BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

Pengembangan dan Implementasi Sistem Deteksi Serangan DDoS Berbasis Algoritma Random Forest

Dedy Kiswanto[1*], Fanny Ramadhani[2], Nurul Maulida Surbakti[3], Nadrah Afiati Nasution[4]
[1,2] Fakultas Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
[3] Fakultas Matematika dan Ilmu Pengetahuan Alam, Matematika, Universitas Negeri Medan, Medan, Indonesia
[4] Fakultas Matematika dan Ilmu Pengetahuan Alam, Pendidikan Matematika, Universitas Negeri Medan, Medan, Indonesia
Email: [1*]dedykiswanto@unimed.ac.id, [2]fannyr@unimed.ac.id, [3]nurulmaulida@unimed.ac.id, [4]nadrahaflati@unimed.ac.id
(* : coressponding author: dedykiswanto@unimed.ac.id)

Abstrak-Serangan Distributed Denial of Service (DDoS) merupakan ancaman serius bagi keamanan jaringan, sementara metode deteksi tradisional seperti threshold-based detection dan signature-based detection memiliki keterbatasan dalam mengenali pola serangan baru maupun anomali lalu lintas yang kompleks. Penelitian ini bertujuan merancang dan mengimplementasikan model prediksi serangan DDoS berbasis algoritma Random Forest yang mampu membedakan trafik normal dan berindikasi serangan secara akurat. Pendekatan Research and Development (R&D) digunakan, meliputi studi literatur, perancangan model, implementasi, serta evaluasi performa menggunakan metrik akurasi, precision, recall, F1-score, confusion matrix, dan learning curve. Berdasarkan hasil evaluasi, model Random Forest menunjukkan kinerja sangat baik dengan akurasi 0,99942 (99,942%). Precision untuk kelas 0 dan 1 masing-masing sebesar 0,99979 dan 0,99884, sedangkan recall mencapai 0,99928 untuk kelas 0 dan 0,99966 untuk kelas 1. Nilai F1-score tinggi, yaitu 0,99953 untuk kelas 0 dan 0,99925 untuk kelas 1, dengan macro average F1-score sebesar 0,99939 dan weighted average sebesar 0,99942, menunjukkan keseimbangan performa pada kedua kelas. Confusion Matrix menunjukkan kesalahan klasifikasi rendah (44 false positive dan 13 false negative dari 99.066 sampel). Analisis learning curve mengungkapkan akurasi pelatihan stabil di atas 0,998, sedangkan akurasi validasi meningkat dari 0,986 pada 10.000 data hingga di atas 0,998 pada 80.000 data, dengan jarak antarkurva semakin kecil. Pola ini menandakan model mampu memanfaatkan data tambahan untuk meningkatkan generalisasi tanpa gejala overfitting atau underfitting. Temuan ini membuktikan bahwa model Random Forest yang dirancang dapat menjadi solusi deteksi dini serangan DDoS yang andal, adaptif, dan berpotensi diintegrasikan dalam sistem keamanan jaringan secara real-time.

Kata Kunci: DDoS, Serangan Siber, Random Forest, keamanan jaringan, *Machine learning*

Abstract- Distributed Denial of Service (DDoS) attacks pose a serious threat to network security, while traditional detection methods such as threshold-based detection and signature-based detection face limitations in identifying novel attack patterns and complex traffic anomalies. This study aims to design and implement a DDoS attack prediction model based on the Random Forest algorithm, capable of accurately distinguishing between normal traffic and traffic indicative of an attack. A Research and Development (R&D) approach was employed, encompassing literature review, model design, implementation, and performance evaluation using accuracy, precision, recall, F1-score, confusion matrix, and learning curve metrics. Evaluation results show that the Random Forest model achieved outstanding performance, with an accuracy of 0.99942 (99.942%). Precision for class 0 and class 1 was 0.99979 and 0.99884, respectively, while recall reached 0.99928 for class 0 and 0.99966 for class 1. High F1-scores were recorded—0.99953 for class 0 and 0.99925 for class 1—with a macro average F1-score of 0.99939 and a weighted average of 0.99942, indicating balanced performance across both classes. The confusion matrix revealed minimal misclassification (44 false positives and 13 false negatives out of 99,066 samples). Learning curve analysis showed training accuracy consistently above 0.998, while validation accuracy improved from 0.986 with 10,000 samples to over 0.998 with 80,000 samples, with narrowing gaps between the curves. This pattern suggests the model effectively leverages additional data to enhance generalization without signs of overfitting or underfitting. These findings confirm that the proposed Random Forest model is a reliable, adaptive early detection solution for DDoS attacks, with strong potential for real-time integration into network security systems.

Keywords: DDoS, Cyber attack detection, Random Forest, Network security, Machine learning

1. PENDAHULUAN

Keamanan jaringan telah menjadi salah satu aspek krusial di era digital yang semakin terhubung. Perkembangan teknologi informasi mendorong pertumbuhan layanan berbasis internet secara masif, mulai dari transaksi perbankan, komunikasi daring, hingga pengelolaan infrastruktur kritis[1]. Namun, kemajuan ini juga diiringi dengan meningkatnya ancaman siber yang dapat mengganggu ketersediaan, kerahasiaan, dan integritas data. Salah satu ancaman yang paling menonjol adalah serangan Distributed Denial of Service (DDoS), yang dapat melumpuhkan layanan dalam hitungan detik dengan membanjiri server atau jaringan target menggunakan lalu lintas berlebihan[2]. Dampak dari serangan ini tidak hanya bersifat teknis, tetapi juga dapat menimbulkan kerugian ekonomi yang signifikan serta merusak reputasi penyedia layanan. Oleh karena itu, kebutuhan akan sistem deteksi dini yang akurat dan andal menjadi semakin mendesak untuk menjaga keberlangsungan operasional di berbagai sektor.

Fenomena serangan DDoS terus menunjukkan tren peningkatan baik dari sisi frekuensi, skala, maupun kompleksitas teknik yang digunakan oleh penyerang. Serangan DdoS tidak hanya menargetkan perusahaan berskala besar, tetapi juga institusi pemerintahan, lembaga pendidikan, hingga usaha kecil dan menengah. Karakteristik terdistribusi dari serangan DDoS—yang memanfaatkan ribuan hingga jutaan perangkat yang terinfeksi untuk mengirimkan lalu lintas berlebihan ke

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

satu target—menjadikannya sulit diantisipasi dengan metode konvensional[3]. Di tengah transformasi digital dan adopsi layanan daring yang masif, kerentanan terhadap serangan DDoS menjadi semakin tinggi. Urgensi untuk mengembangkan metode deteksi yang cepat, akurat, dan adaptif bukan hanya untuk meminimalkan dampak serangan, tetapi juga sebagai bagian dari strategi keamanan siber yang berkelanjutan.

