JTIM : Jurnal Teknologi Informasi dan Multimedia

p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim

Perbandingan *Support Vector Machine, Random Forest Classifier,* dan *K-Nearest Neighbour* dalam Pendeteksian Anomali pada Jaringan DDos

Haeruddin [1], Erick [1], Heru Wijayanto Aripradono [1]

[1] Program Studi Teknologi Informasi, Universitas Internasional Batam

* Korespondensi: haeruddin@uib.ac.id

Sitasi: Haeruddin, H.; Erick, E.; Aripradono, H. W. (**2025**). Perbandingan *Support Vector Machine, Random Forest Classifier*, dan *K-Nearest Neighbour* dalam Pendeteksian Anomali pada Jaringan DDos. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 7(1), 23-33. https://doi.org/10.35746/jtim.v7i1.628

Diterima: 08-11-2024
Direvisi: 28-11-2024
Disetujui: 03-12-2024

Abstract: A Distributed Denial of Service (DDoS) attack poses a serious threat to network security and can disrupt online services by overwhelming the target server with excessive traffic. Effective detection of DDoS attacks requires a system capable of identifying anomalies in network traffic. In this context, Machine Learning (ML) offers an effective approach for classification and anomaly detection. However, different ML algorithms have varying strengths and weaknesses when processing large and complex network data. Therefore, this study aims to evaluate the performance of three ML algorithms: Support Vector Machine (SVM), Random Forest Classifier (RFC), and K-Nearest Neighbors (KNN) in detecting DDoS anomalies. The dataset used consists of 225,745 data points with 85 attributes that describe various characteristics of network traffic, such as destination port, flow duration, packet count, and packet size. This dataset is classified into two classes, BENIGN and DDoS, representing normal traffic and DDoS attacks, respectively. Evaluation is performed using several performance metrics, including accuracy, precision, recall, MCC (Matthews Correlation Coefficient), F-Measure, ROC Area, PRC Area, True Positive Rate (TPR), and False Positive Rate (FPR). The results show that the Random Forest Classifier (RFC) delivers the best performance with an accuracy of 99.99%, precision of 99.98%, recall of 100%, and a very low FPR of 0.02%. This is followed by the Support Vector Machine (SVM) with an accuracy of 99.91%, and the K-Nearest Neighbor (KNN) with an accuracy of 99.98%. All three algorithms demonstrate strong performance in detecting DDoS anomalies, with RFC slightly outperforming others in terms of consistency and higher classification capability. The findings of this study provide valuable insights for selecting the best algorithm to detect DDoS attacks in networks.

Keywords: Classification, DDoS, K-Nearest Neighbors, Random Forest Classifier, Support Vector Machine.

Abstrak: Serangan DDoS (Distributed Denial of Service) menimbulkan ancaman serius terhadap keamanan jaringan dan dapat menyebabkan gangguan layanan online dengan membanjiri server target dengan lalu lintas yang berlebihan. Deteksi serangan DDoS yang efektif memerlukan sistem yang dapat mendeteksi anomali pada lalu lintas jaringan. Dalam konteks ini, Machine Learning (ML) memberikan pendekatan yang efektif untuk klasifikasi dan deteksi anomali. Namun, algoritme ML yang berbeda memiliki kekuatan dan kelemahan yang berbeda saat memproses data jaringan yang besar dan kompleks. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi kinerja tiga algoritma ML: Support Vector Machine (SVM), Random Forest Classifier (RFC), dan K-Nearest Neighbors (KNN) dalam mendeteksi anomali DDoS. Dataset yang digunakan terdiri dari 225.745 buah data dengan 85 atribut yang menggambarkan berbagai karakteristik lalu lintas jaringan, seperti port tujuan, durasi aliran, jumlah paket, dan ukuran paket. Kumpulan data ini dibagi menjadi dua kelas, BENIGN dan DDoS, yang masing-masing mewakili lalu lintas normal dan serangan

DDoS. Evaluasi menggunakan beberapa metrik kinerja termasuk akurasi, precision, recall, MCC (Matthews Correlation Coefficient), F-Measure, ROC Area, PRC Area, True Positive Rate (TPR), dan False Positive Rate (FPR). Hasilnya menunjukkan Random Forest Classifier (RFC) memberikan performa terbaik dengan akurasi 99,99%, presisi 99,98%, recall 100%, dan FPR sangat rendah yaitu 0,02%. Disusul Support Vector Machine (SVM) dengan akurasi 99,91% dan K-Nearest Neighbor (KNN) dengan akurasi 99,98%. Ketiga algoritma tersebut menunjukkan kinerja yang baik dalam mendeteksi anomali DDoS, dengan RFC sedikit lebih unggul dalam hal konsistensi dan kemampuan klasifikasi yang tinggi. Hasil penelitian ini memberikan wawasan berharga dalam memilih algoritma terbaik untuk mendeteksi serangan DDoS pada jaringan.

Kata kunci: DDoS, Klasifikasi, K-Nearest Neighbors, Random Forest Classifier, Support Vector Machine.

