TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

Analisis SelectKBest pada Klasifikasi Trafik VPN Menggunakan Random Forest dan SVM

Andri Nurdiyansah[*], Dwi Robiul R, Sururi, Nana Sujana

Fakultas Teknologi Komputer, Program Studi Teknologi Komputer, Politeknik Pajajaran, Bandung, Indonesia
Email: [1,*]andrinnurdiansyah@email.com, [2]dwirobiul@gmail.com, [3]sururi@poljan.ac.id, [4]nana.sujana@poljan.ac.id
Email Penulis Korespondensi: andrinnurdiansyah@gmail.com

Abstrak–Pertumbuhan penggunaan Virtual Private Network (VPN) pada jaringan modern menimbulkan tantangan dalam pemantauan dan pengelolaan trafik jaringan, khususnya dalam membedakan trafik VPN dan Non-VPN secara akurat dan efisien. Penelitian ini bertujuan untuk menganalisis efektivitas metode seleksi fitur SelectKBest dalam meningkatkan kinerja klasifikasi trafik VPN menggunakan algoritma Random Forest dan Support Vector Machine (SVM). Dataset yang digunakan adalah CIC VPN-NonVPN Traffic Dataset dari Canadian Institute for Cybersecurity (CIC), yang merupakan dataset standar pada penelitian keamanan jaringan. Proses seleksi fitur dilakukan menggunakan SelectKBest dengan fungsi ANOVA (f_classif), yang mereduksi jumlah fitur menjadi 15 fitur paling relevan. Hasil eksperimen menunjukkan bahwa Random Forest mampu mencapai akurasi pengujian sebesar 84,94% dengan nilai F1-score dan ROC-AUC yang tinggi, serta rata-rata akurasi cross-validation sebesar 95,18% dengan variansi yang rendah. Sebaliknya, SVM menunjukkan performa yang relatif rendah dengan akurasi pengujian sekitar 62%, yang mengindikasikan keterbatasan model dalam menangkap kompleksitas pola trafik jaringan. Analisis kurva ROC, Precision–Recall, confusion matrix, dan learning curve mengonfirmasi bahwa Random Forest memiliki kemampuan generalisasi yang lebih baik dibandingkan SVM. Hasil penelitian ini menunjukkan bahwa kombinasi SelectKBest dan Random Forest tidak hanya menghasilkan akurasi yang tinggi, tetapi juga menawarkan efisiensi komputasi melalui reduksi dimensi fitur, sehingga layak digunakan untuk klasifikasi trafik VPN pada lingkungan jaringan skala besar.

Kata Kunci: SelectKBest; Random Forest; Support Vector Machine; VPN Traffic Classification

Abstract–The increasing use of Virtual Private Networks (VPNs) in modern networks poses significant challenges for network monitoring and traffic management, particularly in accurately and efficiently distinguishing VPN and non-VPN traffic. This study aims to analyze the effectiveness of the SelectKBest feature selection method in improving VPN traffic classification performance using Random Forest and Support Vector Machine (SVM) algorithms. The dataset used in this study is the CIC VPN-NonVPN Traffic Dataset provided by the Canadian Institute for Cybersecurity (CIC), which is widely recognized as a standard benchmark in network security research. Feature selection was performed using SelectKBest with the ANOVA (f_classif) scoring function, reducing the original feature set to 15 most relevant features. Experimental results show that the Random Forest classifier achieved a test accuracy of 84.94%, along with high F1-score and ROC-AUC values, and an average cross-validation accuracy of 95.18% with low variance. In contrast, the SVM model demonstrated relatively poor performance, with a test accuracy of approximately 62%, indicating its limitation in capturing the complex patterns of network traffic data. Further analysis using ROC curves, Precision–Recall curves, confusion matrices, and learning curves confirms that Random Forest exhibits superior generalization capability compared to SVM. These findings indicate that the combination of SelectKBest and Random Forest not only delivers high classification performance but also improves computational efficiency through feature dimensionality reduction, making it suitable for large-scale VPN traffic classification scenarios.

Keywords: SelectKBest; Random Forest; Support Vector Machine; VPN Traffic Classification

1. PENDAHULUAN

Pertumbuhan penggunaan Virtual Private Network (VPN) sebagai mekanisme enkripsi dan tunneling pada jaringan modern membuat sebagian besar trafik menjadi tidak dapat dianalisis dengan teknik tradisional seperti port-based identification dan deep packet inspection (DPI) (Balachandran & Amritha, 2022). Hal ini menimbulkan tantangan serius bagi operator jaringan dalam memantau dan mengelola trafik, khususnya untuk membedakan trafik VPN vs Non-VPN pada level flow, bukan sekadar jenis layanan atau aplikasi di dalam VPN(Afuwape et al., 2021a; Lohiya & Bamnote, 2025). Kegagalan mengklasifikasikan trafik secara tepat berpotensi menurunkan kualitas layanan (QoS), menghambat kebijakan keamanan, dan meningkatkan risiko ancaman yang bersembunyi di balik kanal terenkripsi (Afuwape et al., 2021a; Messaoud, 2025).

Sejumlah penelitian terkini menunjukkan bahwa machine learning (ML) dan deep learning mampu meningkatkan akurasi klasifikasi trafik terenkripsi, termasuk VPN/non-VPN, dibandingkan metode rule-based tradisional (Afuwape et al., 2021a; Messaoud, 2025; Telikani et al., 2022). Misalnya, Balachandran & Amritha (2022) menggabungkan estimasi entropi dan fitur waktu untuk klasifikasi VPN dan Non-VPN, serta identifikasi aplikasi, dan melaporkan akurasi di atas 90% menggunakan Random Forest, KNN, dan ANN (Balachandran & Amritha, 2022). Afuwape et al. (2021) mengevaluasi berbagai algoritma ML pada trafik VPN/non-VPN dan menemukan bahwa ensemble seperti Random Forest dan Gradient Boosting memberikan Precision, Recall, dan F1-score lebih tinggi daripada classifier tunggal (Afuwape et al., 2021a).

