilai hilang serta pembentukan representasi kategorikal melalui one-hot encoding dilakukan hanya pada data latih. Parameter transformasi yang diperoleh dari data latih kemudian diterapkan pada data uji tanpa melakukan perhitungan ulang. Pendekatan ini memastikan bahwa informasi dari data uji tidak memengaruhi proses pembelajaran model. Setelah proses pembagian data dan transformasi awal dilakukan, analisis eksploratif (exploratory data analysis) dilakukan pada data latih untuk memahami karakteristik distribusi fitur dan kelas target. Analisis ini mencakup distribusi kelas attack_detected, distribusi fitur-fitur kunci yang berkaitan dengan aktivitas autentikasi seperti failed_logins dan login_attempts, serta evaluasi awal feature importance. Hasil analisis tersebut digunakan sebagai dasar dalam perancangan session-level feature representation dan dalam menentukan strategi optimasi model pada tahap selanjutnya.
3) Pembentukan Session-Level Feature Representation
   Tahap selanjutnya adalah pembentukan session-level feature representation, yang merupakan kontribusi utama dalam penelitian ini. Pada tahap ini, aktivitas jaringan yang terjadi dalam satu sesi komunikasi

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

direpresentasikan dalam bentuk fitur agregat dan fitur turunan yang dirancang untuk menangkap pola perilaku jaringan secara lebih komprehensif. Beberapa fitur yang dibentuk pada level sesi meliputi rasio kegagalan autentikasi terhadap jumlah percobaan login, kepadatan lalu lintas terhadap durasi sesi, serta interaksi antara indikator anomali seperti jumlah koneksi mencurigakan dan waktu akses yang tidak lazim. Pendekatan ini bertujuan untuk mereduksi fragmentasi informasi yang umum terjadi pada representasi fitur level paket (packet-level) atau aliran (flow-level), sekaligus mengurangi kompleksitas dimensi data. Dengan merepresentasikan aktivitas jaringan dalam satuan sesi, model diharapkan mampu mengenali pola serangan secara lebih konsisten, karena setiap data mencerminkan perilaku jaringan yang utuh dalam periode waktu tertentu, bukan sekadar potongan aktivitas jaringan yang terpisah.

4) Pembangunan Model Deteksi Intrusi Menggunakan Boosting
Model deteksi intrusi pada penelitian ini dibangun menggunakan dua algoritma boosting, yaitu XGBoost dan LightGBM. Kedua algoritma ini dipilih karena kemampuannya dalam menggabungkan sejumlah weak learners untuk membentuk model prediksi yang kuat, serta efektif dalam menangani hubungan non-linear antar fitur dan ketidakseimbangan kelas pada data IDS. XGBoost dimanfaatkan karena memiliki mekanisme regularisasi yang efektif dalam mencegah overfitting dan stabil dalam memodelkan hubungan fitur yang kompleks [17], sedangkan LightGBM digunakan karena efisiensi komputasinya yang tinggi melalui pendekatan leaf-wise tree growth, sehingga cocok untuk data berukuran besar dan heterogen [18]. Pada tahap ini, model dilatih dalam dua skenario, yaitu skenario baseline dan skenario optimasi. Pada skenario baseline, model dilatih menggunakan fitur sesi dasar dengan parameter default dan ambang klasifikasi (threshold) sebesar 0,5. Selanjutnya, pada skenario optimasi, dilakukan penerapan feature engineering lanjutan, pembelajaran sensitif biaya (cost-sensitive learning) melalui pengaturan bobot kelas, serta penyesuaian ambang klasifikasi untuk meningkatkan keseimbangan antara presisi dan recall.

5) Pengukuran Performa Model
Tahap akhir penelitian adalah menguji performa model untuk menilai efektivitas pendekatan yang diusulkan dalam mendeteksi intrusi jaringan. Evaluasi dilakukan menggunakan data uji yang tidak terlibat dalam proses pelatihan model. Metrik evaluasi yang digunakan meliputi confusion matrix, akurasi, presisi, recall, F1-score, serta Receiver Operating Characteristic (ROC) Curve dan nilai Area Under the Curve (ROC-AUC). Confusion matrix digunakan untuk menganalisis kesalahan klasifikasi pada masing-masing kelas, sementara presisi dan recall digunakan untuk mengevaluasi ketepatan dan kelengkapan dalam klasifikasi. F1-score dihitung untuk menilai keseimbangan antara presisi dan recall [19]. Selain itu, ROC Curve digunakan untuk menggambarkan performa model pada berbagai nilai ambang klasifikasi, sedangkan nilai ROC-AUC digunakan sebagai indikator kemampuan diskriminatif model secara keseluruhan [20]. Evaluasi dilakukan dengan membandingkan kinerja model pada skenario baseline dan skenario optimasi, sehingga dampak penerapan session-level feature representation, feature engineering, dan pembelajaran sensitif biaya terhadap peningkatan kinerja deteksi intrusi dapat dianalisis secara komprehensif.

