DETEKSI AKTIVITAS MENCURIGAKAN PADA LOG NGINX MENGGUNAKAN ISOLATION FOREST BERBASIS ANALISIS PERILAKU

Rezky Yuranda[1], Delpiah Wahyuningsih[2], Parlia Romadiana[3]

[12,]Program Studi Teknik Informatika, ISB Atma Luhur, [3,]Program Studi Sistem Informasi, ISB Atma Luhur

Jln Selindung Kec Gabek Pangkalpinang
[1]yurandarezky@atmaluhur.ac.id, [2] delphibabel@atmaluhur.ac.id, [3] parliaromadiana@atmaluhur.ac.id

Abstract

Security auditing of web server logs presents significant challenges due to the large volume of data and the limitations of manual analysis. This study proposes a behavior-based anomaly detection approach to identify suspicious activities in Nginx logs using unsupervised learning methods. The main contribution of this work lies in the integration of behavior-based features derived from time window aggregation, including request count, burstiness, error ratio (4xx), and the number of unique URLs, to represent user access patterns without requiring labeled data. Anomaly detection is performed using Isolation Forest, Local Outlier Factor (LOF), and One-Class SVM, with evaluation based on weak labels combined with feature distribution analysis and behavioral validation. Experimental results show that Isolation Forest achieves the best performance with an F1-score of 0.285, followed by One-Class SVM (0.191) and LOF (0.057), indicating the superiority of isolation-based methods in capturing anomalous patterns in time-aggregated data. Further analysis reveals that the detected anomalies exhibit characteristics consistent with suspicious activities such as web scanning and automated bot crawling. These findings demonstrate that behavior-based feature analysis can serve as an effective and scalable initial solution for anomaly detection in web server logs under unlabeled conditions.

Keywords : *Nginx Log, Anomaly Detection, Isolation Forest, Behavioral Analysis, Web Security*

Abstrak

Audit keamanan pada log web server menjadi tantangan akibat volume data yang besar serta keterbatasan proses analisis manual. Penelitian ini mengusulkan pendekatan deteksi anomali berbasis perilaku untuk mengidentifikasi aktivitas mencurigakan pada log Nginx menggunakan metode unsupervised learning. Kontribusi utama penelitian ini terletak pada integrasi fitur perilaku berbasis agregasi waktu (time windowing), meliputi request count, burstiness, rasio kesalahan (4xx), dan jumlah URL unik, untuk merepresentasikan pola akses pengguna tanpa memerlukan data berlabel. Deteksi anomali dilakukan menggunakan Isolation Forest, Local Outlier Factor (LOF), dan One-Class SVM, dengan evaluasi berbasis weak labelyang dikombinasikan dengan analisis distribusi fitur dan validasi perilaku. Hasil eksperimen menunjukkan bahwa Isolation Forest memberikan performa terbaik dengan nilai F1-score sebesar 0.285, diikuti oleh One-Class SVM (0.191) dan LOF (0.057), yang menunjukkan keunggulan metode berbasis isolasi dalam menangkap pola anomali pada data agregasi waktu. Analisis lebih lanjut menunjukkan bahwa data yang terdeteksi sebagai anomali memiliki karakteristik yang konsisten dengan aktivitas mencurigakan seperti web scanning dan automated bot crawling. Hasil ini menunjukkan bahwa pendekatan berbasis fitur perilaku dapat menjadi solusi awal yang efektif dan scalable untuk deteksi anomali pada log web server dalam kondisi tanpa label.

Kata kunci : *Nginx Log, Anomaly Detection, Isolation Forest, Behavioral Analysis, Web Security*

1. PENDAHULUAN

Keamanan infrastruktur web pada sektor pendidikan tinggi menjadi semakin krusial seiring meningkatnya ketergantungan terhadap layanan berbasis web seperti portal akademik dan jurnal ilmiah daring. Sebagai contoh, pengujian keamanan pada aplikasi *Open Journal System* (OJS) menunjukkan adanya puluhan kerentanan kritis yang dapat dieksploitasi oleh pihak tidak bertanggung jawab jika tidak dimonitor dengan baik [1]. Oleh karena itu, server web seperti Nginx yang mencatat setiap aktivitas pengguna dalam bentuk log akses, mengandung informasi historis yang sangat vital untuk dimanfaatkan secara proaktif dalam mendeteksi aktivitas mencurigakan maupun serangan siber [2],[3].

Namun, volume data log yang besar serta kompleksitas pola serangan modern membuat analisis manual menjadi tidak efektif dan rentan terhadap kesalahan. Pendekatan berbasis aturan (rule-based) yang umum digunakan juga memiliki keterbatasan dalam mendeteksi serangan baru yang tidak sesuai dengan pola yang telah ditentukan sebelumnya [4],[5]. Terlebih lagi, deteksi pada anomali log yang bersifat sekuensial memiliki tantangan tersendiri karena sifat datanya yang sangat fluktuatif dan dipengaruhi oleh pola deret waktu (*time series*) [6],[7].

