P-ISSN : 2580-4308
DOI https://doi.org/10.37817/ikraith-teknologi.v10i2 E-ISSN : 2654-8046

Deteksi Anomali Lalu Lintas Jaringan Menggunakan Algoritma K-Means dengan Optimasi Jumlah Klaster Berbasis Elbow Method dan Silhouette Score

(Network Traffic Anomaly Detection Using the K-Means Algorithm with Cluster Number Optimization Based on the Elbow Method and Silhouette Score)

[1]Alvin Rainaldy Hakim
[1]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muria Kudus
[1]Kudus, Jawa Tengah
E-mail: [1]alvin.rainaldy@umk.ac.id

ABSTRAK

Peningkatan volume lalu lintas jaringan komputer menyebabkan deteksi aktivitas anomali secara manual menjadi tidak efisien, sementara pendekatan berbasis aturan (*signature-based*) tidak mampu mengenali pola serangan baru. Penelitian ini bertujuan mengelompokkan pola lalu lintas jaringan menggunakan algoritma K-Means guna mengidentifikasi trafik anomali tanpa memanfaatkan label selama proses pelatihan. *Dataset* yang digunakan terdiri atas 10.005 rekaman koneksi jaringan dengan atribut durasi, jumlah paket, *bytes* terkirim, *bytes* diterima, *bytes* per paket, dan protokol. Tahapan *preprocessing* meliputi pemeriksaan *missing value* dan duplikasi, seleksi fitur, *one-hot encoding* pada atribut protokol, serta standarisasi fitur menggunakan *StandardScaler*. Data kemudian direduksi menjadi 5.000 sampel melalui *random sampling* untuk efisiensi komputasi. Penentuan jumlah klaster optimal dilakukan pada rentang K=1 hingga K=5 menggunakan *Elbow Method* dan *Silhouette Score*. Hasil pengujian menunjukkan nilai K optimal adalah 2 dengan *Silhouette Score* sebesar 0,6347. Klaster pertama berisi 4.807 data yang merepresentasikan trafik normal, sedangkan klaster kedua berisi 193 data dengan karakteristik durasi koneksi dan volume *bytes* terkirim yang jauh lebih tinggi. Validasi terhadap label asli menunjukkan seluruh anggota klaster kedua (100%) merupakan trafik anomali dengan presisi 100% dan *recall* 18,7%. Hasil ini membuktikan bahwa K-Means dengan optimasi jumlah klaster mampu mengisolasi pola trafik anomali bervolume tinggi secara *unsupervised*.

Kata kunci : *anomali, clustering, Elbow Method, K-Means, lalu lintas jaringan, Silhouette Score*

ABSTRACT

The increasing volume of computer network traffic makes manual detection of anomalous activity inefficient, while signature-based approaches are unable to recognize new attack patterns. This study aims to cluster network traffic patterns using the K-Means algorithm to identify anomalous traffic without utilizing labels during training. The dataset consists of 10,005 network connection records with attributes of duration, packet count, bytes sent, bytes received, bytes per packet, and protocol. Preprocessing stages include missing value and duplication checks, feature selection, one-hot encoding of the protocol attribute, and feature standardization using StandardScaler. The data was then reduced to 5,000 samples through random sampling for computational efficiency. The optimal number of clusters was determined in the range K=1 to K=5 using the Elbow Method and Silhouette Score. The results show that the optimal K value is 2 with a Silhouette Score of 0.6347. The first cluster contains 4,807 records representing normal traffic, while the second cluster contains 193 records characterized by much higher connection duration and volume of bytes sent. Validation against the original labels shows that all members of the second cluster (100%) are anomalous traffic, with 100% precision and 18.7% recall. These results

P-ISSN : 2580-4308
E-ISSN : 2654-8046 DOI https://doi.org/10.37817/ikraith-teknologi.v10i2

prove that K-Means with cluster number optimization is able to isolate high-volume anomalous traffic patterns in an unsupervised manner.

