Volume 12 Issue 1 (2027) Pages 179 - 192
Jurnal Mirai Management
ISSN : 2597-4084 (Online)

Deteksi Adware Pada Trafik Jaringan Menggunakan Metode K-Means Clustering Berbasis Data Wireshark

Rinto Dariyanto Putro[✉] , **Anjik Sukmaaji**[2]
[1,2] *Sistem Informasi, Teknologi dan Informatika, Universitas Dinamika*

Abstrak

Analisis trafik jaringan menjadi langkah penting dalam mendeteksi pola komunikasi yang berpotensi mengindikasikan aktivitas adware. Permasalahan utama penelitian ini adalah besarnya jumlah paket jaringan yang harus diperiksa secara manual sehingga pola komunikasi berulang, tujuan komunikasi dominan, dan anomali ukuran paket sulit diidentifikasi secara efisien. Penelitian menggunakan dataset yang diperoleh dari tujuh file hasil ekspor *wireshark* dengan total 135.996 paket jaringan. Data mentah diproses melalui tahapan pembersihan data, transformasi atribut, agregasi aliran komunikasi, pembentukan fitur numerik, dan normalisasi. Fitur yang dianalisis meliputi *packet count*, *total length*, *average length*, *destination frequency*, *source frequency*, *tcp warning count*, *query like count*, *packets persecond*, *bytes persecond*, serta representasi protokol dominan. Hasil agregasi menghasilkan 1.839 aliran komunikasi yang kemudian dikelompokkan menggunakan algoritma *k-means*. Jumlah klaster ditentukan melalui kombinasi *elbow method* dan *silhouette score*, sehingga dipilih K-7 sebagai konfigurasi terbaik dengan *silhouette score* 0,258 dan *davies bouldin index* 1,569. Hasil klasterisasi menunjukkan bahwa Klaster 0, 4, dan 6 memiliki karakteristik komunikasi yang paling berpotensi mencurigakan, ditandai oleh intensitas komunikasi tinggi, frekuensi tujuan yang dominan, ukuran paket tertentu, serta tingginya peringatan TCP. Temuan ini menunjukkan bahwa *k-means* efektif sebagai pendekatan eksploratif untuk penyaringan awal trafik jaringan sehingga dapat membantu proses identifikasi aktivitas adware secara lebih cepat dan terstruktur.

Kata Kunci: K-Means Clustering, Analisis Trafik Jaringan, Wireshark, Deteksi adware, Silhouette Score.

Abstract

Network traffic analysis plays an important role in identifying communication patterns associated with potential adware activity. The primary challenge addressed in this study is the large volume of network packets that must be inspected manually, making it difficult to recognize recurring communication patterns, dominant destinations, and packet-size anomalies efficiently. The dataset consists of seven Wireshark export files representing network traffic captured during internet service usage, with a total of 135,996 packets. The raw data underwent data cleaning, attribute transformation, communication flow aggregation, numerical feature extraction, and normalization. The extracted features included packet count, total length, average length, destination frequency, source frequency, tcp warning count, query like count, packets persecond, bytes persecond, and dominant protocol representation. The aggregation process produced 1,839 communication flows, which were clustered using the K-Means algorithm. The optimal number of clusters was determined through the combined use of the Elbow Method and Silhouette Score, resulting in K = 7 with a Silhouette Score of 0.258 and a Davies–Bouldin Index of 1.569. The clustering results indicate that Clusters 0, 4, and 6 exhibit

the most suspicious traffic characteristics, including intensive communication frequency, dominant destination patterns, specific packet-size behavior, and elevated TCP warning counts. These findings demonstrate that K-Means clustering is an effective exploratory approach for the preliminary filtering of network traffic, supporting faster and more structured identification of potential adware-related activities.

Keywords: *K-Means Clustering, Network Traffic Analysis, Wireshark, Adware Detection, Unsupervised Learning*.

✉ Corresponding author :
Email Address : 20410100072@dinamika.ac.id

PENDAHULUAN

Perkembangan layanan digital membuat aktivitas jaringan meningkat secara signifikan. Perangkat pengguna secara terus-menerus melakukan pertukaran data melalui berbagai protokol, baik untuk akses web, layanan pesan, sinkronisasi aplikasi, maupun komunikasi latar belakang. Kondisi tersebut memperluas risiko keamanan jaringan karena perangkat lunak berbahaya dapat memanfaatkan koneksi internet untuk menjalankan aktivitas tanpa sepengetahuan pengguna, sehingga analisis lalu lintas jaringan menjadi pendekatan penting dalam mendeteksi pola komunikasi yang mencurigakan. Salah satu ancaman yang sering ditemukan adalah adware, yaitu perangkat lunak yang menampilkan iklan secara agresif dan dapat memicu komunikasi berulang ke layanan iklan, domain pelacakan, atau tujuan jaringan tertentu. Aktivitas tersebut umumnya menghasilkan pola trafik yang berulang dan persisten sehingga dapat diidentifikasi melalui karakteristik komunikasi jaringan dan analisis fitur lalu lintas data (Sadeghpour & Vlajic, 2021).

Adware tidak selalu merusak sistem secara langsung, tetapi keberadaannya dapat menurunkan kenyamanan pengguna, meningkatkan konsumsi sumber daya perangkat, serta menimbulkan risiko terhadap privasi pengguna melalui pengumpulan dan pengiriman data secara terus-menerus. Pada level jaringan, aktivitas adware dapat dikenali melalui pola permintaan (*request*) yang berulang, komunikasi intensif menuju alamat tujuan tertentu, koneksi terenkripsi yang berlangsung secara persisten, serta ukuran paket yang relatif konsisten dalam interval waktu tertentu. Karakteristik tersebut merupakan indikator penting dalam analisis anomali lalu lintas jaringan karena mencerminkan perilaku komunikasi otomatis yang berbeda dari aktivitas pengguna normal. Pola komunikasi tersebut dapat diamati menggunakan *wireshark*, yaitu perangkat lunak analisis protokol jaringan yang mampu menangkap dan menampilkan informasi paket secara rinci, termasuk waktu pengiriman, alamat sumber dan tujuan, jenis protokol, panjang paket, serta informasi komunikasi lainnya untuk mendukung proses identifikasi dan investigasi trafik jaringan (Seraj et al., 2024).

