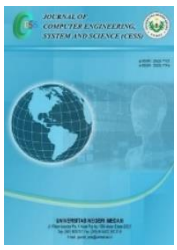
CESS (Journal of Computer Engineering, System and Science)
Vol. 10, No. 1, Januari 2025, pp.335-346

CESS
(Journal of Computer Engineering, System and Science)
Available online: https://jurnal.unimed.ac.id/2012/index.php/cess
ISSN: 2502-714x (Print) | ISSN: 2502-7131 (Online)



Model Data Mining Klasifikasi Serangan Siber untuk Deteksi Dini Serangan Menggunakan Algoritma Random Forest

Cyber Attack Classification Data Mining Model for Early Detection of Attacks Using the Random Forest Algorithm

Zunaida Sitorus[1], Adittya Pratama[2], Oky Adinata Hidayatullah[3], Adi Widarma[4]*

[1,2,3]Program Studi Teknik Informatika, Universitas Asahan

Jl. Jend. Ahmad Yani, Kec. Kota Kisaran Timur, Sumatera Utara 21216, Indonesia
[4]Program Studi Pendidikan Teknologi Informatika dan Komputer (PTIK), Universitas Negeri Medan
Jl. William Iskandar Ps. V, Kec. Percut Sei Tuan, Kab. Deli Serdang, Sumatera Utara 20221, Indonesia
Email: [1]z_sitorus@yahoo.com, [2]adittyaapratama29@gmail.com, [3]rimbiaulia16@gmail.com, [4]adiwidarma10@gmail.com

**Corresponding Author*

A B S T R A K

Serangan siber menjadi ancaman yang serius pada era digital saat ini. Deteksi dini serangan sangat penting untuk meminimalkan dampak dari ancaman siber. Dengan identifikasi yang cepat dan akurat, organisasi dapat mengambil langkah-langkah mitigasi yang diperlukan sebelum kerusakan lebih lanjut terjadi. Salah satu pendekatan yang menjanjikan dalam mendeteksi dan mengklasifikasikan serangan siber adalah penggunaan algoritma data mining. Penelitian ini bertujuan untuk mengembangkan model klasifikasi ancaman siber yang lebih komprehensif dengan menggunakan algoritma Random Forest. Random Forest adalah algoritma ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi. Data yang digunakan dalam penelitian ini diambil dari sumber-sumber terbuka yang menyediakan data serangan siber yang umum digunakan seperti KDD Cup 1999 yaitu *Cybersecurity Intrusion Detection Dataset*. Kumpulan dataset ini dirancang untuk mendeteksi intrusi siber berdasarkan lalu lintas jaringan dan perilaku pengguna yang terdiri dari 9537 baris/record dan 11 atribut. Tahapan yang dilakukan dalam penelitian ini yaitu pengumpulan data, *preprocessing* data, pemodelan, pelatihan dan pengujian model serta evaluasi model. Model dilatih dan diuji sebanyak 3 perbandingan antara data latih dan data uji yaitu 90:10, 80:20 dan 70:30. Dari 3 perbandingan tersebut, perbandingan 90:10 memiliki hasil evaluasi model dengan *Confusion Matrix* yang paling tinggi yaitu akurasi sebesar 88,16%, *precision* sebesar 100%, *recall* sebesar 72,77% dan *F-Measure* sebesar 84,24%. Dengan demikian, Random Forest merupakan pilihan yang sangat baik untuk digunakan dalam sistem deteksi intrusi yang memerlukan deteksi dini dan akurasi tinggi. Sehingga diharapkan penelitian ini dapat memberikan kontribusi yang signifikan dalam pengembangan sistem deteksi intrusi yang lebih efektif dan efisien.

Kata Kunci: *Data mining; klasifikasi; serangan siber, random forest; deteksi dini.*

A B S T R A C T

Cyber attacks are a serious threat in today's digital era. Early detection of attacks is critical to minimizing the impact of cyber threats. With fast and accurate identification, organizations can take the necessary mitigation steps before further damage occurs. One promising approach in detecting and classifying cyber attacks is the use of data mining algorithms. This research aims to develop a more comprehensive cyber threat classification model using the Random Forest algorithm. Random Forest is an ensemble algorithm that combines multiple decision trees to improve prediction accuracy. The data used in this research was taken from open sources that provide commonly used cyber attack data such as the 1999 KDD Cup, namely the Cybersecurity Intrusion Detection Dataset. This dataset is designed to detect cyber intrusions based on network traffic and user behavior which consists of 9537 rows/records and 11 attributes. The stages carried out in this research are data collection, data preprocessing, modeling, model training and testing and model evaluation. The model was trained and tested with 3 comparisons between training data and test data, namely 90:10, 80:20 and 70:30. Of the 3 comparisons, the 90:10 comparison has the highest model evaluation results with the Confusion Matrix, namely accuracy of 88.16%, precision of 100%, recall of 72.77% and F-Measure of 84.24%. Thus, Random Forest is an excellent choice for use in intrusion detection systems that require early detection and high accuracy. So it is hoped that this research can make a significant contribution to the development of a more effective and efficient intrusion detection system.