Seiring dengan perkembangan machine learning, pendekatan berbasis pembelajaran mesin mulai banyak digunakan untuk mengotomatisasi deteksi anomali dalam data log. Pendekatan ini mampu mengidentifikasi pola tidak normal tanpa bergantung pada aturan eksplisit, sehingga lebih adaptif terhadap variasi serangan baru [5],[8]. Beberapa studi terkait telah mengeksplorasi ragam teknik machine learning untuk analisis keamanan dan performa server, mulai dari penggunaan kombinasi clustering K-Means dan DBSCAN pada log server [9], pemanfaatan metode deep learning untuk klasifikasi lalu lintas [10], penerapan Artificial Neural Network (ANN) untuk deteksi serangan pada log web server [11], klasifikasi serangan terdistribusi menggunakan algoritma Gradient Boosting [12], hingga ekstraksi data historis berbasis metrik perangkat keras (low-level hardware) untuk mengidentifikasi eksploitasi sistem [13].

Deteksi anomali sendiri merupakan proses untuk mengidentifikasi data yang menyimpang dari pola normal, yang sering kali mengindikasikan kejadian penting seperti kesalahan sistem, lonjakan beban kerja komputasi yang tidak wajar [14], atau aktivitas berbahaya. Dalam konteks keamanan web, anomali dapat merepresentasikan aktivitas serangan seperti *brute force*, *scanning*, hingga penyimpangan pada infrastruktur jaringan yang ter-virtualisasi [15]. Pendekatan algoritma yang kuat, khususnya *Isolation Forest*, terbukti sangat andal dan efisien dalam menangani kumpulan data berskala besar yang ekstrem (*imbalanced*) tanpa memerlukan pelabelan awal, sebagaimana divalidasi pada kasus deteksi penipuan finansial [16] maupun pengawasan sistem keamanan *web server* [17].

Penelitian ini mengusulkan pendekatan deteksi anomali berbasis *unsupervised learning* dengan memanfaatkan algoritma seperti *Isolation Forest*, *Local Outlier Factor* (LOF), dan *One-Class Support Vector Machine* (One-Class SVM). Pendekatan *unsupervised* dipilih karena data log Nginx yang dianalisis tidak memiliki label, serta sangat relevan ketika diintegrasikan dengan metode analisis perilaku (*behavioral feature analysis*) melalui rekayasa fitur. Pendekatan ini diharapkan lebih sensitif dalam mengidentifikasi pola serangan masa kini yang belum terdeteksi sebelumnya dan dapat diterapkan sebagai landasan mekanisme audit otomatis [5],[18].

2. METODOLOGI PENELITIAN

2.1. Skema Alur Penelitian

Penelitian ini menggunakan pendekatan unsupervised learning untuk mendeteksi anomali pada data log akses web server. Pendekatan ini dipilih karena data log tidak memiliki label, serta mampu mengidentifikasi pola menyimpang tanpa memerlukan data pelatihan berlabel [5].

Secara umum, tahapan penelitian ini terdiri dari beberapa proses utama yaitu:

1) Pengumpulan data log
2) Preprocessing
3) ekstraksi fitur
4) transformasi data
5) Pembentukan dataset akhir

Pada tahap awal, data dikumpulkan dari access log server Nginx yang digunakna pada web jurnal dan portal kampus. Data log ini berisi informasi aktivitas pengguna seperti alamat IP, waktu akses, metode HTTP, URL yang diakses, status respoons, ukuran data, serta user-agent. Data mentah tersebut kemudian diproses melalui tahap parsing menggunakan regular experssion untuk mengekstrak atribut-atribut penting dari setiap baris log.

Selanjutnya dilakukan proses data cleaning untuk menghilangkan data yang tidak valid dan tidak sesuai format (malformated log). Untuk menjaga aspek privasi, alamat IP pengguna dianonimkan menggunakan fungsi hashing berbasis SHA-256 sehingga tidak dapat dilacak kembali ke identitas asli pengguna.

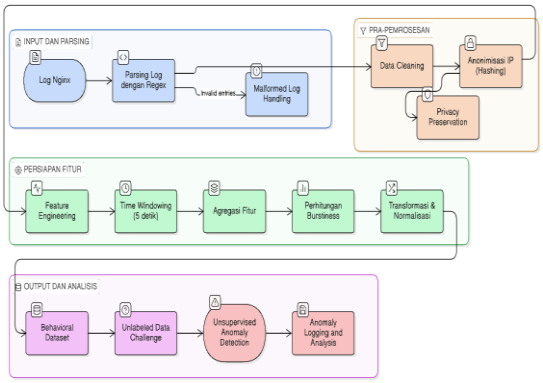
Tahap berikutnya adalah feature engineering, dimana dilakukan transformasi dan penambahan fitur baru yang relevan terhadap deteksi anomali. Fitur yang dihasilkan meliputi klasifikasi user-agent (browser, mobile, bot, CLI, dan unknown), esktraksi path URL, serta berbagai fitur statik berbasis waktu.

Untuk menangkap pola perilaku akses dalam rentang waktu tertentu, digunakna metode time windowing dengan interval 5 detik. Data kemudian dikelompokan berdasarkan IP yang telah di anonimkan dan interval waktu terseubt. Pada setiap kelompok dilakukan agregasi fitur seperti jumalh request, jumalah URL unik, rata-rata ukuran respons, serta rasio kesalahan (status kode >= 400).

Selain itu, penelitian ini juga menghitung burstiness untuk mengukur tingkat lonjakan aktivitas dalam suatu interval waktu. Nilai ini diperoleh dari rasio antara standar deviasi dan rata-rata selisih waktu antar request, sehingga dapat mengidentifikasi pola akses yang tidak normal seperti serangan atau aktivitas bot.