Keywords : *anomaly, clustering, Elbow Method, K-Means, network traffic, Silhouette Score*

1. PENDAHULUAN

Perkembangan teknologi informasi mendorong pertumbuhan lalu lintas jaringan komputer secara masif, baik pada institusi pemerintahan, perguruan tinggi, maupun industri. Di balik pertumbuhan tersebut, ancaman keamanan siber seperti *Distributed Denial of Service* (DDoS), eksfiltrasi data, pemindaian *port*, dan aktivitas *botnet* turut meningkat dan semakin sulit dikenali secara kasat mata. Aktivitas berbahaya tersebut umumnya meninggalkan jejak berupa pola trafik yang menyimpang dari perilaku normal, misalnya durasi koneksi yang tidak wajar, volume data terkirim yang sangat besar, atau rasio *bytes* per paket yang tidak lazim. Oleh karena itu, kemampuan mendeteksi anomali pada lalu lintas jaringan menjadi salah satu komponen penting dalam sistem keamanan jaringan modern [1], [2].

Pendekatan deteksi intrusi secara umum terbagi menjadi dua, yaitu berbasis aturan (*signature-based*) dan berbasis anomali (*anomaly-based*). Pendekatan berbasis aturan bekerja dengan mencocokkan trafik terhadap pola serangan yang telah tercatat sebelumnya, sehingga tidak mampu mendeteksi tipe serangan baru yang belum pernah terdokumentasi (*zero-day*) dan rentan menghasilkan *false-negative* [3]. Sebaliknya, pendekatan berbasis anomali membangun model perilaku normal dari data historis dan menandai setiap penyimpangan sebagai potensi ancaman, sehingga lebih adaptif terhadap serangan baru [2]. Namun, pelabelan data trafik jaringan dalam jumlah besar untuk keperluan pembelajaran terarah (*supervised learning*) membutuhkan biaya dan waktu yang tinggi; pengalaman pembangunan *dataset benchmark* intrusi seperti KDD Cup 99/NSL-KDD dan CICIDS2017 menunjukkan kompleksitas proses pelabelan trafik yang melibatkan verifikasi manual berlapis [4], [5]. Kondisi ini menjadikan metode pembelajaran tak terarah (*unsupervised learning*) sebagai alternatif yang menarik karena tidak memerlukan label pada tahap pelatihan.

Algoritma K-Means merupakan salah satu metode *clustering* yang paling banyak digunakan karena sederhana, efisien secara komputasi, dan mampu menangani *dataset* berukuran besar [6], [7]. Popularitas dan relevansi K-Means bahkan tetap bertahan lebih dari lima dekade sejak pertama kali diperkenalkan, dengan berbagai varian dan perbaikan yang terus dikembangkan hingga era *big data* [8], [9]. K-Means mempartisi data ke dalam K kelompok berdasarkan kedekatan jarak terhadap pusat klaster (*centroid*), sehingga objek dalam satu klaster memiliki karakteristik yang serupa. Dalam konteks keamanan jaringan, klaster berukuran kecil yang terpisah jauh dari mayoritas data dapat diinterpretasikan sebagai kandidat trafik anomali; gagasan mendeteksi intrusi dari data tak berlabel menggunakan *clustering* telah dirintis sejak awal 2000-an [10]. Beberapa penelitian terdahulu telah menerapkan K-Means untuk analisis trafik jaringan. Penelitian pada jaringan LAN dengan 156 paket trafik menghasilkan presisi 100% namun *recall* hanya 61% akibat ketidakseimbangan distribusi data dan pemilihan nilai K yang belum optimal [11]. Penelitian lain mengoptimalkan K-Means dengan *Elbow Method* dan fitur berbasis aturan keamanan pada *log web server* dan memperoleh *Silhouette Score* hingga 0,9136 [12]. Kajian literatur sistematis juga menegaskan bahwa K-Means efektif digunakan untuk memprediksi kemunculan lalu lintas jaringan anomali [13].