Penelitian lain berfokus pada optimalisasi fitur dan arsitektur model. Elnawawy et al. (2020) membuktikan bahwa pemilihan fitur menggunakan Random Forest dan regresi bertahap dapat mempertahankan akurasi hingga 98,5% dengan F-score 0,93, sekaligus menekan waktu pemrosesan melalui reduksi fitur (Elnawawy et al., 2020). Pada sisi lain, Telikani et al. (2022) mengusulkan pendekatan cost-sensitive deep learning pada dataset ISCX VPN-NonVPN (yang

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

kini dikenal sebagai CIC VPN-NonVPN) untuk mengatasi ketidakseimbangan kelas, dan menunjukan peningkatan kinerja pada kelas minoritas tanpa perlu resampling agresif (Telikani et al., 2022).

Lebih spesifik pada deteksi VPN, Almomani (2022) menerapkan stacking ensemble (Random Forest, Neural Network, SVM) untuk klasifikasi VPN dan Non-VPN pada trafik terenkripsi dan mencapai akurasi sekitar 99%, namun penelitian tersebut tidak membahas secara mendalam seleksi fitur berbasis filter serta tidak menyoroti perbedaan performa antar algoritma pada skenario reduksi dimensi (Almomani, 2022). Dener et al. (2023) menggabungkan feature selection berbasis Random Forest dengan GRU dan teknik penyeimbangan data untuk dataset ISCX VPN-NonVPN dan UTMobileNet2021, dan menunjukkan bahwa kombinasi tersebut efektif untuk klasifikasi biner dan multikelas terenkripsi, tetapi fokus utamanya adalah arsitektur deep learning dan penyeimbangan data, bukan analisis komparatif yang terkontrol antara Random Forest dan SVM pada subset fitur yang dipilih (Dener et al., 2023).

Di sisi lain, kajian komparatif tradisional antara Random Forest dan SVM untuk klasifikasi trafik internet menunjukkan bahwa Random Forest seringkali mengungguli SVM dalam hal akurasi dan efisiensi pada data berdimensi tinggi (Lohiya & Bamnote, 2025; Salau & Beyene, 2024). Namun, banyak studi ini menggunakan dataset umum (misalnya UNB/UNIBS, CICIDS) dan tidak secara eksplisit membahas task biner VPN vs Non-VPN berbasis flow dengan layanan beragam maupun pengaruh seleksi fitur sederhana seperti SelectKBest-ANOVA terhadap kinerja kedua algoritma. Selain itu, beberapa penelitian VPN terbaru memaksimalkan akurasi global, tetapi kurang menekankan Recall dan kesalahan klasifikasi (false negative) dari kelas VPN, padahal salah mendeteksi VPN sebagai Non-VPN dapat berdampak serius terhadap kebijakan keamanan dan pemantauan kepatuhan (Afuwape et al., 2021; Telikani et al., 2022).

Berdasarkan celah tersebut, penelitian ini secara khusus memfokuskan pada klasifikasi biner trafik VPN vs Non-VPN pada level aliran (flow) menggunakan CIC VPN-NonVPN Traffic Dataset dari Canadian Institute for Cybersecurity (CIC) sebagai dataset standar, bukan sekadar dataset tidak terverifikasi dari repositori pihak ketiga. Penelitian ini menganalisis efektivitas metode seleksi fitur filter-based SelectKBest dengan fungsi ANOVA (f_classif) untuk mereduksi ratusan fitur menjadi 15 fitur paling relevan, kemudian membandingkan kinerja dua algoritma populer, Random Forest dan Support Vector Machine (SVM), pada skenario reduksi dimensi tersebut.

Berbeda dari banyak penelitian yang melaporkan banyak metrik sekaligus, studi ini menekankan Recall dan false negative (FN) pada kelas VPN sebagai metrik utama, karena FN VPN berarti trafik VPN yang seharusnya terdeteksi justru diklasifikasikan sebagai Non-VPN, yang berimplikasi pada blind-spot dalam monitoring jaringan (Afuwape et al., 2021a; Telikani et al., 2022). Analisis tetap melibatkan metrik pendukung seperti akurasi, F1-score, dan ROC-AUC, tetapi interpretasi utama diarahkan pada kemampuan model mengurangi FN serta menjaga kemampuan generalisasi melalui evaluasi cross-validation.

Secara teoritis, Random Forest sebagai ensemble decision tree dikenal robust untuk data berdimensi tinggi dan mampu menangkap hubungan non-linear yang kompleks antar fitur trafik (Boateng et al., 2020; Elnawawy et al., 2020; Messaoud, 2025), sedangkan SVM memiliki keunggulan pada data berdimensi tinggi dengan margin yang jelas, namun sensitif terhadap pemilihan kernel, parameter, dan skala fitur (Boateng et al., 2020; Cervantes et al., 2020). Kinerja SVM yang menurun pada dataset besar atau distribusi fitur yang kompleks juga telah dilaporkan pada beberapa studi jaringan dan keamanan (Cervantes et al., 2020; Salau & Beyene, 2024). Dengan demikian, pengujian keduanya pada fitur-fitur terpilih dari CIC VPN-NonVPN memberikan landasan empiris untuk mengevaluasi klaim teoretis tersebut dalam konteks trafik VPN modern.

Berdasarkan latar belakang dan celah penelitian yang telah diuraikan, penelitian ini bertujuan untuk mengevaluasi efektivitas metode seleksi fitur *filter-based* SelectKBest dengan fungsi ANOVA (*f_classif*) dalam mereduksi dimensi fitur pada CIC VPN-NonVPN Traffic Dataset tanpa mengorbankan kinerja klasifikasi trafik VPN dan Non-VPN. Dengan menggunakan subset 15 fitur paling relevan, penelitian ini membandingkan kinerja dua algoritma yang banyak digunakan dalam klasifikasi trafik jaringan, yaitu Random Forest dan Support Vector Machine (SVM), dengan penekanan khusus pada metrik *Recall* dan kesalahan *false negative* (FN) pada kelas VPN. Penekanan ini didasarkan pada konteks keamanan jaringan, di mana kegagalan mendeteksi trafik VPN yang sebenarnya dapat menciptakan *blind spot* dalam proses pemantauan dan penerapan kebijakan keamanan jaringan (Afuwape et al., 2021a; Telikani et al., 2022).