2.2 Session-Level Feature Representation

Pada penelitian ini, data lalu lintas jaringan direpresentasikan pada level sesi (session-level feature representation), di mana setiap sesi komunikasi diperlakukan sebagai satu unit observasi. Pendekatan ini bertujuan untuk menangkap pola perilaku jaringan secara lebih komprehensif dibandingkan representasi pada level paket (packet-level) atau aliran (flow-level), yang cenderung terfragmentasi dan berdimensi tinggi [21]. Representasi berbasis sesi memungkinkan peringkasan aktivitas jaringan yang terjadi dalam suatu periode interaksi antar entitas jaringan, sehingga karakteristik serangan dapat dikenali secara lebih konsisten.

Session-level feature representation dibangun melalui proses feature engineering dengan menurunkan fitur-fitur agregat dan rasio dari atribut mentah yang tersedia pada dataset. Beberapa fitur yang digunakan antara lain rasio kegagalan autentikasi terhadap total percobaan login, tingkat kepadatan lalu lintas terhadap durasi sesi, serta interaksi antara indikator reputasi dan waktu akses yang tidak lazim. Secara umum, fitur rasio dan laju aktivitas dihitung menggunakan persamaan (1):

$$f_{ratio} = \frac{x_a}{x_b + \epsilon} \quad (1)$$

di mana x_a dan x_b masing-masing merepresentasikan atribut aktivitas tertentu dalam satu sesi, dan ϵ adalah konstanta kecil untuk menghindari pembagian dengan nol. Selain itu, fitur berbasis laju aktivitas dihitung menggunakan persamaan (2):

$$f_{rate} = \frac{x}{t + \epsilon} \quad (2)$$

di mana x menyatakan volume atau intensitas aktivitas jaringan dan tadalah durasi sesi. Melalui pendekatan ini, setiap sesi jaringan direpresentasikan sebagai vektor fitur yang lebih informatif dan ringkas, sehingga mendukung pembelajaran pola serangan secara lebih efektif oleh model boosting.

2.3 Extreme Gradient Boosting (XGBoost)

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

Extreme Gradient Boosting (XGBoost) merupakan algoritma ensemble learning berbasis gradient boosting yang membangun model prediksi secara iteratif dengan menggabungkan sejumlah pohon keputusan sebagai weak learners [22]. Setiap iterasi bertujuan untuk memperbaiki kesalahan prediksi dari model sebelumnya dengan meminimalkan fungsi kerugian (loss function) yang telah ditentukan [23]. Secara umum, model XGBoost meminimalkan fungsi objektif yang dirumuskan pada persamaan (3).

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

di mana $l(y_i, \hat{y}_i)$ adalah fungsi kerugian antara label aktual y_i dan prediksi $\hat{y}_i$, sedangkan $\Omega(f_k)$ merupakan fungsi regularisasi untuk pohon ke-k yang bertujuan mengontrol kompleksitas model. Fungsi regularisasi pada XGBoost dirumuskan sebagai persamaan (4).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (4)$$

di mana T menyatakan jumlah daun pada pohon dan $w_j$adalah bobot daun ke-j. Mekanisme regularisasi ini memungkinkan XGBoost mengurangi risiko overfitting dan meningkatkan generalisasi model.

Dalam penelitian ini, XGBoost digunakan sebagai model klasifikasi biner untuk mendeteksi intrusi jaringan. Untuk menangani ketidakseimbangan kelas, diterapkan pendekatan cost-sensitive learning melalui parameter scale_pos_weight, yang memberikan penalti lebih besar pada kesalahan klasifikasi kelas serangan. Kombinasi antara regularisasi dan pembobotan kelas menjadikan XGBoost efektif dalam memodelkan hubungan non-linear antar fitur pada data IDS berbasis session-level.

2.4 Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) merupakan algoritma gradient boosting yang dirancang untuk meningkatkan efisiensi komputasi pada data berskala besar dan berdimensi tinggi. Berbeda dengan XGBoost yang menggunakan pendekatan level-wise tree growth, LightGBM menerapkan strategi leaf-wise tree growth, di mana node dengan penurunan loss terbesar dipilih untuk dikembangkan terlebih dahulu [24]. Pendekatan ini memungkinkan pembelajaran yang lebih cepat dan akurat dengan jumlah iterasi yang lebih sedikit [25].

Secara matematis, proses pembelajaran LightGBM tetap didasarkan pada minimisasi fungsi kerugian yang serupa dengan gradient boosting pada umumnya, seperti yang ditunjukkan pada persamaan (5).

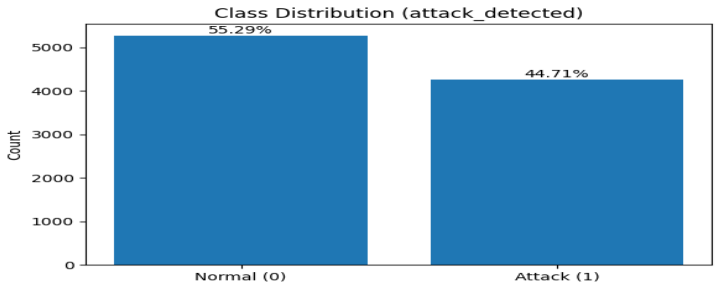
$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (5)$$

di mana $f_k(x_i)$ adalah kontribusi pohon ke-k terhadap prediksi data ke-i. Untuk mengatasi ketidakseimbangan kelas pada data IDS, LightGBM menyediakan mekanisme pembobotan kelas melalui parameter is_unbalance, yang secara otomatis menyesuaikan bobot kelas berdasarkan distribusi data latih.

