JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

Analisis Big Data Metadata CERN Berbasis Cloud Computing Menggunakan Algoritma Z-Score Outlier Detection

Lastri Putri Silaban[1*], Muhammad Tsabat Muhyiyuddin[2], Zevan Irfan Surbakti[3]

[1,2,3]Fakultas Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
Email: [1*]lastrisilaban2006@email.com, [2]izuddintsabat@email.com, [3]zevanirfandisurbakti@email.com
(*Email Corresponding Author: lastrisilaban2006@gmail.com)
Received: 17 Maret 2026 | Revision: 15 Juli 2026 | Accepted: 19 Juli 2026

Abstrak

Penelitian ini bertujuan untuk mengidentifikasi anomali serta mengenali pola tersembunyi pada dataset metadata eksperimen ATLAS berskala *Big Data* yang bersumber dari *CERN Open Data*. Permasalahan utama dalam penelitian ini adalah tingginya variasi jumlah kejadian fisika (*events*) yang menyulitkan proses analisis dan interpretasi data secara akurat. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan pendekatan *Knowledge Discovery in Database* (KDD) berbasis komputasi awan (*Cloud Computing*) yang meliputi agregasi data, komputasi statistik deskriptif, analisis kemiringan distribusi (*skewness*), *Z-Score Outlier Detection*, dan pengenalan pola sekuensial (*pattern recognition*). Hasil analisis terhadap 374 dataset menunjukkan rata-rata jumlah kejadian sebesar 5.075.259,44 *events* dengan standar deviasi 18.689.769,53 *events*, di mana distribusi populasi terkonfirmasi memiliki tingkat kemiringan positif yang ekstrem (*Right-Skewed*). Berdasarkan perhitungan Z-Score dengan ambang batas $|Z| > 3$, ditemukan sebanyak 6 dataset (sekitar 1,6%) yang teridentifikasi sebagai anomali ekstrem. Lebih lanjut, analisis tren sekuensial dan visualisasi spasial membuktikan bahwa anomali tersebut tidak terjadi secara acak, melainkan membentuk klaster sistemik yang terpusat pada rentang *batch* observasi tertentu. Integrasi metode Z-Score dan pengenalan pola ini terbukti efektif dalam memvalidasi kualitas data (*Veracity*) serta mengekstraksi informasi operasional pada dataset saintifik berskala masif.

Kata Kunci: Big Data, Komputasi Awan, Deteksi Anomali, Pengenalan Pola, Z-Score.

Abstract

This study aims to identify anomalies and recognize hidden patterns in the Big Data-scale ATLAS experimental metadata dataset sourced from CERN Open Data. The primary issue addressed is the high variance in the number of physics events, which complicates accurate data analysis and interpretation. To overcome this problem, the research applies a cloud computing-based Knowledge Discovery in Database (KDD) approach, encompassing data aggregation, descriptive statistical computation, distribution skewness analysis, Z-Score Outlier Detection, and sequential pattern recognition. The analysis of 374 datasets reveals a mean of 5,075,259.44 events and a standard deviation of 18,689,769.53 events, with the population distribution confirmed to exhibit an extreme positive skewness (Right-Skewed). Based on the Z-Score calculation with a threshold of $|Z| > 3$, a total of 6 datasets (approximately 1.6%) were identified as extreme anomalies. Furthermore, the sequential trend analysis and spatial visualization prove that these anomalies did not occur randomly but formed a systemic cluster centralized within a specific observation batch range. The integration of the Z-Score method and pattern recognition is proven effective in validating data quality (Veracity) and extracting operational information from massive-scale scientific datasets.

***Keywords**: Big Data, Cloud Computing, Anomaly Detection, Pattern Recognition, Z-Score.*

1. PENDAHULUAN

Fisika merupakan ilmu fundamental yang berperan sebagai bahasa universal dalam memahami hukum-hukum alam, mulai dari fenomena makroskopik seperti pergerakan benda langit hingga perilaku partikel subatomik [1]. Perkembangan teknologi informasi dalam beberapa dekade terakhir telah mendorong peningkatan volume data secara signifikan, khususnya dalam bidang fisika energi tinggi. Data yang dihasilkan dari eksperimen modern kini mencapai skala *Big Data*, yang tidak hanya ditandai oleh besarnya volume, tetapi juga kompleksitas, kecepatan pertumbuhan, serta keragaman data yang dihasilkan [2].