Selain itu, penelitian ini mengevaluasi kemampuan generalisasi kedua model melalui skema *cross-validation* dan analisis kurva pembelajaran untuk menilai stabilitas performa model pada berbagai proporsi data latih. Pendekatan evaluasi ini memberikan gambaran yang lebih komprehensif mengenai konsistensi kinerja model dibandingkan pengujian berbasis *hold-out set* tunggal, serta relevan untuk skenario penerapan pada lingkungan jaringan berskala besar (Elnawawy et al., 2020; Lohiya & Bamnote, 2025). Dari sisi efisiensi, reduksi fitur hingga 15 atribut yang paling informatif diharapkan dapat menurunkan beban komputasi pada tahap ekstraksi fitur dan inferensi, sehingga mendukung implementasi klasifikasi VPN/non-VPN secara *real-time* tanpa beban pemrosesan yang berlebihan (Afuwape et al., 2021a; Elnawawy et al., 2020).

Kontribusi penelitian ini terletak pada penyajian analisis komparatif yang terkontrol antara Random Forest dan SVM pada *task* klasifikasi biner VPN vs Non-VPN berbasis *flow* dengan seleksi fitur eksplisit menggunakan SelectKBest-ANOVA pada dataset CIC VPN-NonVPN. Pendekatan ini masih relatif jarang dibahas secara mendalam dalam penelitian sejenis yang umumnya berfokus pada arsitektur *deep learning* atau teknik penyeimbangan data (Almomani, 2022; Balachandran & Amritha, 2022; Dener et al., 2023; Telikani et al., 2022). Hasil penelitian menunjukkan bahwa kombinasi SelectKBest dan Random Forest mampu mempertahankan kinerja klasifikasi yang

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

tinggi, khususnya dari sisi *Recall* kelas VPN dan stabilitas hasil *cross-validation*, dengan kompleksitas komputasi yang relatif rendah. Di sisi lain, penelitian ini juga memberikan bukti empiris mengenai keterbatasan SVM dalam menangani pola trafik VPN yang kompleks pada konfigurasi fitur yang sama, sebagaimana juga dilaporkan pada beberapa studi klasifikasi trafik dan keamanan jaringan (Cervantes et al., 2020; Salau & Beyene, 2024).

Dengan demikian, penelitian ini memberikan landasan metodologis dan empiris yang relevan bagi pengembangan sistem klasifikasi trafik VPN/non-VPN yang akurat dan efisien untuk mendukung manajemen dan keamanan jaringan modern.

2. METODOLOGI PENELITIAN

2.1 Kerangka Dasar Penelitian

Penelitian ini merupakan penelitian eksperimental kuantitatif di bidang rekayasa jaringan dan keamanan informasi yang bertujuan untuk menganalisis kinerja metode seleksi fitur SelectKBest terhadap performa algoritma Random Forest dan Support Vector Machine (SVM) dalam melakukan klasifikasi biner trafik VPN dan Non-VPN. Fokus utama penelitian ini adalah membedakan trafik VPN vs Non-VPN pada level aliran (flow), bukan mengklasifikasikan jenis layanan atau aplikasi di dalam VPN, sehingga hasil penelitian relevan untuk kebutuhan pemantauan dan pengamanan jaringan modern.

Kerangka penelitian diawali dengan studi literatur terhadap penelitian-penelitian terkait klasifikasi trafik jaringan terenkripsi menggunakan pendekatan machine learning, khususnya Random Forest, SVM, serta teknik seleksi fitur berbasis filter seperti SelectKBest dengan fungsi ANOVA (Afuwape et al., 2021a; Elnawawy et al., 2020; Telikani et al., 2022). Studi literatur ini digunakan untuk mengidentifikasi permasalahan utama, menentukan celah penelitian (research gap), serta merumuskan pendekatan metodologis yang sesuai dengan karakteristik trafik VPN modern.

Dataset yang digunakan dalam penelitian ini adalah CIC VPN-NonVPN Traffic Dataset yang disediakan oleh Canadian Institute for Cybersecurity (CIC), yang merupakan dataset standar dan banyak digunakan dalam penelitian klasifikasi trafik terenkripsi (Gupta, 2021; Gupta et al., 2021; Izadi et al., 2022; Zhengyang Liu et al., 2025; Ziao Liu et al., 2025). Dataset ini berisi data trafik jaringan berbasis flow dengan ratusan fitur statistik yang merepresentasikan karakteristik paket dan aliran komunikasi. Penggunaan dataset standar ini bertujuan untuk meningkatkan validitas dan reprodusibilitas hasil penelitian dibandingkan dataset tidak terverifikasi dari repositori pihak ketiga.

Variabel independen dalam penelitian ini adalah fitur-fitur trafik jaringan hasil ekstraksi flow, sedangkan variabel dependen adalah label kelas trafik, yaitu VPN dan Non-VPN. Untuk meningkatkan efisiensi komputasi dan mengurangi kompleksitas model, dilakukan seleksi fitur menggunakan metode SelectKBest berbasis ANOVA (f_classif) yang memilih sejumlah fitur dengan skor statistik terbaik terhadap label kelas. Dalam penelitian ini, jumlah fitur yang dipilih ditetapkan sebanyak 15 fitur (k = 15), yang ditentukan berdasarkan keseimbangan antara performa klasifikasi dan efisiensi pemrosesan.

Model klasifikasi yang digunakan adalah Random Forest dan Support Vector Machine (SVM) sebagai metode pembanding. Random Forest dipilih karena kemampuannya menangani data berdimensi tinggi serta menangkap hubungan non-linear antar fitur (Boateng et al., 2020; Elnawawy et al., 2020; Khademioureh et al., 2025; Olaniran et al., 2025; Ratnasingam & Muñoz-Lopez, 2023), sedangkan SVM dipilih karena dikenal efektif pada data berdimensi tinggi dengan margin pemisah yang jelas, meskipun sensitif terhadap pemilihan parameter dan skala fitur (Cervantes et al., 2020; Tao et al., 2025; Thakur et al., 2025).

Evaluasi kinerja model dilakukan menggunakan beberapa metrik, yaitu akurasi, precision, recall, F1-score, dan ROC-AUC. Namun, metrik Recall dan False Negative (FN) pada kelas VPN menjadi fokus utama penelitian ini, karena kesalahan mendeteksi trafik VPN sebagai Non-VPN dapat menimbulkan blind-spot dalam pengawasan dan kebijakan keamanan jaringan (Afuwape et al., 2021a; Telikani et al., 2022). Dengan kerangka dasar ini, penelitian diharapkan mampu memberikan gambaran empiris mengenai efektivitas seleksi fitur SelectKBest serta perbandingan kinerja Random Forest dan SVM dalam konteks klasifikasi trafik VPN modern.