Keunggulan LightGBM dalam efisiensi komputasi dan kemampuannya menangani fitur heterogen menjadikannya sesuai untuk diterapkan pada data IDS berbasis session-level feature representation. Dalam penelitian ini, LightGBM digunakan sebagai pembanding terhadap XGBoost untuk mengevaluasi konsistensi dan stabilitas performa model boosting dalam mendeteksi intrusi jaringan.

3. HASIL DAN PEMBAHASAN

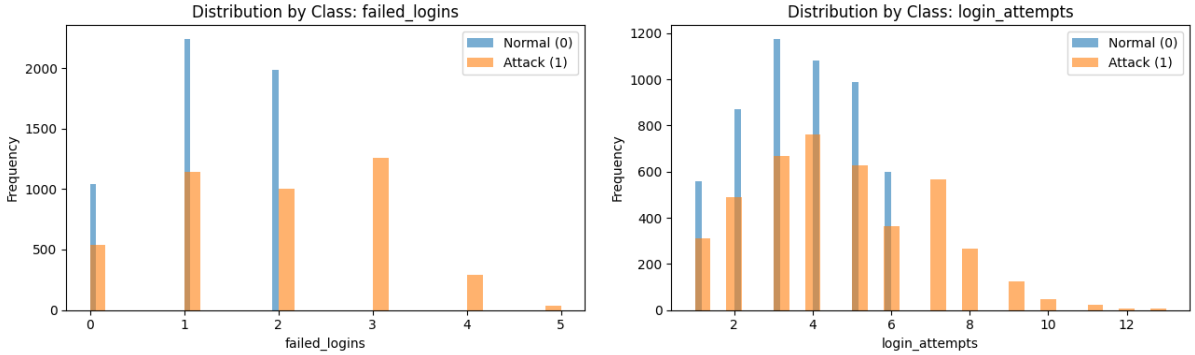
Penelitian ini berfokus pada pembangunan dan evaluasi model deteksi intrusi jaringan berbasis machine learning menggunakan pendekatan boosting dengan session-level feature representation. Dataset yang digunakan merupakan dataset publik dari Kaggle berjudul Cybersecurity Intrusion Detection Dataset [14], yang berisi data aktivitas jaringan dengan label trafik normal dan trafik serangan. Setiap entri data merepresentasikan satu sesi komunikasi jaringan dan memuat atribut numerik serta kategorikal yang menggambarkan karakteristik autentikasi, durasi sesi, intensitas lalu lintas, dan indikator reputasi jaringan. Sebelum membahas performa model, dilakukan analisis awal terhadap distribusi kelas target untuk memahami tingkat ketidakseimbangan data yang digunakan. Distribusi kelas target ditunjukkan pada Gambar 2.



Gambar 2. Distribusi Kelas attack_detected

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

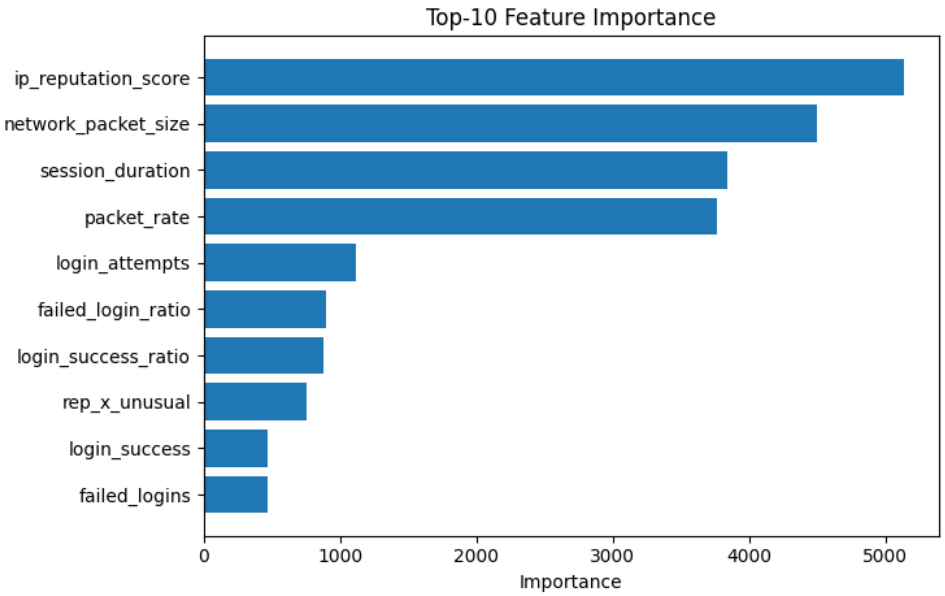
Gambar 2 memperlihatkan bahwa dataset yang digunakan memiliki distribusi kelas yang relatif seimbang, meskipun proporsi kelas Normal (0) sedikit lebih dominan dibandingkan kelas Attack (1). Kelas Normal mencakup 55,29% dari total data, sedangkan kelas Attack sebesar 44,71%. Komposisi ini masih mencerminkan kondisi yang wajar pada data lalu lintas jaringan nyata, di mana aktivitas normal umumnya lebih banyak dibandingkan aktivitas serangan. Untuk memperdalam pemahaman terhadap perilaku fitur, dilakukan analisis distribusi beberapa atribut penting berdasarkan kelas target. Distribusi fitur failed_logins dan login_attempts ditampilkan pada Gambar 3.