Keywords: *Data mining; classification; cyber attacks, random forests; early detection.*

1. PENDAHULUAN

Dalam era digital saat ini, ancaman siber menjadi salah satu tantangan terbesar bagi organisasi dan individu. Menurut laporan *Cybersecurity Ventures*, kerugian global akibat serangan siber diperkirakan mencapai $6 triliun pada tahun 2021 dan diperkirakan akan meningkat menjadi $10,5 triliun pada tahun 2025[1]. Serangan ini tidak hanya mengancam keamanan data, tetapi juga dapat merusak reputasi perusahaan dan mengganggu operasional bisnis. Oleh karena itu, deteksi dini terhadap ancaman siber sangat penting untuk meminimalkan dampak yang ditimbulka dan juga hal ini menunjukkan urgensi untuk mengembangkan sistem deteksi yang lebih efektif dan efisien.

Ancaman siber tidak hanya bervariasi dalam bentuk, tetapi juga dalam teknik dan strategi yang digunakan oleh pelaku kejahatan. Misalnya, serangan *phishing, malware*, dan *ransomware* merupakan beberapa jenis ancaman yang sering terjadi. Data dari Internet Crime Complaint Center (IC3) menunjukkan bahwa pada tahun 2020, terdapat lebih dari 791.000 laporan kejahatan siber di AS, dengan kerugian yang mencapai lebih dari $4,2 miliar[2]. Dengan meningkatnya kompleksitas dan jumlah serangan, pendekatan tradisional dalam deteksi ancaman siber menjadi kurang efektif, sehingga diperlukan model yang lebih canggih.

Salah satu pendekatan yang menjanjikan dalam mendeteksi dan mengklasifikasikan ancaman siber adalah penggunaan algoritma data mining. Model data mining, khususnya klasifikasi menggunakan algoritma Random Forest, telah terbukti efektif dalam

mengidentifikasi pola dan karakteristik dari data yang besar. Random Forest adalah algoritma ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi[3]. Dalam konteks deteksi dini serangan jaringan, Random Forest dapat digunakan untuk menganalisis data lalu lintas jaringan dan mengklasifikasikan jenis ancaman yang mungkin terjadi. Deteksi dini serangan jaringan sangat penting untuk meminimalkan dampak dari ancaman siber. Dengan identifikasi yang cepat dan akurat, organisasi dapat mengambil langkah-langkah mitigasi yang diperlukan sebelum kerusakan lebih lanjut terjadi.

Penelitian mengenai deteksi ancaman siber menggunakan teknik data mining telah berkembang pesat dalam beberapa tahun terakhir. Berbagai algoritma, termasuk *Support Vector Machi*ne (SVM), *K-Nearest Neighbors* (KNN), dan *Neural Networks* telah digunakan untuk mengklasifikasikan serangan siber. Namun, Random Forest muncul sebagai salah satu metode yang paling banyak digunakan karena kemampuannya dalam menangani data besar dan kompleksitas yang tinggi[4].

Sebuah studi oleh Prabhu(2022) menunjukkan bahwa Random Forest dapat mencapai akurasi hingga 95% dalam mengidentifikasi serangan DDoS (*Distributed Denial of Service*) ketika diterapkan pada dataset yang besar[5]. Penelitian lain oleh Hidayat(2020) juga menemukan bahwa Random Forest lebih unggul dibandingkan dengan algoritma lain dalam hal kecepatan dan akurasi deteksi intrusi[6]. Ini menunjukkan bahwa Random Forest memiliki potensi yang besar dalam meningkatkan sistem deteksi intrusi (IDS) yang ada saat ini.

Meskipun terdapat banyak penelitian yang menggunakan algoritma Random Forest untuk klasifikasi ancaman siber, masih ada beberapa kesenjangan yang perlu diatasi. Pertama, banyak studi yang fokus pada satu jenis serangan tertentu, seperti malware atau DDoS, tanpa mempertimbangkan berbagai jenis ancaman yang mungkin terjadi secara bersamaan. Hal ini dapat menyebabkan model yang dikembangkan menjadi kurang fleksibel dan tidak mampu mendeteksi serangan yang lebih kompleks[7].

Kedua, sebagian besar penelitian masih menggunakan dataset yang tidak mencerminkan kondisi dunia nyata, sehingga hasil yang diperoleh mungkin tidak dapat diterapkan secara langsung dalam praktik. Dalam konteks ini, penting untuk mengembangkan dataset yang lebih representatif dan beragam agar model yang dihasilkan lebih akurat dan dapat diandalkan. Penelitian oleh Widyastuti(2021) menunjukkan bahwa penggunaan dataset yang lebih bervariasi dapat meningkatkan kinerja model dalam mendeteksi berbagai jenis serangan[8].

Ketiga, banyak penelitian yang belum mempertimbangkan faktor waktu dalam analisis data. Serangan siber sering kali terjadi dengan pola yang berubah-ubah, sehingga model yang tidak mempertimbangkan dinamika waktu dapat kehilangan kemampuan untuk mendeteksi serangan baru. Oleh karena itu, pendekatan yang lebih adaptif dan responsif terhadap perubahan pola serangan sangat diperlukan.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi deteksi intrusi keamanan siber yang lebih komprehensif dengan menggunakan algoritma Random Forest. Dengan demikian, model yang dihasilkan diharapkan dapat memberikan deteksi dini yang lebih akurat dan efektif. Selain itu, penelitian ini juga akan menggunakan dataset yang lebih representatif dan mencerminkan kondisi dunia nyata.

