KAMPUS AKADEMIK PUBLISING
Jurnal Multidisiplin Ilmu Akademik
Vol.2, No.6 Desember 2025
e-ISSN: 3032-7377; p-ISSN: 3032-7385, Hal 632-640
DOI: https://doi.org/10.61722/jmia.v2i6.7213

PENERAPAN ALGORITMA RANDOM FOREST DAN K-NEAREST NEIGHBOR UNTUK DETEKSI INTRUSI PADA DATASET CICIDS2017

Mujiono
Politeknik Negeri Jember
Devita Ayu Larasati
Universitas Jember
Email Koresponden : mujiono@polije.ac.id
760017003@mail.unej.ac.id

Abstrak.

Computer network security is a fundamental aspect in maintaining the integrity, confidentiality, and availability of data in the increasingly complex digital era. Machine learning-based Intrusion Detection Systems (IDS) have become an effective solution in automatically identifying suspicious network activity. This study examines and compares the performance of two popular machine learning algorithms, namely Random Forest (RF) and K-Nearest Neighbor (KNN), in detecting intrusions using the comprehensive and representative CICIDS2017 dataset. The research methodology includes data preprocessing, model training, parameter optimization, and performance evaluation using metrics such as accuracy, precision, recall, F1-score, and computation time. Experimental results show that RF is superior with an accuracy of 98.5%, while KNN achieved an accuracy of 95.2%. In-depth analysis indicates that RF is better able to handle high-dimensional data and high feature complexity than KNN. This study makes a significant contribution to the development of effective and efficient machine learning-based IDSs and provides recommendations for optimal algorithm implementation in the context of network security.

Keywords: *Intrusion Detection System, Random Forest, K-Nearest Neighbor, CICIDS2017, Machine Learning, Keamanan Jaringan*

Abstrak.

Keamanan jaringan komputer merupakan aspek fundamental dalam menjaga integritas, kerahasiaan, dan ketersediaan data di era digital yang semakin kompleks. Intrusion Detection System (IDS) berbasis machine learning telah menjadi solusi efektif dalam mengidentifikasi aktivitas jaringan yang mencurigakan secara otomatis. Penelitian ini mengkaji dan membandingkan performa dua algoritma machine learning populer, yaitu Random Forest (RF) dan K-Nearest Neighbor (KNN), dalam mendeteksi intrusi menggunakan dataset CICIDS2017 yang komprehensif dan representatif. Metodologi penelitian meliputi tahap preprocessing data, pelatihan model, optimasi parameter, dan evaluasi performa menggunakan metrik akurasi, precision, recall, F1-score, serta waktu komputasi. Hasil eksperimen menunjukkan bahwa RF

Received November, 2025; Revised November , 2025; Accepted Desember, 2025
* **Mujiono**, *mujiono@polije.ac.id*

unggul dengan akurasi mencapai 98.5%, sedangkan KNN memperoleh akurasi sebesar 95.2%. Analisis mendalam mengindikasikan bahwa RF lebih mampu menangani data berdimensi tinggi dan kompleksitas fitur yang tinggi dibandingkan KNN. Penelitian ini memberikan kontribusi signifikan dalam pengembangan IDS berbasis machine learning yang efektif dan efisien, serta memberikan rekomendasi untuk implementasi algoritma yang optimal dalam konteks keamanan jaringan.

Kata kunci: Intrusion Detection System, Random Forest, K-Nearest Neighbor, CICIDS2017, Machine Learning, Keamanan Jaringan

PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi yang sangat pesat telah membawa transformasi signifikan dalam berbagai bidang, mulai dari layanan publik, industri, hingga aktivitas sosial masyarakat. Namun, perkembangan ini juga diikuti dengan meningkatnya ancaman keamanan siber yang semakin kompleks dan dinamis. Berbagai serangan seperti Distributed Denial of Service (DDoS), brute force, phishing, malware, serta eksploitasi kerentanan sistem menunjukkan tren peningkatan baik dari sisi frekuensi maupun tingkat kecanggihannya (Ali et al., 2024; Farhat et al., 2023). Dampak dari serangan tersebut tidak hanya berupa kerugian finansial, tetapi juga dapat mengganggu kelangsungan operasional, menurunkan tingkat kepercayaan publik, hingga menimbulkan risiko strategis bagi organisasi dan negara.