1. Pendahuluan

Sistem deteksi intrusi menjadi senjata utama dalam mengamankan jaringan, bertugas mengidentifikasi dan mencatat apakah suatu paket data merupakan serangan atau tidak. Meskipun telah ada aplikasi seperti IDS (Intrusion Detection System), sebagian besar masih menggunakan aturan dan tanda tangan, sedangkan hanya sedikit yang mengandalkan pendekatan anomali [1]. Anomali diartikan sebagai aktivitas yang tidak lazim dalam data yang normal. Salah satu jenis anomali yang banyak dijumpai berupa DDoS. DDoS merupakan jenis serangan siber yang menuntut sejumlah besar perangkat ataupun sistem komputer terhubung secara bersamaan ke suatu jaringan internet [2]. Sistem tersebut umumnya melakukan penyerangan terhadap suatu target, misalnya suatu jaringan, dengan tujuannya untuk memberikan suatu gangguan atas ketersediaan layanan bagi pengguna yang sah. Serangannya ditujukan untuk membua sistem target tidak dapat diakses para pengguna sah dengan membanjiri target dengan lalu lintas data yang tidak biasa atau tidak diinginkan [3].

DDos memiliki karakteristik utama, berupa penggunaan jaringan terdistribusi. Dalam konteks ini, serangan tersebut banyak diselenggarakan dari berbagai sumber yang tersebar di berbagai lokasi [4]. Hal ini menyebabkan jenis serangan DDos cenderung sulit untuk dicegah dan dideteksi, mengingat lalu lintas dapat datang dari berbagai alamat IP yang berbeda [5][6]. Serangannya dapat dilakukan menggunakan berbagai cara, misalnya serangan dengan volume tinggi (misalnya, melalui UDP flood atau ICMP flood), serangan yang memanfaatkan kelemahan protokol atau aplikasi (misalnya, HTTP flood), maupun bentuk serangan yang memanfaatkan perangkat yang telah terinfeksi malware untuk menjadi bagian dari jaringan penyerang yang terdistribusi (botnet) [7].

Deteksi DdoS merupakan proses yang penting dalam mengupayakan penjagaan atas kemanan-keamanan jaringan, khususnya dalam menghadapi serangan yang melibatkan banyak perangkat dan komputer disaat yang sama dalam penyerangan terhadap suatu target dengan tujuan mengganggu ketersediaan layanan [5]. Metode deteksi DDoS melibatkan pengawasan lalu lintas jaringan untuk mengindentifikasi pola atau perilaku yang tidak biasa yang dapat menunjukan adanya serangannya. Adapun metode data mining ditandai sebagai metode yang cukup efektif untuk mendeteksi intrusi berbasis anomali DDoS [7]. Dalam konteks ini, keunggulan dari pendekatan data mining terletak pada kemampuannya dalam memproses data logs dan auditing dalam jumlah besar, serta efektifitasnya dalam mengintergrasikan teknik pendeteksian anomali. Aplikasi data mining juga memungkinkan identifikasi pola rutin dalam dataset besar, serta menyediakan solusi untuk mengatasi masalah reduksi dataset, sehingga memperkecil beban kerja analisis dalam mengidentifikasi data dan menyampaikan analisis [8].

Salah satu metode data mining yang terkenal dengan tingkat akurasi tinggi adalah Support Vector Machine (SVM). SVM merupakan algoritma klasifikasi yang mampu bekerja baik untuk data linear maupun non-linear, dengan menerapkan pemetaan non-linear untuk mengubah data latih ke dimensi yang lebih tinggi [9]. Pendekatan tersebut ditandai efektif untuk mengidentifikasi pola paket data jaringan, seperti yang dibuktikan oleh penelitian sebelumnya, sehingga dinilai efektif dan akurat dalam mendeteksi anomali DDoS pada jaringan. SVM memiliki kemampuan untuk menemukan hyperplane terbaik yang memisahkan dua kelas data dengan jarak maksimum, sehingga cocok untuk menangani data yang kompleks [10].

Kemudian, terdapat metode lainnya berupa Random Forest Classifier (RFC). Pendekatan ini melibatkan atas penggabungan berbagai prediksi dari pohon keputusan berbeda dalam membentuk suatu hasil yang lebih akurat dan stabil [11]. Setiap pohon keputusan dalam Random Forest Classifier pembentukannya terselenggara secara independen menggunakan sampel acak dari data pelatihan dan subset acak dari fitur. Selanjutnya, hasil prediksi dari setiap pohon digabungkan melalui suara mayoritas atau rata-rata untuk menentukan kelas yang paling mungkin [12]. Keunggulan utama dari Random Forest Classifier terletak pada kemampuannya untuk menangani data yang kompleks dan beragam dengan baik, termasuk data dengan fitur yang tidak terstruktur atau kurangnya pola yang jelas. Selain itu, Random Forest Classifier juga memiliki kemampuan untuk mengatasi overfitting, karena penggunaan banyak pohon keputusan yang berbeda mengurangi risiko overfitting pada model [13]. Dengan adanya kombinasi prediksi dari banyak pohon keputusan, serta dapat memberikan akurasi yang tinggi dan skalabilitas yang baik, maka metode ini dinilai efektif dalam mengatasi dataset DDoS besar.