(a) (b)

Gambar 3. (a) Distribusi Failed Logins dan (b) Login Attempts Berdasarkan Kelas Target

Gambar 3 (a) menunjukkan bahwa pada kelas Attack (1) terdapat kecenderungan jumlah failed_logins yang lebih tinggi dibandingkan kelas Normal (0). Pola ini mengindikasikan bahwa aktivitas login gagal yang berulang merupakan salah satu indikator penting dalam mendeteksi potensi serangan. Sedangkan Gambar 3 (b) terlihat bahwa kelas Attack (1) memiliki sebaran login_attempts yang lebih luas dan nilai maksimum yang lebih tinggi dibandingkan kelas normal. Hal ini menunjukkan bahwa serangan cenderung melibatkan percobaan login berulang dalam satu sesi, sehingga fitur ini berkontribusi signifikan dalam membedakan perilaku normal dan anomali. Untuk mengidentifikasi fitur yang paling berpengaruh dalam proses klasifikasi, dilakukan analisis feature importance menggunakan model LightGBM. Hasil peringkat sepuluh fitur teratas ditampilkan pada Gambar 4.



Gambar 4. Sepuluh Fitur Terpenting

Gambar 4 menunjukkan bahwa fitur ip_reputation_score memiliki kontribusi paling dominan dalam proses klasifikasi, diikuti oleh network_packet_size, session_duration, dan packet_rate. Dominasi fitur reputasi IP menunjukkan bahwa karakteristik sumber koneksi memiliki peran krusial dalam mendeteksi serangan jaringan. Selain itu, fitur berbasis perilaku sesi seperti durasi sesi dan laju paket juga memberikan informasi penting terkait pola aktivitas yang mencurigakan. Temuan ini memperkuat asumsi bahwa kombinasi fitur reputasi dan fitur perilaku jaringan efektif untuk membangun sistem deteksi intrusi berbasis machine learning.

Pada penelitian ini, data mentah yang diperoleh berasal dari aktivitas jaringan dan autentikasi pengguna yang bersifat event-level, seperti percobaan login, paket jaringan, dan durasi koneksi. Agar pola serangan dapat direpresentasikan secara lebih komprehensif, seluruh data mentah tersebut ditransformasikan ke dalam bentuk session-level feature representation. Untuk memperjelas proses pembentukan fitur berbasis sesi, disajikan sebuah

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

studi kasus sederhana yang diambil dari pola data pada dataset yang digunakan. Misalkan terdapat satu sesi komunikasi jaringan dengan karakteristik yang ditunjukkan pada Tabel 1.

Tabel 1. Sampel Data Satu Sesi Jaringan

No	Atribut	Nilai	Satuan
1	Login Attempts	5	kali
2	Failed Logins	2	kali
3	Network Packet Size	4	byte
4	Session Duration	40	detik

Berdasarkan data sesi jaringan yang disajikan pada Tabel 1, selanjutnya dilakukan perhitungan fitur turunan berbasis rasio dan laju aktivitas untuk membentuk session-level feature representation. Rasio kegagalan login dihitung menggunakan Persamaan (1) dengan hasil sebagai berikut:

$$f_{ratio} = \frac{2}{5+10^{-6}} = 0{,}40$$

Nilai rasio sebesar 0,40 menunjukkan bahwa 40% dari total percobaan login pada sesi tersebut mengalami kegagalan, yang dapat mengindikasikan adanya aktivitas autentikasi yang mencurigakan, seperti brute-force attack. Selanjutnya, laju aktivitas lalu lintas jaringan dihitung menggunakan persamaan (2) dengan hasil sebagai berikut:

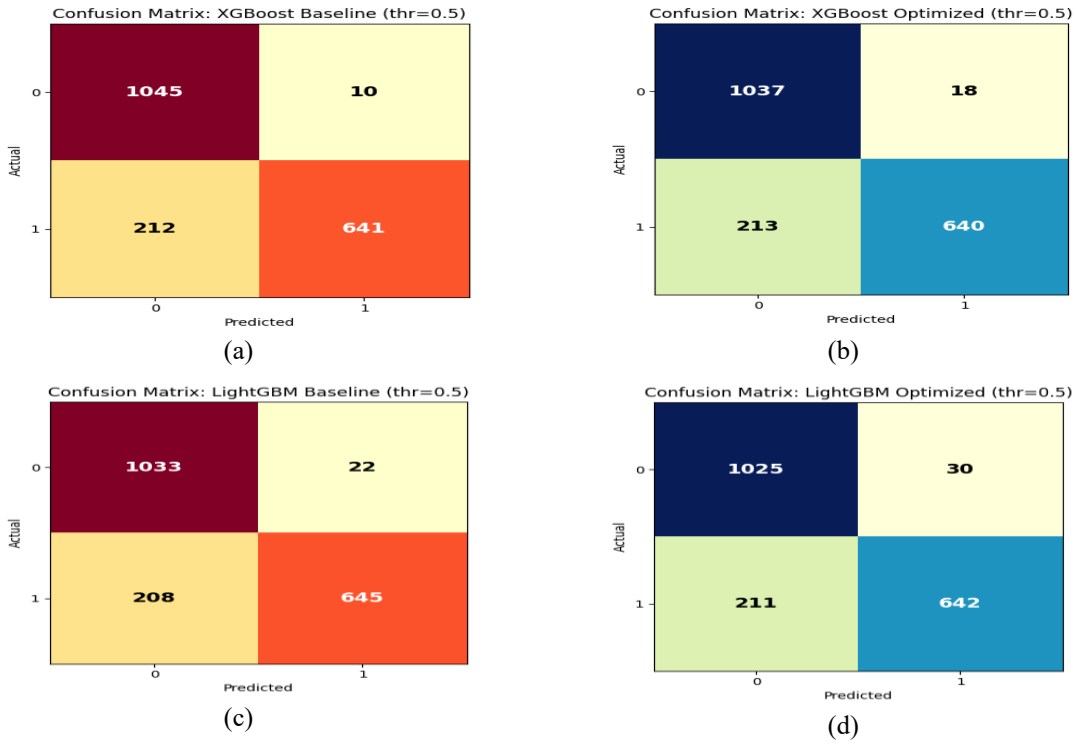
$$f_{rate} = \frac{4000}{40+10^{-6}} = 100 \text{ byte/detik}$$