Dalam kondisi tersebut, para peneliti menghadapi tantangan dalam mengolah dan menganalisis data secara efektif. Meskipun data tersedia dalam jumlah yang sangat besar, informasi yang dapat diekstraksi tidak selalu sebanding. Fenomena ini dikenal dengan istilah *drowning in data but starving for knowledge*, yang menggambarkan kondisi di mana data melimpah tetapi pengetahuan yang diperoleh masih terbatas [2]. Oleh karena itu, diperlukan pendekatan baru yang mampu mengelola dan menganalisis data secara efisien.

Eksperimen fisika partikel yang dilakukan di CERN melalui *ATLAS Collaboration* menghasilkan dataset dalam jumlah yang sangat besar, yang terdiri dari jutaan hingga miliaran kejadian (*events*) dalam setiap eksperimen. Berdasarkan pengamatan terhadap ratusan dataset metadata CERN yang mencakup lebih dari satu miliar kejadian fisika, diperlukan suatu sistem yang mampu mengelola, menyimpan, dan menganalisis data tersebut secara efisien. Tanpa sistem yang terstruktur, data dalam jumlah besar justru dapat menjadi hambatan dalam proses penemuan pengetahuan baru.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

Secara tradisional, pengelolaan *Big Data* sering kali bergantung pada klaster fisik seperti *Hadoop Distributed File System* (HDFS). Namun, seiring dengan evolusi teknologi, pendekatan arsitektur *Decoupled Storage and Compute* berbasis komputasi awan (*Cloud Computing*) menawarkan fleksibilitas dan efisiensi yang lebih tinggi bagi peneliti. Melalui platform seperti *Google Colab*, proses data mentah dari *Application Programming Interface* (API) dapat langsung diolah secara *in-memory* menggunakan pustaka *data science* seperti *Pandas* dan *NumPy* tanpa memerlukan instalasi infrastruktur klaster lokal yang rumit.

Selain aspek penyimpanan, proses analisis data juga menjadi komponen penting dalam pengolahan *Big Data*. Pendekatan analisis data tradisional dinilai tidak lagi memadai untuk menangani data yang besar dan kompleks, sehingga diperlukan metode analisis yang lebih adaptif dan efisien [10]. Dalam kerangka *Knowledge Discovery in Database* (KDD), pemanfaatan *Big Data* memungkinkan identifikasi pola tersembunyi, prediksi tren, serta evaluasi risiko secara lebih akurat [9]. Salah satu teknik analisis krusial dalam domain ini adalah *anomaly detection*, yaitu metode untuk mengidentifikasi data yang menyimpang dari pola umum. Dalam bidang fisika partikel, deteksi anomali sangat penting karena dapat membantu menemukan kejadian langka maupun mendeteksi kesalahan kalibrasi dalam sistem pengukuran [12].

Berbagai penelitian sebelumnya telah dilakukan untuk mendukung pengolahan dan analisis data dalam skala besar. Ramadhani [2] membahas karakteristik *Big Data* serta tantangan dalam ekstraksi informasi. Pebralia [3] menunjukkan bahwa infrastruktur terdistribusi mampu meningkatkan keandalan sistem penyimpanan data. Guerrieri [5] menekankan pentingnya *anomaly detection* dalam mengidentifikasi pola tidak normal, sedangkan Ramdhan dan Oktova [4] menunjukkan efektivitas metode Z-Score dalam standarisasi data [13]. Selain itu, Marshall dkk. [6] mengembangkan format data spesifik untuk meningkatkan efisiensi pengolahan data eksperimen, dan Karamoy dkk. [7] membuktikan bahwa pemrosesan paralel dalam sistem terdistribusi mampu mempercepat analisis data secara signifikan. Penelitian lain juga menunjukkan potensi penemuan fenomena baru melalui analisis data eksperimen ATLAS di CERN [14].

Meskipun berbagai penelitian tersebut telah memberikan kontribusi yang signifikan, masih terdapat celah penelitian (*research gap*) yang perlu dikaji lebih lanjut. Sebagian besar penelitian cenderung berfokus secara terpisah antara arsitektur penyimpanan dan metode analisis dasar, tanpa mengintegrasikannya ke dalam ekosistem komputasi awan yang komprehensif. Selain itu, penerapan metode *Z-Score* pada dataset metadata fisika partikel berskala miliaran kejadian—yang dikombinasikan dengan analisis kemiringan distribusi (*skewness*) dan pengenalan pola sekuensial (*pattern recognition*)—masih belum banyak dieksplorasi secara mendalam.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk mengimplementasikan *Z-Score Outlier Detection* dipadukan dengan analisis *Pattern Recognition* pada metadata eksperimen ATLAS CERN berbasis *Cloud Computing*. Penelitian ini diharapkan mampu membuktikan efisiensi platform komputasi awan dalam pengolahan data masif, mempercepat proses analisis filtrasi anomali, serta menghasilkan informasi yang lebih akurat mengenai perilaku sistemik eksperimen pada data skala besar. Lebih lanjut, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode KDD di bidang sains komputasi serta menjadi referensi bagi penelitian interdisipliner selanjutnya.