2.2 Tahapan Penelitian

Tahapan penelitian ini disusun secara sistematis untuk memastikan bahwa proses klasifikasi trafik VPN dan Non-VPN dilakukan secara terstruktur, dapat direplikasi, serta terhindar dari kebocoran data (*data leakage*). Penelitian diawali dengan pengumpulan dataset CIC VPN-NonVPN Traffic Dataset yang digunakan sebagai sumber data utama. Dataset tersebut selanjutnya melalui tahap prapemrosesan yang mencakup pembersihan data, penanganan nilai kosong, serta pengkodean fitur kategorikal agar seluruh atribut dapat diproses oleh algoritma pembelajaran mesin.

Data yang telah dipraproses kemudian dibagi menjadi data latih dan data uji menggunakan metode *stratified train-test split* dengan rasio 80:20. Pendekatan ini bertujuan untuk menjaga proporsi kelas VPN dan Non-VPN tetap konsisten pada kedua subset data. Analisis rasio kelas dilakukan untuk memastikan bahwa penelitian ini tidak menerapkan teknik *oversampling* maupun *undersampling*, melainkan menggunakan distribusi data asli. Oleh karena itu, evaluasi kinerja model lebih menekankan pada metrik Recall dan F1-score, yang dinilai lebih representatif dalam kondisi distribusi kelas yang tidak sepenuhnya seimbang (Afuwape et al., 2021a).

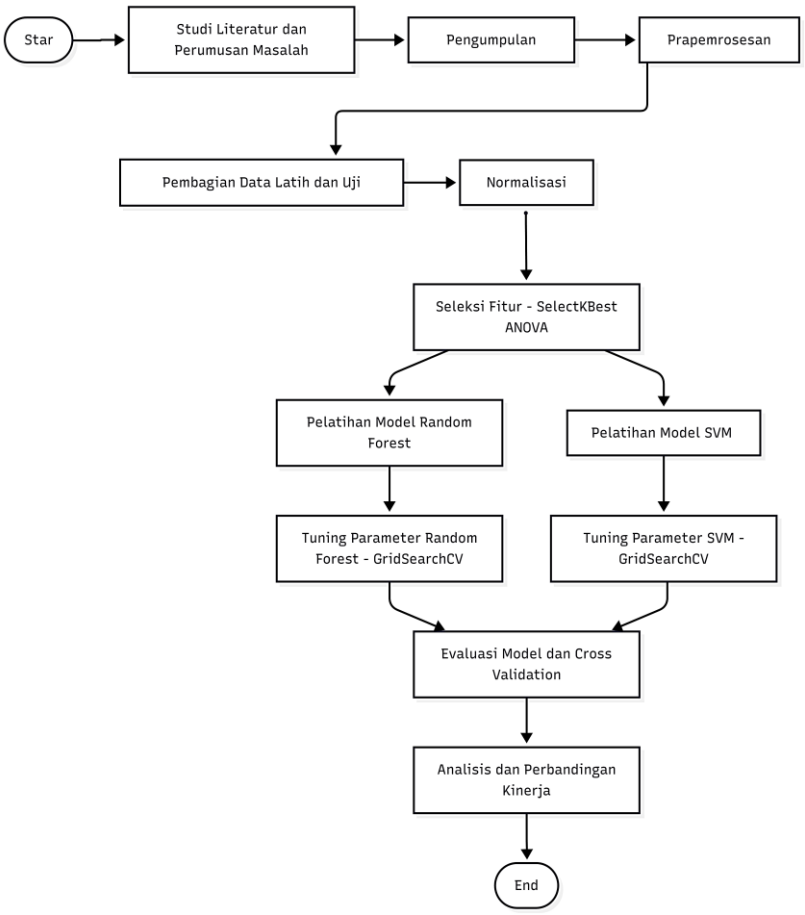
Tahap selanjutnya adalah penerapan *machine learning pipeline* yang mengintegrasikan proses standardisasi fitur menggunakan StandardScaler, seleksi fitur menggunakan SelectKBest dengan fungsi ANOVA (f_classif), serta

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

algoritma klasifikasi Random Forest dan Support Vector Machine (SVM) dalam satu alur terpadu. Penggunaan pipeline ini bertujuan untuk memastikan bahwa seluruh tahapan prapemrosesan dan seleksi fitur dilakukan di dalam skema validasi silang, sehingga potensi terjadinya *data leakage* dapat dihindari secara menyeluruh.

Untuk memperoleh konfigurasi model yang optimal, dilakukan proses *hyperparameter tuning* menggunakan GridSearchCV dengan skema validasi silang 5-fold. Pada algoritma Random Forest, parameter yang dituning meliputi jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), serta jumlah minimum sampel untuk proses pemisahan (*min_samples_split*). Sementara itu, pada algoritma SVM dengan kernel Radial Basis Function (RBF), proses tuning difokuskan pada parameter C dan gamma. Rentang parameter yang digunakan ditentukan berdasarkan praktik umum dan temuan pada penelitian sebelumnya dalam klasifikasi trafik jaringan (Cervantes et al., 2020). Model terbaik yang diperoleh dari proses tuning selanjutnya dievaluasi menggunakan data uji dengan berbagai metrik kinerja, serta dianalisis melalui *confusion matrix* untuk mengidentifikasi kesalahan klasifikasi, khususnya *false negative* pada kelas VPN. Selain itu, dilakukan evaluasi tambahan menggunakan *k-fold cross-validation* dan analisis *learning curve* untuk menilai stabilitas model serta kemampuan generalisasi terhadap variasi ukuran data latih. Tahapan penelitian diakhiri dengan analisis feature importance pada algoritma Random Forest untuk mengidentifikasi fitur-fitur yang memiliki kontribusi paling signifikan dalam membedakan trafik VPN dan Non-VPN. Seluruh hasil dari tahapan ini kemudian dianalisis secara komprehensif sebagai dasar penarikan kesimpulan mengenai efektivitas metode SelectKBest serta perbandingan kinerja antara Random Forest dan SVM dalam konteks klasifikasi trafik VPN modern.

Alur dan tahapan penelitian yang digunakan dalam klasifikasi trafik VPN dan Non-VPN secara keseluruhan dirangkum dalam sebuah diagram alir untuk memperjelas urutan proses yang dilakukan.