Analisis manual terhadap paket jaringan memiliki keterbatasan ketika volume data yang diperiksa sangat besar. Dalam penelitian ini, dataset yang dianalisis terdiri atas 135.996 paket jaringan, sehingga proses pemeriksaan setiap paket secara individual menjadi tidak efisien dan berpotensi mengurangi ketelitian analisis. Deteksi anomali pada trafik jaringan memerlukan pendekatan analitik yang mampu mengelompokkan data berdasarkan karakteristik komunikasi, mengidentifikasi pola perilaku normal, serta menemukan lalu lintas yang menyimpang dari pola umum

secara otomatis. Pendekatan berbasis analisis fitur dan pembelajaran data terbukti efektif dalam meningkatkan kemampuan identifikasi anomali pada lingkungan jaringan dengan volume trafik yang tinggi (essa tayeb et al., 2025).

Dalam bidang keamanan jaringan, pendekatan unsupervised menjadi penting karena sebagian besar data trafik tidak memiliki label normal maupun berbahaya. Kondisi tersebut menyebabkan metode klasifikasi konvensional kurang efektif ketika menghadapi pola serangan baru (*zero day attack*) atau aktivitas yang belum pernah dikenali sebelumnya. K-Means mampu mengidentifikasi kelompok komunikasi yang memiliki karakteristik serupa tanpa memerlukan data pelatihan berlabel, sehingga sesuai digunakan sebagai tahap eksplorasi dalam proses deteksi anomali trafik jaringan. Penelitian terbaru menunjukkan bahwa pengelompokan berbasis K-Means efektif dalam memisahkan pola trafik normal dan trafik yang memiliki karakteristik menyimpang berdasarkan atribut komunikasi jaringan (Krajewska & Niewiadomska-Szynkiewicz, 2024).

Pada penelitian ini, proses klasterisasi dilakukan menggunakan sejumlah fitur yang diekstraksi dari hasil packet capture *wireshark,* yaitu *packet count, total length, average length, destination frequency, source frequency, tcp warning count, query like count, packets persecond, dan bytes persecond.* Kombinasi fitur tersebut merepresentasikan intensitas komunikasi, ukuran paket, frekuensi tujuan, distribusi sumber, serta laju pertukaran data yang menjadi indikator penting dalam menggambarkan perilaku suatu aliran trafik. Dengan memanfaatkan fitur-fitur tersebut, K-Means dapat membentuk kelompok komunikasi yang memiliki pola statistik serupa sehingga mempermudah identifikasi trafik dominan maupun trafik yang berbeda secara signifikan dari mayoritas komunikasi jaringan (Budiati et al., 2022).

Aktivitas adware umumnya tidak langsung merusak sistem operasi, namun menghasilkan komunikasi jaringan yang berlangsung secara terus-menerus menuju server iklan (*advertising server*) atau layanan pelacakan (*tracking server*). Pola tersebut ditandai oleh frekuensi koneksi yang tinggi, pengiriman paket berukuran relatif kecil namun berulang, peningkatan jumlah permintaan DNS atau http/https, serta komunikasi periodik dengan tujuan yang sama. Karakteristik ini menyebabkan trafik adware memiliki distribusi fitur yang berbeda dibandingkan aktivitas pengguna normal, sehingga berpotensi membentuk klaster tersendiri pada hasil pengelompokan. Dengan demikian, hasil klasterisasi tidak hanya memberikan representasi visual terhadap struktur trafik jaringan, tetapi juga membantu peneliti menentukan kelompok yang layak dianalisis lebih lanjut sebagai kandidat trafik mencurigakan (Zhukabayeva et al., 2025).

Secara konseptual, penggunaan K-Means pada penelitian ini berfungsi sebagai metode analisis eksploratif untuk menemukan pola komunikasi tersembunyi (hidden traffic patterns) pada dataset yang terdiri dari 135.996 paket jaringan. Pendekatan ini diharapkan mampu mengurangi ketergantungan terhadap pemeriksaan manual, meningkatkan efisiensi analisis data dalam jumlah besar, serta memberikan dasar objektif dalam proses identifikasi anomali yang berkaitan dengan aktivitas adware pada lingkungan jaringan.

Penelitian ini bertujuan untuk menganalisis pola trafik jaringan yang diperoleh dari hasil packet capture Wireshark menggunakan metode *K-Means Clustering* sebagai pendekatan *unsupervised learning*. Analisis dilakukan untuk mengidentifikasi karakteristik komunikasi jaringan berdasarkan kemiripan fitur, sehingga dapat mengungkap pola trafik dominan maupun kelompok trafik yang berpotensi

merepresentasikan aktivitas adware tanpa memerlukan data berlabel. Pendekatan ini relevan untuk mendukung deteksi dini terhadap anomali jaringan yang sulit dikenali melalui pemeriksaan manual, terutama pada dataset dengan jumlah paket yang besar (Krajewska & Niewiadomska-Szynkiewicz, 2024).