Di sisi lain, K Nearest Neighbor (KNN) merupakan metode klasifikasi yang sederhana namun efektif dalam bidang data mining dan machine learning. Konsep dasar dari KNN adalah dengan upaya membandingkan data baru yang akan diprediksi dengan data pelatihan yang sudah ada, lalu menentukan kelas atau label data baru berdasarkan mayoritas kelas dari tetangga terdekatnya. Jumlah tetangga yang digunakan dalam KNN (ditentukan oleh parameter k) dapat menyatakan pengaruh atas tingkat akurasi dan sensitivitas model [14]. Dalam konteks ini, KNN cenderung bergantung pada jarak antara titik data dan tetangga terdekatnya, dan dapat memberikan akurasi yang baik terutama untuk data dengan struktur cluster yang jelas. KNN cenderung membutuhkan waktu komputasi yang lebih tinggi dan kurang efektif dalam menangani dataset besar [15].

Beberapa studi telah meneliti pendeteksian anomali pada data, beberapa studi-studi tersebut melakukan metode algoritma SVM, KNN, dan RFC. Penelitian studi tersebut dilakukan diberbagai daerah seperti German, China, Iran, India, Turkey, Malaysia, Arabia, Thailand dan Croatia. [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

Kebanyakan studi berfokus pada pendeteksian anomaly pada data [16], menggunakan metode RFC untuk mendeteksi anomaly pada jaringan dan hasil akurasi menunjukan di angka 78.69%. Dan [21] menggunakan algoritma KNN sebagai pendeteksiannya dan hasil yang dimiliki adalah KNN akurasi 90.913%, recall 91.283%, precision 90.302%, F-measure 90.790 [25] menggunakan 2 metode algoritma yaitu RFC dan SVM dengan hasil akhirnya yaitu SVM akurasi 92.5%, F1-scores 85.2%, precision 78.2%, recall 93.6%, dan hasil akurasi RFC 99.84%. Begitupun juga dengan penelitian [26]. yang menyatakan bahwa SVM menghasilkan tingkat akurasi 90.0% dan F1-scores 95.0%. Selanjutnya juga terdapat penelitian yang memberikan temuan kesempurnaan dan efektivitas RFC dalam mendeteksi intrusi pada dataset penyakit, dimana memberikan tingkat akurasi 100.0%, recall 100.0%, precision 100.0%, dan ROC-AUC 100.0% [27].

Terdapat salah satu peneliti hanya menggunakan log database yaitu [19] yang menjelaskan bahwa hasil akhir Improved KNN lebih signifikan dari pada KNN tradisional, hasil menunjukan improved KNN akurasi 92.63% dan 91.8%, KNN akurasi 91.03%. Model

ML dapat mengidentifikasi pola serangan DDoS secara efektif. Hasil penelitian menunjukkan bahwa algoritma K-Nearest Neighbor memiliki performa yang lebih baik dibandingkan model pembelajaran konvensional dalam mendeteksi serangan DDoS, dan hasil akhirnya adalah KNN akurasi 98.51%, recall 97.8%, precision 98.9%, f-measure 1.0005, error rate 1.50, efficiency 98.48%.

Dalam konteks ini, dapat dinyatakan bahwa sebagian besar dari sistem deteksi intrusi yang tersedia saat ini masih mengandalkan metode berbasis aturan ataupun tanda tangan yang terbatas dalam kemampuannya untuk mendeteksi serangan baru yang belum dikenal [28], [29]. Maka dari itu, lahir urgensi untuk melaksanakan pengembangan metode deteksi berbasis anomali, yang lebih adaptif dan efektif dalam menangani serangan DDoS, sangat dibutuhkan.

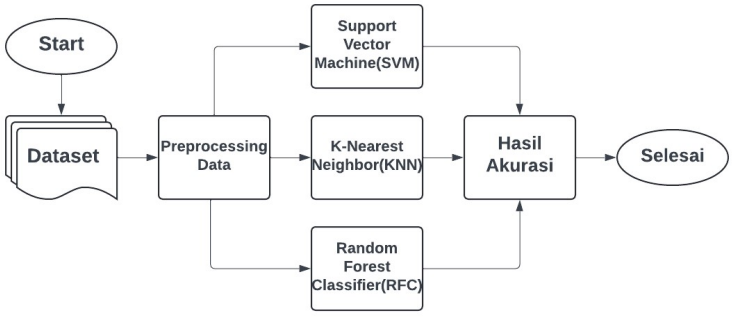
Kemudian, meskipun penelitian terdahulu telah menunjukkan efektivitas algoritma SVM, KNN, serta RFC, masih tersedia ruang untuk meningkatkan akurasi deteksi, khususnya dalam mengidentifikasi serangan DDoS dengan tingkat keakuratan yang lebih tinggi dan mengurangi kemungkinan kesalahan dalam klasifikasi [21]. Oleh sebabnya, dibutuhkan penelitian yang lebih mandalam untuk dapat mengoptimalkan model-model ini, memperbaiki hasil deteksi, serta menguji metode-metode baru yang dapat meningkatkan performa deteksi secara keseluruhan. Dengan menginput metrik evaluasi tambahan, seperti halnya dengan F1-Score, Recall, Precision, ROC-Area, FPR, dan TPR, maka penelitian ini dapat menyediakan pemahaman secara komprehensif terkait dengan kinerja model deteksi DDoS. Metrik-metrik diatas sangat penting untuk dapat menilai sejauh mana model dapat mengidentifikasi serangan dengan akurat, serta meminimalkan kesalahan yang dapat terjadi dalam pengklasifikasian data [30], [31]. Evaluasi secara menyeluruh ini tidak hanya memungkinkan perbandingan antara berbagai algoritma, tetapi juga memberikan *insight* yang lebih mendalam mengenai kelebihan dan kekurangan masing-masing model, yang juga dapat digunakan untuk menyempurnakan sistem deteksi intrusi.