Nilai packet rate sebesar 100 byte/detik menunjukkan kepadatan lalu lintas jaringan dalam satu sesi. Nilai yang relatif tinggi pada durasi singkat dapat mengindikasikan pola anomali, terutama jika dikombinasikan dengan indikator lain seperti kegagalan login atau reputasi IP yang rendah. Berdasarkan perhitungan tersebut, sesi jaringan ini direpresentasikan dalam bentuk vektor fitur tingkat sesi sebagai berikut:

$$S = [f_{ratio} = 0{,}40;\ f_{rate} = 100; \text{fitur sesi lainnya}]$$

Vektor fitur ini kemudian digunakan sebagai input bagi model boosting untuk proses pelatihan dan pengujian. Dengan pendekatan ini, informasi perilaku jaringan yang semula tersebar pada level paket atau aliran berhasil dirangkum ke dalam representasi tingkat sesi yang lebih kompak dan informatif, sehingga memudahkan model XGBoost dan LightGBM dalam mempelajari pola serangan secara lebih konsisten.

Evaluasi kinerja model dilakukan dengan membandingkan skenario baseline dan optimized (dengan session-level feature representation) pada algoritma XGBoost dan LightGBM. Confusion matrix untuk masing-masing model ditampilkan pada Gambar 5.

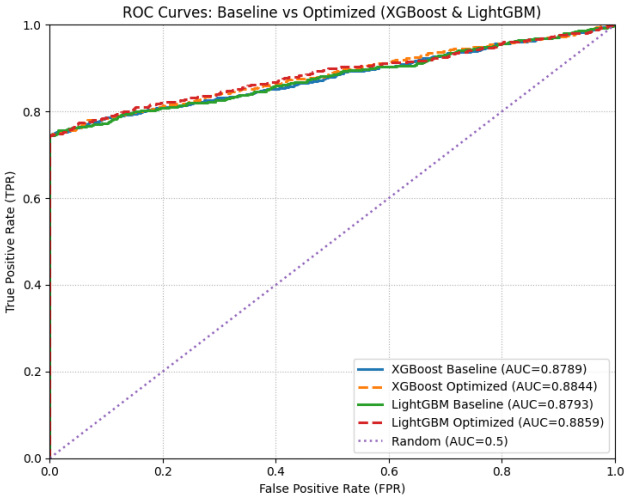


Gambar 5. Confusion Matrix Model (a) XGBoost Baseline, (b) XGBoost Optimized, (c) LightGBM Baseline, dan (d) LightGBM Optimized

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

Pada Gambar 5 (a), model XGBoost baseline menunjukkan jumlah false negative yang relatif tinggi, yaitu 212 kasus serangan yang salah diklasifikasikan sebagai normal. Setelah dilakukan optimasi hiperparameter, sebagaimana ditunjukkan pada Gambar 5 (b), terjadi peningkatan dalam ketepatan prediksi kelas serangan, meskipun masih terdapat trade-off berupa peningkatan false positive. Hal ini menunjukkan bahwa optimasi lebih menekankan peningkatan sensitivitas terhadap serangan. Hasil pada Gambar 5 (c) dan 5 (d) menunjukkan pola yang serupa pada LightGBM. Model baseline mampu mengklasifikasikan kelas normal dengan baik, namun masih memiliki keterbatasan dalam mendeteksi serangan. Setelah optimasi, jumlah prediksi benar pada kelas serangan meningkat, meskipun disertai peningkatan false positive pada kelas normal.

Kemampuan diskriminatif model dievaluasi secara menyeluruh melalui analisis Receiver Operating Characteristic (ROC). Perbandingan kurva ROC antara model baseline dan model yang telah dioptimasi disajikan pada Gambar 6.