Kontribusi penelitian ini diharapkan dapat memberikan dampak positif bagi pengembangan teknologi keamanan siber khususnya dalam hal deteksi intrusi keamanan siber. Dengan pemahaman yang lebih baik tentang ancaman ini, organisasi dapat mengambil langkah-langkah pencegahan yang lebih efektif[9]. Dengan pengembangan model klasifikasi

menggunakan algoritma Random Forest diharapkan juga dapat meningkatkan akurasi dan kecepatan deteksi serangan. Dengan demikian, organisasi dapat merespons serangan dengan lebih cepat dan efisien, meminimalkan kerugian yang mungkin ditimbulkan.

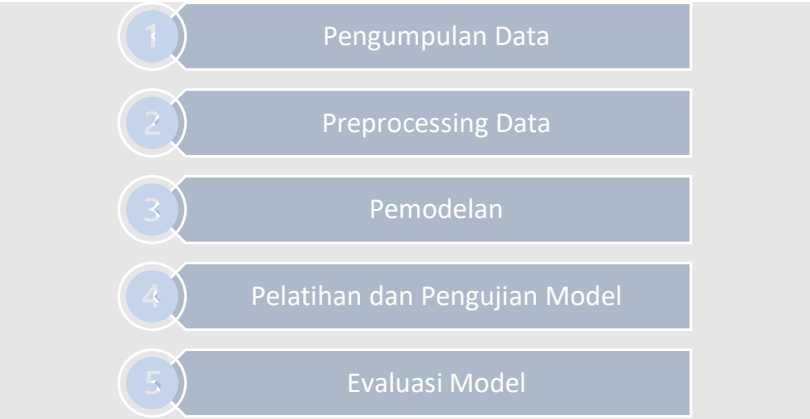
2. METODE PENELITIAN

2.1. Pendekatan Penelitian

Dalam penelitian ini, pendekatan yang digunakan adalah pendekatan kuantitatif dengan metode eksperimen. Pendekatan kuantitatif memungkinkan peneliti untuk mengukur dan menganalisis data numerik secara statistik, sehingga dapat memberikan gambaran yang jelas tentang efektivitas model yang dikembangkan. Penelitian ini bertujuan untuk mengembangkan model klasifikasi menggunakan algoritma Random Forest dalam mendeteksi intrusi keamanan siber serta untuk menganalisis akurasi dan performa model tersebut.

2.2. Tahapan Penelitian

Metode penelitian ini terdiri dari beberapa tahapan yang sistematis, dimulai dari pengumpulan data hingga evaluasi model. Adapun tahapan yang dilakukan dalam penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Tahapan Penelitian

2.2.1. Pengumpulan Data

Sebagai langkah awal, peneliti melakukan pengumpulan data yang relevan. Pengumpulan data adalah langkah penting dalam penelitian ini. Data yang digunakan dalam penelitian ini diambil dari sumber-sumber terbuka yang menyediakan data serangan siber yang umum digunakan seperti KDD Cup 1999 yaitu *Cybersecurity Intrusion Detection Dataset*. Kumpulan dataset ini dirancang untuk mendeteksi intrusi siber berdasarkan lalu lintas jaringan dan perilaku pengguna yang terdiri dari 9537 baris/*record* dan 11 atribut. Dengan menggunakan dataset ini, peneliti dapat melatih model untuk mendeteksi intrusi keamanan siber.

2.2.2. *Preprocessing* Data

Setelah pengumpulan data, tahap selanjutnya adalah *preprocessing* data. Proses ini melibatkan pembersihan data, transformasi data, penghapusan nilai yang hilang, dan normalisasi fitur untuk memastikan bahwa model dapat berfungsi dengan baik. Menurut Liu(2020), kualitas data yang baik dapat meningkatkan akurasi model hingga 30%[10]. Dalam penelitian ini, teknik pengurangan dimensi seperti PCA (*Principal Component Analysis*) juga

diterapkan untuk mengurangi kompleksitas data yang dapat membantu meningkatkan kecepatan dan akurasi model[11].

2.2.3. Pemodelan

Dalam penelitian ini, algoritma yang dipilih adalah Random Forest, yang merupakan salah satu algoritma *machine learning* yang paling populer untuk klasifikasi. Random Forest bekerja dengan membangun sejumlah pohon keputusan dan menggabungkan hasilnya untuk meningkatkan akurasi prediksi. Keunggulan lain dari Random Forest adalah kemampuannya untuk menangani data yang tidak seimbang, yang sering terjadi dalam konteks serangan siber.