Berbeda dengan penelitian-penelitian sebelumnya yang umumnya menerapkan supervised learning dan bergantung pada ketersediaan dataset berlabel, penelitian ini mengusulkan pendekatan deteksi awal (*early detection*) menggunakan *K-Means Clustering*. Melalui pengelompokan trafik berdasarkan fitur yang diekstraksi dari hasil packet capture Wireshark, metode ini mampu menemukan pola komunikasi tersembunyi (*hidden traffic patterns*) serta mengidentifikasi kandidat trafik mencurigakan tanpa proses pelabelan sebelumnya. Dengan demikian, penelitian ini diharapkan menjadi referensi awal bagi pengembangan sistem penyaringan dan pemantauan trafik jaringan yang lebih adaptif terhadap kemunculan aktivitas adware maupun anomali jaringan lainnya (Quirumbay Yagual et al., 2025).

METODOLOGI

Penelitian ini menggunakan pendekatan kuantitatif yang berfokus pada analisis trafik jaringan dan penerapan *machine learning* tanpa pengawasan (*unsupervised learning*). Metode utama yang digunakan adalah *K-Means Clustering*, yaitu algoritma clustering yang bertujuan mengelompokkan data berdasarkan tingkat kemiripan karakteristiknya. Dalam penelitian ini, *K-Means* digunakan untuk mengidentifikasi dan mengelompokkan pola komunikasi jaringan berdasarkan karakteristik flow komunikasi yang diperoleh dari hasil *capture* trafik menggunakan Wireshark. Pendekatan ini memungkinkan pola trafik dianalisis secara objektif tanpa memerlukan label kelas pada setiap data. Tahapan penelitian secara keseluruhan meliputi pengumpulan dataset, preprocessing data, pembentukan flow komunikasi, feature engineering, normalisasi data, penentuan jumlah klaster optimal, proses clustering, evaluasi hasil klaster, dan interpretasi karakteristik trafik pada setiap klaster. Setiap tahapan dilakukan secara berurutan untuk memastikan bahwa data yang digunakan dalam proses clustering telah memenuhi kebutuhan analisis dan mampu merepresentasikan karakteristik komunikasi jaringan secara optimal.

Dataset

Dataset penelitian diperoleh dari tujuh file hasil ekspor Wireshark dalam format spreadsheet (.csv) dengan jumlah paket jaringan yang berbeda pada setiap dataset. Seluruh dataset merepresentasikan hasil capture trafik jaringan selama aktivitas browsing, streaming, komunikasi aplikasi berbasis web , serta penggunaan layanan internet berlangsung.

Tabel 1. Dataset

No	Nama
1.	*No*
2.	*Time*
3.	*Source*
4.	*Destination*
5.	*Protocol*
6.	*Length*
7.	*Info*

Ketujuh dataset kemudian digabungkan pada tahap preprocessing untuk membentuk dataset utama yang digunakan dalam proses analisis clustering. Total keseluruhan data yang dianalisis mencapai 135.996 paket jaringan.

Tabel 2. Ringkasan Dataset Penelitian

DATASET	**JUMLAH PAKET**	**KETERANGAN**
DATASET 1	XXX	CAPTURE WIRESHARK
DATASET 2	XXX	CAPTURE WIRESHARK
DATASET 3	XXX	CAPTURE WIRESHARK
DATASET 4	XXX	CAPTURE WIRESHARK
DATASET 5	XXX	CAPTURE WIRESHARK
DATASET 6	XXX	CAPTURE WIRESHARK
DATASET 7	XXX	CAPTURE WIRESHARK
TOTAL		**135.996**

Setelah seluruh dataset digabungkan, diperoleh ringkasan data seperti pada Tabel 2.

KETERANGAN	**NILAI**
JUMLAH FILE DATASET	7
JUMLAH PAKET AWAL	135.996
JUMLAH ATRIBUT AWAL	7 ATRIBUT
RENTANG WAKTU CAPTURE	14.654,97 DETIK
JUMLAH ALIRAN KOMUNIKASI HASIL AGREGASI	1.839
JUMLAH DUPLIKASI BARIS TERDETEKSI	15.764

Preprocessing Data

Tahap preprocessing dilakukan untuk memastikan data hasil ekspor Wireshark dapat digunakan pada proses machine learning. Data mentah hasil capture masih mengandung format atribut yang tidak konsisten, data kosong, serta paket jaringan yang belum tersusun dalam bentuk flow komunikasi. Tahapan preprocessing pada penelitian ini meliputi:

1. Penggabungan Dataset. Seluruh file hasil ekspor Wireshark digabungkan menjadi satu dataset utama menggunakan library pandas pada Python.
2. Pembersihan Data. Proses cleaning dilakukan dengan: menghapus atribut yang tidak digunakan, memperbaiki format data, menyesuaikan tipe data numerik, mengisi nilai kosong pada atribut Info, serta mendeteksi data duplikat.
3. Transformasi Tipe Data. diubah ke format numerik agar dapat digunakan pada proses perhitungan statistik dan clustering.
4. Penanganan Nilai Kosong. Nilai kosong pada atribut Info diisi menggunakan string kosong untuk menghindari error pada proses feature engineering.
5. Deteksi Duplikasi. Data duplikat dideteksi untuk mengetahui kemungkinan paket berulang pada hasil capture jaringan.
6. Pembentukan Flow Komunikasi. Data paket jaringan diagregasi menjadi flow komunikasi berdasarkan kombinasi:

$$\text{Flow} = (\textit{Source, Destination, Protocol})$$

Pendekatan flow-based digunakan karena lebih mampu merepresentasikan pola komunikasi jaringan dibandingkan analisis paket tunggal.