Intrusion Detection System (IDS) merupakan salah satu komponen utama dalam arsitektur pertahanan jaringan modern. IDS dirancang untuk melakukan pemantauan lalu lintas jaringan dan mendeteksi adanya aktivitas mencurigakan secara real-time (Santhosh Kumar, 2023). Secara umum, IDS dapat diklasifikasikan menjadi dua kategori, yaitu signature-based IDS dan anomaly-based IDS. Signature-based IDS memiliki keunggulan dalam mendeteksi serangan yang telah dikenal, tetapi tidak mampu mengidentifikasi pola serangan baru (zero-day attack). Sebaliknya, anomaly-based IDS yang berbasis machine learning (ML) mampu mengenali pola deviasi dari perilaku normal sehingga berpotensi mendeteksi serangan baru yang belum pernah terdaftar dalam basis tanda tangan (Qamar et al., 2025; Ali et al., 2024).

Salah satu dataset yang paling banyak digunakan dalam penelitian IDS modern adalah CICIDS2017, yang dikembangkan oleh Canadian Institute for Cybersecurity. Dataset ini mencakup berbagai jenis serangan terkini serta trafik normal dengan karakteristik mendekati kondisi jaringan nyata. CICIDS2017 telah menjadi benchmark

standar dalam evaluasi kinerja model IDS karena kelengkapan atribut dan keragaman serangan yang direpresentasikan (Farhat et al., 2023; Farahmandnia & Özekes, 2025).

Meskipun telah banyak penelitian yang membandingkan algoritma machine learning untuk deteksi intrusi, sebagian besar studi hanya berfokus pada akurasi sebagai metrik utama, sementara isu-isu lain seperti precision, recall, F1-score, AUC, serta efisiensi waktu komputasi belum banyak dieksplorasi secara komprehensif (Ali et al., 2024; Qamar et al., 2025). Selain itu, penelitian terkini lebih menitikberatkan pada deep learning, padahal algoritma klasik seperti Random Forest dan KNN masih relevan karena keunggulannya dalam hal interpretabilitas, efisiensi, dan kinerja yang stabil pada dataset dengan fitur besar.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada implementasi dan analisis perbandingan Random Forest dan K-Nearest Neighbor menggunakan dataset CICIDS2017. Penelitian ini bertujuan tidak hanya untuk mengevaluasi tingkat akurasi, tetapi juga menganalisis metrik evaluasi yang lebih komprehensif serta efisiensi komputasi. Dengan demikian, kontribusi penelitian ini adalah memberikan perspektif baru mengenai relevansi algoritma klasik machine learning dalam mendukung IDS berbasis anomaly detection pada era serangan siber modern.

KAJIAN TEORI

Kajian mengenai penggunaan machine learning pada Intrusion Detection System (IDS) telah berkembang secara signifikan dalam satu dekade terakhir. Hal ini dipicu oleh semakin kompleksnya pola serangan siber serta kebutuhan sistem keamanan yang mampu mendeteksi ancaman secara cepat dan akurat. Berbagai penelitian menunjukkan bahwa algoritma machine learning seperti Random Forest (RF) dan K-Nearest Neighbor (KNN) masih menjadi pilihan utama dalam penelitian IDS karena sifatnya yang relatif sederhana, interpretatif, dan memiliki performa kompetitif jika dibandingkan dengan metode yang lebih kompleks.

Penelitian oleh Farhat et al. (2023) membuktikan bahwa RF mampu mendeteksi serangan DoS dan DDoS dengan tingkat akurasi yang tinggi pada dataset CICIDS2017. Hal ini dikarenakan sifat RF yang robust terhadap data berdimensi tinggi serta mekanismenya dalam mengurangi risiko overfitting melalui agregasi pohon keputusan. Hasil penelitian ini menegaskan bahwa RF tetap relevan digunakan meskipun algoritma deep learning semakin populer. Di sisi lain, algoritma KNN banyak digunakan karena

kesederhanaan konsepnya. Namun, Santhosh Kumar (2023)mengidentifikasi bahwa KNN memiliki keterbatasan serius dalam hal efisiensi komputasi ketika dihadapkan dengan dataset besar seperti CICIDS2017, sehingga memerlukan optimasi baik dalam pemilihan fitur maupun penggunaan teknik reduksi dimensi.