Meskipun berbagai solusi telah dikembangkan untuk mendeteksi serangan DDoS, sebagian besar metode tradisional seperti threshold-based detection atau signature-based detection memiliki keterbatasan signifikan. Pendekatan berbasis ambang batas sering kali menghasilkan tingkat false positive yang tinggi ketika lalu lintas jaringan meningkat secara tiba-tiba namun masih dalam batas wajar, misalnya saat terjadi lonjakan pengguna secara mendadak[4]. Sementara itu, metode berbasis tanda tangan hanya mampu mengidentifikasi pola serangan yang telah diketahui sebelumnya, sehingga kurang efektif dalam menghadapi variasi serangan baru atau serangan yang menggunakan teknik penyamaran[5].

Kompleksitas pola lalu lintas jaringan modern, ditambah dengan volume data yang sangat besar, membuat metode deteksi DdoS konvensional sulit beradaptasi terhadap dinamika ancaman[6]. Kondisi ini menegaskan perlunya pendekatan yang lebih cerdas, adaptif, dan mampu mengenali pola anomali secara real time untuk meningkatkan efektivitas sistem deteksi DDoS. Untuk mengatasi keterbatasan metode deteksi tradisional, penelitian ini menawarkan solusi berbasis machine learning dengan memanfaatkan algoritma Random Forest sebagai model prediksi serangan DDoS. Pendekatan ini dirancang untuk menganalisis pola lalu lintas jaringan secara otomatis dengan mempertimbangkan berbagai fitur yang merepresentasikan karakteristik paket data. Random Forest, sebagai salah satu metode ensemble learning, mampu menggabungkan kekuatan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil, sekaligus meminimalkan risiko overfitting[7]. Dengan kemampuan dalam menangani data berskala besar dan beragam tipe fitur, model ini diharapkan dapat mengenali perbedaan halus antara lalu lintas normal dan lalu lintas yang mengindikasikan serangan DDoS. Solusi ini dirancang agar adaptif terhadap perubahan pola serangan, sehingga tetap relevan dalam menghadapi ancaman siber yang terus berkembang.

Berbagai penelitian dalam lima tahun terakhir menunjukkan tren peningkatan penggunaan algoritma machine learning untuk deteksi ancaman siber, termasuk serangan DDoS dan jenis serangan lainnya. Penelitian [8] mengkaji serangan phishing melalui email dengan membandingkan kinerja Decision Tree, Random Forest, dan SVM. Hasilnya, Random Forest mencapai akurasi tertinggi sebesar 96% dan terbukti unggul dalam menangani data yang tidak seimbang serta menurunkan false positives. Meski demikian, penelitian ini fokus pada phishing berbasis email dan belum menguji performa model pada serangan DDoS atau lalu lintas jaringan berskala besar. Studi [9] mengembangkan sistem deteksi serangan siber pada sistem informasi akademik menggunakan Decision Tree dan Random Forest, dengan hasil akurasi 92% untuk Random Forest. Model ini mampu mendeteksi aktivitas anomali seperti login mencurigakan dan query basis data abnormal, namun penelitian ini tidak menampilkan confusion matrix sehingga distribusi kesalahan klasifikasi tidak dapat dianalisis secara rinci. Sementara itu, penelitian [10] juga menegaskan potensi metode ensemble seperti Random Forest dalam menghasilkan performa klasifikasi yang tinggi. Namun, hasil classification report pada penelitian tersebut menunjukkan nilai precision, recall, dan f1-score yang sempurna (100%) untuk semua kelas, yang mengindikasikan kemungkinan terjadinya overfitting akibat model terlalu menyesuaikan diri dengan data latih.

Penelitian [11] memfokuskan pada pengembangan sistem deteksi serangan berbasis machine learning untuk meningkatkan keamanan jaringan. Metode yang digunakan melibatkan pemrosesan data secara sistematis dan penerapan algoritma klasifikasi, dengan hasil evaluasi yang menunjukkan tingkat akurasi yang cukup tinggi. Meskipun demikian, penelitian ini tidak menampilkan confusion matrix maupun learning curve, sehingga distribusi kesalahan klasifikasi dan tren performa model terhadap ukuran data pelatihan tidak dapat dianalisis secara mendalam. Sementara itu, studi [12] mengusulkan kebijakan keamanan adaptif berbasis machine learning pada firewall Software-Defined Networking (SDN) menggunakan algoritma Random Forest dan dataset CICIDS2017. Model yang dikembangkan mencapai akurasi 99,9978%, precision dan recall 99,996%, serta hanya dua kesalahan klasifikasi dari 45.149 data uji. Meskipun hasil ini menunjukkan performa yang sangat tinggi, penelitian ini tidak menampilkan learning curve, sehingga kemampuan generalisasi model terhadap variasi ukuran data pelatihan belum dapat dievaluasi secara menyeluruh. Dari kedua penelitian ini, terlihat bahwa meskipun performa klasifikasi dapat mencapai tingkat sangat tinggi, tantangan utama yang tersisa adalah memastikan generalisasi model pada kondisi nyata, khususnya pada deteksi DDoS di lalu lintas jaringan real-time yang memiliki pola dinamis dan tidak sepenuhnya terprediksi.

Penelitian ini difokuskan pada perancangan dan implementasi model deteksi serangan DDoS berbasis algoritma Random Forest yang diharapkan mampu memberikan prediksi akurat terhadap lalu lintas jaringan, baik yang bersifat normal maupun yang mengandung serangan. Pendekatan ini disertai evaluasi komprehensif menggunakan metrik seperti akurasi, precision, recall, F1-score, confusion matrix, dan learning curve untuk memperoleh gambaran menyeluruh mengenai performa dan kemampuan generalisasi model. Harapannya, model yang dikembangkan dapat menjadi solusi deteksi dini yang andal, adaptif, dan dapat diintegrasikan ke dalam sistem keamanan jaringan secara real-time. Dengan demikian, hasil penelitian ini diharapkan tidak hanya memberikan kontribusi pada pengembangan teknologi deteksi DDoS, tetapi juga memperkuat perlindungan infrastruktur digital dari ancaman serangan siber yang terus berkembang.