Gambar 1. Tahapan penelitian klasifikasi trafik VPN dan Non-VPN

Alur penelitian diawali dengan studi literatur dan perumusan masalah, kemudian dilanjutkan dengan pengumpulan dataset CIC VPN-NonVPN sebagai sumber data utama. Data yang diperoleh selanjutnya melalui tahap prapemrosesan dan pengkodean fitur agar dapat diproses oleh algoritma pembelajaran mesin, sebelum dibagi menjadi data latih dan data uji secara terstratifikasi. Proses normalisasi fitur menggunakan StandardScaler dan seleksi fitur dengan metode SelectKBest berbasis ANOVA diterapkan untuk menghasilkan subset fitur yang paling relevan. Subset fitur tersebut kemudian digunakan pada dua skema pelatihan model, yaitu Random Forest dan Support Vector Machine dengan kernel RBF, yang masing-masing dituning menggunakan GridSearchCV. Tahap akhir penelitian meliputi evaluasi model menggunakan berbagai metrik kinerja, validasi silang, serta analisis perbandingan kinerja untuk menentukan model yang paling efektif dalam membedakan trafik VPN dan Non-VPN, sebagaimana dirangkum dalam alur penelitian pada Gambar 1.

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen dan pembahasan berdasarkan rancangan metodologi penelitian yang digunakan dalam studi ini. Rancangan metodologi meliputi integrasi tahapan prapemrosesan data, seleksi fitur berbasis ANOVA, optimasi parameter menggunakan GridSearchCV, serta evaluasi performa yang menekankan kemampuan generalisasi model melalui cross-validation. Eksperimen dilakukan pada CIC VPN–NonVPN Traffic Dataset dengan pembagian data latih dan uji secara terstratifikasi (80:20), sehingga proporsi kelas VPN dan Non-VPN tetap seimbang.

Dataset hasil prapemrosesan memiliki ukuran data latih sebanyak 4.776 sampel dan data uji sebanyak 1.195 sampel dengan distribusi kelas yang relatif seimbang (Non-VPN 50,11% dan VPN 49,89%). Seleksi fitur menggunakan SelectKBest (f_classif) mereduksi fitur menjadi 15 atribut paling relevan, yang bertujuan menurunkan kompleksitas komputasi tanpa mengorbankan performa klasifikasi.

3.1 Hasil

3.1.1 Hasil Optimasi Hyperparameter

Optimasi hyperparameter menggunakan GridSearchCV dengan skema validasi silang 5-fold menghasilkan konfigurasi terbaik untuk masing-masing model. Random Forest memperoleh parameter optimal berupa jumlah pohon (*n_estimators*) sebesar 150, kedalaman maksimum pohon (*max_depth*) sebesar 20, dan *min_samples_split* sebesar 2. Konfigurasi ini menunjukkan bahwa model membutuhkan jumlah pohon yang relatif banyak dengan kedalaman terkontrol untuk menangkap pola kompleks trafik VPN tanpa mengalami overfitting. Pada SVM dengan kernel RBF, parameter terbaik yang diperoleh adalah nilai $C = 10$ dan *gamma* = scale. Nilai *C* yang cukup besar menunjukkan upaya model untuk meminimalkan kesalahan klasifikasi pada data latih, sementara *gamma* berskala otomatis menyesuaikan pengaruh setiap sampel terhadap fungsi keputusan.

3.1.2 Evaluasi Kinerja pada Data Uji

Tabel 1 menyajikan perbandingan kinerja Random Forest dan SVM berdasarkan metrik utama.

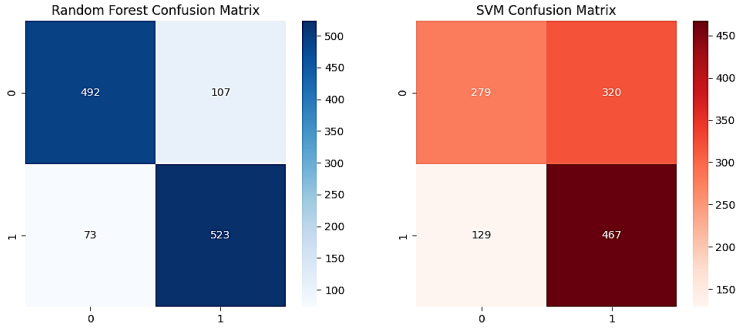
Tabel 1. Perbandingan Kinerja Model

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0,8494	0,8302	0,8775	0,8532	0,9193
SVM (RBF)	0,6243	0,5934	0,7836	0,6753	0,6797

Hasil pada Tabel 1 menunjukkan bahwa Random Forest secara konsisten mengungguli SVM pada seluruh metrik evaluasi. Nilai akurasi Random Forest sebesar 84,94% menunjukkan kemampuan model dalam mengklasifikasikan trafik VPN dan Non-VPN secara tepat. Nilai *recall* kelas VPN yang tinggi (87,75%) mengindikasikan bahwa sebagian besar trafik VPN berhasil terdeteksi dengan baik, sehingga potensi *false negative*—yang berbahaya dalam konteks keamanan jaringan, dapat ditekan Sebaliknya, SVM hanya mencapai akurasi 62,43% dengan ROC-AUC sebesar 0,68, yang menunjukkan kemampuan pemisahan kelas yang relatif terbatas. Meskipun *recall* VPN pada SVM tergolong cukup tinggi, nilai *precision* yang rendah mengindikasikan banyaknya kesalahan prediksi, khususnya *false positive*.

3.1.3 Analisis Confusion Matrix

Analisis *confusion matrix* memperlihatkan bahwa Random Forest mampu mengklasifikasikan kedua kelas secara relatif seimbang. Jumlah *true positive* dan *true negative* yang tinggi disertai *false positive* dan *false negative* yang rendah menunjukkan bahwa model tidak bias terhadap salah satu kelas. Hal ini penting untuk implementasi nyata, karena kesalahan mendeteksi VPN sebagai Non-VPN dapat menciptakan *blind spot* dalam pemantauan jaringan. Pada SVM, kesalahan klasifikasi Non-VPN sebagai VPN relatif lebih tinggi. Kondisi ini berpotensi menimbulkan alarm palsu dan menurunkan efisiensi sistem pengawasan jaringan apabila diterapkan secara operasional. Hal tersebut ditunjukan pada Gambar 2.