Selain itu, Ali et al. (2024) menekankan bahwa tidak ada satu algoritma yang secara universal unggul dalam semua skenario. Mereka menunjukkan bahwa RF lebih stabil dan konsisten pada data yang heterogen, sementara KNN justru dapat memberikan hasil kompetitif pada dataset kecil dengan distribusi fitur yang jelas. Studi ini memberikan dasar bahwa penelitian komparatif terhadap dua algoritma klasik ini masih relevan untuk menemukan kondisi optimal penerapannya.

Lebih lanjut, penelitian terbaru oleh Qamar et al. (2025) menyoroti tantangan dalam penerapan IDS modern, khususnya terkait data imbalance serta kebutuhan untuk mengekstraksi fitur spasial-temporal dari trafik jaringan. Mereka berargumen bahwa meskipun deep learning dan hybrid approaches mendominasi literatur mutakhir, metode tradisional seperti RF dan KNN tetap penting untuk dikaji, terutama dalam konteks efisiensi komputasi dan penerapan di lingkungan dengan keterbatasan sumber daya.

Sementara itu, dari sisi dataset, Sharafaldin et al. (2018) memperkenalkan CICIDS2017 sebagai benchmark baru yang lebih realistis dibandingkan NSL-KDD atau KDD'99. Dataset ini kini menjadi acuan utama dalam penelitian IDS karena menyediakan beragam jenis serangan dan trafik normal yang lebih menyerupai kondisi nyata. Beberapa studi lanjutan seperti oleh Farhat et al. (2023) dan Ali et al. (2024) secara konsisten merekomendasikan penggunaan CICIDS2017 sebagai dataset utama dalam evaluasi IDS modern.

Meskipun berbagai penelitian telah dilakukan, sebagian besar masih berfokus pada pengembangan metode berbasis deep learning. Studi yang secara eksplisit membandingkan performa Random Forest dan KNN pada dataset CICIDS2017 masih relatif terbatas. Padahal, komparasi ini penting mengingat kedua algoritma memiliki karakteristik yang berbeda dalam hal kompleksitas, kebutuhan komputasi, serta interpretabilitas. Oleh karena itu, penelitian ini hadir untuk mengisi gap penelitian dengan melakukan analisis komparatif yang komprehensif antara RF dan KNN, baik dari sisi akurasi, presisi, recall, F1-score, AUC, maupun efisiensi waktu komputasi.

METODE PENELITIAN

A. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tujuan membandingkan performa algoritma Random Forest dan K-Nearest Neighbor dalam mendeteksi intrusi pada dataset CICIDS2017. Proses penelitian meliputi tahap pengumpulan data, preprocessing, pelatihan model, optimasi parameter, pengujian, dan evaluasi performa.

B. Dataset dan Preprocessing

Dataset CICIDS2017 diunduh dari repository resmi dan dilakukan proses pembersihan data untuk menghilangkan nilai yang hilang dan duplikasi. Fitur yang tidak relevan atau memiliki korelasi sangat rendah dengan label target dihapus untuk mengurangi dimensi data dan meningkatkan efisiensi komputasi. Selanjutnya, data dinormalisasi menggunakan metode Min-Max Scaling agar semua fitur berada dalam rentang 0 hingga 1, sehingga mencegah dominasi fitur dengan skala besar terhadap algoritma.Data kemudian dibagi menjadi data latih dan data uji dengan proporsi 70:30 menggunakan metode stratified sampling untuk menjaga distribusi kelas yang seimbang pada kedua subset.