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

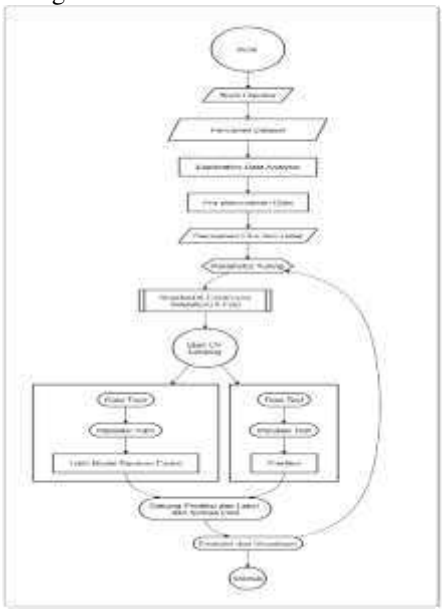
2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan Research and Development (R&D) yang dirancang untuk mengembangkan sistem prediksi serangan Distributed Denial of Service (DDoS) berbasis algoritma Random Forest. Pendekatan ini dipilih karena mampu mengakomodasi dua tujuan utama penelitian, yaitu melakukan analisis komprehensif terhadap data lalu lintas jaringan serta menghasilkan produk akhir berupa model prediktif yang dapat diintegrasikan ke dalam sistem deteksi dini serangan siber. Metode R&D memadukan unsur analisis data, perancangan, implementasi, dan pengujian, sehingga sesuai untuk membangun solusi yang dapat langsung diimplementasikan pada lingkungan nyata[13].

Penelitian diawali dengan studi literatur yang difokuskan pada tiga aspek utama, pertama adalah pemahaman mengenai karakteristik dan teknik serangan DDoS, kedua peninjauan metode deteksi anomali berbasis machine learning, dan terakhir adalah identifikasi fitur-fitur lalu lintas jaringan yang relevan untuk klasifikasi. Sumber literatur diambil dari artikel jurnal bereputasi, prosiding konferensi internasional, serta dokumentasi teknis dari platform keamanan jaringan. Studi literatur ini memberikan landasan teoretis yang menjadi acuan dalam perancangan arsitektur model, pemilihan parameter, dan strategi evaluasi[14].

Dataset penelitian diperoleh dari platform Kaggle dengan judul "DDoS Attack" yang dikembangkan oleh Shayal Vaghasiya. Dataset ini memuat data lalu lintas jaringan yang telah dilabeli sebagai normal atau DDoS, sehingga sesuai untuk digunakan dalam tugas klasifikasi biner. Setelah dataset diperoleh, dilakukan tahap Exploratory Data Analysis (EDA) untuk memahami karakteristik dan struktur data secara umum. Selanjutnya, data diproses melalui tahap pra-pemrosesan agar layak digunakan dalam pelatihan model. Proses pengembangan model dimulai dengan pelatihan menggunakan algoritma Random Forest, yaitu metode ensemble berbasis pohon keputusan yang menggabungkan hasil dari banyak model untuk menghasilkan prediksi yang lebih akurat dan stabil. Algoritma ini dipilih karena kemampuannya dalam menangani dataset berskala besar, toleransinya terhadap noise, serta kemampuannya mengurangi risiko overfitting melalui teknik pengacakan fitur (feature bagging). Penyesuaian hiperparameter dilakukan untuk mengoptimalkan kinerja model, diikuti dengan pelatihan menggunakan pendekatan cross-validation guna memastikan hasil evaluasi yang lebih reliabel dan tidak bergantung pada pembagian data tertentu.

Model yang telah dilatih dievaluasi menggunakan beberapa metrik untuk menilai kualitas prediksi. Metrik yang digunakan meliputi akurasi, precision, recall, dan F1-score sebagai indikator utama performa klasifikasi. Evaluasi lanjutan dilakukan menggunakan confusion matrix untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas secara rinci. Selain itu, learning curve digunakan untuk menganalisis tren performa model terhadap variasi jumlah data pelatihan. Kombinasi penggunaan metrik numerik dan visualisasi ini memungkinkan penilaian yang lebih komprehensif terhadap performa model, baik dari sisi akurasi prediksi maupun konsistensinya pada berbagai skenario pengujian. Pendekatan evaluasi seperti ini memberikan kejelasan tidak hanya mengenai hasil akhir, tetapi juga proses pembelajaran model selama pelatihan. Hal ini penting untuk memastikan bahwa model yang dihasilkan tidak hanya bekerja optimal pada data latih, tetapi juga mampu beradaptasi pada data baru yang belum pernah dilihat sebelumnya secara keseluruhan, tahapan penelitian dimulai dari studi literatur, dilanjutkan dengan pencarian dan pemahaman data, pra-pemrosesan data, pelatihan model, dan diakhiri dengan evaluasi model. Alur lengkap penelitian ini divisualisasikan pada Gambar 1 untuk memberikan gambaran yang lebih jelas mengenai urutan dan keterkaitan antar tahapan.



Gambar 1. Alur Penelitian

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

3. HASIL DAN PEMBAHASAN

3.1 Sumber Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari platform *Kaggle* dengan nama *"DDoS Attack"*, yang disusun oleh Shayal Vaghasiya. Dataset ini terdiri atas berbagai fitur yang merepresentasikan karakteristik lalu lintas jaringan, serta telah dilengkapi dengan label klasifikasi, sehingga mempermudah proses pelatihan dan evaluasi model. Pemilihan dataset ini didasarkan pada pertimbangan kelengkapan atribut, kejelasan anotasi label, serta ketersediaannya secara terbuka (*open access*), menjadikannya relevan dan representatif untuk keperluan eksperimen dalam pengembangan model deteksi serangan *Distributed Denial of Service* (DDoS).

3.1 EDA(*Exploratory Data Analysis)*

Tahap awal dalam penelitian ini adalah melakukan *Exploratory Data Analysis* (EDA), yaitu proses eksplorasi terhadap data guna memperoleh pemahaman mengenai struktur, pola, dan karakteristik distribusi data sebelum dilakukan proses pemodelan[15]. EDA memiliki peran penting dalam mengidentifikasi potensi masalah pada data, seperti keberadaan nilai ekstrem (*outlier*), kesalahan entri data, dan ketidakseimbangan distribusi kelas, serta dalam menentukan strategi *preprocessing* yang sesuai[16]. Pada penelitian ini, EDA diawali dengan pemeriksaan jumlah fitur, jenis tipe data, serta identifikasi rentang nilai (minimum dan maksimum) untuk setiap atribut numerik. Langkah-langkah ini memberikan pemahaman awal yang esensial terhadap cakupan, variasi, dan potensi tantangan yang terdapat pada dataset yang digunakan. Ringkasan hasil analisis awal ini disajikan pada Tabel 1.