P-ISSN : 2580-4308
DOI https://doi.org/10.37817/ikraith-teknologi.v10i2 E-ISSN : 2654-8046

Permasalahan utama dalam penerapan K-Means adalah penentuan jumlah klaster K yang harus ditetapkan di awal. Nilai K yang terlalu kecil berpotensi menggabungkan pola yang seharusnya terpisah, sedangkan K yang terlalu besar dapat memecah pola normal menjadi klaster-klaster yang tidak bermakna. *Elbow Method* dan *Silhouette Score* merupakan dua teknik validasi internal yang umum digunakan untuk menentukan K optimal; *Elbow Method* mengamati titik belok pada kurva *inertia* [14], sedangkan *Silhouette Score* mengukur kohesi dan separasi antarklaster [15]. Integrasi *Elbow Method* dengan K-Means terbukti mampu menghasilkan jumlah klaster yang konsisten pada volume data yang berbeda [16] sekaligus menurunkan *Sum of Squared Error* sehingga kualitas klaster meningkat [17]. Adapun kombinasi *Elbow Method* dengan evaluasi tambahan seperti *purity* juga telah divalidasi pada studi penentuan jumlah klaster [18]. Penggunaan kedua metode secara bersamaan memberikan dasar pemilihan K yang lebih kuat dibandingkan hanya mengandalkan satu metode.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk: (1) menerapkan tahapan *preprocessing* yang tepat pada data lalu lintas jaringan, meliputi seleksi fitur, *one-hot encoding*, dan standarisasi; (2) melakukan reduksi data menjadi 5.000 sampel untuk efisiensi komputasi; (3) menentukan jumlah klaster optimal pada rentang K=1 hingga K=5 menggunakan *Elbow Method* dan *Silhouette Score*; serta (4) mengevaluasi kemampuan klaster yang terbentuk dalam mengisolasi trafik anomali melalui validasi terhadap label asli yang tidak digunakan selama proses *clustering*. Kontribusi penelitian ini adalah tersedianya kerangka kerja deteksi anomali trafik jaringan berbasis *unsupervised learning* yang dapat direplikasi tanpa memerlukan data berlabel pada tahap pelatihan.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan alur: (1) pengumpulan data, (2) *preprocessing* data, (3) reduksi data, (4) penentuan K optimal, (5) pemodelan K-Means, dan (6) evaluasi serta visualisasi hasil. Seluruh eksperimen diimplementasikan menggunakan bahasa pemrograman Python 3.12 dengan pustaka *pandas*, *scikit-learn* versi 1.8 [20], *matplotlib*, dan *seaborn* pada lingkungan *Jupyter Notebook*. Nilai *random_state* ditetapkan sebesar 42 pada seluruh proses agar hasil dapat direproduksi.

2.1 Dataset

Dataset yang digunakan adalah data lalu lintas jaringan (*network traffic*) berformat CSV yang terdiri atas 10.005 rekaman koneksi dengan 12 atribut. Setiap rekaman merepresentasikan satu sesi koneksi jaringan yang dilengkapi label biner (0 = normal, 1 = anomali) sebagai *ground truth*. Rincian atribut *dataset* disajikan pada Tabel 1. Label hanya digunakan pada tahap validasi akhir dan sama sekali tidak dilibatkan dalam proses *clustering*, sehingga skenario penelitian tetap bersifat *unsupervised*.

Tabel 1. Deskripsi atribut dataset lalu lintas jaringan

Atribut	Tipe Data	Keterangan
time	Datetime	Waktu kejadian koneksi
source_ip_int	Integer	Alamat IP sumber
destination_ip_int	Integer	Alamat IP tujuan
source_port	Integer	Port sumber
destination_port	Integer	Port tujuan
protocol	Kategori	Jenis protokol (0–5)
duration	Float	Durasi koneksi (detik)
packet_count	Integer	Jumlah paket
bytes_sent	Integer	Total bytes dikirim

P-ISSN : 2580-4308
E-ISSN : 2654-8046 DOI https://doi.org/10.37817/ikraith-teknologi.v10i2