Tabel 3. Atribut awal hasil *capture Wireshark*

ATRIBUT	**KETERANGAN**
NO	NOMOR URUT PAKET HASIL CAPTURE
TIME	WAKTU KEMUNCULAN PAKET DALAM SATUAN DETIK
SOURCE	ALAMAT SUMBER PAKET
DESTINATION	ALAMAT TUJUAN PAKET
PROTOCOL	PROTOKOL KOMUNIKASI
LENGTH	PANJANG PAKET DALAM BYTE
INFO	INFORMASI TAMBAHAN DARI WIRESHARK

Feature engineering dilakukan untuk membentuk atribut numerik yang dapat merepresentasikan perilaku komunikasi jaringan. Atribut mentah Wireshark belum secara langsung menunjukkan karakteristik trafik sehingga diperlukan pembentukan fitur turunan berbasis flow komunikasi.

HASIL DAN PEMBAHASAN

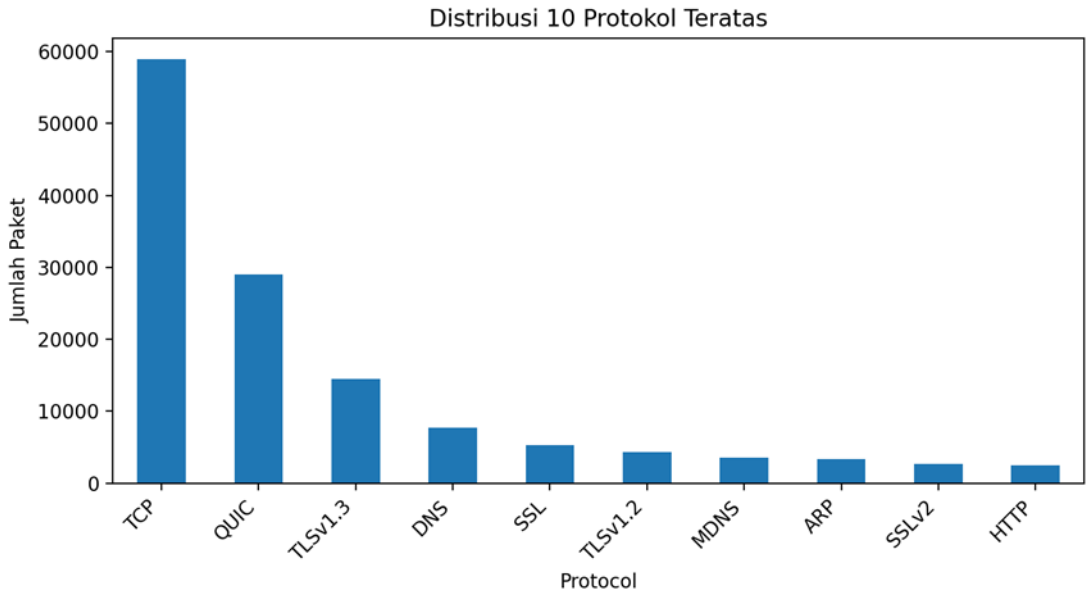
Distribusi Protokol Trafik Jaringan

Analisis awal dilakukan untuk mengetahui distribusi protokol yang muncul pada dataset hasil capture Wireshark. Distribusi protokol digunakan untuk menggambarkan karakteristik umum komunikasi jaringan selama proses pengambilan data berlangsung. Berdasarkan hasil analisis, trafik jaringan didominasi oleh protokol TCP, QUIC, TLSv1.3, DNS, dan SSL. Dominasi protokol berbasis web dan komunikasi terenkripsi menunjukkan bahwa sebagian besar aktivitas jaringan berasal dari layanan internet modern seperti browsing, streaming, media sosial, dan komunikasi aplikasi berbasis cloud.

Tabel 5 Distribusi Sepuluh Protokol Paling Dominan Pada Dataset Penelitian.

Protocol	Jumlah Paket
TCP	58.947
QUIC	28.987
TLSv1.3	14.482
DNS	7.740
SSL	5.257
TLSv1.2	4.272
MDNS	3.524
ARP	3.372
SSLv2	2.643
HTTP	2.483

Dominasi protokol TCP menunjukkan bahwa sebagian besar komunikasi masih menggunakan mekanisme transport berbasis koneksi. Sementara itu, tingginya trafik QUIC dan TLSv1.3 menunjukkan meningkatnya penggunaan komunikasi terenkripsi pada aplikasi modern. Kondisi tersebut menyebabkan inspeksi payload secara langsung menjadi lebih sulit sehingga pendekatan berbasis flow analysis menjadi lebih relevan dibandingkan analisis isi paket. Selain itu, kemunculan DNS dalam jumlah cukup tinggi menunjukkan aktivitas resolusi domain yang intensif selama komunikasi berlangsung. Dalam konteks analisis trafik jaringan, komunikasi DNS yang berulang dapat digunakan sebagai salah satu indikator aktivitas komunikasi latar belakang (background communication) atau koneksi otomatis aplikasi tertentu. Keberadaan protokol MDNS dan ARP menunjukkan adanya komunikasi lokal antarperangkat dalam jaringan yang umumnya berkaitan dengan proses discovery perangkat dan layanan jaringan lokal. Trafik tersebut cenderung merepresentasikan komunikasi normal dan tidak secara langsung menunjukkan indikasi aktivitas mencurigakan.



Gambar 1. Distribusi 10 protokol teratas pada dataset.

Hasil Penentuan Jumlah Klaster

Penentuan jumlah klaster dilakukan menggunakan Elbow Method, Silhouette Score, dan Davies-Bouldin Index. Pengujian dilakukan pada rentang jumlah klaster K=2hingga K=10.