Gambar 6. Perbandingan ROC Curve XGBoost dan LightGBM

Gambar 6 memperlihatkan bahwa seluruh model yang diuji memiliki kinerja yang secara konsisten lebih baik dibandingkan klasifikasi acak. Model XGBoost pada skenario optimasi mencapai nilai ROC-AUC sebesar 0,8844, meningkat dari 0,8789 pada kondisi baseline. Sementara itu, LightGBM pada kondisi optimasi menunjukkan performa terbaik dengan nilai ROC-AUC sebesar 0,8859, melampaui LightGBM baseline (0,8793) maupun XGBoost. Peningkatan nilai ROC-AUC ini menunjukkan bahwa strategi optimasi dan pengayaan fitur pada level sesi meningkatkan kemampuan model dalam membedakan trafik normal dan serangan pada berbagai ambang batas keputusan.

Gambaran kuantitatif yang lebih komprehensif terkait kinerja masing-masing model diperoleh melalui perbandingan metrik evaluasi yang dirangkum dalam Tabel 2.

Tabel 2. Perbandingan Kinerja Model Baseline dan Optimized

Model	Setting	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	Baseline	0.8779	0.9755	0.7456	0.8452	0.8789
XGBoost	Optimized	**0.8847**	0.9953	0.7456	**0.8525**	0.8844
LightGBM	Baseline	0.8768	0.9626	**0.7538**	0.8455	0.8793
LightGBM	Optimized	**0.8847**	**0.9969**	0.7444	0.8523	**0.8859**

Berdasarkan hasil yang dirangkum pada Tabel 2, LightGBM pada skenario optimized (dengan session-level feature representation) menunjukkan kinerja terbaik secara keseluruhan, ditandai dengan nilai ROC-AUC tertinggi sebesar 0,8859 dan akurasi sebesar 0,8847. Capaian ini mengindikasikan bahwa LightGBM optimized memiliki kemampuan diskriminatif paling stabil dalam membedakan trafik normal dan serangan pada berbagai ambang keputusan. Posisi berikutnya ditempati oleh XGBoost optimized, yang menghasilkan akurasi 0,8847 dengan nilai ROC-AUC sebesar 0,8844, diikuti oleh LightGBM baseline (0,8793) dan XGBoost baseline (0,8789).

Keunggulan LightGBM optimized terutama disebabkan oleh karakteristik algoritma leaf-wise tree growth yang lebih adaptif dalam mengeksplorasi hubungan non-linear antar fitur. Pendekatan ini memungkinkan LightGBM menangkap pola kompleks pada data IDS yang bersifat heterogen dan mengandung interaksi fitur yang saling bergantung. Selain itu, penggunaan session-level feature representation berperan penting dalam meningkatkan kinerja kedua model boosting. Representasi berbasis sesi mampu merangkum aktivitas jaringan secara lebih kontekstual melalui fitur agregat dan rasio, seperti rasio kegagalan login, kepadatan lalu lintas terhadap durasi sesi, serta interaksi indikator reputasi dan pola akses. Pendekatan ini mengurangi fragmentasi informasi yang umum terjadi pada representasi packet-level atau flow-level, sehingga pola serangan dapat dikenali secara lebih konsisten. Peningkatan ini tercermin pada kenaikan nilai F1-score dan ROC-AUC pada skenario optimasi.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489–498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

Menariknya, pada LightGBM skenario optimized terlihat adanya trade-off yang jelas antara precision dan recall. Precision meningkat secara signifikan dari 0,9626 menjadi 0,9969, sementara recall sedikit menurun dari 0,7538 menjadi 0,7444. Fenomena ini menunjukkan bahwa model menjadi lebih konservatif dalam memprediksi kelas serangan, yakni hanya memberikan label "Attack" ketika tingkat keyakinan prediksi sangat tinggi. Akibatnya, jumlah false positive berkurang secara signifikan, namun terdapat risiko peningkatan false negative dalam skala kecil. Dalam konteks sistem Intrusion Detection System (IDS), trade-off semacam ini dapat diterima tergantung pada kebutuhan operasional sistem. Peningkatan precision yang sangat tinggi berarti sistem menghasilkan alarm palsu yang jauh lebih sedikit, sehingga mengurangi beban analisis lanjutan dan meningkatkan efisiensi operasional. Meskipun terjadi sedikit penurunan recall, nilai F1-score tetap meningkat, yang menunjukkan bahwa keseimbangan keseluruhan antara precision dan recall tetap membaik. Oleh karena itu, konfigurasi optimized pada LightGBM lebih sesuai untuk lingkungan operasional yang memprioritaskan stabilitas sistem dan minimisasi false alarm, sementara konfigurasi baseline dapat dipertimbangkan pada skenario yang lebih menekankan sensitivitas terhadap deteksi serangan.