Setelah proses ekstraksi fitur selesai, dilakukan transformasi data berupa log transformastion pada jumlah request untuk mengurangi skweness distribusi, serta normalisasi fitur numerik menggunakan metode Min-Max Scalling. Tahapan ini berutujan untuk memastikan bahwa seluruh fitur berada dalam sekali yang seragam sebelum digunakan dalam proses analisis.

Hasil akhir dari seluruh tahapan ini adalah dataset tersetruktur yang merepresentasikan perilaku akses web dalam bentuk fitur numerik. Dataset ini kemudian digunakan sebagai input dalam proses deteksi anomali menggunakan metode unsupervised learning.



Gambar 1. Alur Penelitian

2.2. Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan menggunakan metode observasi langsung terhadap sistem, yaitu dengan mengambil data dari access log server web berbasis Nginx tanpa melakukan intervensi terhadap sistem.

a. Data Primer

Data primer yang digunakan dalam penelitian ini berupa access log dari server web berbasis Nginx yang digunakan pada:

1) Web Jurnal ISB Atma Luhur
2) Web Portal Akademik ISB Atma Luhur

Log tersebut memuat informasi aktivitas HTTP request seperti alamat IP, waktu akses, metode HTTP, URL, status kode, ukuran respons, dan user-agent. Data yang diperoleh masih dalam bentuk mentah (raw data) sehingga memerlukan proses pengolahan lebih lanjut.

b. Data Sekunder

Data sekunder dalam penelitian ini digunakansebagai pendukung dalam proses analisis dan pengembangan sistem, yang meliputi:

1) Literatur ilmiah terkait anomaly detection pada web log
2) Dokumentasi format access log Nginx
3) Referensi metode unsupervised learning untuk deteksi anomali.
4) Dokumentasi pustaka Python seperti Pandas, NumPy, dan Scikit-learn.

Data sekunder ini berperan dalam menentukan metode pengolahan data, pemilihan

fitur, serta pendekatan analisis yang digunakan dalam penelitian.

c. Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini dilakukan melalui observasi sistem, studi dokumentasi dan preprocessing awal. Observasi sistem dilakukan dengan mengambil data secara langsung dari file access log server Nginx. Data yang dikumpulkan merupakan data histori yang mempersentasikan aktivitas pengguna dalam periode tertentu tanpa ada nya intervensi terhadap sistem yang berjalan.

Selain itu, penelitian ini juga menggunakan pendektatan studi dokumentasi yang digunakan untuk memahami struktur dan format log Nginx, termasuk element-element penting yang dapat di ekstraksi untuk kebutuhan analisis. Pemahaman ini diperlukan agar proses pengolahan data dapat dilakukan secara tepat sesuai dengan format log yang digunakan.

Setelah data dikumpulkan, dilakukan tahap preprocessing awal menggunakan aplikasi berbasis Python. Proses ini meliputi ekstraksi informasi penting dari log, pembersihan data yang tidak valid (malformed entries), anonimisasi data sensitif seperti alamat IP, serta pembentukan dataset yang siap digunakan untuk tahap analisis. Seluruh proses dilakukan secara offline untuk memastikan kualitas dasta sebelum digunakan dalam proses deteksi anomali.

Data yang diperoleh pada tahap awal berjumlah 42.566 baris log. Setelah melalui proses preprocessing yang meliputi pembersihan data, anonimisasi, dan agregasi berbasis waktu, jumlah data yang digunakan dalam penelitian ini menjadi 20.021 data observasi

2.3. Analisa Data

Analisa data dalam penelitian ini bertujuan untuk mengidentifikasi pola akses norma dan tidak normal apda server web berdasarkan data log yang telah di proses. Mengingat data yang digunakan tidak memiliki label, pendekatan unsupervised anomali detection digunakan untuk mendeteksi penyimpangan perilaku tanpa bergantung pada data pelatihan berlabel.

Data yang dianalisis merupkana hasil preprocessing dari access log nginx yang telah ditransformasikan menjadi dataset tersetruktur berbasis fitur. Fitur yang digunakan mencerminkan karakteristik perilaku akses pengguna, atara lain jumlah request, jumlah URL unik, rata-rata ukuran respons, rasio kesalahan (status kode >= 400), tipe user-agent, serta fitur temporal seperti burstiness yang menggambarkan pola intensitas akses dalam interval waktu tertentu.

Analisis difokuskan pada perilaku entitas akses yang dirpresentasikan dalam bentuk IP yang telah di anonimkan. Dengan demikian, penelitian ini tidak berfokus poada identitias pengguna, melainkan pada pola aktivitas yang dihasilkan oleh setiap entitas dalam rentang waktu tertentu.

Data yang digunakan berasal dari lingkungan operasional server web ISB Atma Luhur, khususnya pada web jurnal dan portal akademik, sehingga mencerminkan kondisi nyata penggunaan sistem. Proses analisis dilakukan secara offline terhadap data historis yang telah melalui tahapan agregasi berbasis waktu untuk menangkap dinamika kativitas dalam interval tertentu.

Pendekatan analisis yang dilakukan bersifat berbasis fitur (feature-based analysis), dimana setiap entitas akses direpresentasikan dalam bentuk vektor numerik. Dataset ini kemudian digunakan sebagai input dalam metode deteksi anomali untuk mengidentifikasi pola yang menyimpang dari perilaku normal. Pendekatan ini memungkinakan identifikasi aktivitas tidak wajar seperti lonjakan trafik, pola akses berulang, maupun indikasi aktivitas otomatis seperti bot dan scanning.