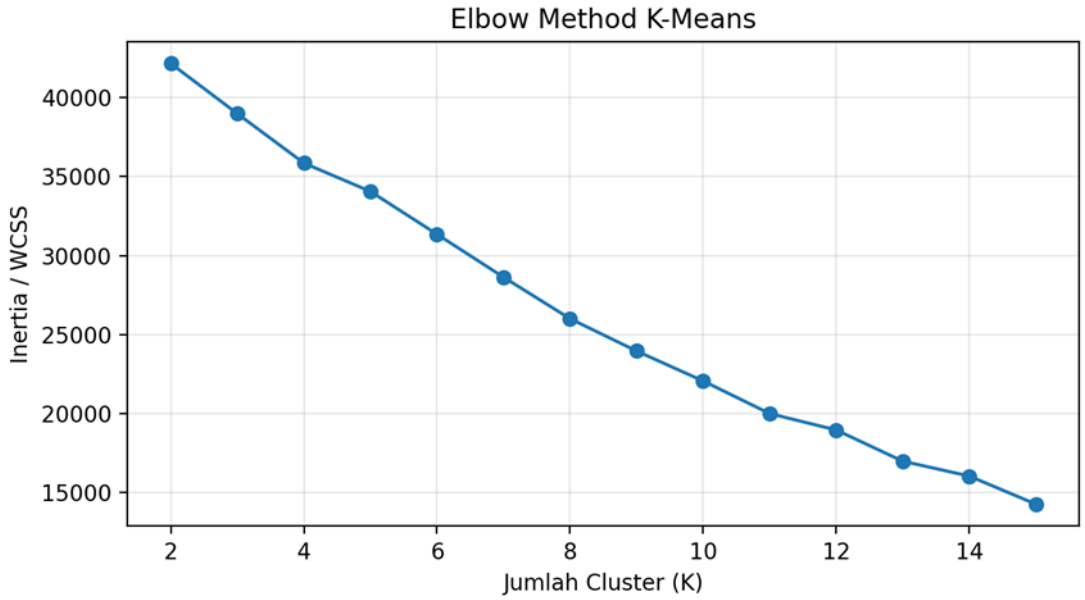
Tabel 6. menunjukkan hasil evaluasi jumlah klaster.

K	Inertia	Silhouette Score	Davies-Bouldin Index
2	42.160,67	0,169	2,633
3	38.986,24	0,172	2,307
4	35.858,25	0,221	1,853
5	34.063,41	0,254	1,764
6	31.356,02	0,237	1,645
7	28.638,91	0,297	1,389
8	26.012,04	0,286	1,308
9	23.963,63	0,326	1,208
10	22.070,24	0,351	1,191

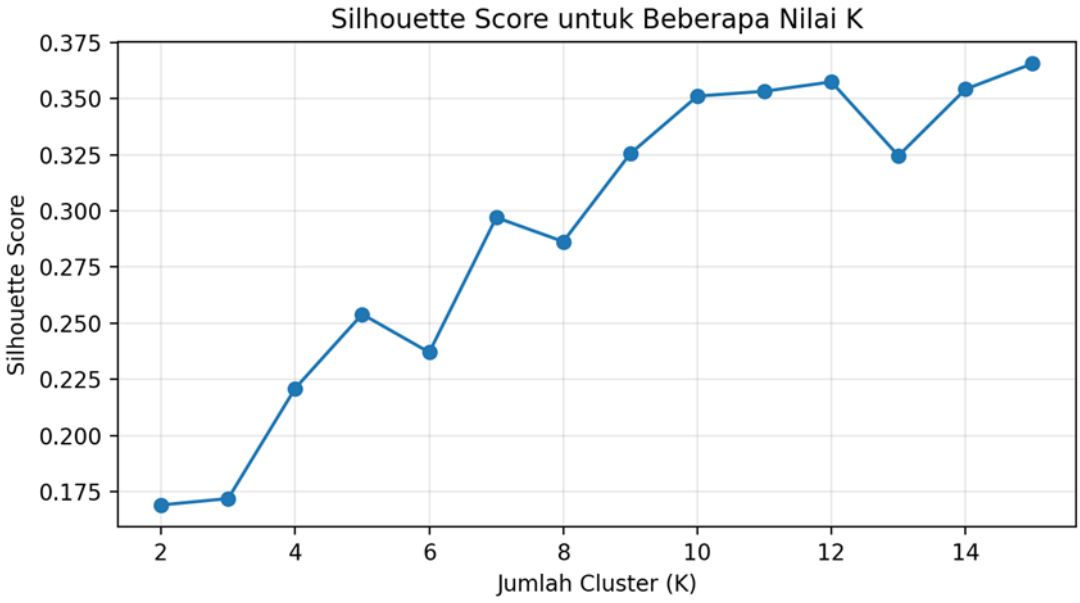
Berdasarkan hasil pengujian, peningkatan jumlah klaster menyebabkan nilai inertia dan Davies-Bouldin Index mengalami penurunan. Kondisi tersebut menunjukkan bahwa cluster menjadi semakin kompak dan memiliki tingkat pemisahan yang lebih baik. Nilai Silhouette Score tertinggi diperoleh pada K=10dengan nilai sebesar 0,351. Namun, penelitian ini menggunakan K=7karena

mempertimbangkan keseimbangan antara kualitas pemisahan klaster dan kemudahan interpretasi pola komunikasi jaringan. Penggunaan jumlah klaster yang terlalu besar berpotensi menyebabkan fragmentasi trafik sehingga interpretasi cluster menjadi lebih sulit dilakukan.

Nilai Silhouette Score sebesar 0,297 menunjukkan bahwa pemisahan klaster berada pada kategori sedang. Kondisi tersebut umum ditemukan pada trafik jaringan nyata karena komunikasi normal, background traffic, dan komunikasi aplikasi modern sering memiliki pola yang saling tumpang tindih, terutama ketika menggunakan protokol dan layanan internet yang serupa. Sementara itu, nilai Davies-Bouldin Index sebesar 1,389 menunjukkan bahwa cluster yang terbentuk memiliki tingkat keterpisahan yang cukup baik meskipun masih terdapat overlap antar kelompok komunikasi.