Tabel 1. EDA Jumlah Kolom dan Range Nilai Setiap Kolom

No	Nama Fitur di Dataset	Nama Fitur Aslinya	Nilai Terkecil	Nilai Terbesar	Keterangan (Fungsi Fitur)
1	dt	datetime	2488.00	42935.00	Waktu saat aliran data tercatat
2	switch	switch	1.00	10.00	ID switch OpenFlow tempat aliran lewat
3	src	source_ip	167772161.00	167772180.00	IP sumber aliran (format integer)
4	dst	destination_ip	167772161.00	167772178.00	IP tujuan aliran (format integer)
5	pktcount	packet_count	0.00	260006.00	Jumlah paket yang lewat dalam satu aliran
6	bytecount	byte_count	0.00	147128002.00	Total byte yang ditransmisikan dalam aliran
7	dur	duration_seconds	0.00	1881.00	Durasi aliran dalam detik
8	dur_nsec	duration_nanoseconds	0.00	999000000.00	Tambahan durasi dalam nanodetik
9	tot_dur	total_duration_ nanoseconds	0.00	1880000000000 .00	Durasi total dalam nanodetik
10	flows	number_of_flows	2.00	17.00	Jumlah total aliran aktif
11	packetins	packet_in_count	4.00	25224.00	Banyaknya packet-in yang diterima controller
12	pktperflow	packets_per_flow	-130933.00	19190.00	Rata-rata jumlah paket per aliran
13	byteperflow	bytes_per_flow	-146442594.00	14953872.00	Rata-rata byte per aliran
14	pktrate	packet_rate	-4365.00	639.00	Laju kirim paket per detik
15	Pairflow	bidirectional_flow_ flag	0.00	1.00	Indikator apakah aliran bersifat dua arah

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

16	Protocol	network_protocol	1.00	64.00	Protokol jaringan yang digunakan
17	port_no	port_number	1.00	5.00	Nomor port tempat paket masuk
18	tx_bytes	transmitted_bytes	2527.00	1270000000.00	Byte yang dikirim lewat switch port
19	rx_bytes	received_bytes	856.00	991000000.00	Byte yang diterima oleh switch port
20	tx_kbps	transmit_kilobits_per_sec	0.00	20580.00	Kecepatan transmisi data (Kbps)
21	rx_kbps	receive_kilobits_per_sec	0.00	16577.00	Kecepatan penerimaan data (Kbps)
22	A1	auxiliary_feature_1	0.00	0.00	Fitur dummy/placeholder petama
23	A2	auxiliary_feature_2	0.00	0.00	Fitur dummy/placeholder kedua.
24	tot_kbps	total_throughput_kbps	0.00	20580.00	Total throughput (tx + rx dalam Kbps)
25	label	attack_label	0.00	1.00	Kelas data (0 = normal l, = serangan)

Tahap lanjutan dalam proses Exploratory Data Analysis (EDA) dilakukan untuk mengevaluasi kualitas dan integritas data sebelum memasuki proses pemodelan. Pemeriksaan awal difokuskan pada deteksi nilai kosong (missing values) pada setiap fitur. Hasil analisis menunjukkan bahwa sebagian besar fitur tidak mengandung nilai kosong, kecuali fitur rx_kbps dan tot_kbps, yang masing-masing memiliki 506 nilai kosong. Adapun fitur target label terkonfirmasi tidak memiliki nilai kosong. Selain itu, analisis terhadap data duplikat mengidentifikasi sebanyak 5.091 baris yang merupakan duplikasi, yang apabila tidak ditangani, dapat menimbulkan bias dalam proses pelatihan model dan memengaruhi akurasi prediksi.

Selanjutnya, distribusi kelas pada variabel target menunjukkan adanya ketidakseimbangan data, dengan 63.561 sampel termasuk dalam kelas 0 dan 40.784 sampel dalam kelas 1. Ketimpangan ini dapat berdampak pada kinerja algoritma klasifikasi, khususnya dalam mengenali kelas minoritas. Lebih lanjut, ditemukan nilai-nilai negatif pada beberapa fitur numerik, seperti packetperflow, byteperflow, dan packetrate, yang masing-masing mengandung 188 nilai negatif. Nilai-nilai ini dinilai tidak logis dalam konteks lalu lintas jaringan, dan oleh karena itu perlu dipertimbangkan untuk diatasi pada tahap data cleaning. Temuan-temuan ini menjadi landasan dalam penentuan strategi preprocessing yang sesuai sebelum pelatihan model dilakukan.

3.2 Preprocessing

Tahapan preprocessing merupakan salah satu proses penting dalam perancangan model machine learning yang bertujuan untuk menyiapkan data mentah menjadi data yang layak digunakan untuk pelatihan model. Preprocessing adalah tahapan sistematis yang mencakup pembersihan, transformasi, dan rekayasa fitur, guna memastikan kualitas data yang optimal sebelum digunakan dalam proses pemodelan[17]. Tujuan dari tahap ini adalah untuk memperoleh data yang bersih, bebas dari anomali, serta memiliki format yang sesuai, sehingga dapat meningkatkan akurasi dan kinerja model prediktif.

Langkah pertama yang dilakukan adalah menghilangkan baris data yang bersifat duplikat penghapusan data duplikat bertujuan untuk menghindari bias akibat pengulangan informasi yang dapat memengaruhi proses pembelajaran model secara negatif. Selanjutnya, seluruh baris yang mengandung nilai negatif pada fitur pktperflow, byteperflow, dan pktrate, dihapus. Nilai negatif pada fitur-fitur tersebut tidak sesuai secara logis dalam konteks lalu lintas jaringan, dan berpotensi mengganggu performa model. Selain itu, dua fitur yaitu A1 dan A2 dihapus dari dataset karena hanya berisi nilai nol tanpa variasi, sehingga dianggap sebagai fitur non-informatif (noise) yang tidak memberikan kontribusi terhadap proses pembelajaran. Kolom waktu (dt), yang semula tersimpan dalam format numerik khas Excel, dikonversi menjadi format timestamp untuk memungkinkan ekstraksi fitur temporal. Dari timestamp tersebut, diambil informasi waktu yang lebih