Oleh karena itu, penelitian ini dilakukan untuk meningkatkan hasil akurasi pada pendeteksian anomaly pada jaringan dari para komuniti yang ada di kaggle dan menambahkan Precision, Recall, F1-Score, MCC, ROC-Area, TPR, dan FPR sebagai memperjelas hasil akhirnya.

2. Bahan dan Metode

2.1. Perancangan Penelitian

Jenis penelitian yang akan dilakukan adalah untuk melihat kemampuan klasifikasi mana yang level adaptasinya lebih baik dalam mendeteksi anomali pada jaringan. Data yang dipakai dalam penelitian adalah dataset yang diambil dari alamat website kaggle.



Gambar 1. Flowchart ML

Diagram yang ditampilkan merepresentasikan alur kerja proses evaluasi performa algoritma pembelajaran mesin untuk klasifikasi. Proses dimulai dengan dataset sebagai input, yang kemudian melalui tahap preprocessing data untuk memastikan data siap digunakan, misalnya dengan normalisasi, penghapusan nilai hilang, atau pembagian data menjadi set pelatihan dan pengujian. Data yang telah diproses kemudian digunakan sebagai input untuk tiga algoritma pembelajaran mesin, yaitu Support Vector Machine (SVM), K-Nearest Neighbor (KNN), dan Random Forest Classifier (RFC), yang dijalankan secara paralel. SVM bertujuan untuk memisahkan kelas data dengan hyperplane terbaik, KNN memprediksi kelas berdasarkan kedekatan data dengan tetangga terdekat, sedangkan RFC menggunakan pendekatan ensemble learning dengan membangun beberapa pohon keputusan untuk menghasilkan prediksi akhir. Hasil performa dari setiap algoritma kemudian dievaluasi menggunakan berbagai metrik, seperti akurasi, precision, recall, dan F-Measure, untuk menentukan algoritma yang paling optimal.

2.2. Dataset

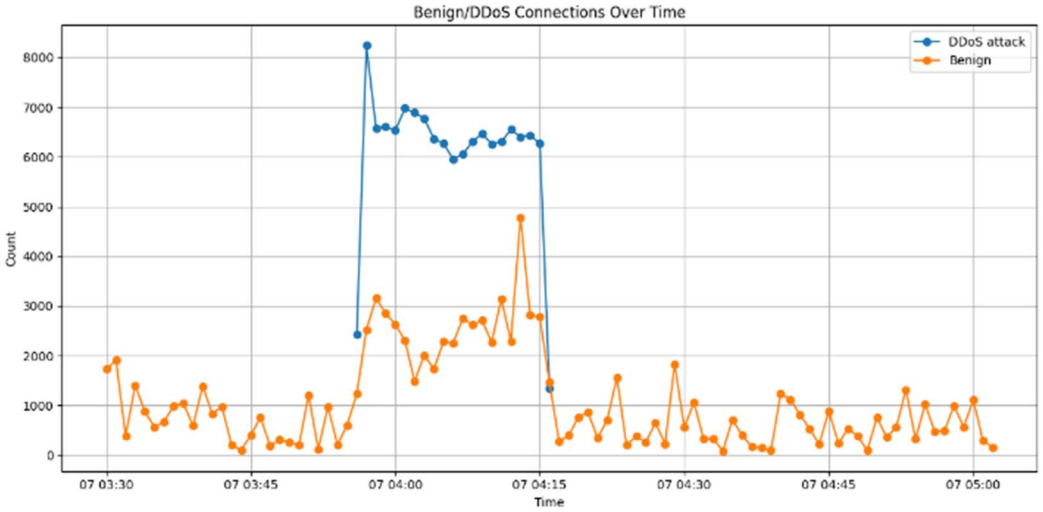
Pengumpulan dataset yang ambil merupakan dataset yang telah disediakan oleh website yang alamatnya Kaggle. Kaggle merupakan tempat situs scientist untuk berbagi ide dan bersaing. Kemudian dataset tersebut akan dipakai sebagai klasifikasi pada algoritma. Alamat website dataset diambil dari https://www.kaggle.com/code/aymenabb/ml-anomaly-training/input yang diakses pada tanggal 5 april 2024.

Dataset yang digunakan pada penelitian ini terdiri dari 225.745 data dengan 85 atribut yang mewakili berbagai karakteristik lalu lintas jaringan. Atribut ini mencakup karakteristik seperti Destination Port, Flow Duration, Total Fwd Packets, Flow Bytes/s, Average Packet Size, Idle Mean, dan Label. Data ini dibagi menjadi dua kelas: BENIGN, yang mewakili lalu lintas jaringan normal, dan DDOS, yang mewakili lalu lintas jaringan yang mewakili DDoS. Dataset ini disimpan dalam format CSV (comma Separated Values), sehingga memudahkan dalam mengolah dan menganalisis data.