Gambar 2. Grafik Elbow Method.



Gambar 3. Grafik Silhouette Score.

Hasil K-Means Clustering

Proses K-Means pada 1.839 aliran komunikasi menghasilkan tujuh klaster. Setiap klaster menggambarkan pola trafik yang berbeda berdasarkan intensitas paket, ukuran data, dominasi tujuan komunikasi, frekuensi sumber, dan peringatan TCP.

Tabel 7. Ringkasan hasil klaster

Cluster	Jumlah Flow	Total Paket	Persentase Paket	Rata-rata Paket/Flow	Rata-rata Length	TCP Warning
0	17	47.913	35,23%	2.818,41	1.271,68	15.904
1	606	35.769	26,30%	59,02	230,54	3.943
2	8	27	0,02%	3,38	1.833,82	16
3	64	3.524	2,59%	55,06	184,71	0
4	504	17.484	12,86%	34,69	586,39	4.992
5	444	10.193	7,50%	22,96	1.029,07	1.422
6	196	21.086	15,50%	107,58	555,57	0

Berdasarkan hasil clustering, Cluster 0 menjadi kelompok dengan intensitas komunikasi tertinggi. Meskipun hanya terdiri dari 17 flow, cluster ini mencakup 35,23% dari total paket jaringan. Tingginya rata-rata paket per flow dan jumlah TCP warning menunjukkan adanya komunikasi yang berlangsung sangat aktif pada sejumlah koneksi tertentu. Cluster 1 memiliki jumlah flow terbesar, yaitu 606 flow, dengan rata-rata paket relatif rendah. Karakteristik tersebut menunjukkan pola komunikasi yang lebih mendekati trafik umum pengguna seperti browsing dan akses layanan web biasa. Cluster 4 menunjukkan kombinasi trafik aktif dan jumlah TCP warning yang cukup tinggi. Kondisi tersebut mengindikasikan adanya komunikasi yang relatif intensif dan berlangsung secara berulang pada beberapa koneksi tertentu. Sementara itu, Cluster 6 memiliki dominasi trafik QUIC dengan jumlah paket cukup tinggi namun tanpa warning TCP. Karakteristik tersebut cenderung merepresentasikan komunikasi aplikasi modern berbasis HTTP/3 atau layanan streaming.

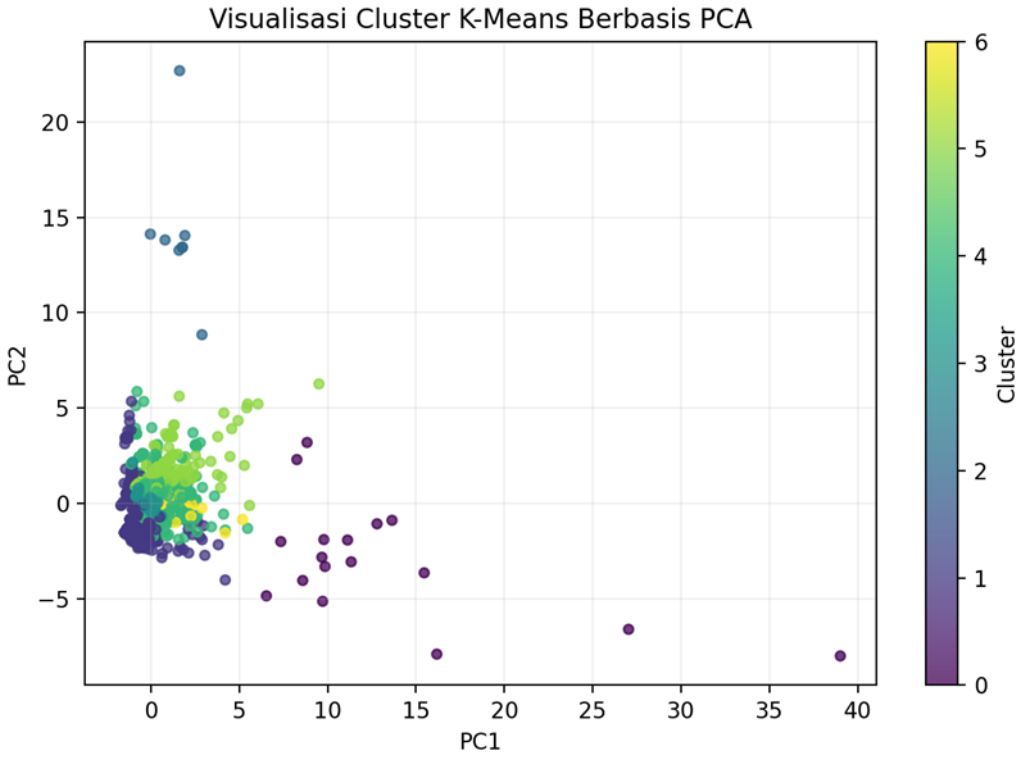
Cluster 2 memiliki jumlah flow dan paket yang sangat kecil tetapi memiliki ukuran paket rata-rata paling tinggi. Cluster tersebut lebih tepat diperlakukan sebagai outlier cluster yang memerlukan verifikasi manual dibandingkan langsung diasumsikan sebagai komunikasi mencurigakan.