2.4. Tahapan Pengelolaan dan Implementasi Sistem

Tahapan pengelolaan data dalam penelitian ini dilakukan secara sistemasis untuk mengahasilkan dataset yang repsentatif terhadap perilaku akses pengguna. Proses ini mencakup parsing data, anonimisasi, agregasi berbasis waku, ekstraksi fitur, serta transformasi data sebelum digunakna dalam proses deteksi anomai.

a. Parsing Log

Tahap awal dilakukan dengan mengekstraki informasi dari file access log Nginx menggunakn pendekatan regular expression. Informasi yang diperoleh meliputi alamat IP, waktu akses (timestamp), metode HTTP, URL, status kode, ukuran respons, referrer, dan user-agent. Data yang tidaak sesuai dengan format akan diabaikan untuk menjaga kualitas dan konsistensi dataset.

b. Anomisasi Data

Untuk menjaga privasi pengguna, alamat IP dianonimkan menggunakan fungsi hashing berbasis SHA-256. Proses ini memungkinkan analisis perilakuk tetap dilakukan menggunakan identiatas asli pengguna.

c. Time Windowing

Data kemudian dikelompokkan berdasarkan interval waktu menggunakan metode time windowing dengan rentang waktu sebesar 5 detik. Pendekatan berbasis time series memungkinkan identifikasi pola akses yang bersifat temporal serta perubahan perilaku pengguna dari waktu ke waktu [7]. Pemilihan interval waktu 5 detik didasarkan pada kebutuhan untuk menangkap pola aktivitas jangka pendek seperti burst traffic dan scanning yang umumnya terjadi dalam rentang waktu singkat. Interval yang terlalu besar berpotensi mengaburkan pola anomali, sedangkan interval yang terlalu kecil dapat menyebabkan fragmentasi data yang berlebihan. Oleh karena itu, interval 5 detik dipilih sebagai kompromi antara resolusi temporal dan stabilitas agregasi data.

d. Ekstraksi dan Agregasi Fitur

Pada setiap kelompok data berdasarkan IP dan interval waktu, dilakukan ekstraksi fitur berbasis perilaku (behavioral features), yaitu:

1) Jumlah request (request count)
2) Jumlah URL unik (unique path)
3) rata-rata ukuran respons (average byte)
4) rasio kesalahan (status kode >= 400)
5) tipe user-agent

Fitur-fitur ini digunakan untuk merepresentasikan karakteristik aktivitas pengguna dalam bentuk numerik.

e. Perhitungan Burstines

Untuk mengidentifikasi pola lonjakan aktivitas, digunakan fitur burstiness yang dihitung berdasarkan selisih waktu antara request dalam satu interval. Nilai burstiness dihitung menggunakan Persamaan (1).

$$B = \frac{\sigma}{\mu + \epsilon} \qquad (1)$$

dengan σ sebagai standar deviasi dan μ sebagai rata-rata selisih waktu antar request. Nilai ini merepresentasikan tingkat ketidakstabilan pola akses dalam suatu interval waktu.

f. Transformasi Data

Untuk mengatasi distibusi data yang tidak seimbang dilakukan transformasi logaritmik pada fitur jumlah request menggunakan fingsi log(1+x). Selain itu, normalisasi dilakukan menggunakan metode Min-Max Scaling untuk menyamakan skala atar fitur

g. Pembentukan Dataset Akhir

Hasil dari seluruh tahapan ini adalah dataset terstruktur yang merepresentasikan prilaku akses pengguna dalam bentuk fitur numerik. Dataset ini kemudian digunakan sebagai input dalam proses deteksi anomali.

2.5. Metode Anomaly Detection

Penelitian ini menggunakan pendekatan unsupervised learning untuk mendeteksi anomali pada log akses web, mengingat data yagn digunakan tidak memiliki label. Tiga algoritma yang digunakan dalam peneltiia nini adalah Isolation Forest, Local Outlier Factor (LOF), dan One-Class Support Vector Machines (One-Class SVM). Penggunaan beberapa algrotitma bertujuan untuk memperoleh hasil deteksi yang lebih komprehensif serta memungkinkan perbandingan performa dalam mengidentifikasi anomali.

h. Isolation Forest

Isolation Forest merupakan algoritma deteksi anomali yang bekerja dengan cara mengisolasi data melalui proses partisi acak. Data yang lebih mudah diisolasi (memiliki path lebih pendek) dianggap sebagai anomali. Metode ini efektif untuk dataset berdimensi tinggi dan tidak memerlukan proses labeling [5].

i. Local Outline Factor (LOF)

Local Outlier Factor (LOF) merupakan metode berbasis kepadatan yang mengukur tingkat kepadatan lokal suatu data dibandingkan dengan tetangganya. Data dengan kepadatan yang lebih rendah dibandingkan lingkungan sekitarnya akan dikategorikan sebagai anomali [18].

j. One-Class Support Vector Machine

One-Class Support Vector Machine digunakan untuk memodelkan distribusi data normal dan mengidentifikasi data yang berada di luar batas distribusi tersebut sebagai anomali. Metode ini

banyak digunakan dalam kasus deteksi intrusi dan keamanan sistem [18].

k. Evaluasi

Evaluasi dalam penelitian ini dilakukan dengan pendekatan semi-kuantitatif menggunakan kombinasi weak label, analisis distribusi, serta validasi perilaku. Pendekatan ini dipilih karena dataset yang digunakan tidak memiliki ground truth label yang pasti, sehingga tidak memungkinkan penggunaan evaluasi supervised secara konvensional.