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

relevan, seperti jam (hour) dan hari dalam seminggu (dayofweek), yang kemudian digunakan sebagai variabel baru dalam pemodelan. Setelah proses ekstraksi dilakukan, kolom dt dan timestamp dihapus karena tidak lagi diperlukan.
Tahapan preprocessing juga mencakup proses ekstraksi subnet dari alamat IP sumber (src) dan tujuan (dst), yang awalnya tersimpan dalam format bilangan bulat. Ekstraksi ini dilakukan untuk memperoleh representasi topologi jaringan yang lebih bermakna, yang diharapkan dapat meningkatkan kemampuan model dalam mengenali pola-pola serangan DDoS berdasarkan asal dan tujuan trafik jaringan. Proses ekstraksi subnet dilakukan dengan mengambil tiga oktet pertama dari alamat IP sumber dan tujuan, yang kemudian dikodekan menggunakan metode Label Encoding agar dapat direpresentasikan dalam format numerik yang kompatibel dengan algoritma machine learning. Setelah proses encoding selesai, kolom asli src dan dst dihapus dari dataset untuk menghindari redundansi informasi. Selanjutnya, beberapa fitur numerik yang sebenarnya merepresentasikan kategori, yaitu fitur switch, port_no, dan Protocol, diubah ke tipe data kategorikal. Perubahan ini bertujuan agar algoritma machine learning dapat memperlakukan fitur-fitur tersebut sesuai dengan sifat aslinya, serta menghindari kesalahan interpretasi dalam proses pelatihan model. Sebagai langkah akhir dalam tahap preprocessing, dataset dipisahkan menjadi dua bagian utama, yaitu fitur prediktor (X) dan label target (y). Pemisahan ini penting untuk memastikan bahwa proses pelatihan model berlangsung secara terstruktur, dengan fokus pada relasi antara fitur masukan dan variabel keluaran yang ingin diprediksi, dalam hal ini adalah keberadaan serangan DDoS.

3.3 Parameter Random Forest

Dalam pengembangan model machine learning, salah satu tahapan penting yang sangat memengaruhi performa akhir model adalah parameter tuning atau penyesuaian hiperparameter. Parameter tuning merupakan proses pencarian kombinasi nilai terbaik dari hiperparameter yang mengatur perilaku algoritma pembelajaran[18]. Tidak seperti parameter model yang dipelajari langsung dari data, hiperparameter ditentukan terlebih dahulu sebelum proses pelatihan dimulai. Setiap algoritma memiliki konfigurasi hiperparameter yang berbeda, baik dari segi fungsi maupun pengaruhnya terhadap proses pelatihan dan prediksi. Oleh karena itu, dibutuhkan pendekatan khusus dalam proses penyesuaiannya. Dalam penelitian ini digunakan algoritma Random Forest, dengan konfigurasi parameter sebagaimana disajikan pada Tabel 2. Proses penyesuaian dilakukan secara manual, yaitu dengan menguji beberapa kombinasi nilai secara bertahap dan mengevaluasi kinerjanya menggunakan data validasi. Tujuan utama dari proses ini adalah untuk mengoptimalkan performa model, meningkatkan akurasi prediksi, serta meminimalkan risiko terjadinya *overfitting* yaitu kondisi ketika model terlalu menyesuaikan diri terhadap data pelatihan sehingga kehilangan kemampuan generalisasi dan *underfitting* yaitu kondisi ketika model gagal menangkap pola penting dari data karena terlalu sederhana[19].

Tabel 2. EDA Jumlah Kolom dan Range Nilai Setiap Kolom

No	Nama Parameter	Nilai Parameter	Fungsi dan Kegunaan
1	n_estimators	300	Jumlah pohon kecil yang digabung menjadi hutan untuk membuat keputusan akhir[20].
2	max_depth	None	Batas tinggi pohon. None berarti pohon bisa tumbuh setinggi yang diperlukan[20].
3	min_samples_split	7	Minimal jumlah data di satu cabang sebelum pohon memecahnya menjadi cabang yang lebih kecil[20].
4	min_samples_leaf	5	Minimal jumlah data di daun (bagian paling ujung pohon) saat pohon selesai tumbuh[20].
5	max_leaf_nodes	300	Jumlah maksimal daun (bagian ujung pohon) yang boleh dibuat[20].
6	ccp_alpha	0.0	Nilai pemangkasan pohon untuk membuang cabang yang kurang penting, nilai 0 berarti tidak memangkas[20].
7	random_state	42	Angka acuan supaya hasil selalu sama saat hutan pohon dibuat ulang[20].
8	class_weight	balanced	Menyeimbangkan bobot kelas agar data yang jarang tetap punya pengaruh dalam keputusan[20].

3.4 Validasi Silang(*Cross Validation*)

Cross-validation merupakan teknik evaluasi model yang digunakan untuk mengukur kemampuan generalisasi model pembelajaran mesin terhadap data yang tidak terlihat sebelumnya[21]. Dalam penelitian ini, digunakan metode Stratified K-Fold Cross Validation dengan 5 lipatan (fold), yang menjaga distribusi proporsi label (kelas) tetap seimbang di setiap fold. Proses ini membagi dataset menjadi lima bagian yang proporsional, kemudian secara bergantian menggunakan empat bagian untuk melatih model dan satu bagian untuk mengujinya. Setiap fold melibatkan proses pembentukan model

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

Random Forest yang dibangun ulang dari awal, sehingga tidak terjadi kebocoran informasi antara data pelatihan dan data pengujian. Di setiap iterasi, dilakukan juga imputasi nilai hilang menggunakan strategi mean pada data pelatihan dan diterapkan secara konsisten ke data pengujian. Hasil prediksi dari kelima fold digabungkan untuk menghasilkan evaluasi keseluruhan model. Dengan pendekatan ini, performa model dapat dinilai secara adil dan menyeluruh, mencerminkan bagaimana model akan bekerja pada data baru yang belum pernah dilihat sebelumnya.

3.5 Evaluasi dan Visualisasi

Bagian ini membahas proses evaluasi dan visualisasi terhadap kinerja model Random Forest yang telah dibangun. Evaluasi dilakukan untuk menilai seberapa baik model dalam mengklasifikasikan trafik jaringan sebagai serangan DDoS atau bukan, serta untuk memastikan bahwa model memiliki performa yang stabil dan dapat diandalkan. Evaluasi yang dilakukan dalam penelitian ini antara lain meliputi classification report, confusion matrix, dan learning curve.

a. *Classification Report*

Classification report merupakan salah satu metode evaluasi yang digunakan untuk mengukur kinerja model klasifikasi secara lebih rinci melalui sejumlah metrik penting, yaitu precision, recall, f1-score, accuracy, macro average, dan weighted average[22]. Precision mengukur sejauh mana prediksi positif yang dihasilkan oleh model benar adanya, sedangkan recall mengukur sejauh mana model mampu mengenali seluruh kasus aktual yang tergolong dalam kelas positif[23]. F1-score merupakan rata-rata harmonik dari precision dan recall, dan sering digunakan sebagai ukuran utama dalam kasus klasifikasi, terutama ketika terdapat ketidakseimbangan kelas[24]. Accuracy menunjukkan persentase total prediksi yang benar terhadap seluruh data yang diuji[25]. Macro average menghitung rata-rata dari precision, recall, dan f1-score masing-masing kelas tanpa memperhitungkan jumlah data di tiap kelas, sementara weighted average menghitung rata-rata yang sama namun dengan mempertimbangkan proporsi jumlah data pada masing-masing kelas[26].