Visualisasi Klaster Menggunakan PCA

Visualisasi cluster dilakukan menggunakan Principal Component Analysis (PCA) untuk mereduksi dimensi fitur numerik menjadi dua komponen utama. Tujuan visualisasi ini adalah untuk membantu pengamatan distribusi dan pemisahan cluster secara visual. Hasil visualisasi menunjukkan bahwa sebagian besar cluster berada

pada area yang saling berdekatan. Kondisi tersebut menunjukkan adanya overlap karakteristik antar flow komunikasi. Hal ini umum ditemukan pada trafik jaringan nyata karena berbagai aplikasi modern menggunakan protokol dan pola komunikasi yang serupa. Meskipun demikian, beberapa cluster terlihat memiliki distribusi yang lebih terpisah dibanding cluster lainnya. Cluster dengan distribusi paling jauh umumnya memiliki karakteristik komunikasi yang lebih dominan, seperti jumlah paket tinggi, ukuran paket besar, atau intensitas komunikasi yang berbeda secara signifikan dibanding flow lainnya.

Visualisasi PCA juga menunjukkan bahwa pendekatan clustering masih mampu membedakan kelompok komunikasi tertentu meskipun data jaringan memiliki tingkat kompleksitas yang tinggi.



Gambar 4. Visualisasi klaster menggunakan reduksi dimensi PCA.

Interpretasi Klaster

Interpretasi klaster dilakukan berdasarkan karakteristik numerik masing-masing kelompok komunikasi.

Cluster 0

Cluster 0 merupakan cluster dengan intensitas komunikasi tertinggi. Jumlah flow yang sedikit tetapi total paket sangat besar menunjukkan adanya komunikasi yang berlangsung terus-menerus pada koneksi tertentu. Nilai TCP warning yang tinggi mengindikasikan kemungkinan retransmission, duplicate ACK, out-of-order packet, atau komunikasi tidak stabil. Pada konteks penelitian ini, pola tersebut relevan untuk diperiksa lebih lanjut karena menunjukkan komunikasi dominan yang berbeda dibandingkan cluster lainnya.

Cluster 1

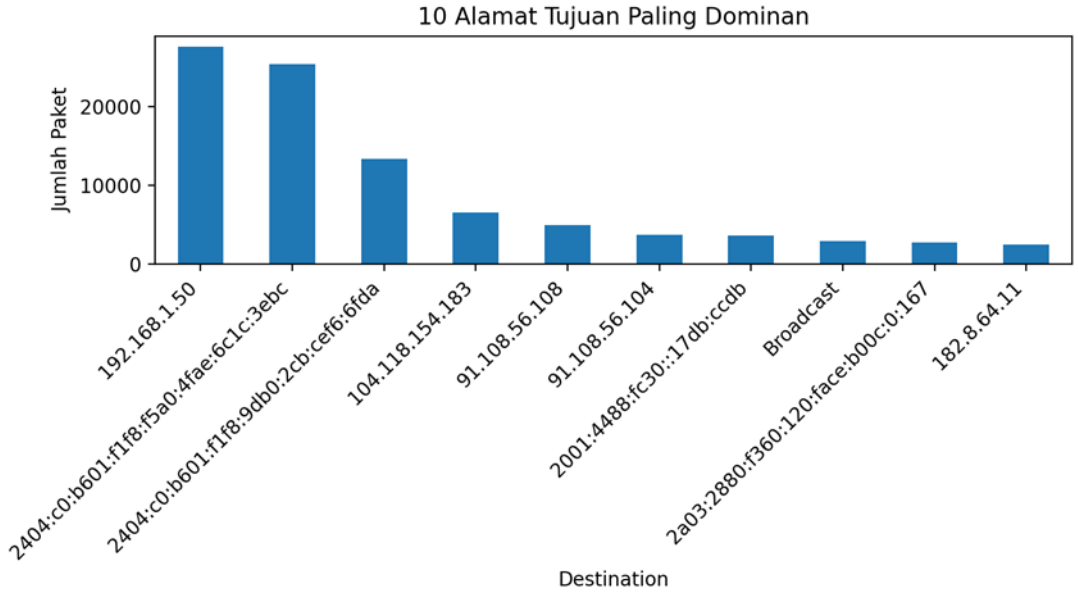
Cluster 1 memiliki jumlah flow terbesar dengan rata-rata paket relatif kecil. Pola tersebut lebih mendekati karakteristik komunikasi normal pengguna seperti browsing dan akses layanan web umum.

Cluster 4
Cluster 4 menunjukkan kombinasi trafik aktif dan jumlah TCP warning yang cukup tinggi. Cluster ini dapat dikategorikan sebagai kelompok komunikasi yang perlu dianalisis lebih lanjut karena menunjukkan pola koneksi berulang dan intensitas komunikasi yang cukup dominan.

Cluster 6
Cluster 6 didominasi komunikasi berbasis QUIC dengan volume paket cukup besar. Trafik pada cluster ini kemungkinan berasal dari layanan modern seperti video streaming, media sosial, atau layanan berbasis HTTP/3.

Cluster 2, 3, dan 5
Ketiga cluster tersebut memiliki karakteristik yang lebih kecil dan cenderung merepresentasikan komunikasi spesifik tertentu. Cluster 3 didominasi trafik lokal seperti MDNS, sedangkan Cluster 5 menunjukkan komunikasi terenkripsi dengan ukuran paket relatif besar.

Pendekatan tersebut penting untuk menghindari over-interpretation terhadap hasil clustering.



Gambar 5. Sepuluh alamat tujuan paling dominan pada dataset.

Pembahasan

Hasil penelitian menunjukkan bahwa pendekatan flow-based clustering mampu membantu proses eksplorasi trafik jaringan dalam jumlah besar secara lebih terstruktur dibandingkan analisis manual paket jaringan. Pendekatan clustering memberikan keuntungan karena mampu mengelompokkan komunikasi berdasarkan pola karakteristik tanpa memerlukan label serangan. Hal tersebut menjadi penting pada kondisi nyata ketika dataset jaringan belum memiliki ground truth yang jelas. Dibandingkan pemeriksaan manual menggunakan Wireshark, metode ini mampu

membantu administrator jaringan dalam memprioritaskan kelompok komunikasi yang memiliki pola paling dominan atau paling berbeda dibanding trafik umum. Dengan demikian, proses investigasi jaringan dapat dilakukan secara lebih efisien.