l. Pembentukan Weak Label

Weak label dibentuk menggunakan pendekatan rule-based heuristic berbasis karakteristik perilaku akses yang secara umum diasosiasikan dengan aktivitas mencurigakan. Suatu data dikategorikan sebagai anomali apabila memiliki nilai yang melebihi ambang batas tertentu pada salah satu fitur utama, yaitu:

1) Jumlah request dalam interval waktu tertentu melebihi ambang batas T_r
2) Jumlah URL unik yang diakses melebihi ambang batas T_u
3) Nilai burstiness yang tinggi melebihi ambang batas T_b

Ambang batas ditentukan berdasarkan distribusi data menggunakan pendekatan percentile (P95) untuk menangkap perilaku ekstrem.

m. Evaluasi Berbasis Konsistensi Model

Hasil prediksi dari masing-masing algoritma dibandingkan dengan weak label menggunakan metrik:

1) Precision
2) Recall
3) F1-Score

Pendekatan ini berutujan kesesuaian relatif antara model dan pola heuristik yang digunakan sebagai referensi.

n. Analisis Distribusi Anomali

Selain metrik kuantitatif, dilakukan analisis distribusi untuk membandingkan karakteristik fitur antara data normal dan anomali, meliputi:

1) Distribusi request count
2) Distribusi burstiness
3) Distribusi rasio kesalahan (4xx)

Tujuan analisis ini adalah untuk memastikan bahwa data yang terdeteksi sebagai anomali yang memiliki pola berbeda secara signifikan

d. Validasi Perilaku (Behavioral Validation)

Sebagai tahap akhir evaluasi, dilakukan validasi manual terhadap sampel data yang terdeteksi sebagai anomali. Validasi dilakukan dengan mengamati:

1) pola akses URL
2) frekuensi request
3) rasio error
4) jenis user-agent

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Dataset yang digunakan dalam peneltian ini merupakan hasil preprocessing dari data access log server web berbasis Nginx pada web jurnal dan portal akademik ISB Atma Luhur. Data awal berjumalh 42.566 baris log, yang setelah melalui proses pembersihan data, anonimisasi, serta agregasi berbasis waktu mengasilkan 20.021 data observasi.

Setiap data observasi merepresentasikan aktivitas akses dalam interval waktu tertentu yang telah ditransformasikan menjadi fitur numerik berbasis perilaku (behavioral features). Fitur-fitur tersebut digunakan untuk merepresentasikan karakteristik akses pengguna dan menjadi input dalam proses deteksi anomali menggunakan metode unsupervised learning. Deskripsi fitur yang digunakan disajikan pada Tabel I.

TABEL I. DESKRIPSI FITUR DATASET

No	Fitur	Deskripsi
1	req_count	Jumlah request dalam interval waktu
2	uniq_path	Jumlah URL unik yang diakses
3	avg_bytes	Rata-rata ukuran respons
4	status_4xx_ratio	Rasio error client
5	ua_type	Kategori user-agent
6	Burstiness	Variasi waktu antar request

3.2 Hasil Pengujian

Pengujian dilakukan menggunakan algoritma Isolation Forest terhadap dataset hasil preprocessing yang telah direpresentasikan dalam bentuk fitur numerik berbasis perilaku

akses. Evaluasi dilakukan menggunakan pendekatan weak label sebagai referensi awal, serta didukung dengan analisis distribusi fitur dan pola temporal dari data yang terdeteksi sebagai anomali..

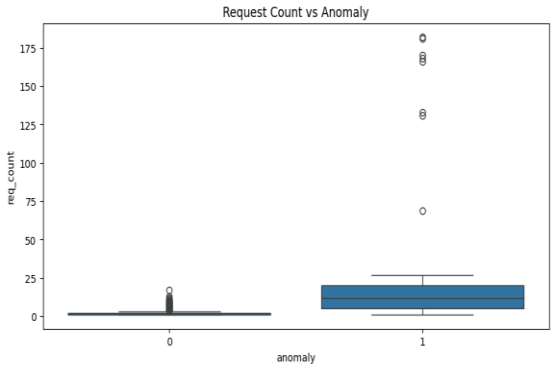
Hasil evaluasi model disajikan pada Table II.

TABEL II. EVALUASI MODEL ANOMALY DETECTION

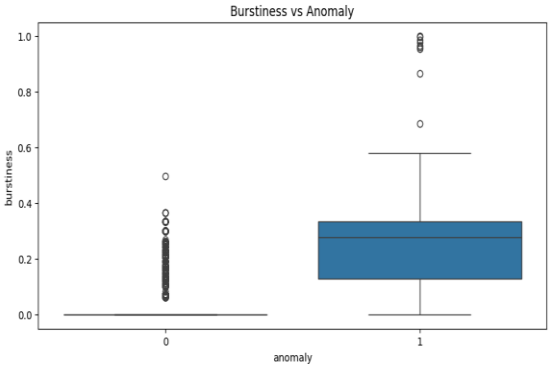
No	Algorima	Precision	Recall	F1-Score	waktu
1	Isolation Forest	0.166389	1.0	0.285307	0.03
2	LOF	0.105708	1.0	0.191205	0.02
3	One-Class SVM	0.033445	0.2	0.057307	0.04

Berdasarkan Table II, model menunjukkan tingkat kesesuaian tertentu terhadap pola anomali berbasis heuristic yang digunakan sebagai referensi. Namun demikian, hasil ini tidak dimaksudkan sebagai ukuran performa absolut, melainkan sebagai indikasi awal dalam mengevaluasi konsistensi model terhadap pola perilaku yang dianggap menyimpang. Nilai recall yang tinggi dipengaruhi oleh penggunaan weak label, sehingga hasil evaluasi lebih merefleksikan kesesuaian terhadap aturan heuristik dibandingkan performa terhadap ground truth yang sebenarnya.