Berikut adalah beberapa rumus dan konsep kunci dalam algoritma Random Forest:

1. *Bootstrap Sampling*
   Random Forest dimulai dengan mengambil sampel acak dari dataset asli menggunakan teknik bootstrap. Setiap subset yang diambil digunakan untuk melatih satu pohon keputusan.
2. Pembentukan *Decision Trees*
   Untuk setiap subset, sebuah decision tree dibentuk dengan memilih fitur secara acak di setiap node. Hal ini memastikan bahwa setiap pohon memiliki variasi, yang mengurangi kemungkinan overfitting.
3. *Voting/Averaging*
   Setelah semua pohon dibentuk, prediksi akhir dihasilkan dengan cara:
   - Klasifikasi: Menggunakan voting mayoritas dari hasil prediksi setiap pohon.
   - Regresi: Menghitung rata-rata hasil prediksi dari semua pohon.
4. Rumus Prediksi
   Untuk klasifikasi, rumus yang digunakan adalah:

$$\text{I(y)}=\text{argmax}_c \left(\sum_{n=0}^{n} Ihn(y) = c \right) \quad (1)$$

   Di mana:
   - I(y) adalah prediksi untuk input *y*.
   - I adalah fungsi indikator yang bernilai 1 jika kondisi terpenuhi dan 0 jika tidak.
   - h_n adalah pohon ke-n dari Random Forest.
   - N adalah jumlah total pohon dalam hutan.
5. Estimasi Kesalahan *Out-of-Bag* (OOB)
   Random Forest juga menggunakan estimasi kesalahan *out-of-bag*, yang dihitung sebagai rata-rata dari kesalahan prediksi untuk setiap kasus pelatihan yang tidak termasuk dalam sampel bootstrap untuk pohon tersebut.
6. Pemilihan Fitur
   Dalam proses pembentukan pohon, hanya sebagian kecil dari fitur yang dipilih secara acak (misalnya, $\sqrt{m}$, di mana *m* adalah jumlah total fitur). Ini membantu meningkatkan keragaman antar pohon.

Selanjutnya, model Random Forest dibangun menggunakan RapidMiner. Parameter model, seperti jumlah pohon dan kedalaman maksimum, dioptimalkan menggunakan teknik *grid search* untuk mendapatkan kombinasi terbaik. Penelitian oleh Zhang et al. (2022) menunjukkan bahwa pengaturan parameter yang tepat dapat meningkatkan kinerja model secara signifikan[12].

2.2.4. Pelatihan dan Pengujian Model

Setelah model Random Forest dibangun, langkah selanjutnya adalah melakukan pengujian. Pengujian dilakukan dengan membagi dataset menjadi dua bagian yaitu data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk mengukur kinerja model. Data dibagi menjadi dua bagian: data pelatihan dan data pengujian dengan rasio 70:30, 80:20 dan 90:10. Teknik validasi silang (*cross-validation*) juga diterapkan untuk memastikan bahwa model tidak overfit pada data pelatihan. Penggunaan *k-fold cross-validation* juga diterapkan untuk memastikan bahwa model tidak mengalami overfitting, yang sering menjadi masalah dalam model pembelajaran mesin[13].

2.2.5. Evaluasi Model

Setelah model dilatih dan diuji, selanjutnya evaluasi model dilakukan menggunakan metrik seperti akurasi, *precision, recall*, dan *F-Measure*. Hasil evaluasi ini akan dianalisis untuk menentukan apakah model Random Forest efektif dalam mendeteksi intrusi keamanan siber. Selain itu, analisis pentingnya fitur juga dilakukan untuk mengidentifikasi faktor-faktor yang paling berkontribusi terhadap klasifikasi yang tepat.

Hasil akurasi menunjukkan tingkat ketepatan model dalam mengklasifikasikan data dengan benar. *Confusion Matrix* digunakan untuk menguji akurasi model yang sudah dibangun dengan memanfaatkan rumus:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \qquad (2)$$

Keterangan:
TN (*True Negative*) = Jumlah prediksi negatif yang benar
FN (*False Negative*) = Jumlah prediksi positif yang salah diprediksi sebagai negatif
TP (*True Positive*) = Jumlah prediksi positif yang benar
FP (*False Positive*) = Jumlah prediksi negatif yang salah diprediksi sebagai positif

Precision digunakan untuk mengukur tingkat akurasi model dalam mengenali instance positif tanpa menghasilkan kesalahan tambahan (*false positive*). Sementara itu, *recall* berfungsi untuk menilai seberapa efektif model dalam mendeteksi semua instance positif yang relevan, meskipun hal ini dapat mengakibatkan peningkatan jumlah *false negatives*. *F-Measure* digunakan untuk mengevaluasi kinerja keseluruhan model dengan memberikan bobot yang setara pada presisi dan recall, sehingga menciptakan skala yang lebih universal untuk menilai model klasifikasi.

Dengan pendekatan yang sistematis dan berbasis data ini, diharapkan penelitian dapat memberikan wawasan yang berharga mengenai efektivitas algoritma Random Forest dalam konteks keamanan siber.

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini yaitu data sekunder tentang dataset deteksi intrusi keamanan siber yang berasal dari situs Kaggle.com dengan alamat yaitu https://www.kaggle.com/datasets/dnkumars/cybersecurity-intrusion-detection-dataset.
Dataset ini digunakan untuk mendeteksi intrusi siber berdasarkan lalu lintas jaringan dan

perilaku pengguna yang terdiri dari 9537 *record* /baris dan 11 atribut. Adapun rincian dan penjelasan dari setiap atribut dapat dilihat pada tabel 1.