Atribut	Tipe Data	Keterangan
bytes_received	Integer	Total bytes diterima
bytes_per_packet	Float	Rasio bytes per paket
label	Biner (0/1)	Ground truth (validasi)

2.2 Preprocessing Data

Tahapan *preprocessing* diawali dengan pemeriksaan kualitas data. Hasil pemeriksaan menunjukkan *dataset* tidak mengandung *missing value* maupun baris duplikat, sehingga seluruh 10.005 rekaman dapat digunakan. Tahap berikutnya adalah seleksi fitur: atribut identitas yang tidak mencerminkan perilaku trafik, yaitu *time*, *source_ip_int*, *destination_ip_int*, *source_port*, dan *destination_port*, dikeluarkan dari himpunan fitur karena berpotensi menimbulkan bias jarak tanpa memberikan informasi pola perilaku. Atribut *label* juga dikeluarkan agar proses *clustering* murni bersifat *unsupervised*. Enam fitur yang dipertahankan adalah *protocol*, *duration*, *packet_count*, *bytes_sent*, *bytes_received*, dan *bytes_per_packet*.

Atribut *protocol* bertipe kategorikal dengan enam nilai (0–5), sehingga ditransformasikan menggunakan *one-hot encoding* menjadi enam kolom biner agar tidak terjadi asumsi urutan (ordinalitas) yang keliru pada perhitungan jarak *Euclidean*. Setelah *encoding*, jumlah fitur menjadi 11 kolom. Karena K-Means sensitif terhadap perbedaan skala antarfitur, seluruh fitur distandarisasi menggunakan *StandardScaler* sehingga setiap fitur memiliki rata-rata mendekati 0 dan simpangan baku mendekati 1, sebagaimana ditunjukkan pada Persamaan (1).

$$z = (x - \mu) / \sigma \quad (1)$$

dengan z merupakan nilai hasil standarisasi, x nilai asli fitur, μ rata-rata fitur, dan σ simpangan baku fitur.

2.3 Reduksi Data

Untuk meningkatkan efisiensi komputasi sekaligus menjaga keterbacaan visualisasi diagram titik, data hasil *preprocessing* direduksi dari 10.005 menjadi 5.000 sampel menggunakan teknik *random sampling* tanpa pengembalian dengan *random_state* tetap. *Random sampling* dipilih karena mempertahankan distribusi statistik data asli secara proporsional tanpa memihak kelas tertentu, sehingga struktur klaster alami pada data tetap terjaga.

2.4 Algoritma K-Means

K-Means mempartisi n objek data ke dalam K klaster dengan meminimalkan jumlah kuadrat jarak antarobjek terhadap *centroid* klasternya, yang dikenal sebagai *inertia* atau *Within-Cluster Sum of Squares* (WCSS) sebagaimana Persamaan (2). Algoritma bekerja secara iteratif: (1) inisialisasi K *centroid*, (2) penugasan setiap objek ke *centroid* terdekat berdasarkan jarak *Euclidean*, (3) pembaruan posisi *centroid* sebagai rata-rata anggota klaster, dan (4) pengulangan langkah 2–3 hingga posisi *centroid* konvergen [6], [7]. Pada penelitian ini digunakan inisialisasi *k-means++* [19] dengan *n_init* sebanyak 10 kali untuk mengurangi sensitivitas terhadap inisialisasi awal.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

2.5 Penentuan Nilai K Optimal

Jumlah klaster diuji pada rentang K=1 hingga K=5. *Elbow Method* menentukan K optimal dengan mengamati titik belok (*elbow*) pada kurva penurunan *inertia*; setelah titik tersebut, penambahan klaster tidak lagi memberikan penurunan *inertia* yang signifikan [14], [16]. *Silhouette Score* s(i) mengukur seberapa mirip suatu objek dengan klasternya sendiri dibandingkan klaster lain, dihitung menggunakan Persamaan (3), dengan a(i) adalah rata-rata jarak objek i ke seluruh anggota klaster yang sama dan b(i) adalah rata-rata jarak terkecil objek i ke anggota klaster lain [15]. Nilai *Silhouette* berkisar antara −1 hingga 1; semakin mendekati 1,