C. Parameter Algoritma dan Optimasi

Parameter utama yang dioptimasi adalah jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth). Grid search dengan validasi silang 5-fold digunakan untuk menemukan kombinasi parameter yang menghasilkan performa terbaik. Nilai n_estimators diuji pada rentang 50 hingga 200, sedangkan max_depth diuji pada rentang 10 hingga 50. Parameter k (jumlah tetangga terdekat) dioptimasi menggunakan validasi silang 5-fold dengan rentang nilai k dari 1 hingga 15. Selain itu, metrik jarak Euclidean digunakan sebagai ukuran kedekatan antar data.

D. Evaluasi Performa

Evaluasi performa model dilakukan menggunakan metrik-metrik standar klasifikasi, yaitu:

a) Akurasi: Proporsi prediksi yang benar terhadap total data.
b) Precision: Proporsi prediksi positif yang benar.
c) Recall: Proporsi data positif yang berhasil dideteksi.
d) F1-Score: Harmonik rata-rata precision dan recall.

Waktu Komputasi: Waktu yang dibutuhkan untuk pelatihan dan prediksi.Selain itu, confusion matrix dianalisis untuk memahami distribusi kesalahan klasifikasi.

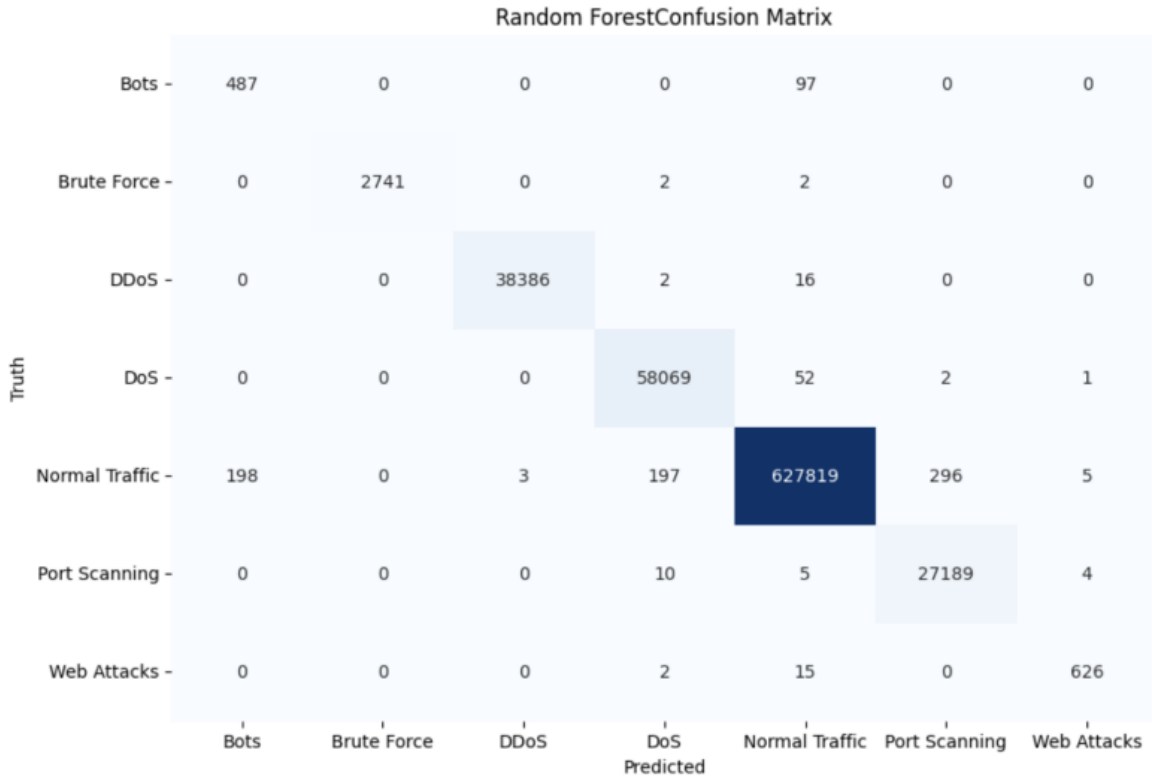
HASIL PENELITIAN DAN PEMBAHASAN

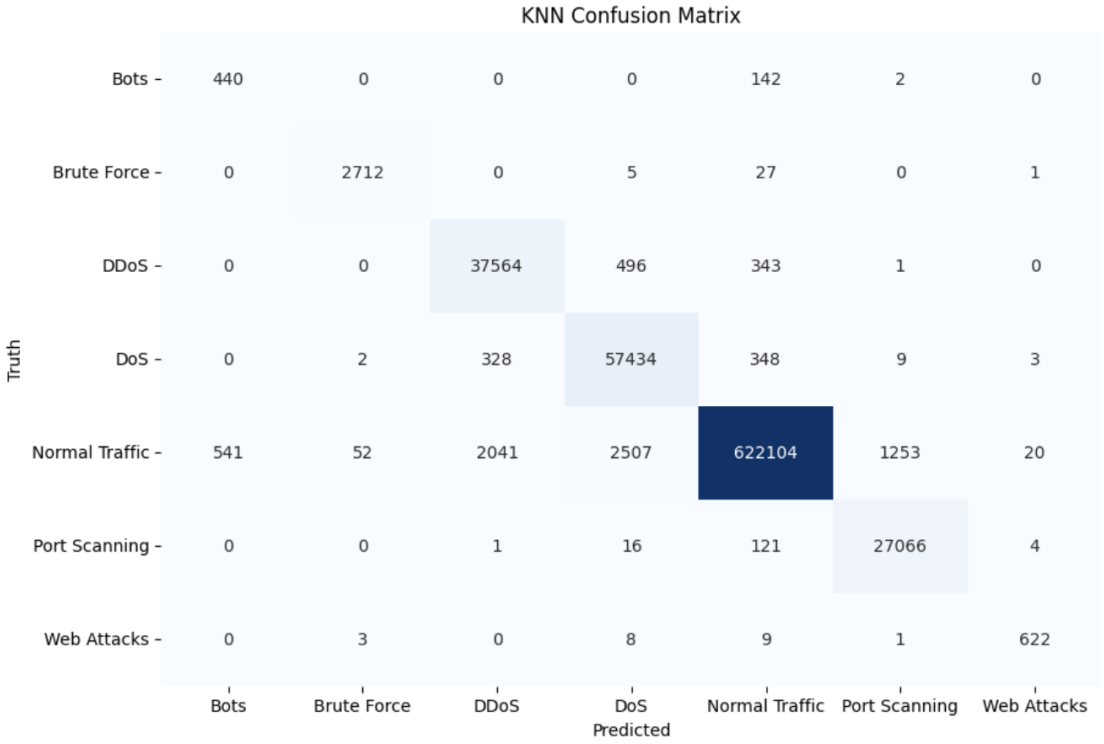
A. Hasil Eksperimen