Tabel 1. Atribut dataset penelitian

No.	Atribut	Keterangan
1	session_id	Nomor unik sesi ID
2	network_packet_size	Ukuran paket jaringan berkisar antara 64 hingga 1500 byte
3	protocol_type	Protokol yang digunakan dalam sesi: TCP, UDP, atau ICMP
4	login_attempts	Banyaknya Login dimana nilai yang tinggi mungkin mengindikasikan serangan brute force (upaya login berulang kali).
5	session_duration	Durasi Sesi dalam Detik dimana Sesi yang sangat panjang mungkin menunjukkan akses tidak sah atau kegigihan penyerang.
6	encryption_used	Protokol enkripsi yang digunakan: AES, DES, None
7	ip_reputation_score	Kepercayaan Alamat IP dimana skor dari 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan aktivitas mencurigakan.
8	failed_logins	Upaya Masuk Gagal dimana jumlah login yang gagal tinggi menunjukkan isian kredensial atau serangan kamus.
9	browser_type	Browser yang digunakan: Chrome, Firefox, Edge, Safari.
10	unusual_time_access	Anomali Waktu Masuk yang menunjukkan apakah akses terjadi pada waktu yang tidak biasa.
11	attack_detected	Serangan terdeteksi dimana klasifikasi biner: 1 berarti serangan terdeteksi, 0 berarti aktivitas normal.

3.2. *Preprocessing Data*

Pra-pemrosesan data merupakan tahap pengolahan di mana data awal dievaluasi untuk menentukan apakah semua atribut akan digunakan dalam penelitian. Berdasarkan dataset yang terdapat pada tabel 1, atribut session_ID tidak akan dimanfaatkan karena tidak memberikan kontribusi terhadap penelitian. Selain itu, proses transformasi data juga dilakukan terhadap atribut *attack_detected*, di mana kode 1 diubah menjadi "*Attack was Detected*" dan kode 0 menjadi "*Normal Activity*". Hasil dari proses transformasi data ini dapat dilihat pada tabel 2.

Tabel 2. Dataset hasil *data preparation*

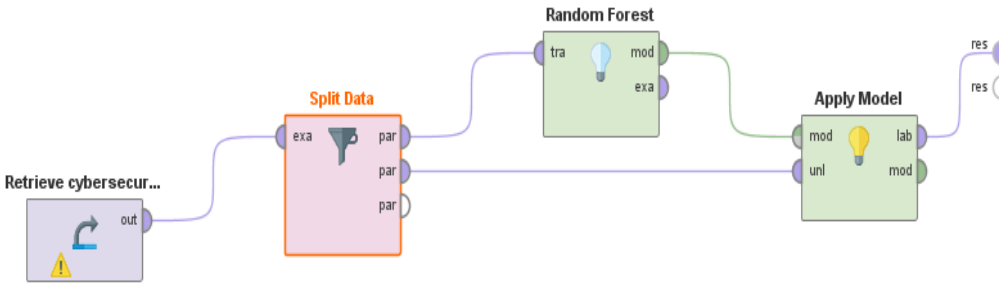
No.	network_packet_size	protocol_type	login_attempts	session_duration	encryption_used	ip_reputation_score	failed_logins	browser_type	unusual_time_access	attack_detected
1	599	TCP	4	492.9832634426563	DES	0.606818	1	Edge	0	Attack was Detected
2	472	TCP	3	1557.9964611204384	DES	0.30156896759608937	0	Firefox	0	Normal Activity

3	629	TCP	3	75.044261664207741	DES	0.7391643279163831	2	Chrome	0	Attack was Detected
4	804	UDP	4	601.2488351708328	DES	0.12326717575248465	0	Unknown	0	Attack was Detected
5	453	TCP	5	532.5408884201419	AES	0.054874	1	Firefox	0	Normal Activity
..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..
9535	664	TCP	5	35.17025	AES	0.3592001659927287 5	1	Firefox	0	Normal Activity
9536	406	TCP	4	86.664702538688668	AES	0.5374167250372589	1	Chrome	1	Normal Activity
9537	340	TCP	6	86.876744036561241	None	0.2770686858095288 7	4	Chrome	1	Attack was Detected

Persiapan data juga dilakukan untuk memisahkan dataset menjadi data pelatihan (*training data*) dan data pengujian (*testing data*). Dalam studi ini, pembagian akan dilakukan dengan rasio 90:10, 80:20, dan 70:30. Proses pemisahan dataset dilakukan menggunakan aplikasi RapidMiner, di mana data dipisahkan secara otomatis melalui operator Split Data.