2. METODOLOGI PENELITIAN

2.1 Jenis Penelitian dan Sumber Data

Penelitian ini menggunakan jenis penelitian sains terapan dengan pendekatan kuantitatif. Pendekatan ini dipilih karena penelitian berfokus pada pengolahan data numerik berskala besar untuk menghasilkan informasi yang objektif dan terukur. Metode yang digunakan adalah kerangka kerja *Knowledge Discovery in Database* (KDD) yang difokuskan pada *descriptive task*, yaitu proses mengekstraksi pola dan informasi tersembunyi dari dataset [2]. Metode ini dipilih karena mampu menangani karakteristik *Big Data* yang memiliki volume besar dan kompleksitas tinggi.

Sumber data berasal dari portal *CERN Open Data* yang menyediakan metadata eksperimen fisika partikel. Populasi penelitian mencakup seluruh metadata eksperimen CERN, namun penelitian ini dibatasi pada eksperimen ATLAS (*A Toroidal LHC ApparatuS*) karena representatif dalam menghasilkan data berskala besar. Teknik pengambilan sampel menggunakan *purposive sampling* dengan kriteria dataset yang berkaitan dengan produksi Boson Z pada energi pusat massa 13 TeV. Berdasarkan kriteria tersebut, diperoleh 373 dataset metadata dalam format JSON sebagai sampel penelitian.

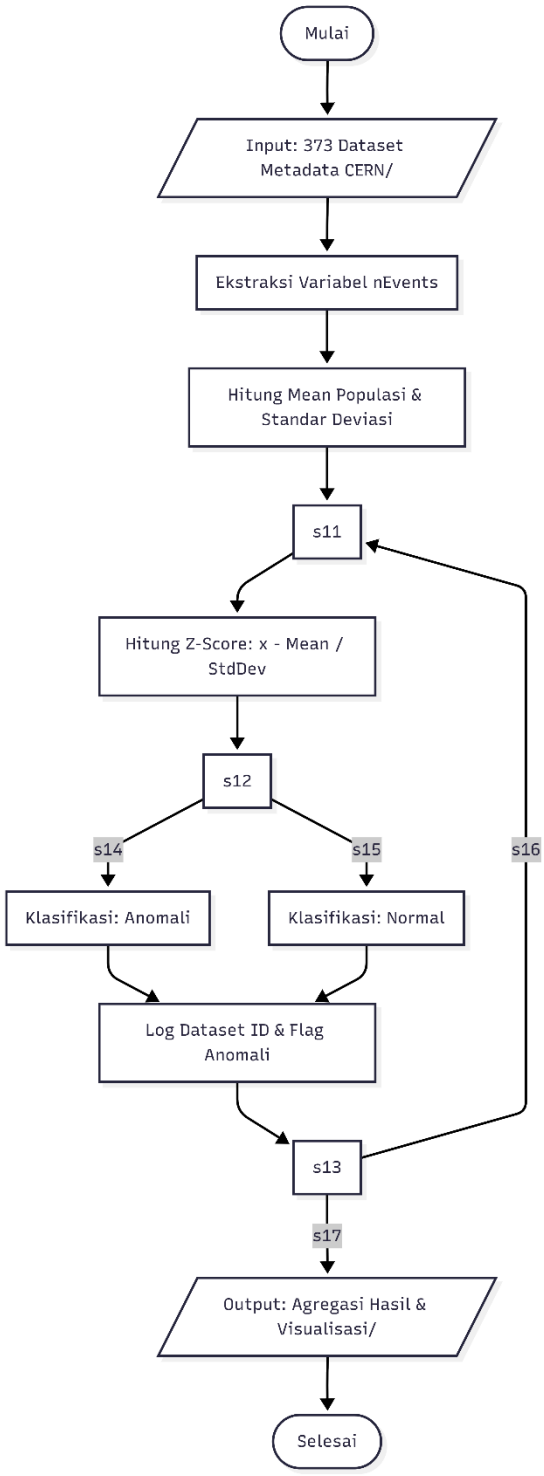
Secara keseluruhan, dataset tersebut merepresentasikan 1.893.071.770 kejadian fisika (*events*). Unit analisis mencakup setiap kejadian dalam metadata, sehingga memungkinkan pengujian terhadap kemampuan sistem dalam menangani data berskala besar, khususnya dalam aspek skalabilitas dan *fault tolerance*. Besarnya volume data juga menjadi dasar penggunaan metode *unsupervised learning* berbasis Z-Score untuk mendeteksi anomali.