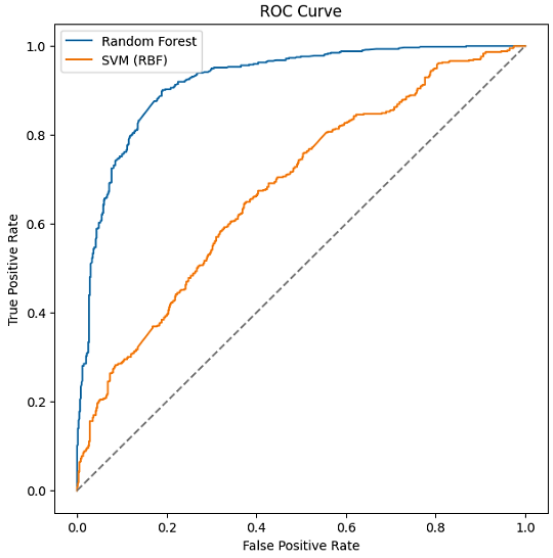


Gambar 2. Confusion matrix hasil klasifikasi trafik VPN dan Non-VPN menggunakan (a) Random Forest dan (b) Support Vector Machine.

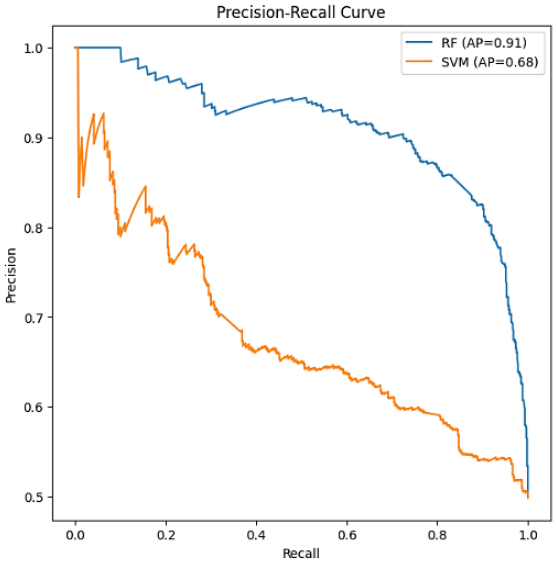
TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

3.1.4 Analisis ROC-AUC dan Precision–Recall

Kurva ROC Random Forest berada jauh di atas garis diagonal dengan nilai ROC-AUC sebesar 0,9193, yang menandakan kemampuan pemisahan kelas yang sangat baik pada berbagai ambang keputusan. Kurva Precision–Recall juga menunjukkan *average precision* yang tinggi, menandakan bahwa prediksi VPN yang dihasilkan memiliki tingkat keandalan yang kuat. Hal tersebut ditunjukan pada **Gambar 3**.



Gambar 3. Kurva ROC (Receiver Operating Characteristic) untuk model Random Forest dan SVM dalam membedakan trafik VPN dan Non-VPN.



Gambar 4. Kurva Precision–Recall untuk model Random Forest dan SVM sebagai evaluasi kinerja klasifikasi pada berbagai nilai ambang keputusan

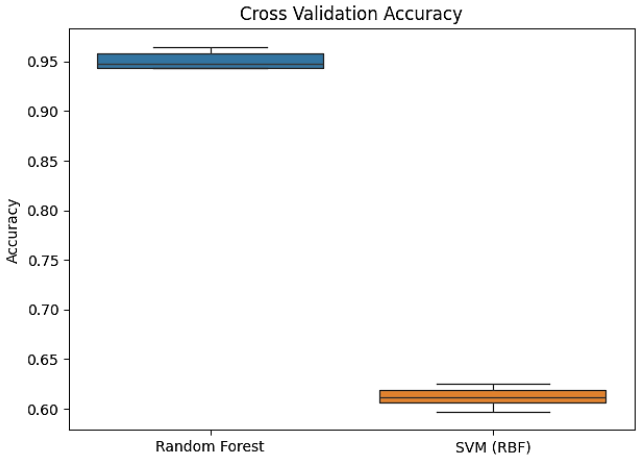
Sebaliknya, kurva ROC dan Precision–Recall milik SVM lebih mendekati garis diagonal. Hal ini menunjukkan bahwa peningkatan *recall* pada SVM diikuti oleh penurunan *precision*, sehingga model kurang konsisten dalam membedakan trafik VPN dan Non-VPN seperti yang ditunjukan pada Gambar 4.

3.1.5 Evaluasi Kestabilan Model dengan Cross-Validation

Evaluasi kestabilan menggunakan 5-Fold Cross-Validation menunjukkan bahwa Random Forest mencapai *mean accuracy* sebesar 0,9518 dengan deviasi ±0,0084. Nilai ini menunjukkan bahwa performa Random Forest sangat stabil dan konsisten pada berbagai subset data. Sebaliknya, SVM hanya mencapai *mean accuracy* sebesar 0,6118 ± 0,0097, yang menandakan performa yang lebih rendah dan kurang stabil. Hal tersebut ditunjukan pada Gambar 5.

Sebaran nilai akurasi yang sempit pada Random Forest mengindikasikan bahwa model tidak sensitif terhadap variasi data latih dan data uji, sehingga memiliki kemampuan generalisasi yang baik.

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136



Gambar 5. Evaluasi Kestabilan Model dengan Cross Validation

3.2 Pembahasan

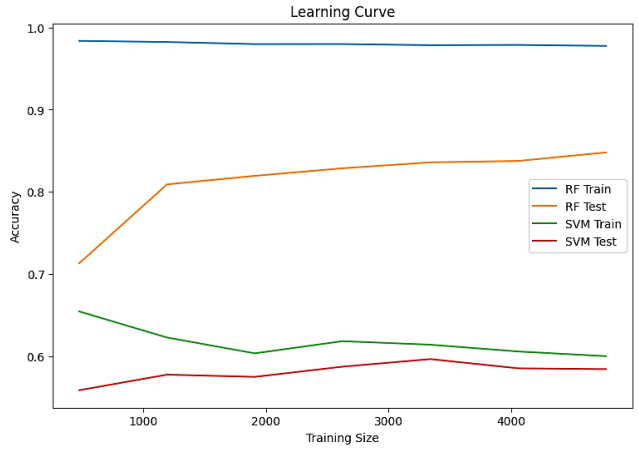
Hasil penelitian ini menunjukkan adanya perbedaan kinerja yang jelas antara algoritma Random Forest dan Support Vector Machine (SVM) dalam melakukan klasifikasi trafik VPN dan Non-VPN berbasis fitur aliran. Perbedaan tersebut tercermin secara konsisten pada berbagai metrik evaluasi kuantitatif maupun analisis visual yang disajikan pada bagian hasil.