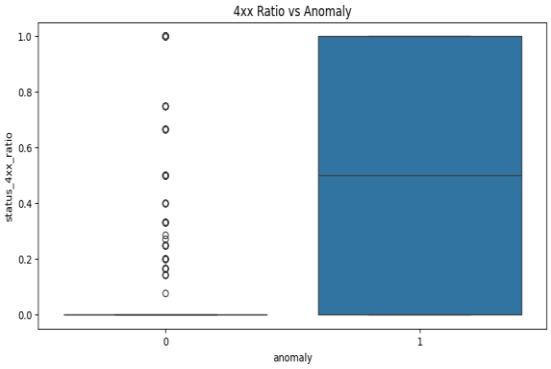
Untuk memahami karakteristik data yang terdeteksi sebagai anomali, dilakukan analisis distribusi fitur sebagaimana ditunjukkan pada Gambar 2, Gambar 3, dan Gambar 4 yang dihasilkan dari hasil deteksi menggunakan Isolation Forest.



Gambar 2. Perbandingan Distribusi Request Count antara Data Normal dan Anomali



Gambar 3. Perbandingan Distribusi Burstiness antara Data Normal dan Anomali

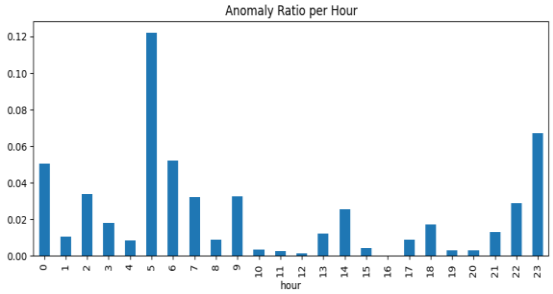


Gambar 4. Perbandingan Distribusi Rasio 4xx antara Data Normal dan Anomali

Gambar 2 menunjukkan distribusi request count antara data normal dan anomali. Terlihat bahwa sebagian data anomali memiliki jumlah request yang relatif lebih tinggi dibandingkan data normal dalam interval waktu yang sama.

Gambar 3 memperlihatkan distribusi nilai burstiness. Data yang terdeteksi sebagai anomali cenderung memiliki variasi waktu antar request yang lebih besar, yang mengindikasikan pola akses yang tidak stabil dalam interval tertentu.

Selanjutnya, pada Gambar 4 terlihat bahwa rasio kesalahan (status code 4xx) pada data anomali cenderung lebih tinggi dibandingkan data normal. Hal ini dapat mengindikasikan adanya percobaan akses terhadap endpoint yang tidak valid atau tidak tersedia, yang umum ditemukan pada aktivitas probing atau scanning terhadap sistem web [2],[5].



Gambar 5. Distribusi Rasio Anomali Berdasarkan Waktu (jam)

Selain analisis berbasis fitur, dilakukan juga analisis terhadap distribusi anomali berdasarkan waktu, sebagaimana ditunjukkan pada Gambar 5. Hasilnya menunjukkan bahwa kemunculan anomali tidak tersebar secara merata, melainkan cenderung meningkat pada periode waktu tertentu. Pola ini dapat berkaitan dengan aktivitas otomatis seperti bot crawling atau web scanning yang berjalan secara periodic [5].

Secara umum, hasil pengujian menunjukkan bahwa data yang terdeteksi sebagai anomali memiliki karakteristik yang berbeda dibandingkan data normal, baik dari sisi jumlah aktivitas, pola waktu akses, maupun tingkat kesalahan. Temuan ini memberikan indikasi bahwa pendekatan berbasis unsupervised anomaly detection mampu menangkap pola perilaku yang menyimpang dalam data log akses web tanpa memerlukan label eksplisit [5],[10].

Sebagai validasi tambahan, dilakukan pemeriksaan manual terhadap beberapa sampel data yang terdeteksi sebagai anomali. Hasil pemeriksaan menunjukkan bahwa salah satu alamat IP memiliki pola akses dengan jumlah request tinggi dalam waktu singkat serta percobaan akses berulang terhadap endpoint tertentu, yang mengindikasikan aktivitas brute force. Temuan ini memberikan indikasi bahwa hasil deteksi memiliki relevansi terhadap aktivitas mencurigakan pada log akses web.

3.3 Pembahasan

Hasil penelitian menunjukkan bahwa pendekatan berbasis fitur perilaku (behavior-based features) efektif dalam mendeteksi anomali pada data log akses web. Fitur seperti jumlah request (request count), burstiness, serta rasio kesalahan (4xx ratio) terbukti mampu membedakan pola akses normal dan tidak normal secara signifikan, sebagaimana ditunjukkan pada Gambar 2, Gambar 3, dan Gambar 4.