Tabel 1. Contoh Dataset

Destination Port	Flow Duration	Total Fwd Packets	Flow Bytes/s	Flow Packets/s	Average Packet Size	Idle Mean	Label
54865	3	2	4000000	6.666.666.667	9	0	BENIGN
55054	109	1	1.100.917.431	1.834.862.385	9	0	BENIGN
55055	52	1	2.307.692.308	3.846.153.846	9	0	BENIGN
46236	34	1	3.529.411.765	5.882.352.941	9	0	BENIGN
80	8000148	7	1.457.098.044	1.624.969.938	8.971.538.462	6052010	DDoS
80	10590742	4	2.266.130.173	0.377688362	7.5	10600000	DDoS
80	10584710	5	283.427.699	0.472379498	7.2	10600000	DDoS
80	10583724	5	2.834.541.037	0.472423506	7.2	10600000	DDoS

Pada Tabel 1 berikut merupakan beberapa contoh atribut yang ada pada dataset dan merupakan memiliki peran penting dalam melatih ML. Beberapa atribut utama meliputi Destination Port, yang menunjukkan port tujuan komunikasi; Flow Duration, yang merepresentasikan durasi aliran data dalam jaringan; Total Fwd Packets, jumlah paket yang diteruskan selama koneksi; serta Flow Bytes/s dan Flow Packets/s, yang mencatat rata-rata jumlah byte dan paket yang ditransfer per detik. Selain itu, atribut seperti Average Packet Size dan Idle Mean memberikan informasi tambahan mengenai ukuran rata-rata paket data dan jeda rata-rata antar aliran. Dataset ini diklasifikasikan ke dalam dua label, yaitu BENIGN, yang mewakili lalu lintas jaringan normal, dan DDoS, yang mengindikasikan serangan terhadap jaringan. Data ini mencakup variasi pola aliran data yang signifikan antara lalu lintas normal dan anomali, seperti perbedaan durasi aliran dan kecepatan transfer data, yang menjadi dasar bagi model pembelajaran mesin dalam mengenali dan membedakan kedua kelas tersebut.



Gambar 2. Visualisasi Dataset

Menunjukkan visualisasi data yang menggambarkan jumlah jaringan (DDoS attack) dan koneksi normal (benign) dalam suatu jaringan selama periode waktu tertentu. Sumbu X mewakili waktu, sedangkan sumbu Y menunjukkan jumlah jaringan. Garis biru mewakili jumlah koneksi DDoS, sementara garis oranye mewakili jumlah koneksi benign.

Visualisasi data menunjukkan pola fluktuasi yang signifikan pada jumlah koneksi berbahaya (DDoS) dan koneksi normal (benign) dalam jaringan selama periode pengamatan. Terdapat beberapa puncak yang mengindikasikan terjadinya serangan DDoS dalam skala besar pada waktu-waktu tertentu. Hal ini menunjukkan bahwa jaringan tersebut sering menjadi target serangan siber.

2.3. Preprocessing

Preprocessing merupakan jantung dari machine learning itu sendiri hal yang harus dilakukan untuk membuat algoritma machine learning tersebut untuk berjalan. Proses tersebut contohnya seperti memperbaiki error, menghapus koma, dll. Berikut adalah tahapan-tahapan preprocessing yang dilakukan.

1) Cleaning data : proses identifikasi dan perbaikan atau penghapusan data yang tidak akurat, tidak lengkap, duplikat, atau tidak relevan dari dataset untuk meningkatkan kualitas dan konsistensi data yang akan digunakan dalam analisis atau pemodelan.
2) Transformasi Data: : untuk menyamakan skala data, encoding variabel kategorikal menjadi bentuk numerik menggunakan teknik seperti label encoding, serta penerapan transformasi logaritmik untuk mengurangi skewness pada distribusi data, contoh BENIGN = [0] dan DDoS = [1].
3) Split Data: merupakan langkah penting untuk memastikan evaluasi model dilakukan secara objektif dan menghindari bias. Dataset dibagi menjadi tiga subset utama, yaitu data pelatihan (training set), data validasi (validation set), dan data pengujian (test set).

2.4. Analisa Algoritma

2.4.1 SVM

SVM adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini mencari hyperplane optimal dalam ruang fitur yang memisahkan dua kelas (misalnya, BENIGN dan DDoS) dengan margin terbesar. SVM menggunakan

fungsi kernel untuk memetakan data ke ruang dimensi lebih tinggi, memungkinkan pemisahan non-linear yang efektif. Algoritma ini sangat efektif untuk dataset dengan banyak fitur dan sangat baik dalam mengatasi overfitting.

- Kernel Functions: Beberapa jenis kernel yang digunakan adalah linear, polynomial, dan Radial Basis Function (RBF). Pemilihan kernel yang tepat sangat penting untuk kinerja model.
- Margin Maximization: SVM berusaha untuk memaksimalkan margin antara kelas-kelas dalam data pelatihan untuk meningkatkan kemampuan generalisasi model.