DOI https://doi.org/10.37817/ikraith-teknologi.v10i2
P-ISSN : 2580-4308
E-ISSN : 2654-8046

semakin baik kualitas pemisahan klaster. *Silhouette Score* tidak terdefinisi untuk K=1.

$$s(i) = (b(i) - a(i)) / max\{a(i), b(i)\} \quad (3)$$

2.6 Evaluasi dan Visualisasi

Hasil *clustering* divisualisasikan dalam bentuk diagram titik (*scatter plot*) dua dimensi. Karena data memiliki 11 fitur, digunakan *Principal Component Analysis* (PCA) untuk memproyeksikan data ke dua komponen utama yang menangkap variansi terbesar [21]. Sebagai pelengkap, disajikan pula diagram titik pada dua fitur asli yang paling diskriminatif. Kualitas klaster dievaluasi secara internal menggunakan *Silhouette Score*, dan secara eksternal melalui tabulasi silang antara label klaster dengan label *ground truth* untuk menghitung presisi dan *recall* deteksi anomali.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Preprocessing dan Reduksi Data

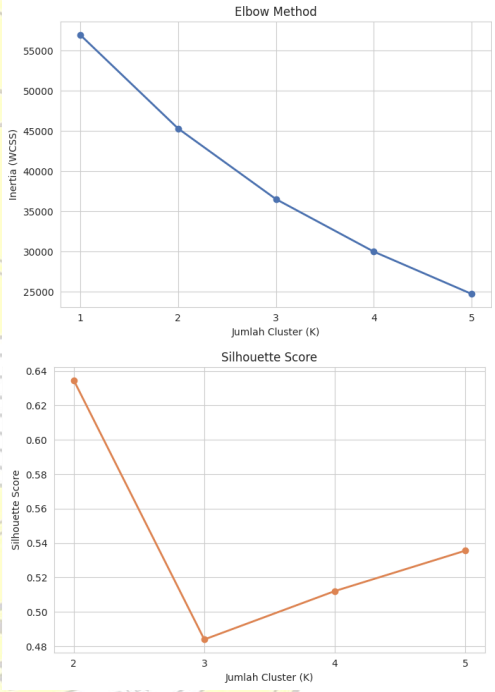
Pemeriksaan kualitas data memastikan tidak terdapat *missing value* maupun duplikasi pada 10.005 rekaman. Setelah seleksi fitur dan *one-hot encoding* atribut *protocol*, diperoleh 11 fitur numerik yang selanjutnya distandarisasi. Verifikasi statistik pascastandarisasi menunjukkan seluruh fitur memiliki rata-rata mendekati 0 dan simpangan baku mendekati 1, yang menandakan proses *scaling* berjalan dengan benar. Reduksi data melalui *random sampling* menghasilkan 5.000 sampel yang digunakan pada seluruh proses *clustering*. Proporsi label anomali pada sampel sebesar 20,6% (1.032 rekaman), relatif konsisten dengan proporsi pada data asli sebesar 20,0%, yang mengindikasikan *random sampling* tidak mengubah karakteristik distribusi data.

3.2 Penentuan Nilai K Optimal

Pengujian K-Means pada rentang K=1 hingga K=5 menghasilkan nilai *inertia* dan *Silhouette Score* sebagaimana dirangkum pada Tabel 2 dan divisualisasikan pada Gambar 1.