Secara keseluruhan, hasil penelitian menunjukkan bahwa kombinasi session-level feature representation dengan model boosting memberikan peningkatan kemampuan diskriminatif yang konsisten. Pemilihan model akhir tetap perlu mempertimbangkan kebutuhan operasional IDS, apakah lebih menekankan pada pengurangan alarm palsu atau peningkatan sensitivitas terhadap serangan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan model boosting berbasis machine learning dengan session-level feature representation mampu meningkatkan kinerja deteksi intrusi jaringan secara kuantitatif dan konsisten. Hasil pengujian memperlihatkan bahwa skenario optimasi menghasilkan peningkatan accuracy dari 0,8779 menjadi 0,8847, F1-score dari 0,8452 menjadi 0,8525 pada XGBoost dan dari 0,8455 menjadi 0,8523 pada LightGBM, serta ROC-AUC dari 0,8789 menjadi 0,8844 pada XGBoost dan dari 0,8793 menjadi 0,8859 pada LightGBM. LightGBM pada kondisi optimasi memberikan performa terbaik secara keseluruhan dengan nilai ROC-AUC tertinggi sebesar 0,8859, yang menunjukkan kemampuan pemisahan kelas paling stabil pada berbagai ambang keputusan. Sementara itu, XGBoost pada kondisi optimasi menghasilkan nilai precision tertinggi sebesar 0,9953, yang merefleksikan kemampuan menekan false positive secara signifikan. Nilai recall pada XGBoost tetap berada pada 0,7456, sedangkan pada LightGBM berubah dari 0,7538 menjadi 0,7444, yang mengindikasikan adanya trade-off antara peningkatan precision dan sensitivitas deteksi serangan. Meskipun recall pada LightGBM sedikit menurun, peningkatan ROC-AUC menunjukkan bahwa stabilitas kemampuan diskriminatif model secara keseluruhan tetap mengalami perbaikan. Peningkatan performa ini dipengaruhi oleh penggunaan session-level feature representation yang merangkum perilaku jaringan dalam satuan sesi melalui fitur agregat dan rasio, sehingga mengurangi fragmentasi informasi yang umum terjadi pada representasi packet-level atau flow-level. Pendekatan ini memungkinkan model menangkap pola serangan secara lebih kontekstual dan konsisten. Meskipun demikian, penelitian ini masih memiliki keterbatasan karena menggunakan dataset statis dan belum mengakomodasi skenario deteksi intrusi secara real-time maupun threshold adaptif. Penelitian selanjutnya disarankan untuk mengintegrasikan mekanisme threshold dinamis, melakukan evaluasi lintas dataset, serta menerapkan pendekatan stream-based atau online learning guna meningkatkan sensitivitas model terhadap pola serangan yang terus berkembang.

REFERENCES

[1] P. A. Khairunnisa, N. Annisa, Y. Yukandri, and J. Parhusip, "Perancangan Sistem Keamanan Jaringan Berbasis Cybersecurity untuk Mitigasi Ancaman Siber pada Infrastruktur TI: Studi Kasus di Indonesia," Tek. J. Ilmu Tek. dan Inform., vol. 4, no. 2, pp. 9–16, 2024, doi: https://doi.org/10.51903/teknik.v3i1.570.

[2] T. Widodo and A. S. Aji, "Pemanfaatan Network Forensic Investigation Framework untuk Mengidentifikasi Serangan Jaringan Melalui Intrusion Detection System (IDS)," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 7, no. 1, pp. 46–55, 2022, doi: https://doi.org/10.14421/jiska.2022.7.1.46-55.

[3] D. Firdaus, F. Fahira, and R. Rianti, "Deteksi Anomali dan Serangan Low Rate Ddos Dalam Lalu Lintas Jaringan Menggunakan Naive Bayes," NARATIF J. Ilm. Nas. Ris. Apl. dan Tek. Inform., vol. 05, no. 02, pp. 140–148, 2023, doi: https://doi.org/10.53580/naratif.v5i2.208.

[4] C. Chandra, D. Prima, and F. Faradika, "Deteksi Serangan Siber Menggunakan Machine Learning: Studi Pada Sistem Informasi Akademik," JISKA J. Sist. Inf. Dan Inform., vol. 3, no. 2, pp. 106–110, 2025, doi: https://doi.org/10.47233/jiska.v3i2.2139.

[5] A. T. Zy, A. T. Sasongko, and A. Z. Kamalia, "Penerapan Naïve Bayes Classifier, Support Vector Machine, dan Decision Tree untuk Meningkatkan Deteksi Ancaman Keamanan Jaringan," KLIK Kaji. Ilm. Inform. dan Komput., vol. 4, no. 1, pp. 610–617, 2023, doi: https://doi.org/10.30865/klik.v4i1.1134.

[6] R. Rio and K. Handoko, "Analisis Perbandingan Kinerja Algoritma Machine Learning Berbasis Feature Selection Dalam Deteksi Serangan Botnet," J. Comasie, vol. 12, no. 2, pp. 139–148, 2025, doi: https://doi.org/10.33884/comasiejournal.v12i2.9778.

[7] D. P. Sari, Z. Halim, B. Waseso, and S. Saromah, "Implementasi Machine Learning untuk Deteksi Intrusi pada Jaringan Komputer," J. Minfo Polgan, vol. 13, no. 2, pp. 1389–1394, 2024, doi: https://doi.org/10.33395/jmp.v13i2.14074.