3.3. *Pemodelan*

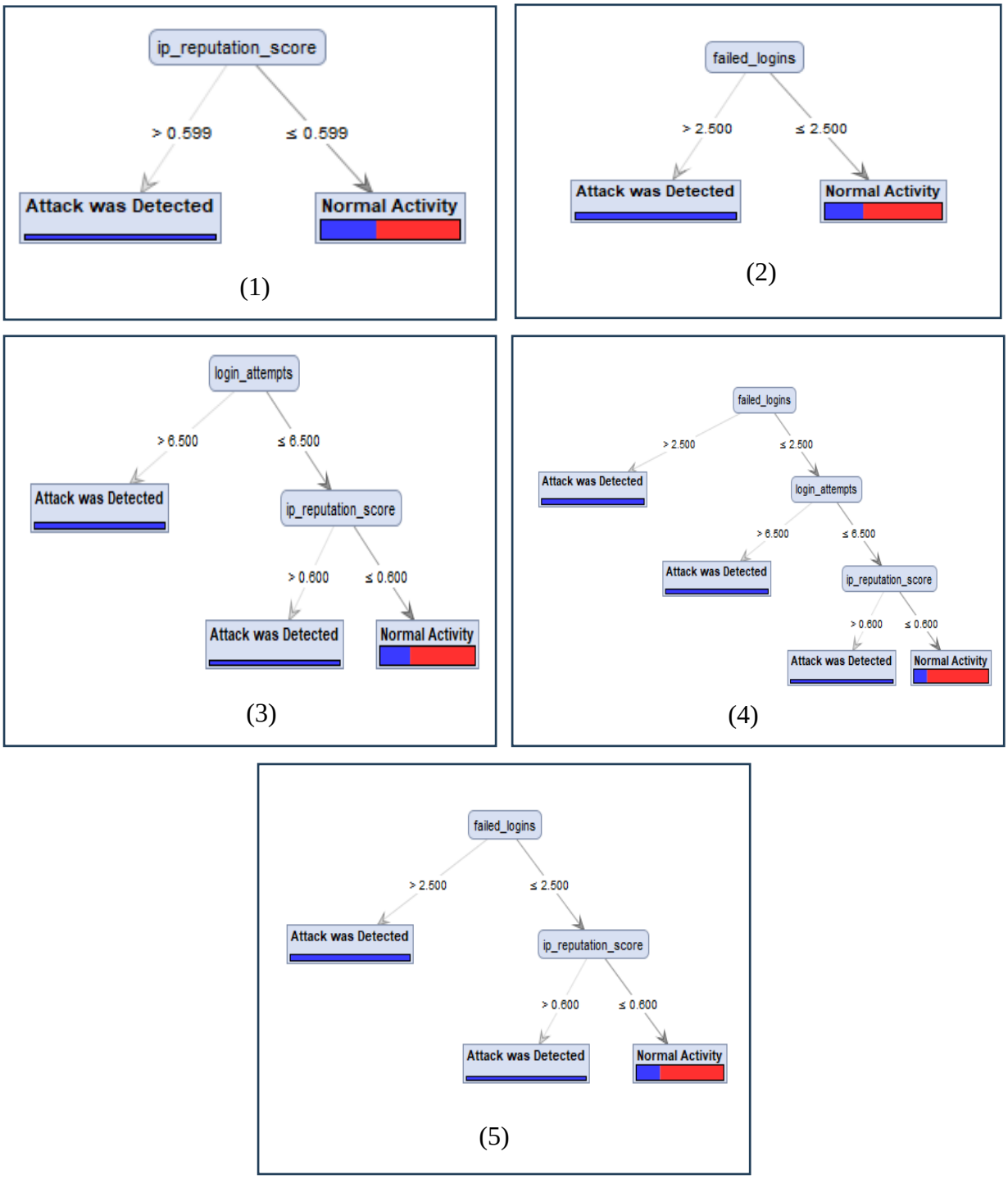
Dari 10 atribut yang ada dalam dataset yang digunakan dalam penelitian ini, dapat disimpulkan bahwa dataset tersebut tergolong dalam kategori pembelajaran terawasi (*Supervised Learning*). Dalam pendekatan ini, algoritma melakukan proses pembelajaran dengan mempertimbangkan nilai dari variabel target yang berhubungan dengan nilai variabel prediktor. Pembelajaran terawasi adalah teknik yang memanfaatkan metode klasifikasi, di mana data dikelompokkan berdasarkan model keterikatan antara data yang satu dengan yang lainnya[14]. Yang menjadi target/ label pada dataset yang digunakan yaitu atribut *attack_detected*. Algoritma untuk klasifikasi deteksi serangan siber dalam penelitian ini digunakan algoritma *Random Forest* dan setelah itu dilakukan pemodelan dengan menggunakan aplikasi RapidMiner. Adapun model yang dibangun dapat dilihat pada Gambar 2 berikut.



Gambar 2. Model klasifikasi deteksi serangan keamanan siber

Algoritma Random Forest mengambil sampel acak dari dataset asli menggunakan teknik bootstrap. Dari model yang dibangun dihasilkan pengetahuan berupa pola dalam bentuk pohon keputusan dengan memilih jumlah pohon (*number of tree*) adalah 5. Setiap subset

dibentuk pohon keputusan dengan memilih fitur secara acak di setiap node kemudian setiap subset yang diambil akan digunakan untuk melatih satu pohon keputusan. Adapun hasil 5 pohon keputusan yang dipilih dapat dilihat pada Gambar 3.