Tabel 2. Nilai inertia dan Silhouette Score untuk K=1 sampai K=5

K	Inertia (WCSS)	Silhouette Score
1	56.998,67	–
2	45.337,57	0,6347
3	36.542,48	0,4840
4	30.001,93	0,5121
5	24.733,74	0,5356



Gambar 1. Kurva Elbow Method (atas) dan Silhouette Score (bawah) untuk K=1 sampai K=5

Berdasarkan Tabel 2, *Silhouette Score* tertinggi sebesar 0,6347 dicapai pada K=2, jauh melampaui nilai pada K=3 (0,4840), K=4 (0,5121), maupun K=5 (0,5356). Kurva *Elbow* pada Gambar 1 memperlihatkan penurunan *inertia* paling tajam terjadi dari K=1 ke K=2 (sebesar 11.661,10 atau 20,5%), yang kemudian melandai pada nilai K berikutnya. Konsistensi kedua metode ini memberikan dasar yang kuat untuk menetapkan K=2 sebagai jumlah klaster optimal. Temuan ini selaras dengan

P-ISSN : 2580-4308
E-ISSN : 2654-8046 DOI https://doi.org/10.37817/ikraith-teknologi.v10i2

struktur alami data trafik jaringan yang pada dasarnya terpolarisasi menjadi dua perilaku, yaitu trafik normal dan trafik menyimpang, yang juga menjadi asumsi dasar pendekatan deteksi intrusi berbasis *clustering* pada data tak berlabel [10].

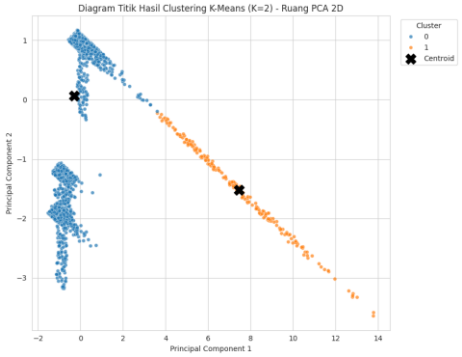
3.3 Hasil Clustering dan Visualisasi Diagram Titik

Model K-Means final dengan K=2 menghasilkan dua klaster dengan komposisi yang sangat timpang: klaster 0 berisi 4.807 rekaman (96,1%) dan klaster 1 berisi 193 rekaman (3,9%). Ketimpangan ukuran ini merupakan indikasi awal bahwa klaster 1 menangkap pola minoritas yang menyimpang dari perilaku umum. Karakteristik rata-rata fitur asli pada masing-masing klaster disajikan pada Tabel 3.

Tabel 3. Karakteristik rata-rata fitur pada tiap klaster

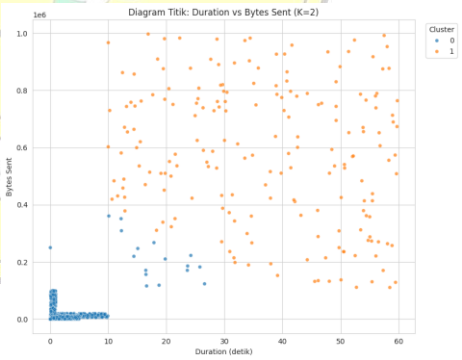
Fitur (rata-rata)	Klaster 0	Klaster 1
duration (detik)	2,53	36,96
packet_count	297,50	302,50
bytes_sent	8.050,61	583.812,32
bytes_received	6.941,72	2.983,22
bytes_per_packet	147,60	2.291,21

Tabel 3 memperlihatkan perbedaan karakteristik yang sangat kontras. Klaster 1 memiliki rata-rata durasi koneksi 36,96 detik atau sekitar 14,6 kali lebih lama dibandingkan klaster 0, dengan volume *bytes_sent* mencapai 583.812 *bytes* atau sekitar 72,5 kali lebih besar. Rasio *bytes_per_packet* klaster 1 juga sangat tinggi (2.291,21 berbanding 147,60). Sebaliknya, *bytes_received* klaster 1 justru lebih rendah, membentuk pola komunikasi satu arah bervolume besar. Pola durasi panjang, pengiriman data masif, dan respons minim semacam ini konsisten dengan karakteristik aktivitas eksfiltrasi data atau serangan berbasis volume. Seluruh anggota klaster 1 juga menggunakan protokol tipe 0, berbeda dengan klaster 0 yang tersebar pada enam tipe protokol.