Berdasarkan nilai akurasi dan ROC-AUC pada Tabel 1, Random Forest menunjukkan kemampuan klasifikasi yang lebih unggul dibandingkan SVM. Keunggulan ini mengindikasikan bahwa Random Forest lebih efektif dalam memodelkan hubungan non-linear dan kompleks antar fitur statistik trafik jaringan. Dalam konteks klasifikasi trafik terenkripsi, kemampuan ini menjadi penting karena pola trafik VPN dan Non-VPN sering kali tidak dapat dipisahkan secara linear.

Analisis kurva ROC dan Precision–Recall memperkuat temuan tersebut. Kurva ROC Random Forest yang berada jauh di atas garis diagonal menunjukkan kemampuan diskriminasi kelas yang tinggi pada berbagai nilai ambang keputusan. Selain itu, nilai average precision yang lebih baik mencerminkan keseimbangan yang lebih optimal antara tingkat deteksi trafik VPN dan kesalahan klasifikasi. Kondisi ini sangat relevan untuk sistem pemantauan jaringan, karena kesalahan false negative pada trafik VPN berpotensi menciptakan celah pemantauan, sementara false positive yang berlebihan dapat meningkatkan beban operasional sistem.

Hasil confusion matrix juga menunjukkan bahwa Random Forest mampu mengklasifikasikan trafik VPN dan Non-VPN secara relatif seimbang dengan jumlah kesalahan yang lebih rendah pada kedua kelas. Sebaliknya, SVM cenderung menghasilkan kesalahan klasifikasi yang lebih tinggi, khususnya pada trafik Non-VPN yang diklasifikasikan sebagai VPN. Hal ini mengindikasikan bahwa meskipun SVM memiliki nilai recall yang cukup tinggi, kualitas prediksi secara keseluruhan masih kurang optimal karena tingginya false positive.

Dari sisi kemampuan generalisasi, analisis learning curve menunjukkan bahwa Random Forest memiliki jarak yang relatif kecil antara kinerja data latih dan data uji. Pola ini menandakan bahwa model mampu memanfaatkan penambahan data latih secara efektif tanpa mengalami overfitting. Sebaliknya, performa SVM cenderung stagnan meskipun jumlah data latih meningkat, yang mengindikasikan keterbatasan model dalam menangkap kompleksitas pola trafik VPN modern berbasis fitur aliran. Hal tersebut ditunjukan pada **Gambar 6.**

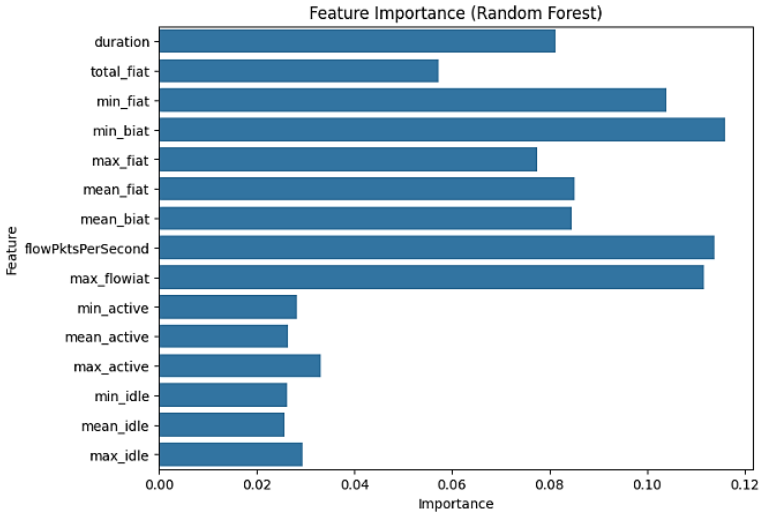


Gambar 6. *Learning curve* model Random Forest dan Support Vector Machine pada klasifikasi trafik VPN dan Non-VPN

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

Evaluasi kestabilan menggunakan skema 5-fold cross-validation semakin menegaskan keunggulan Random Forest. Nilai akurasi rata-rata yang tinggi dengan deviasi yang kecil menunjukkan bahwa performa model relatif konsisten pada berbagai subset data. Stabilitas ini merupakan indikator penting bagi penerapan model pada lingkungan jaringan nyata yang bersifat dinamis. Sebaliknya, SVM menunjukkan variasi performa yang lebih besar, yang menandakan sensitivitas model terhadap perubahan distribusi data latih dan data uji.

Analisis feature importance memberikan wawasan tambahan mengenai faktor-faktor yang berkontribusi terhadap keberhasilan klasifikasi. Fitur-fitur berbasis karakteristik temporal dan intensitas aliran, seperti *flowPktsPerSecond*, *min_biat*, dan *max_flowiat*, terbukti memiliki peranan dominan dalam membedakan trafik VPN dan Non-VPN. Temuan ini selaras dengan karakteristik trafik VPN yang umumnya memiliki pola interval waktu dan kecepatan paket yang berbeda akibat mekanisme enkripsi dan tunneling. Hal ini menunjukkan bahwa fitur statistik berbasis aliran masih efektif untuk mendeteksi trafik terenkripsi tanpa perlu melakukan inspeksi payload. Hal tersebut ditunjukan oleh Gambar 7.



Gambar 7. *Feature importance* hasil model Random Forest pada klasifikasi trafik VPN dan Non-VPN

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pendekatan klasifikasi berbasis Random Forest dengan seleksi fitur yang tepat mampu menghasilkan performa yang lebih andal dan stabil dibandingkan SVM pada tugas klasifikasi trafik VPN dan Non-VPN. Pendekatan ini memberikan evaluasi yang lebih representatif terhadap kemampuan generalisasi model serta relevan untuk diterapkan pada sistem pemantauan jaringan berskala besar.

4. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa pendekatan klasifikasi trafik jaringan berbasis *machine learning* mampu membedakan trafik VPN dan Non-VPN secara efektif dengan tingkat akurasi dan stabilitas yang memadai. Berdasarkan hasil eksperimen, model Random Forest menunjukkan kinerja yang lebih unggul dibandingkan Support Vector Machine pada hampir seluruh metrik evaluasi, termasuk akurasi, *recall*, *F1-score*, dan ROC-AUC. Keunggulan tersebut tidak hanya tercermin dari nilai numerik pada data uji, tetapi juga diperkuat oleh analisis visual melalui kurva ROC, Precision–Recall, *learning curve*, evaluasi *cross-validation*, serta analisis *feature importance*. Temuan ini menjawab permasalahan utama penelitian, yaitu pemilihan model klasifikasi yang paling andal untuk mendeteksi trafik VPN pada lingkungan jaringan modern yang semakin kompleks dan didominasi oleh trafik terenkripsi. Meskipun demikian, penelitian ini masih memiliki keterbatasan, terutama pada ruang lingkup dataset yang digunakan dan ketergantungan pada fitur statistik aliran tanpa mempertimbangkan dinamika trafik secara real-time. Selain itu, evaluasi dilakukan pada skenario klasifikasi biner, sehingga belum merepresentasikan kondisi jaringan nyata yang melibatkan berbagai jenis aplikasi dan protokol VPN. Oleh karena itu, penelitian selanjutnya dapat diarahkan pada penggunaan dataset yang lebih beragam, pengujian pada skenario multi-kelas, serta eksplorasi pendekatan *deep learning* atau *online learning* untuk meningkatkan adaptabilitas model terhadap perubahan pola trafik. Dengan demikian, hasil penelitian ini diharapkan dapat menjadi landasan yang kuat bagi pengembangan sistem deteksi trafik VPN yang lebih akurat, stabil, dan siap diterapkan pada lingkungan operasional jaringan yang dinamis.

REFERENCES

Afuwape, A., Xu, Y., Anajemba, J., & Srivastava, G. (2021). Performance evaluation of secured network traffic classification using a machine learning approach. *Comput. Stand. Interfaces*, *78*, 103545. https://doi.org/10.1016/j.csi.2021.103545

TIN: Terapan Informatika Nusantara
Vol 6, No 8, January 2026, page 1533-1541
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i8.9136

Almomani, A. (2022). Classification of Virtual Private networks encrypted traffic using ensemble learning algorithms. *Egyptian Informatics Journal*. https://doi.org/10.1016/j.eij.2022.06.006

Balachandran, A., & Amritha, P. (2022). VPN Network Traffic Classification Using Entropy Estimation and Time-Related Features. *IOT with Smart Systems*. https://doi.org/10.1007/978-981-16-3945-6_50

Boateng, E., Otoo., J., & Abaye, D. (2020). *Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review*. *08*, 341–357. https://doi.org/10.4236/jdaip.2020.84020

Cervantes, J., García, F., Rodríguez-Mazahua, L., & López-Chau, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Dener, M., Al, S., & Ok, G. (2023). RFSE-GRU: Data Balanced Classification Model for Mobile Encrypted Traffic in Big Data Environment. *IEEE Access*, *11*, 21831–21847. https://doi.org/10.1109/access.2023.3251745

Elnawawy, M., Sagahyroon, A., & Shanableh, T. (2020). FPGA-Based Network Traffic Classification Using Machine Learning. *IEEE Access*, *8*, 175637–175650. https://doi.org/10.1109/access.2020.3026831

Gupta, A. (2021). VPN-nonVPN Traffic Classification Using Deep Reinforced Naive Bayes and Fuzzy K-means Clustering. *2021 IEEE 41st International Conference on Distributed Computing Systems Workshops (ICDCSW)*, 1–6. https://doi.org/10.1109/icdcsw53096.2021.00008

Gupta, N., Jindal, V., & Bedi, P. (2021). Encrypted Traffic Classification Using eXtreme Gradient Boosting Algorithm. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-16-3071-2_20

Izadi, S., Ahmadi, M., & Rajabzadeh, A. (2022). Network Traffic Classification Using Deep Learning Networks and Bayesian Data Fusion. *Journal of Network and Systems Management*, *30*. https://doi.org/10.1007/s10922-021-09639-z

Khademioureh, S., Dinu, I., & Peignier, S. (2025). GSHAPA: Gene Set Analysis for Single-Cell RNAseq Using Random Forest and SHAP Values. *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. https://doi.org/10.1145/3672608.3707901

Liu, Zhengyang, Wei, Q., Song, Q., & Duan, C. (2025). Fine-Grained Encrypted Traffic Classification Using Dual Embedding and Graph Neural Networks. *Electronics*. https://doi.org/10.3390/electronics14040778

Liu, Ziao, Xie, Y., Luo, Y., Wang, Y., & Ji, X. (2025). TransECA-Net: A Transformer-Based Model for Encrypted Traffic Classification. *Applied Sciences*. https://doi.org/10.3390/app15062977

Lohiya, P., & Bamnote, G. (2025). Internet Traffic Classification through Supervised Learning: Exploring Machine Learning Techniques. *Intelligent Methods in Engineering Sciences*. https://doi.org/10.58190/imiens.2025.119

Messaoud, M. (2025). Classification Of Network Traffic Using Machine Learning Models On The Netml Dataset. *International Journal of Computer Networks & Communications*. https://doi.org/10.5121/ijcnc.2025.17307

Olaniran, O., Olaniran, S., Alzahrani, A., Alharbi, N. M., & Alzahrani, A. A. (2025). Random Forest Adaptation for High-Dimensional Count Regression. *Mathematics*. https://doi.org/10.3390/math13183041

Ratnasingam, S., & Muñoz-Lopez, J. (2023). Distance Correlation-Based Feature Selection in Random Forest. *Entropy*, *25*. https://doi.org/10.3390/e25091250

Salau, A. O., & Beyene, M. M. (2024). Software defined networking based network traffic classification using machine learning techniques. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-70983-6

Tao, Y., Yan, J., Niu, E., Zhai, P., & Zhang, S. (2025). An SVM-Based Anomaly Detection Method for Power System Security Analysis Using Particle Swarm Optimization and t-SNE for High-Dimensional Data Classification. *Processes*. https://doi.org/10.3390/pr13020549

Telikani, A., Gandomi, A., Choo, K., & Shen, J. (2022). A Cost-Sensitive Deep Learning-Based Approach for Network Traffic Classification. *IEEE Transactions on Network and Service Management*, *19*, 661–670. https://doi.org/10.1109/tnsm.2021.3112283

Thakur, S., Tiwari, V. K., & Agrawal, J. (2025). Performance Analysis of Linear Kernel Support Vector Machine Models on Real-World Datasets. *International Journal of Advanced Networking and Applications*. https://doi.org/10.35444/ijana.2025.17106