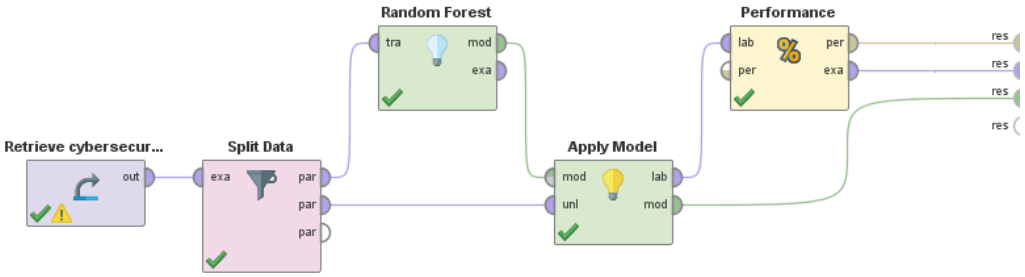


Gambar 3. Hasil Tree yang dibentuk dengan fitur secara acak

3.4. Evaluasi Model

Hasil dari pemodelan yang telah dilakukan akan dievaluasi melalui pengujian *performance* model. Pengujian *performance* bertujuan untuk menilai tingkat akurasi model yang telah

dibangun dengan menggunakan aplikasi RapidMiner. Struktur dari pengujian *performance* ini dapat dilihat pada Gambar 4.



Gambar 4. Model struktur pengujian performance

Pengujian kinerja dilakukan dengan cara membagi dataset menjadi dua bagian, yaitu data pelatihan dan data pengujian, dengan tiga perbandingan yang berbeda. Salah satu perbandingan yang digunakan adalah 90:10, di mana 90% dari data digunakan untuk pelatihan dan 10% sisanya untuk pengujian. Hasil akurasi untuk perbandingan 90:10 dapat dilihat pada Gambar 5.

accuracy: 88.16%

	true Attack was Detected	true Normal Activity	class precision
pred. Attack was Detected	302	0	100.00%
pred. Normal Activity	113	539	82.67%
class recall	72.77%	100.00%	

Gambar 5. Hasil akurasi perbandingan 90:10

Pembagian dataset dilakukan dengan rasio 80:20, di mana 80% digunakan untuk pelatihan data dan 20% untuk data pengujian. Hasil dari akurasi yang diperoleh berdasarkan pembagian ini dapat dilihat pada Gambar 6.

accuracy: 82.17%

	true Attack was Detected	true Normal Activity	class precision
pred. Attack was Detected	510	0	100.00%
pred. Normal Activity	340	1057	75.66%
class recall	60.00%	100.00%	

Gambar 6. Hasil akurasi perbandingan 80:20

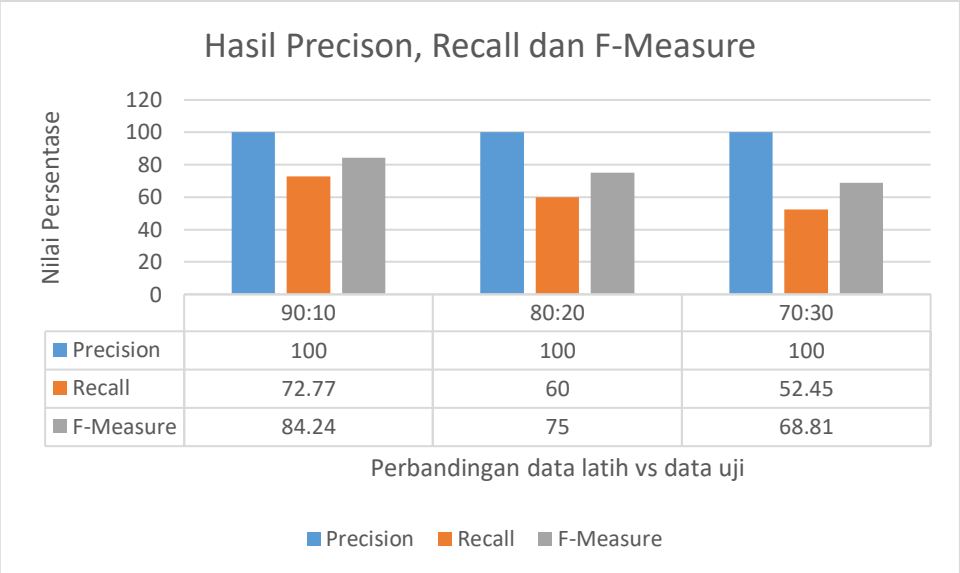
Pembagian dataset dilakukan dengan proporsi 70:30, di mana 70% dialokasikan untuk data pelatihan dan 30% untuk data pengujian. Hasil akurasi yang diperoleh dengan proporsi tersebut dapat dilihat pada Gambar 7.

accuracy: 78.33%

	true Attack was Detected	true Normal Activity	class precision
pred. Attack was Detected	684	0	100.00%
pred. Normal Activity	620	1557	71.52%
class recall	52.45%	100.00%	

Gambar 7. Hasil akurasi perbandingan 70:30

Hasil dari akurasi yang diperoleh dari tiga perbandingan antara data latih dan data uji menunjukkan bahwa perbandingan 90:10 memberikan nilai akurasi tertinggi, yaitu sebesar 88,16%. Selain itu, evaluasi kinerja model klasifikasi juga dilakukan dengan menghitung metrik *Precision, Recall*, dan *F-Measure*. Hasil dari perhitungan metrik evaluasi ini dapat dilihat pada Gambar 8.