Berdasarkan hasil evaluasi, model Random Forest menunjukkan kinerja yang sangat baik dengan nilai accuracy sebesar 0,99942 atau 99,942%. Precision untuk kelas 0 dan kelas 1 masing-masing sebesar 0,99979 dan 0,99884, sedangkan recall-nya mencapai 0,99928 untuk kelas 0 dan 0,99966 untuk kelas 1. Nilai f1-score yang tinggi, yaitu 0,99953 untuk kelas 0 dan 0,99925 untuk kelas 1, mengindikasikan bahwa model mampu mempertahankan keseimbangan antara precision dan recall. Nilai macro average f1-score tercatat sebesar 0,99939, sedangkan weighted average f1-score mencapai 0,99942, yang menunjukkan bahwa model tidak hanya konsisten dalam menangani dua kelas secara adil, tetapi juga mampu mengakomodasi distribusi data yang tidak sepenuhnya seimbang. Secara keseluruhan, metrik-metrik tersebut mengindikasikan bahwa model memiliki akurasi yang sangat tinggi dan kesalahan klasifikasi yang sangat rendah dalam mendeteksi serangan DDoS maupun trafik normal.

Classification Report:

	precision	recall	f1-score	support
0	0.99979	0.99928	0.99953	61214
1	0.99884	0.99966	0.99925	37852
accuracy			0.99942	99066
macro avg	0.99931	0.99947	0.99939	99066
weighted avg	0.99942	0.99942	0.99942	99066

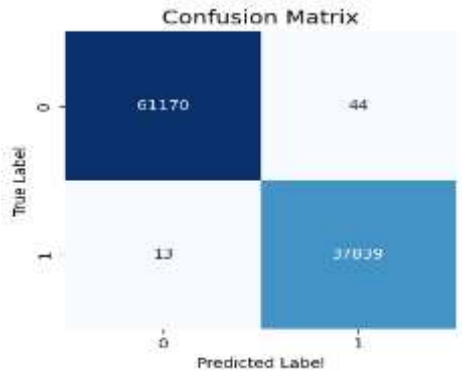
Gambar 2. Classification Report Model Random Forest

b. *Confusion Matrix*

Confusion matrix merupakan salah satu metode evaluasi yang umum digunakan dalam klasifikasi biner untuk menggambarkan performa model dalam bentuk matriks yang menunjukkan jumlah prediksi benar dan salah pada masing-masing kelas[27]. Matriks ini menyajikan perbandingan antara label yang sebenarnya (*true label*) dan label yang diprediksi oleh model, sehingga memudahkan identifikasi terhadap jenis kesalahan yang dilakukan model, seperti *false positives* dan *false negatives*. Tujuan dari evaluasi menggunakan confusion matrix adalah untuk mengetahui sejauh mana model mampu mengklasifikasikan masing-masing kelas dengan benar dan mengukur ketepatan serta kesalahan prediksi yang terjadi secara eksplisit. Berdasarkan hasil confusion matrix yang ditampilkan, diketahui bahwa model berhasil mengklasifikasikan 61.170 sampel kelas 0 (non-DDoS) dengan benar dan 37.839 sampel kelas 1 (DDoS) dengan benar. Sementara itu, terdapat 44 kasus kelas 0 yang salah diklasifikasikan sebagai kelas 1 (*false positives*), dan 13 kasus kelas 1 yang salah diklasifikasikan sebagai kelas 0 (*false negatives*). Hasil ini menunjukkan bahwa kesalahan klasifikasi yang dilakukan oleh model sangat minim, dengan mayoritas prediksi berada pada posisi diagonal utama dalam matriks, yang mengindikasikan prediksi yang benar. Dengan jumlah keseluruhan data yang relatif besar, jumlah kesalahan yang sangat kecil tersebut menunjukkan bahwa model Random

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

Forest memiliki tingkat ketepatan yang sangat tinggi dan mampu membedakan antara trafik normal dan serangan DDoS secara efektif dan konsisten.

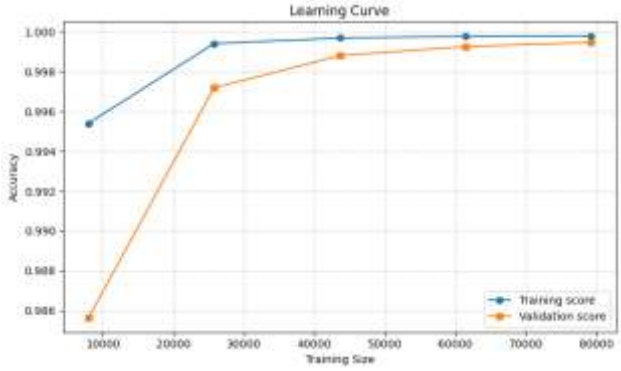


Gambar 3. Confusion Matrix

c. *Learning Curve*

Learning curve merupakan grafik yang menggambarkan hubungan antara ukuran data pelatihan dan performa model, baik pada data pelatihan maupun data validasi. Tujuan dari evaluasi ini adalah untuk memahami bagaimana kinerja model berubah seiring bertambahnya jumlah data yang digunakan selama proses pelatihan, serta untuk mendeteksi potensi masalah seperti overfitting atau underfitting[28]. Dalam grafik learning curve yang ditampilkan, terdapat dua garis utama, yaitu *training score* dan *validation score*, yang masing-masing merepresentasikan akurasi model pada data pelatihan dan data validasi.

Berdasarkan hasil grafik, terlihat bahwa akurasi model pada data pelatihan tetap sangat tinggi dan stabil di atas 0,998, menunjukkan bahwa model mampu mempelajari data pelatihan dengan sangat baik. Di sisi lain, akurasi pada data validasi juga mengalami peningkatan yang signifikan seiring bertambahnya ukuran data pelatihan, dari sekitar 0,986 pada ukuran 10.000 data hingga mencapai lebih dari 0,998 saat mendekati 80.000 data. Perbedaan antara kurva pelatihan dan validasi semakin kecil pada ukuran data yang lebih besar, yang menandakan bahwa model tidak mengalami overfitting maupun underfitting secara signifikan. Pola konvergen antara kedua kurva tersebut menunjukkan bahwa model memiliki generalisasi yang baik dan bahwa jumlah data pelatihan yang digunakan sudah cukup untuk mencapai kinerja optimal. Dengan demikian, learning curve ini mengonfirmasi bahwa model Random Forest yang dibangun memiliki performa yang stabil dan andal dalam mendeteksi serangan DDoS seiring peningkatan ukuran data pelatihan.