Namun demikian, penelitian ini masih memiliki beberapa keterbatasan. Pertama, dataset penelitian hanya terdiri dari tujuh file capture Wireshark sehingga variasi trafik jaringan masih terbatas. Kedua, penelitian belum menggunakan label validasi yang menyatakan komunikasi normal atau adware secara pasti. Ketiga, sebagian besar trafik menggunakan protokol terenkripsi sehingga inspeksi payload tidak dapat dilakukan secara langsung. Meskipun memiliki keterbatasan, hasil penelitian menunjukkan bahwa pendekatan clustering berbasis flow analysis tetap relevan digunakan sebagai metode exploratory analysis pada trafik jaringan modern yang kompleks dan terenkripsi.

SIMPULAN

Penelitian ini berhasil menerapkan metode K-Means Clustering pada data trafik jaringan berbasis hasil capture Wireshark untuk membantu proses analisis dan identifikasi indikasi aktivitas adware. Dataset penelitian terdiri dari tujuh file hasil ekspor Wireshark dengan total 135.996 paket jaringan yang diproses melalui tahapan preprocessing, agregasi flow komunikasi, feature engineering, dan normalisasi data.

Proses agregasi menghasilkan 1.839 flow komunikasi yang dianalisis menggunakan algoritma K-Means Clustering. Penentuan jumlah klaster dilakukan menggunakan Elbow Method, Silhouette Score, dan Davies-Bouldin Index. Berdasarkan hasil evaluasi, penelitian menggunakan K=7karena memberikan keseimbangan antara kualitas pemisahan klaster dan kemudahan interpretasi pola komunikasi jaringan. Hasil evaluasi menghasilkan nilai Silhouette Score sebesar 0,297 dan Davies-Bouldin Index sebesar 1,389.

Hasil clustering menunjukkan bahwa setiap klaster memiliki karakteristik trafik yang berbeda berdasarkan jumlah paket, ukuran komunikasi, intensitas trafik, dominasi alamat tujuan, serta jumlah warning TCP. Cluster 0, Cluster 4, dan Cluster 6 menjadi kelompok yang paling relevan untuk diperiksa lebih lanjut karena menunjukkan komunikasi dominan, koneksi berulang, intensitas trafik tinggi, serta pola komunikasi yang berbeda dibandingkan trafik umum. Pendekatan clustering berbasis flow analysis mampu membantu proses eksplorasi dan penyaringan awal trafik jaringan secara lebih terstruktur dibandingkan pemeriksaan manual paket jaringan satu per satu. Meskipun demikian, hasil clustering pada penelitian ini belum dapat digunakan untuk menyimpulkan aktivitas adware secara pasti karena metode yang digunakan bersifat unsupervised learning dan tidak menggunakan label *ground truth.*

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada jurnal mirai yang memproses arikel ini. Penulis juga menyampaikan terima kasih kepada Universitas Dinamika atas arahan, masukan, dan dukungan selama proses penelitian dan penyusunan artikel ini. Ucapan terima kasih juga disampaikan kepada seluruh pihak yang telah membantu dalam proses pengumpulan dataset, pengolahan data trafik jaringan, serta analisis menggunakan metode K-Means Clustering. Dukungan dan kontribusi berbagai pihak telah membantu penelitian ini dapat diselesaikan dengan baik.

Referensi

Budiati, H., Wijaya, A. B. M., Zebua, B. S. V., Jatmika, & Sumihar, Y. P. (2022). Implementation of K-Means Clustering Method for Network Traffic Anomaly Detection. *Jurnal Mantik*, *6*(3).

essa tayeb, mesud, Chebbi, T., Ali Toumi, J., Badawi, A., & LOUAIL, B. (2025). The impact of Ad overloads perception in social media on Ad avoidance behavior: the mediating effect of social media fatigue and goal impediment. *Management*, *28*(2). https://doi.org/10.58691/man/197329

Krajewska, A., & Niewiadomska-Szynkiewicz, E. (2024). Clustering Network Traffic Using Semi-Supervised Learning. *Electronics (Switzerland)*, *13*(14). https://doi.org/10.3390/electronics13142769

Quirumbay Yagual, D., Fernández Iglesias, D., & Nóvoa, F. J. (2025). A Hybrid Deep Learning-Based Architecture for Network Traffic Anomaly Detection via EFMS-Enhanced KMeans Clustering and CNN-GRU Models. *Applied Sciences (Switzerland)*, *15*(20). https://doi.org/10.3390/app152010889

Sadeghpour, S., & Vlajic, N. (2021). Ads and Fraud: A Comprehensive Survey of Fraud in Online Advertising. *Journal of Cybersecurity and Privacy*, *1*(4). https://doi.org/10.3390/jcp1040039

Seraj, S., Pavlidis, M., Trovati, M., & Polatidis, N. (2024). MadDroid: malicious adware detection in Android using deep learning. *Journal of Cyber Security Technology*, *8*(3). https://doi.org/10.1080/23742917.2023.2247197

Zhukabayeva, T., Ahmad, Z., Adamova, A., Karabayev, N., & Abdildayeva, A. (2025). An Edge-Computing-Based Integrated Framework for Network Traffic Analysis and Intrusion Detection to Enhance Cyber–Physical System Security in Industrial IoT. *Sensors*, *25*(8). https://doi.org/10.3390/s25082395