Instrumen penelitian meliputi perangkat keras dan perangkat lunak berbasis komputasi awan. Perangkat keras berupa laptop digunakan sebagai terminal klien, sedangkan proses komputasi dilakukan melalui *Google Colab* sebagai *Integrated Development Environment* (IDE). Bahasa pemrograman yang digunakan adalah Python 3.x dengan pustaka Pandas dan NumPy untuk pengolahan data, Requests untuk pengambilan data, serta Matplotlib untuk visualisasi.

Pengambilan data dilakukan melalui *CERN Open Data API*, yang memungkinkan akses langsung ke repositori metadata eksperimen ATLAS. Melalui API ini, variabel seperti *nEvents* dapat diperoleh secara otomatis. Penggunaan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

API memastikan data yang digunakan bersifat valid, konsisten, dan berasal dari sumber resmi, sehingga mendukung keandalan proses analisis dalam penelitian ini.



Gambar 1. Alur Proses Analisis Dataset Metadata CERN

Logika algoritma serta seluruh rangkaian tahapan eksekusi sistem dalam penelitian ini dipetakan secara terperinci melalui diagram alir (*flowchart*) yang disajikan pada Gambar 1. Representasi visual ini mengadopsi standarisasi simbol ISO 5807 untuk memberikan distingsi yang jelas bagi setiap entitas operasional dalam jalur pemrosesan data (*data pipeline*). Penggunaan simbol-simbol tersebut bertujuan untuk menjamin ketepatan interpretasi terhadap setiap komponen prosedural, mulai dari batasan instruksi awal (*start*) dan akhir (*end*), mekanisme masukan (*input*) dan keluaran (*output*), hingga instruksi pemrosesan matematis dan logika percabangan.

Berdasarkan struktur diagram pada Gambar 1, tahapan operasional sistem diinisiasi dengan mekanisme ingestasi data berupa metadata mentah berformat JSON yang diperoleh secara otomatis dari portal *CERN Open Data*. Setelah

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

variabel deterministik seperti nEvents berhasil diekstraksi melalui proses *parsing* data, sistem kemudian mengeksekusi kalkulasi statistik deskriptif terhadap seluruh populasi data guna menentukan parameter rata-rata dan standar deviasi sebagai acuan *baseline* distribusi normal.

2.2 Teknik Pengumpulan Data dan Analisis Data

Teknik pengumpulan data dalam penelitian ini dilakukan melalui prosedur *automated data acquisition* (akuisisi data otomatis) menggunakan antarmuka pemrograman aplikasi (API) yang disediakan oleh *CERN Open Data Portal*. Proses ini diawali dengan tahap *filtering* untuk memastikan bahwa metadata yang diambil relevan dengan variabel penelitian, yaitu produksi Boson Z pada eksperimen ATLAS dengan tingkat energi 13 TeV. Pengambilan data dilakukan secara bertahap (*batching*) melalui skrip Python guna menghindari beban berlebih pada *server* penyedia data sekaligus memastikan stabilitas koneksi selama proses transmisi metadata.

Setiap dataset metadata ditarik dalam format JSON (*JavaScript Object Notation*), yang mencakup parameter teknis seperti jumlah kejadian (*nEvents*) dan probabilitas interaksi (*cross section*). Sebanyak 373 file metadata berhasil dikumpulkan dan dipindahkan ke dalam *Local Runtime Storage* pada lingkungan Google Colab. Proses pengumpulan ini juga melibatkan validasi awal untuk memastikan integritas data, di mana skrip secara otomatis melakukan pemeriksaan terhadap ketersediaan kunci-kunci data (*data keys*) yang diperlukan. Dengan total akumulasi mencapai 1.893.071.770 kejadian fisika, teknik pengumpulan ini membuktikan efisiensi penggunaan API dalam menarik informasi berskala besar untuk kepentingan riset saintifik tanpa harus melakukan pengunduhan data fisik yang berukuran petabyte.