Algoritma Isolation Forest menunjukkan performa terbaik dalam penelitian ini. Hal ini disebabkan oleh kemampuannya dalam mengisolasi data yang memiliki karakteristik berbeda secara efisien, terutama pada dataset dengan distribusi tidak merata dan jumlah anomali yang relatif kecil (rare events). Pendekatan ini sesuai dengan karakteristik data log yang cenderung memiliki pola anomali yang sporadis namun signifikan.

One-Class SVM juga menunjukkan performa yang tinggi dalam mendeteksi anomali. Metode ini mampu membentuk batas (decision boundary) terhadap data normal dengan baik, sehingga dapat mengidentifikasi penyimpangan secara efektif. Namun, waktu komputasi yang lebih tinggi menunjukkan bahwa metode ini memiliki kompleksitas yang lebih besar dibandingkan Isolation Forest.

Sebaliknya, Local Outlier Factor (LOF) menunjukkan performa yang rendah. Hal ini kemungkinan disebabkan oleh pendekatan berbasis kepadatan (*density-based*) yang kurang sesuai dengan karakteristik data yang telah melalui proses agregasi berbasis waktu. Proses time windowing dapat mengubah distribusi data sehingga pola kepadatan lokal menjadi kurang representatif, yang berdampak pada penurunan kinerja metode LOF.

Selain itu, analisis temporal pada Gambar 5 menunjukkan bahwa distribusi anomali tidak merata sepanjang waktu. Terdapat periode tertentu dengan peningkatan rasio anomali yang signifikan, yang mengindikasikan adanya pola aktivitas tidak normal yang kemungkinan bersifat terjadwal, seperti aktivitas bot atau scanning otomatis.

Untuk memperkuat hasil deteksi, dilakukan validasi manual terhadap beberapa sampel data yang diklasifikasikan sebagai anomali maupun normal. Hasil pemeriksaan menunjukkan bahwa salah satu alamat IP yang terdeteksi sebagai anomali memiliki pola akses dengan jumlah request yang tinggi dalam waktu singkat, mengakses berbagai endpoint yang berbeda, serta disertai dengan rasio kesalahan (4xx) yang meningkat. Pola ini mengindikasikan adanya aktivitas scanning, di mana sistem mencoba mengakses berbagai resource untuk menemukan celah atau endpoint yang valid.

Sebaliknya, pada data yang diklasifikasikan sebagai normal, pola akses yang diamati cenderung stabil dengan jumlah request yang rendah, variasi URL yang terbatas, serta rasio kesalahan yang minimal. Aktivitas ini mencerminkan perilaku pengguna yang wajar dalam mengakses sistem tanpa adanya indikasi aktivitas mencurigakan.

Hasil penelitian ini sejalan dengan studi sebelumnya yang menunjukkan bahwa Isolation Forest efektif dalam mendeteksi anomali pada

data log berskala besar dan tidak seimbang [5],[17]. Dibandingkan dengan pendekatan berbasis kepadatan seperti LOF, metode Isolation Forest lebih robust terhadap perubahan distribusi data akibat agregasi waktu.

Selain itu, penelitian ini memperkuat temuan bahwa fitur berbasis perilaku seperti burstiness dan request intensity merupakan indikator penting dalam mendeteksi aktivitas mencurigakan, sebagaimana juga dilaporkan dalam studi deteksi anomali berbasis time-series [7].

4. Kesimpulan dan Saran

Penelitian ini berkontribusi dalam pengembangan pendekatan deteksi anomali berbasis perilaku pada log Nginx dalam kondisi tanpa label, melalui integrasi fitur berbasis agregasi waktu (*time windowing*), analisis distribusi data, serta validasi berbasis perilaku akses. Pendekatan ini memungkinkan proses evaluasi yang lebih komprehensif, tidak hanya bergantung pada metrik kuantitatif, tetapi juga mempertimbangkan konteks aktivitas yang terdeteksi.

Hasil eksperimen menunjukkan bahwa algoritma Isolation Forest dan One-Class SVM memiliki konsistensi yang tinggi dalam mendeteksi pola anomali, ditunjukkan oleh nilai recall yang optimal terhadap weak label serta kesesuaian dengan pola distribusi data. Sebaliknya, Local Outlier Factor (LOF) menunjukkan performa yang relatif rendah, yang mengindikasikan bahwa pendekatan berbasis kepadatan kurang sesuai untuk data yang telah melalui proses agregasi berbasis waktu (time windowing), karena cenderung mengurangi sensitivitas terhadap variasi lokal.

Lebih lanjut, analisis distribusi fitur serta validasi manual terhadap sampel data mengindikasikan bahwa data yang terdeteksi sebagai anomali memiliki karakteristik yang selaras dengan aktivitas mencurigakan, seperti scanning dan automated bot crawling. Hal ini memperkuat bahwa pendekatan yang digunakan tidak hanya mampu mendeteksi anomali secara statistik, tetapi juga relevan secara kontekstual dalam domain keamanan web.

Meskipun demikian, penggunaan weak label sebagai referensi evaluasi masih memiliki keterbatasan, terutama karena tidak sepenuhnya merepresentasikan ground truth yang akurat. Oleh karena itu, penelitian selanjutnya disarankan untuk:

1) Menggunakan dataset dengan label yang lebih akurat
2) Mengintegrasikan sistem deteksi dengan mekanisme logging secara real-time untuk mendukung deteksi dini (early detection)
3) Mengembangkan pendekatan evaluasi berbasis semi-supervised learning atau active learning guna meningkatkan kualitas pelabelan secara bertahap

5. UCAPAN TERIMA KASIH

Ucapan terima kasih kami sampaikan kepada ISB Atma Luhur yang telah membantu dalam penelitian ini serta seluruh civitas Fakultas Teknik Informasi, Institut Sains dan Bisnis Atma Luhur.