2.4.2 RFC

Random Forest adalah algoritma ensemble yang menggunakan banyak pohon keputusan (decision trees) untuk membuat prediksi. Setiap pohon dalam hutan memberikan hasil klasifikasi, dan hasil akhir diperoleh melalui mayoritas suara. Salah satu kekuatan utama RF adalah kemampuannya untuk mengurangi overfitting yang sering terjadi pada pohon keputusan tunggal. RFC juga dapat menangani data dengan banyak fitur dan nilai yang hilang (missing values).

- Bagging: RF menggunakan teknik bootstrap aggregation (bagging), di mana data pelatihan diambil secara acak dan dengan penggantian untuk membangun beberapa pohon keputusan.
- Feature Selection: Setiap pohon dalam RF hanya menggunakan subset fitur acak, yang meningkatkan keberagaman pohon dan mengurangi korelasi antar pohon.

2.4.3 KNN

KNN adalah algoritma pembelajaran yang berbasis instance-based learning, yang berarti model tidak membangun model eksplisit. KNN mengklasifikasikan data baru berdasarkan kedekatannya dengan data yang sudah ada (tetangga terdekat). Nilai k menentukan jumlah tetangga terdekat yang digunakan untuk membuat keputusan klasifikasi.

- Distance Metric: KNN mengandalkan metrik jarak seperti Euclidean atau Manhattan untuk mengukur kedekatan antara data baru dengan data pelatihan.
- Lazy Learning: KNN tidak memerlukan fase pelatihan eksplisit, sehingga sangat bergantung pada kecepatan dan efisiensi dalam menghitung jarak antar titik data saat klasifikasi.

Tabel 2. Tabel Perbandingan

Algoritma	Keunggulan	Kelemahan	Aplikasi Umum
SVM	Akurasi tinggi, efektif untuk data non-linear, bagus untuk data besar	Waktu pelatihan lambat, sulit menangani dataset besar tanpa tuning	Pendeteksian pola, klasifikasi teks
RFC	Robust terhadap overfitting, mudah diinterpretasikan, mengatasi data hilang	Model lebih kompleks, lebih lambat dalam pelatihan dibandingkan pohon keputusan tunggal	Deteksi anomali, prediksi regresi
KNN	Sederhana, mudah diimplementasikan, efektif untuk dataset kecil	Tidak efisien pada dataset besar, sangat tergantung pada jarak	Klasifikasi citra, prediksi pola

3. Hasil

3.1. Hasil Pengujian Machine learning Secara Offline

Pada tahap pengujian, sistem akan menguji hasil dari tiga metode machine learning. Model ini dibangun selama fase pelatihan data dan digunakan selama pengujian data. Sistem ini akan diuji untuk melihat kinerjanya dari segi kinerja. Data pelatihan dan proses data menggunakan bahasa Python yang berjalan di Google Colab.

Pengujian yang akan dilakukan terhadap tiga algoritma machine learning tersebut masing-masingnya akan di uji setiap akurasinya, contoh penjelasan setiap akurasinya seperti berikut:

1) Akurasi Model : Metrik yang menunjukkan seberapa baik model algoritma tersebut memprediksi hasil yang benar.
2) Precision : Mengukur proporsi prediksi positif yang benar-benar positif.
3) Recall : Mengukur proporsi data positif asli yang diidentifikasi dengan benar.
4) MCC (Matthews Correlation Coefficient) : Mengukur keseimbangan antara Precision dan Recall, dengan mempertimbangkan True Positive, True Negative, False Positive, dan False Negative.
5) F-Measure : Rata-rata harmonis antara precision dan recall.
6) ROC Area (Receiver Operating Characteristic Area) : Mengukur kemampuan model membedakan antara data positif dan negatif di semua ambang batas klasifikasi.
7) PRC Area (Precision-Recall Curve Area) : Mengukur kemampuan model memprioritaskan data positif di semua ambang batas klasifikasi.
8) FPR (False Positive Rate) : Mengukur proporsi data negatif yang salah diprediksi sebagai positif.
9) TPR (True Positive Rate) : Mengukur proporsi data positif yang benar diprediksi sebagai positif.Teks berlanjut di sini.

3.2. Evaluasi Perbandingan

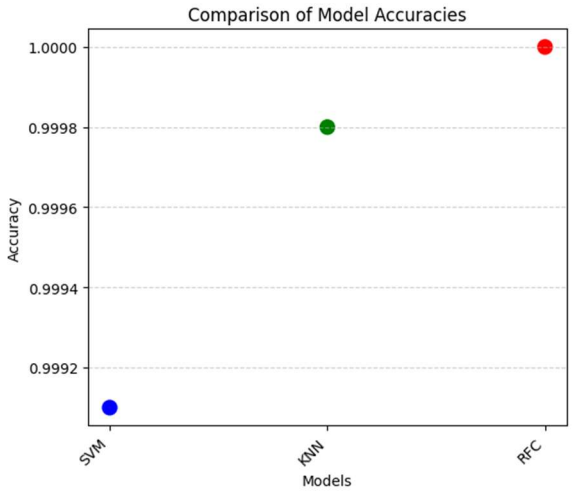
Berikut hasil evaluasi dari masing-masing algoritma yang dipakai. Dengan sembilan indikator evaluasi tersebut.