Gambar 4. Learning Curve

4. KESIMPULAN

Penelitian ini bertujuan untuk merancang dan mengevaluasi model prediksi serangan Distributed Denial of Service (DDoS) dengan menggunakan algoritma Random Forest. Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa model yang dikembangkan mampu memberikan performa klasifikasi yang sangat tinggi dalam membedakan antara trafik jaringan normal dan serangan DDoS. Hal ini dibuktikan melalui hasil evaluasi menggunakan classification report, confusion matrix, dan learning curve yang menunjukkan nilai akurasi mencapai 99,94%, serta precision, recall, dan f1-score yang tinggi dan seimbang pada kedua kelas. Evaluasi lebih lanjut melalui confusion matrix mengungkapkan bahwa jumlah kesalahan klasifikasi sangat rendah, dengan hanya 44 false positive dan 13 false negative dari total 99.066 sampel

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

yang diuji. Sementara itu, analisis learning curve menunjukkan bahwa model memiliki kemampuan generalisasi yang baik dan tidak menunjukkan indikasi overfitting maupun underfitting yang signifikan.

Meskipun hasil yang diperoleh sangat baik, penelitian ini memiliki beberapa keterbatasan. Salah satunya adalah proses tuning hiperparameter yang masih dilakukan secara manual, yang memungkinkan adanya konfigurasi lain yang lebih optimal namun belum dieksplorasi. Selain itu, eksperimen dilakukan hanya pada satu dataset yang bersumber dari platform Kaggle, sehingga validitas model terhadap variasi jenis serangan atau trafik dari lingkungan yang berbeda belum dapat dipastikan secara menyeluruh. Oleh karena itu, untuk pengembangan penelitian selanjutnya disarankan untuk mengintegrasikan teknik optimasi parameter secara otomatis, seperti grid search atau random search, serta menguji model pada berbagai dataset yang lebih beragam agar dapat meningkatkan generalisasi dan keandalannya di lingkungan nyata. Dengan pendekatan tersebut, diharapkan model prediksi serangan DDoS berbasis Random Forest dapat menjadi solusi yang lebih robust dalam sistem keamanan jaringan.

REFERENCES

[1] A. N. R. N. D. D. A. R. Rahmawati, *Pengantar Teknologi Informasi*, vol. 3, no. 1. Payakumbuh: Serasi Media Teknologi, 2025.

[2] Jovanka Daryl Ruindungan, Sherwin R. U. A. Sompie, and Xaverius B. N. Najoan, "Analisis Kerentanan terhadap Serangan Denial of Service pada Website Universitas Sam RatulangiNgrok," vol. 20, no. 1, pp. 39–50, 2025.

[3] L. D. Samsumar *et al.*, *Keamanan Sistem Informasi: Perlindungan Data dan Privasi di Era Digital*. Bekasi: HADLA MEDIA INFORMASI, 2025. doi: 978-623-10-9466-7.

[4] P. A. C. Setiawan, I. A. D. Giriantari, and N. I. ER, "Tinjauan Literatur: Deteksi Anomali Berbasis Analisis Waktu pada CAN Bus Kendaraan Listrik," *ELECTRON J. Ilm. Tek. Elektro*, vol. 6, no. 1, pp. 72–84, 2025, doi: 10.33019/electron.v6i1.282.

[5] Slamet, "Taksonomi Pertahanan Cyber Security Menggunakan Model Cyber Kill Chain," *Spirit*, vol. 16, no. 1, pp. 232–245, 2024, doi: 10.53567/spirit.v16i1.332.

[6] A. R. Syujak, K. Diantoro, V. Yuni T, A. Soderi, and P. A. Sucipto, "Integrasi Deep Packet Inspection dengan Intrusion Detection System (IDS) untuk Identifikasi Serangan DDoS dalam Jaringan Skala Besar," *J. Minfo Polgan*, vol. 13, no. 2, pp. 1971–1975, 2024, doi: 10.33395/jmp.v13i2.14324.

[7] R. Irfannandhy, L. B. Handoko, and N. Ariyanto, "Analisis Performa Model Random Forest dan CatBoost dengan Teknik SMOTE dalam Prediksi Risiko Diabetes," *Edumatic J. Pendidik. Inform.*, vol. 8, no. 2, pp. 714–723, 2024, doi: 10.29408/edumatic.v8i2.27990.

[8] M. Irdi, R. Aditya, A. Ramdhani, D. D. Nugraha, and M. F. Ari, "Analisis masalah serangan phising pada penggunaan email," 2025.

[9] C. Chandra and D. Prima, "Deteksi Serangan Siber Menggunakan Machine Learning : Studi Pada Sistem Informasi Akademik," vol. 3, no. 2, pp. 106–110, 2025.

[10] T. Yuliswar, I. Elfitri, and O. W. Purbo, "Optimization of Intrusion Detection System With Machine Learning for Detecting Distributed Attacks on Server Optimalisasi Sistem Deteksi Intrusi Menggunakan Machine Learning Untuk Deteksi Serangan Terdistribusi Pada Server," *J. Inovtek Polbeng -Seri Inform.*, vol. 10, no. 1, pp. 1–10, 2025.

[11] Sunardi and Suyahman, "Analisis Komparasi Prediksi Serangan DDoS Menggunakan Machine Learning," pp. 84–91, 2025, [Online]. Available: https://www.kaggle.com/datasets/oktayrdeki/ddos-traffic-dataset/

[12] N. Surojudin, A. Turmudi Zy, D. Maulana, and A. Halim Anshor, "Pengembangan Kebijakan Keamanan Adaptif Berbasis Machine Learning pada Firewall SDN," *J. Pustaka AI (Pusat Akses Kaji. Teknol. Artif. Intell.*, vol. 5, no. 1, pp. 45–49, 2025, doi: 10.55382/jurnalpustakaai.v5i1.919.

[13] "Metodologi Research and Development: Teori dan Penerapan Metodologi RnD - Loso Judijanto, Mas'ud Muhammadiah, Rahmawati Ning Utami, Lalu Suhirman, Laurensius Laka, Yoseb Boari, Suri Toding Lembang, Fegie Yoanti Wattimena, Ningrum Astriawati, Rudy Dwi Laksono, Muhammad Yunus - Google Buku." Accessed: Aug. 09, 2025. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=y3INEQAAQBAJ&oi=fnd&pg=PA1&dq=Metode+R%26D+memadukan+unsur+analisis+data,+perancangan,+implementasi,+dan+pengujian,+sehingga+sesuai+untuk+membangun+solusi+yang+dapat+langsung+diimplementasikan+pada+lingkungan+nyata.&ots=LJGoozBu7O&sig=L7J3uuINFMKQVOLTJ_0V2KI6FHI&redir_esc=y#v=onepage&q&f=false

[14] N. Phan, A. Kristianto, J. Kendrico, and W. J. Alexander, "Perencanaan Enterprise Architecture Sistem Informasi pada Akademik: Studi Literatur," *JDMIS J. Data Min. Inf. Syst.*, vol. 2, no. 2, pp. 50–58, 2024, doi: 10.54259/jdmis.v2i2.1877.