Daftar Pustaka:

[1] Y. W, D. T. Yuwono, Rodianto, and Yuliadi, "Deteksi Serangan Vulnerability Pada Open Jurnal System Menggunakan Metode Black-Box," *Jurnal Informatika & Rekayasa Elektronika*, vol. 4, no. 1, 2021, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900

[2] H. Bobde, A. Aglawe, S. Lakhamapure, D. Ukey, and K. Dhakate, "Log Alert System Server Log Recognition and Alert System the Creative Commons Attribution License (CC BY 4.0)," *International Journal of Trend in Scientific Research and Development (IJTSRD)*, vol. 8, no. 6, pp. 69–78, 2024, [Online]. Available: www.ijtsrd.com/papers/ijtsrd70555.pdf

[3] A. R. Purwidyantoro and E. S. Pramukantoro, "Sistem Klasifikasi Serangan Pada Website Berbasis Wordpress Menggunakan Machine Learning," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, pp. 2548–964, Mar. 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[4] N. Alamsyah and V. Claudia Jennifer Kaunang, "Anomaly detection in walking data using isolation forest: an unsupervised learning approach," *Journal Of Information Systems & Artificial Intelligence*, vol. 6, no. 1, Jan. 2025.

[5] W. Chua *et al.*, "Web Traffic Anomaly Detection Using Isolation Forest," *Informatics*, vol. 11, no. 4, Dec. 2024, doi: 10.3390/informatics11040083.

[6] D. Delvita aulia artika, D. R. Rumahorbo, M. haikal A.-M. haikal, and D. Kiswanto, "Implementasi Sistem Keamanan Website Dengan Analisis Log Dan Deteksi Aktivitas

Anomali Menggunakan Isolation Forest," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, Oct. 2025, doi: 10.23960/jitet.v13i3S1.8133.

[7] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," Apr. 30, 2022, *Association for Computing Machinery*. doi: 10.1145/3444690.

[8] F. Rahman, T. E. Sutanto, and N. Fitriyati, "Web Traffic Anomaly Detection using Stacked Long Short-Term Memory," *InPrime: Indonesian Journal of Pure and Applied Mathematics*, vol. 3, no. 2, pp. 112–121, Nov. 2021, doi: 10.15408/inprime.v3i2.21879.

[9] R. P. Irawan, "Kombinasi Algoritma K-Means dan DBSCAN dalam Identifikasi Anomali pada Data Log Server," *Media Informasi Analisa dan Sistem*, vol. 9, no. 2, Dec. 2024.

[10] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series," *IEEE Access*, vol. 7, pp. 1991–2005, Jan. 2019, doi: 10.1109/ACCESS.2018.2886457.

[11] A. Wicahyanto, Nurchim, and Wijiyanto, "Penerapan Artificial Neural Network Dalam Deteksi Serangan Pada Web Server Apache," *Jurnal Informatika & Rekayasa Elektronika*, vol. 8, no. 1, 2025, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900

[12] C. A. Lesmana and L. Hakim, "Klasifikasi Serangan Ddos Menggunakan Reqursive Feature Elimination Dan Gradient Boosting," *Jurnal Informatika & Rekayasa Elektronika)*, vol. 8, no. 1, 2025, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire ISSN.2620-6900

[13] C. Liu, Z. Yang, Z. Blasingame, G. Torres, and J. Bruska, "Detecting data exploits using low-level hardware information: A short time series approach," in *RESEC 2018 - Proceedings of the 1st Workshop on Radical and Experiential Security, Co-located with ASIA CCS 2018*, Association for Computing Machinery, Inc, May 2018, pp. 41–47. doi: 10.1145/3203422.3203433.

[14] S. Kardani-Moghaddam, R. Buyya, and K. Ramamohanarao, "Performance anomaly detection using isolation-trees in heterogeneous workloads of web applications in computing clouds," *Concurr. Comput.*, vol. 31, no. 20, Oct. 2019, doi: 10.1002/cpe.5306.

[15] Girish L and S. K. N Rao, "Anomaly Detection in NFV Using Tree-Based Unsupervised Learning Method," *International Journal of Engineering Sciences and Management-A Multidisciplinary Publication of VTU*, vol. 1, no. 2, pp. 27–31, 2019.

[16] I. Waspada, N. Bahtiar, P. W. Wirawan, B. Dwi, and A. Awan, "Performance Analysis of Isolation Forest Algorithm in Fraud Detection of Credit Card Transactions," *Khazanah Informatika*, vol. 6, no. 2, 2020.

[17] D. B. Santoso and Y. Wahyuni, "Sestem Log Web Server Sebagai Pendeteksi Anomali Menggunakan Isolation Forest Web Server Log System As An Anomaly Detector Using Isolation Forest," *Jurnal Aplikasi Bisnis dan Komputer*, vol. 4, no. 3, pp. 2807–5986, Oct. 2024, [Online]. Available: http://www.jubikom.unpak.ac.id

[18] J. Fan, "Analyzing the applicability of isolation forest for detecting anomalies in time series data," University of Oulu, 2025.