Gambar 2. Diagram titik hasil clustering K-Means (K=2) pada ruang PCA dua dimensi beserta posisi centroid

Gambar 2 menyajikan diagram titik hasil *clustering* pada ruang PCA dua dimensi dengan total variansi yang dijelaskan sebesar 39,63%. Kedua klaster tampak terpisah dengan jelas: klaster 0 terkonsentrasi pada wilayah kiri sedangkan klaster 1 membentang di sepanjang sumbu komponen utama pertama ke arah kanan, menegaskan bahwa pemisahan didominasi oleh fitur volume trafik. Untuk interpretasi yang lebih intuitif, Gambar 3 menampilkan diagram titik pada dua fitur asli yang paling diskriminatif, yaitu *duration* dan *bytes_sent*.



Gambar 3. Diagram titik duration terhadap bytes_sent berdasarkan klaster

Pada Gambar 3, anggota klaster 0 mengelompok padat pada wilayah durasi singkat dan volume kecil (pojok kiri bawah), sedangkan anggota klaster 1 tersebar pada wilayah durasi 10–60 detik dengan *bytes_sent* antara 100 ribu hingga 1 juta *bytes*. Visualisasi ini memperkuat

DOI https://doi.org/10.37817/ikraith-teknologi.v10i2
P-ISSN : 2580-4308
E-ISSN : 2654-8046

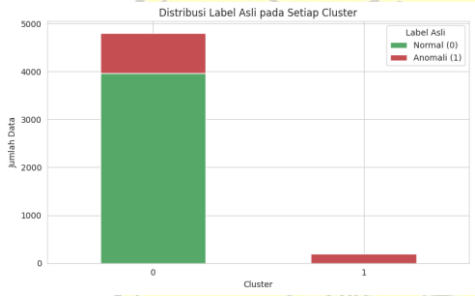
interpretasi bahwa klaster 1 merepresentasikan sesi koneksi berdurasi panjang dengan pengiriman data bervolume sangat besar.

3.4 Validasi terhadap Label Ground Truth

Sebagai validasi eksternal, label klaster ditabulasi silang dengan label asli *dataset* yang tidak digunakan selama *clustering*. Hasilnya disajikan pada Tabel 4 dan Gambar 4.

Tabel 4. Tabulasi silang label klaster terhadap label ground truth (0 = normal, 1 = anomali)

Klaster	Normal	Anomali	Total
Klaster 0	3.968	839	4.807
Klaster 1	0	193	193
Total	3.968	1.032	5.000



Gambar 4. Distribusi label ground truth pada setiap klaster

Tabel 4 menunjukkan seluruh 193 anggota klaster 1 berlabel anomali tanpa satu pun trafik normal yang salah masuk, sehingga presisi deteksi anomali pada klaster ini mencapai 100%. Artinya, setiap koneksi yang dipetakan model ke klaster 1 dapat dipastikan merupakan trafik anomali. Namun, dari total 1.032 trafik anomali pada sampel, hanya 193 yang berhasil terisolasi ke klaster 1, sehingga *recall* sebesar 18,7%. Sebanyak 839 anomali lainnya masih bercampur di dalam klaster 0, mengindikasikan keberadaan anomali bervolume rendah (*low-rate anomaly*) yang karakteristik fiturnya menyerupai trafik normal sehingga tidak terpisahkan oleh jarak *Euclidean* pada ruang fitur yang digunakan.

Pola presisi tinggi dengan *recall* terbatas ini konsisten dengan temuan penelitian K-Means pada jaringan LAN yang memperoleh presisi 100% dengan *recall* 61% akibat ketidakseimbangan kelas dan kemiripan fitur antara trafik normal dan anomali [11]. Tingginya tingkat *false alarm* memang dikenal sebagai tantangan utama sistem deteksi anomali [2], [3]; dari perspektif operasional keamanan jaringan, presisi 100% pada penelitian ini tetap bernilai praktis tinggi karena setiap alarm yang dihasilkan bukan alarm palsu (*zero false positive*), sehingga analis keamanan dapat langsung menindaklanjuti seluruh anggota klaster anomali tanpa membuang waktu memverifikasi *false alarm*. Adapun keterbatasan *recall* dapat ditangani dengan pendekatan berlapis, misalnya menerapkan *clustering* bertingkat pada klaster mayoritas, menambah fitur berbasis aturan keamanan sebagaimana dilakukan penelitian terdahulu yang berhasil menaikkan *Silhouette Score* hingga 0,9136 [12], atau mengombinasikan K-Means dengan metode deteksi *outlier* berbasis jarak ke *centroid*.