Journal of Information System Research (JOSH)
Volume 7, No. 2, Januari 2026, pp 489−498
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i2.8775

[8] W. Alexander and D. Jollyta, "Sistem Deteksi Intrusi Pada Jaringan Komputer Menggunakan Algoritma XGBoost," J. Mhs. Apl. Teknol. Komput. dan Inf., vol. 7, no. 2, pp. 128–134, 2025, doi: https://doi.org/10.35145/jmapteksi.v7i2.4969.
[9] F. Pratama, E. Ali, R. Rahmaddeni, and W. Agustin, "Perbandingan Kinerja XGBoost dan LightGBM Dalam Klasifikasi Depresi Pada Mahasiswa Berdasarkan Faktor Demografi dan Akademik," J. Algoritm., vol. 22, no. 2, pp. 53–64, 2025, doi: https://doi.org/10.33364/algoritma/v.22-2.2439.
[10] R. Gunawan, E. S. Handika, and E. Ismanto, "Pendekatan Machine Learning Dengan Menggunakan Algoritma XGBoost (Extreme Gradient Boosting) Untuk Peningkatan Kinerja Klasifikasi Serangan Syn," J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 453–463, 2022, doi: https://doi.org/10.37859/coscitech.v3i3.4356.
[11] T. Wulandari, Y. F. Yudisthira, K. Chika, M. A. Wibowo, and C. Andrew, "Algoritma LightGBM untuk Deteksi Aktivitas Cyber Espionage Melalui Dataset Serangan Siber," JUSIFOR J. Sist. Inf. dan Inform., vol. 4, no. 1, pp. 213–219, 2025, doi: https://doi.org/10.70609/jusifor.v4i2.8489.
[12] A. R. Kamila, F. Adikara, S. Sutrisno, and C. Herdian, "Analisa Pengaruh Penambahan Fitur dengan Perbandingan Algoritma berbasis Bagging dan Boosting pada Deteksi Phishing Link," JEPIN (Jurnal Edukasi dan Penelit. Inform., vol. 10, no. 3, pp. 460–467, 2024, doi: https://doi.org/10.26418/jp.v10i3.83366.
[13] I. Ahmad, Y. Rahmanto, R. I. Borman, F. Rossi, Y. Jusman, and A. D. Alexander, "Identification of Pineapple Disease Based on Image Using Neural Network Self-Organizing Map (SOM) Model," in International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), 2022, pp. 12–17. doi: 10.1109/ICE3IS56585.2022.10010110.
[14] D. N. K. Samudrala, "Cybersecurity Intrusion Detection Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/dnkumars/cybersecurity-intrusion-detection-dataset
[15] Parjito, I. Ahmad, R. I. Borman, A. D. Alexander, and Y. Jusman, "Combining Extreme Learning Machine and Linear Discriminant Analysis for Optimized Apple Leaf Disease Classification," in International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), IEEE, 2024, pp. 138–143. doi: 10.1109/ICE3IS62977.2024.10775844.
[16] K. Shah, S. Shah, V. Shah, and A. Godbole, "Optimization of Data Splitting Methods For Machine Learning," in International Conference on Innovative Computing & Communication (ICICC), 2025. doi: https://dx.doi.org/10.2139/ssrn.5190348.
[17] S. E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBoost) pada Klasifikasi Nasabah Kartu Kredit," J. Math. Theory Appl., vol. 4, no. 1, pp. 21–26, 2022, doi: https://doi.org/10.31605/jomta.v4i1.1792.
[18] I. Z. A. Illah, W. S. J. Sapu, and A. T. Damaliana, "Implementasi Metode Klasifikasi LightGBM dan Analisis Survival dalam Memprediksi Pelanggan Churn," J. Komtika (Komputasi dan Inform., vol. 8, no. 1, pp. 43–53, 2024, doi: https://doi.org/10.31603/komtika.v8i1.11194.
[19] Z. Abidin, R. I. Borman, F. B. Ananda, P. Prasetyawan, F. Rossi, and Y. Jusman, "Classification of Indonesian Traditional Snacks Based on Image Using Convolutional Neural Network (CNN) Algorithm," in International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), 2021, pp. 18–23. doi: 10.1109/ICE3IS54102.2021.9649707.
[20] Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks," J. Stat. Comput. Simul., vol. 94, no. 1, pp. 183–203, Jan. 2024, doi: 10.1080/00949655.2023.2238235.
[21] T. Chen and R. C.-W. Wong, "Improving Representation Learning for Session-based Recommendation," in IEEE International Conference on Big Data (Big Data), 2022, pp. 854–863. doi: 10.1109/BigData55660.2022.10020851.
[22] D. T. Murdiansyah, "Prediksi Stroke Menggunakan Extreme Gradient Boosting," JIKO (Jurnal Inform. dan Komputer), vol. 8, no. 2, p. 419, 2024, doi: 10.26798/jiko.v8i2.1295.
[23] D. R. Arrayyan, R. G. Guntara, and M. R. Nugraha, "Deteksi Komentar Spam Judi Online Berbahasa Indonesia Menggunakan XGBoost dan TF-IDF," J. Algoritm., vol. 22, no. 2, pp. 2066–2075, 2025, doi: 10.33364/algoritma/v.22-2.3012.
[24] N. I. Sulistiyani, R. A. Purnomo, N. Rohmatunisa, R. M. Pujo, R. B. Bahaweres, and N. Hakiem, "Prediksi Jenis Kebocoran Data Kesehatan di AS Berdasarkan Laporan HIPAA Menggunakan LightGBM dan Kerangka OSEMN," INDEXIA Inform. Comput. Intell. J., vol. 07, no. 02, pp. 110–119, 2025, doi: https://doi.org/10.30587/indexia.v7i2.10220.
[25] S. Y. Nailendra, W. Witanti, and G. Abdillah, "Optimasi Prediksi Penjualan Retail Online Menggunakan LightGBM dan Hyperparameter Tuning," J. Algoritm., vol. 22, no. 2, pp. 1931–1942, 2025, doi: 10.33364/algoritma/v.22-2.2551.