Gambar 8. Hasil *Precision, Recall* dan *F-Measure*

Dari hasil perhitungan didapatkan nilai Precision untuk semua perbandingan memiliki nilai yang sama yaitu 100%. Nilai Recall untuk perbandingan 90:10 yaitu 72,77%, perbandingan 80:20 yaitu 60% dan perbandingan 70:30 yaitu 52,45%. Nilai F-Measure untuk perbandingan 90:10 yaitu 84,24%, perbandingan 80:20 yaitu 72% dan perbandingan 70:30 yaitu 68,81%.

4. KESIMPULAN

Penelitian ini telah berhasil mengembangkan model data mining klasifikasi menggunakan algoritma Random Forest untuk mendeteksi ancaman siber. Keunggulan dari model ini terletak pada kemampuannya untuk menangani data yang tidak seimbang dan mengurangi angka *false positives*. Hal ini sangat penting dalam konteks keamanan siber, di mana kesalahan dalam deteksi dapat berakibat fatal bagi organisasi. Dengan penerapan teknik *preprocessing* yang tepat, model ini dapat memberikan hasil yang lebih baik dibandingkan dengan metode lain. Setelah dilakukan tahapan evaluasi model dengan menggunakan *Confusion Matrix* didapatkan nilai akurasi, *Precision, Recall* dan *F-Measure* yang paling besar yaitu pada perbandingan data latih (90) dan data uji (10) dengan akurasi sebesar 88,16%, nilai *Precision* sebesar 100%, nilai *Recall* sebesar 72,77% dan nilai *F-Measure* sebesar 84,24%. Hasil

yang diperoleh menunjukkan bahwa model ini memiliki akurasi tinggi, serta nilai presisi dan recall yang memuaskan. Dengan demikian, Random Forest merupakan pilihan yang sangat baik untuk digunakan dalam sistem deteksi intrusi yang memerlukan deteksi dini dan akurasi tinggi. Namun, penelitian ini juga menyadari bahwa meskipun Random Forest menunjukkan performa yang baik, tantangan dalam dunia siber terus berkembang. Oleh karena itu, penelitian lebih lanjut diperlukan untuk mengeksplorasi teknik-teknik baru dan meningkatkan kemampuan model dalam menghadapi ancaman yang semakin kompleks. Integrasi dengan teknologi terbaru, seperti pembelajaran mendalam dan analisis perilaku, dapat menjadi arah penelitian selanjutnya. Dengan demikian, diharapkan penelitian ini dapat memberikan kontribusi yang signifikan dalam pengembangan sistem deteksi intrusi keamanan siber yang lebih efektif dan efisien. Penelitian lebih lanjut di bidang ini akan sangat penting untuk menjaga keamanan informasi dan melindungi aset digital di era yang semakin terhubung.

DAFTAR PUSTAKA

[1] C. Ventures, “2021 Cybercrime Report,” *https://cybersecurityventures.com.*, 2021.
[2] I. C. C. C. (IC3), “2020 Internet Crime Complaint Center Report,” *Diakses dari [IC3](https://www.ic3.gov/)*, 2020.
[3] N. Alexander, R. B. Widodo, and W. Swastika, “Penggunaan Machine Learning Dalam Klasifikasi Bahasa Isyarat BISINDO Menggunakan Kamera,” in *Prosiding Seminar Nasional Informatika & Sistem Informasi*, 2023, pp. 11–26.
[4] L. Zhang, “Random Forest for Cyber Threat Detection: A Review,” *J. Int. Keamanan Siber*, vol. 9, no. 4, pp. 101–115, 2021.
[5] R. Prabhu, “Detection of DDoS Attacks Using Random Forest Algorithm,” *J. Teknol. Inf. dan Komun.*, vol. 5, no. 2, pp. 123–130, 2022.
[6] Hidayat, “Comparison of Machine Learning Algorithms for Intrusion Detection Systems,” *J. Keamanan Siber*, vol. 4, no. 1, pp. 45–60, 2020.
[7] D. P. Sari, “Challenges in Cyber Threat Detection: A Review,” *J. Ilmu Komput.*, vol. 8, no. 1, pp. 15–28, 2023.
[8] S. Widyastuti, “Enhancing Cybersecurity with Diverse Datasets for Machine Learning Models,” *J. Keamanan Jar.*, vol. 6, no. 3, pp. 78–89, 2021.
[9] T. G. Laksana and S. Mulyani, “Pengetahuan Dasar Identifikasi Dini Deteksi Serangan Kejahatan Siber Untuk Mencegah Pembobolan Data Perusahaan,” *JUKIM J. Ilm. Multidisiplin*, vol. 3, no. 1, pp. 109–122, 2024.
[10] H. Liu, H., Motoda and B. Krawczyk, *Data Mining and Knowledge Discovery Handbook*. Springer, 2020.
[11] X. Liu, Y., Wang, Y., & Zhang, “A novel approach for cyber attack detection based on PCA and SVM,” in *Journal of Information Security and Applications*, 2020. doi: DOI: 10.1016/j.jisa.2020.102603.
[12] L. Zhang, Y., Li, Y., & Wang, “An improved random forest algorithm for network intrusion detection,” in *Computers & Security*, 2022. doi: DOI: 10.1016/j.cose.2022.102696.
[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2019.
[14] A. Surahmat, N. Ahmad, H. Fitri, and A. Widarma, *Kecerdasan buatan dalam Data Mining*. Bandung: Widina Bhakti Persada, 2023.