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Means untuk mendeteksi anomali pada data lalu lintas jaringan sebanyak 5.000 sampel hasil reduksi dari 10.005 rekaman. Tahapan *preprocessing* yang meliputi seleksi fitur, *one-hot encoding* atribut protokol, dan standarisasi fitur terbukti menghasilkan ruang fitur yang layak untuk *clustering*. Kombinasi *Elbow Method* dan *Silhouette Score* secara konsisten menetapkan K=2 sebagai jumlah klaster optimal dengan *Silhouette Score* 0,6347. Klaster minoritas yang terbentuk (193 rekaman; 3,9%) memiliki karakteristik durasi koneksi 14,6 kali lebih lama dan volume *bytes* terkirim 72,5 kali lebih besar dibandingkan klaster mayoritas, dan validasi eksternal membuktikan seluruh anggotanya (100%) merupakan trafik anomali dengan presisi

100% dan *recall* 18,7%. Hasil ini menunjukkan bahwa K-Means dengan optimasi jumlah klaster mampu mengisolasi anomali bervolume tinggi secara *unsupervised* tanpa memerlukan data berlabel. Penelitian selanjutnya disarankan menerapkan *clustering* bertingkat atau hibrida untuk menangkap anomali bervolume rendah, membandingkan kinerja dengan algoritma lain seperti DBSCAN [22] dan *Isolation Forest* [23], serta menguji kerangka kerja ini pada data trafik *real-time*.

5. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muria Kudus atas dukungan fasilitas dan bimbingan selama pelaksanaan penelitian ini.

DAFTAR PUSTAKA

[1] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016, doi: 10.1016/j.jnca.2015.11.016.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.

[3] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Ottawa, Canada, 2009, pp. 1–6, doi: 10.1109/CISDA.2009.5356528.

[5] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP)*, Funchal, Portugal, 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, vol. 1, pp. 281–297.

[7] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[8] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

[9] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[10] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," in *Proc. ACM CSS Workshop on Data Mining Applied to Security (DMSA)*, Philadelphia, PA, USA, 2001, pp. 5–8.

[11] [Penulis — lengkapi dari sumber], "Model clustering anomali detection pada jaringan LAN dengan algoritma K-Means," *Jurnal Sistem Komputer dan Informatika (JSON)*, 2025. [Daring]. Tersedia: https://ejurnal.stmik-budidarma.ac.id/index.php/JSON/article/view/9173

[12] [Penulis — lengkapi dari sumber], "Optimized K-Means clustering for web server anomaly detection using Elbow Method and security-rule enhancements," *Journal of Information Systems and Informatics*, vol. 7, 2025. [Daring]. Tersedia: https://journal-isi.org/index.php/isi/article/view/1391

[13] [Penulis — lengkapi dari sumber], "Penerapan machine learning untuk mendeteksi serangan anomali dalam jaringan komputer: Systematic literature review," 2025.

[14] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953, doi: 10.1007/BF02289263.

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[16] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means clustering method and Elbow Method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, p. 012017, 2018, doi: 10.1088/1757-899X/336/1/012017.

[17] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-Means cluster using the sum of squared error (SSE) optimized by using the Elbow Method," *Journal of Physics: Conference Series*, vol. 1361, p. 012015, 2019, doi: 10.1088/1742-6596/1361/1/012015.

[18] D. Marutho, S. H. Handaka, and E. Wijaya, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *Proc. International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, 2018, pp. 533–538, doi: 10.1109/ISEMANTIC.2018.8549751.

[19] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, USA, 1996, pp. 226–231.

[23] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. Eighth IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.