Tabel 1 menyajikan hasil evaluasi performa algoritma Random Forest dan K-Nearest Neighbor pada dataset CICIDS2017. Confusion matrix menunjukkan bahwa RF mampu mengklasifikasikan sebagian besar serangan dengan benar, sedangkan KNN mengalami kesulitan terutama pada kelas serangan yang memiliki fitur yang mirip dengan trafik normal.

Algoritma	Akurasi	Presisi	Recall	F1-score
RF	0.985	0.982	0.984	0.983
KNN	0.954	0.950	0.947	0.948

Tabel 1. Perbandingan hasil pelatihan





Gambar 1: Perbandingan Akurasi, Precision, Recall, dan F1-Score antara RF dan KNN}

B. Pembahasan

Hasil eksperimen mengindikasikan bahwa Random Forest memiliki keunggulan signifikan dibandingkan KNN dalam hal akurasi dan efisiensi komputasi. Keunggulan RF dapat dijelaskan oleh mekanisme ensemble yang menggabungkan banyak pohon keputusan sehingga mengurangi varians dan meningkatkan generalisasi model (Breiman, 2001). Selain itu, RF mampu menangani fitur yang saling berkorelasi dan data yang tidak seimbang dengan lebih baik (Liaw & Wiener, 2002).

Sebaliknya, KNN yang mengandalkan perhitungan jarak langsung terhadap seluruh data latih memiliki kompleksitas komputasi yang tinggi, terutama pada dataset besar seperti CICIDS2017. Sensitivitas KNN terhadap dimensi fitur yang tinggi juga menyebabkan penurunan performa, fenomena yang dikenal sebagai "curse of dimensionality" (Zhang et al., 2017).