Tabel 2. Performansi model

Metode	Akurasi	Precision	Recall	MCC	F-Measure	ROC Area	PRC Area	TPR	FPR
SVM	99,91%	99,90%	99.94%	99,81%	99,92%	99,90%	99,87%	99,92%	0,10%
RFC	99,99%	99,98%	100%	99,98%	99,99%	99,99%	99,98%	100%	0,02%
KNN	99,98%	99,98%	99,98%	99,95%	99,98%	99,98%	99,97%	99,98%	0,02%

Dari hasil di atas, Random Forest Classifier (RFC) memberikan performa terbaik dengan akurasi 100% di semua metrik utama. SVM dan KNN juga menunjukkan kinerja sangat baik tetapi sedikit di bawah RFC. RFC cocok digunakan untuk masalah klasifikasi kompleks dengan data yang beragam, karena stabilitasnya dalam memisahkan kelas positif dan negatif. Sementara itu:

- SVM bisa dipilih untuk kasus dengan kebutuhan lebih cepat karena kompleksitas model lebih rendah dibanding RFC.
- KNN cocok untuk dataset yang lebih kecil, mengingat performanya mendekati RFC namun biasanya membutuhkan lebih banyak waktu komputasi untuk dataset besar.
- Studi ini menunjukkan pentingnya memilih model yang sesuai dengan kebutuhan aplikasi, mempertimbangkan trade-off antara akurasi, kompleksitas, dan interpretasi hasil. RFC dapat direkomendasikan sebagai model andalan berdasarkan hasil ini, tetapi SVM dan KNN tetap relevan tergantung konteks penggunaan.



Gambar 3. Komparasi Model

Pada gambar 3 menunjukan perbandingan akurasi dari tiga model ML, yaitu SVM, RFC, dan KNN. Sumbu X mewakili nama-nama model, sedangkan sumbu Y menunjukkan nilai akurasi yang berisar antara 0,9991 hingga 1. Setiap model diwakili oleh sebuah titik berwarna, dengan warna yang berbeda untuk membedakan masing-masing model. Evaluasi perbandingan kinerja model menunjukkan bahwa Random Forest Classifier (RFC) memberikan hasil akurasi tertinggi sebesar hampir 1, diikuti oleh K-Nearest Neighbors (KNN). Model Support Vector Machine (SVM) menunjukkan akurasi yang sedikit lebih rendah dibandingkan dua model lainnya. Hasil ini mengindikasikan bahwa RFC merupakan model yang paling optimal untuk permasalahan klasifikasi pada dataset yang digunakan dalam penelitian ini.

4. Kesimpulan

Dari hasil penelitian, algoritma Random Forest Classifier (RFC) menunjukkan performa terbaik dengan akurasi 99,99%, precision 99,98%, recall 100%, dan MCC 99,98%. SVM dan KNN juga menunjukkan performa tinggi dengan akurasi masing-masing 99,91% dan 99,98%. Namun, RFC memiliki keunggulan utama pada recall sempurna (100%) dan tingkat kesalahan (FPR) yang paling rendah (0,02%). Hal ini menjadikan RFC sebagai algoritma yang paling efektif dalam mendeteksi anomali jaringan DDoS.

Penelitian ini menunjukkan bahwa RFC adalah pilihan optimal untuk deteksi anomali pada jaringan, khususnya serangan DDoS, karena memberikan keseimbangan yang baik antara akurasi tinggi dan efisiensi deteksi.

Dataset yang digunakan hanya berasal dari satu sumber saja, membuat penelitian ini terdapat keterbatasan. Karena itu perlu dilakukan tindak lanjutan seperti melakukan penelitian dengan dataset dari berbagai sumber berbeda sehingga model algoritma tersebut dapat memprediksi banyak macam teknik-teknik lainnya.

Referensi

[1] Haeruddin"Analisa Dan Perancangan Keamanan Jaringan Lokal Menggunakan Security Onion Dan Mikrotik" Dec. 2020, doi: 10.37253/joint.v1i2.4309.

[2] Anna University and IEEE Aerospace and Electronic Systems Society, *2019 International Carnahan Conference on Security Technology (ICCST) : ICCST 2019 : IEEE 53rd International Carnahan Conference on Security Technology : October 01-03, 2019, Anna University, Chennai, India,* doi: 10.1109/CCST.2019.8888419.

[3] X. Ma *et al.*, "A Comprehensive Survey on Graph Anomaly Detection with Deep Learning," Jun. 2021, doi: 10.1109/TKDE.2021.3118815.

[4] A. Singh and B. B. Gupta, "Distributed Denial-of-Service (DDoS) Attacks and Defense Mechanisms in Various Web- Enabled Computing Platforms: Issues, Challenges, and Future Research Directions," *Int J Semant Web Inf Syst*, vol. 18, no. 1, 2022, doi: 10.4018/IJSWIS.297143.

[5] S. Sambangi and L. Gondi, "A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression," MDPI AG, Dec. 2020, p. 51. doi: 10.3390/proceedings2020063051.