[15] "Buku Ajar Dasar Exploratory Data Analysis (EDA) - Febie Elfaladonna, Indra Griha Tofik Isa, Devi Sartika, Yusniarti, Andre Mariza Putra - Google Buku." Accessed: Aug. 10, 2025. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=n7ofEQAAQBAJ&oi=fnd&pg=PR1&dq=Exploratory+Data+Analysis+(EDA),+yaitu+proses+eksplorasi+terhadap+data+guna+memperoleh+pemahaman+mengenai+struktur,+pola,+dan+karakteristik+distribusi+data+sebelum+dilakukan+proses+pemodelan&ots=2R4JE4lCvf&sig=bLIAbminHdPANv8pYuZ3YG5Xzqw&redir_esc=y#v=onepage&q&f=false

[16] "Decoding Intelligence Algoritma Machine Learning dalam Aksi dan Bisnis - Muhamad Malik Mutoffar, ST., MM, Endang Retnoningsih, S.Kom., M.Kom, Dr. Ir. Yudi Limbar Yasik, M.Sc, Eliza, S.T - Google Buku." Accessed: Aug. 10, 2025. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=c6xZEQAAQBAJ&oi=fnd&pg=PA1&dq=EDA+memiliki+peran+penting+dalam+mengidentifikasi+potensi+masalah+pada+data,+seperti+keberadaan+nilai+ekstrem+(outlier),+kesalahan+entri+data,+dan+ketidakseimbangan+distribusi+kelas,+serta+dalam+menentukan+strategi+preprocessing+yang+sesuai&ots=fECrX3

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 6, No 3, September 2025, Hal. 247-256
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2203
https://journal.fkpt.org/index.php/BIT

-62k&sig=hY-KdVGn6nv5U7jbqCbUSjpPEcw&redir_esc=y#v=onepage&q&f=false

[17] L. Nurlaela, Y. Suhanda, A. Sopian, C. S. Dewi, and R. Syahrial, "Pengembangan Framework Data Mining Berbasis Deep Neural Network Dengan Eksplorasi Teknik Transfer Learning Untuk Prediksi Dan Klasifikasi Data," *JRIS J. Rekayasa Inf. Swadharma*, vol. 5, no. 1, pp. 132–141, 2025, doi: 10.56486/jris.vol5no1.723.

[18] B. Ramadhan and S. F. Pane, "Pengaruh Hyperparameter Tuning untuk Efektivitas pada Pendekatan Hybrid dalam Mendiagnosis Stres dan Depresi : Tinjauan Studi Literatur," *J. Tekno Insentif*, vol. 18, no. 2, pp. 104–118, 2024, doi: 10.36787/jti.v18i2.1516.

[19] "Dasar dan Konsep Machine Learning - Feri Sulianta - Google Buku." Accessed: Aug. 10, 2025. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=l5BwEQAAQBAJ&oi=fnd&pg=PA1&dq=overfitting+yaitu+kondisi+ketika+model+terlalu+menyesuaikan+diri+terhadap+data+pelatihan+sehingga+kehilangan+kemampuan+generalisasi+dan+underfitting+yaitu+kondisi+ketika+model+gagal+menangkap+pola+penting+dari+data+karena+terlalu+sederhana&ots=0vkdabgIKo&sig=wXj2LentHi1gxtqCoTB_4-ROwIE&redir_esc=y#v=onepage&q&f=false

[20] "RandomForestClassifier — scikit-learn 1.7.1 documentation." Accessed: Aug. 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[21] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," *Comput. Sci.*, vol. 4, no. 1, pp. 62–70, 2024, doi: 10.31294/coscience.v4i1.2389.

[22] R. Fauzan, A. V. Vitianingsih, D. Cahyono, A. L. Maukar, and Y. A. B. Suprio, "Penerapan Algoritma Klasifikasi pada Machine Learning untuk Deteksi Phishing," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 2, pp. 531–540, 2025, doi: 10.57152/malcom.v5i2.1968.

[23] R. Alfarez, R. Rianto, and V. Purwayoga, "Penerapan Naïve Bayes untuk Prediksi Customer Churn (Studi Kasus: PT Hutchison 3 Indonesia)," *J. Ris. dan Apl. Mhs. Inform.*, vol. 5, no. 2, pp. 301–307, 2024, doi: 10.30998/jrami.v5i2.8556.

[24] A. Arasy and S. Agustian, "Sentiment Classification Using Multilayer Perceptron Algorithm with TF-IDF Features Klasifikasi Sentimen Menggunakan Metode Multilayer Perceptron dengan Fitur TF-IDF," vol. 5, no. July, pp. 908–919, 2025.

[25] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[26] F. M. Natsir, R. Y. Bakti, and T. Wahyuni, "Analisis Deteksi Dini Penyakit Jantung dengan Pendekatan Support Vector Machine pada Data Pasien," *Arus J. Sains dan Teknol.*, vol. 2, no. 2, pp. 437–446, 2024, doi: 10.57250/ajst.v2i2.669.

[27] F. Reynaldi Valerian, M. Syarief, and D. Abdul Fatah, "Klasifikasi Tingkat Obesitas Menggunakan Metode Gbm Dan Confusion Matrix," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 2242–2249, 2025, doi: 10.36040/jati.v9i2.13062.

[28] M. A. Saputra and T. Sugihartono, "Evaluasi Kinerja Model LSTM Untuk Prediksi Risiko Penyakit Jantung Menggunakan Dataset Evaluasi Kinerja Model LSTM Untuk Prediksi Risiko Penyakit Jantung Menggunakan Program Studi Teknik Informasi , Fakultas Teknologi Informasi , ISB Atma Luhur , Indonesia Performance Evaluation of the LSTM Model for Heart Disease Risk Prediction Using a Dataset," no. July, 2025, doi: 10.52436/1.jpti.821.

[29] Karim, A., Bangun, B., Prayetno, S., & Afrendi, M. (2025). Optimasi Prediksi Harga Sawit Menggunakan Teknik Stacking Algoritma Machine Learning dan Deep Learning dengan SMOTE.