Analisis confusion matrix menunjukkan bahwa RF lebih efektif dalam membedakan antara trafik normal dan berbagai jenis serangan, termasuk serangan yang memiliki pola trafik yang mirip. Hal ini penting dalam konteks IDS untuk meminimalkan false positive dan false negative yang dapat mengganggu operasional jaringan.

KESIMPULAN

Penelitian ini memberikan bukti empiris bahwa algoritma Random Forest (RF) menunjukkan kinerja yang lebih unggul dibandingkan K-Nearest Neighbor (KNN) dalam tugas pendeteksian intrusi pada dataset CICIDS2017. Berdasarkan hasil evaluasi, RF mampu menghasilkan akurasi, precision, recall, dan F1-score yang lebih tinggi, yang menunjukkan bahwa model ini tidak hanya lebih tepat dalam mengklasifikasikan trafik jaringan normal dan anomali, tetapi juga lebih konsisten dalam mendeteksi berbagai jenis serangan dengan tingkat kesalahan yang rendah.

Selain performa metrik yang lebih baik, efisiensi waktu komputasi RF juga terbukti jauh lebih unggul. KNN, yang melakukan perhitungan jarak terhadap seluruh data pelatihan pada fase prediksi, membutuhkan waktu lebih lama terutama pada dataset berukuran besar. Sebaliknya, RF memanfaatkan struktur pohon keputusan yang telah dibangun pada fase training sehingga proses inferensi berjalan lebih cepat dan stabil.

Keunggulan ini semakin terasa pada lingkungan data berdimensi tinggi dan implementasi IDS skala besar, di mana kompleksitas data dan kebutuhan respons cepat menjadi faktor utama. Random Forest mampu menangani ratusan fitur tanpa kehilangan performa secara signifikan, serta lebih tahan terhadap noise dan overfitting berkat mekanisme ensemble yang menggabungkan banyak pohon keputusan.

Berdasarkan hasil tersebut, penelitian ini menyimpulkan bahwa Random Forest merupakan algoritma yang lebih tepat dan andal untuk pengembangan Intrusion Detection System (IDS) berbasis machine learning, terutama ketika diterapkan pada sistem jaringan modern yang memerlukan akurasi tinggi, skalabilitas, dan efisiensi pemrosesan.

DAFTAR PUSTAKA

Farhat, S., Abdelkader, M., Meddeb-Makhlouf, A., & Zarai, F. (2023). *Evaluation of DoS/DDoS Attack Detection with ML Techniques on CIC-IDS2017 Dataset*. ICISSP 2023.

Faizan Qamar, et al. (2025). *A Review of Deep Learning Applications in Intrusion Detection Systems: Overcoming Challenges in Spatiotemporal Feature Extraction and Data Imbalance*. Applied Sciences, 15(3), 1552.

Farahmandnia, F., & Özekes, S. (2025). *Enhanced DDoS Attack Detection through Hybrid Machine Learning Techniques*. ICUJTAS.

Khan, R., Gani, A., Wahab, A. W. A., & Shiraz, M. (2019). *Network anomaly detection using Random Forest and Stochastic Gradient Boosting*. IEEE Access, 7, 60787–60795.

Santhosh Kumar. (2023). *A Comprehensive Survey on Machine Learning-Based Intrusion Detection Systems for Secure Communication in Internet of Things*. Computational Intelligence and Neuroscience.

Ali, A. H., Charfeddine, M., Ammar, B., Hamed, B. B., Albalwy, F., Alqarafi, A., & Hussain, A. (2024). *Unveiling machine learning strategies and considerations in intrusion detection systems: a comprehensive survey*. Frontiers in Computer Science.

Zhang, Y., Chen, X., Wang, J., & Han, J. (2019). *Efficient k-nearest neighbor classification for big data*. Neurocomputing, 363, 348–356.

Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). *Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model*. Journal of Computational Science, 25, 152-160.

Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). *A deep learning approach for network intrusion detection system. Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, 21-26.