[6] N. Mamuriyah, S. E. Prasetyo, and A. O. Sijabat, "Rancangan Sistem Keamanan Jaringan dari serangan DDoS Menggunakan Metode Pengujian Penetrasi," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 1, pp. 162–167, Jan. 2024, doi: 10.47233/jteksis.v6i1.1124.

[7] R. R. Brooks, L. Yu, I. Ozcelik, J. Oakley, and N. Tusing, "Distributed Denial of Service (DDoS): A History," *IEEE Annals of the History of Computing*, vol. 44, no. 2, pp. 44–54, 2022, doi: 10.1109/MAHC.2021.3072582.

[8] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep Learning for Anomaly Detection: A Review," Apr. 01, 2021, *Association for Computing Machinery*. doi: 10.1145/3439950.

[9] İ. Güven and F. Şimşir, "Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods," *Comput Ind Eng*, vol. 147, Sep. 2020, doi: 10.1016/j.cie.2020.106678.

[10] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning: Methods and Applications to Brain Disorders*, Elsevier, 2019, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[11] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 219–244. doi: 10.1007/978-981-16-2164-2_19.

[12] T. Noi Phan, V. Kuch, and L. W. Lehnert, "Land cover classification using google earth engine and random forest classifier-the role of image composition," *Remote Sens (Basel)*, vol. 12, no. 15, Aug. 2020, doi: 10.3390/RS12152411.

[13] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.

[14] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples)," Apr. 2020, doi: 10.1145/3459665.

[15] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 326–338, Feb. 2020, doi: 10.11591/eei.v9i1.1464.

[16] H. HAJIALIAN and C. TOMA, "Network Anomaly Detection by Means of Machine Learning: Random Forest Approach with Apache Spark," *Informatica Economica*, vol. 22, no. 4/2018, pp. 89–98, Dec. 2018, doi: 10.12948/issn14531305/22.4.2018.08.

[17] I. A. Khan, H. Birkhofer, D. Kunz, D. Lukas, and V. Ploshikhin, "A Random Forest Classifier for Anomaly Detection in Laser-Powder Bed Fusion Using Optical Monitoring," *Materials*, vol. 16, no. 19, Oct. 2023, doi: 10.3390/ma16196470.

[18] D. Saraswat, P. Bhattacharya, M. Zuhair, A. Verma, and A. Kumar, "AnSMart: A SVM-based anomaly detection scheme via system profiling in Smart Grids," in *Proceedings of 2021 2nd International Conference on Intelligent Engineering and Management, ICIEM 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 417–422. doi: 10.1109/ICIEM51511.2021.9445353.

[19] B. Wang, S. Ying, G. Cheng, R. Wang, Z. Yang, and B. Dong, "Log-Based Anomaly Detection with the Improved K-Nearest Neighbor," *International Journal of Software Engineering and Knowledge Engineering*, vol. 30, no. 2, pp. 239–262, Feb. 2020, doi: 10.1142/S0218194020500114.

[20] M. Akpinar, M. F. Adak, and G. Guvenc, "SVM-based anomaly detection in remote working: Intelligent software SmartRadar," *Appl Soft Comput*, vol. 109, Sep. 2021, doi: 10.1016/j.asoc.2021.107457.

[21] *Proceeding, 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC 2019) : 13th-14th December 2019, Malaysia*. IEEE, 2019, doi: 10.1109/ICSPC47137.2019.9068081.

[22] S. S. Aljameel *et al.*, "An Anomaly Detection Model for Oil and Gas Pipelines Using Machine Learning," *Computation*, vol. 10, no. 8, Aug. 2022, doi: 10.3390/computation10080138.

[23] *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*. IEEE, 2020, doi: 10.1109/ECTI-CON49241.2020.9158222.

[24] *2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) : conference proceedings : 5-7 March, 2020*. IEEE, 2020, doi: 10.1109/ICIMIA48430.2020.9074929.

[25] S. D. D. Anton, S. Sinha, and H. D. Schotten, "Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.10374

[26] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[27] G. N. Ahmad *et al.*, "Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features," *Applied Sciences (Switzerland)*, vol. 12, no. 15, Aug. 2022, doi: 10.3390/app12157449.

[28] B. Dash, M. F. Ansari, P. Sharma, and A. Ali, "Threats and Opportunities with AI-based Cyber Security Intrusion Detection: A Review," *International Journal of Software Engineering & Applications*, vol. 13, no. 5, pp. 13–21, Sep. 2022, doi: 10.5121/ijsea.2022.13502.

[29] A. A. Salih and A. M. Abdulazeez, "Evaluation of Classification Algorithms for Intrusion Detection System: A Review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 31–40, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.004.

[30] *2021 International Conference on Artificial Intelligence and Big Data Analytics : 27-29 Oct. 2021*. IEEE, 2021.
[31] S. Desmalia, A. Mutoi Siregar, K. A. Baihaqi, and T. Rohana, "Comparison Model Optimal Machine Learning Model With Feature Extraction for Heart Attack Disease Classification," *Scientific Journal of Informatics*, vol. 11, no. 2, 2024, doi: 10.15294/sji.v11i2.4561.