Tabel 1. Atribut Dataset yang Digunakan dalam Penelitian

Nama Variabel	Tipe Data	Deskripsi
dataset_id	String	Identitas unik dataset
n-Events	Integer	Jumlah kejadian dalam dataset
Z_score	Float	Nilai hasil perhitungan Z-Score
Klasifikasi	String	Kategori hasil deteksi

Tabel 1 menyajikan definisi variabel yang digunakan dalam proses analisis dataset pada penelitian ini. Variabel *dataset_id* berfungsi sebagai identitas unik yang merepresentasikan setiap dataset, sehingga mempermudah proses identifikasi dan pengelolaan data. Variabel *nEvents* menunjukkan jumlah kejadian (events) yang terdapat dalam masing-masing dataset, yang selanjutnya digunakan sebagai parameter utama dalam analisis statistik.

Variabel *z_score* merupakan hasil transformasi data menggunakan metode Z-Score, yang digunakan untuk mengukur tingkat penyimpangan suatu nilai terhadap rata-rata distribusi dalam satuan standar deviasi. Nilai ini berperan penting dalam proses deteksi anomali, karena mampu mengidentifikasi data yang memiliki karakteristik berbeda secara signifikan dibandingkan dengan mayoritas data lainnya.

Adapun variabel *klasifikasi* merupakan hasil akhir dari proses analisis, yang digunakan untuk mengelompokkan data ke dalam kategori *anomali* atau *anomali ekstrem*. Penentuan klasifikasi ini didasarkan pada nilai ambang batas tertentu dari Z-Score yang telah ditetapkan sebelumnya. Dengan demikian, keempat variabel tersebut membentuk suatu kerangka analisis yang sistematis dalam mendukung proses identifikasi anomali pada dataset yang diteliti. Lebih lanjut, metode analisis diperluas dengan memanfaatkan modul *SciPy* untuk menghitung kemiringan distribusi (*Skewness*) dan *Matplotlib* untuk memvisualisasikan tren sekuensial berdasarkan urutan ID dataset. Hal ini bertujuan untuk mengevaluasi sifat persebaran observasi dan mengenali pola spasial dari anomali yang terdeteksi.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data Metadata

Pada bagian ini dipaparkan hasil temuan empiris yang diperoleh dari proses eksekusi algoritma Z-Score Outlier Detection dalam lingkungan komputasi awan Google Colab. Pengujian dilakukan dengan menerapkan tahapan *Knowledge Discovery in Database* (KDD) secara terotomasi terhadap kumpulan metadata eksperimen ATLAS yang bersumber dari pangkalan data CERN Open Data. Proses analisis ini tidak hanya bertujuan untuk mengidentifikasi pola distribusi data, tetapi juga untuk memahami karakteristik statistik, mengevaluasi kualitas data, serta mendeteksi potensi keberadaan anomali dalam dataset berskala besar.Dalam konteks Big Data, analisis metadata memiliki peran yang sangat penting karena metadata berfungsi sebagai representasi ringkas dari data utama yang dapat digunakan untuk memahami struktur, ukuran, dan karakteristik dataset tanpa harus mengakses keseluruhan data mentah. Oleh karena itu, analisis terhadap metadata menjadi langkah awal yang strategis dalam proses eksplorasi data, khususnya pada dataset eksperimen fisika partikel yang memiliki kompleksitas tinggi.

Selain itu, analisis terhadap distribusi data juga dapat dilakukan dengan mempertimbangkan rasio antara nilai rata-rata dan standar deviasi. Dalam penelitian ini, nilai standar deviasi yang lebih besar dibandingkan rata-rata

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

menunjukkan bahwa tingkat penyebaran data sangat tinggi. Kondisi ini mengindikasikan bahwa data memiliki variabilitas yang signifikan dan tidak terpusat pada satu nilai tertentu. Dalam konteks analisis Big Data, fenomena ini sering kali menjadi indikator bahwa dataset mengandung banyak variasi kondisi atau skenario yang berbeda.

Lebih lanjut, karakteristik distribusi yang tidak merata ini dapat memberikan dampak terhadap proses pengambilan keputusan berbasis data. Dataset dengan variasi yang tinggi cenderung menghasilkan analisis yang lebih kompleks karena adanya ketidakkonsistenan dalam pola data. Oleh karena itu, diperlukan metode yang mampu menyederhanakan kompleksitas tersebut tanpa menghilangkan informasi penting yang terkandung di dalamnya.

Dalam kaitannya dengan metode Z-Score, penggunaan standar deviasi sebagai parameter utama menjadikan metode ini sensitif terhadap perubahan distribusi data. Ketika terdapat nilai ekstrem dalam jumlah yang cukup signifikan, nilai rata-rata dapat bergeser sehingga memengaruhi hasil perhitungan Z-Score. Hal ini menunjukkan bahwa interpretasi hasil deteksi anomali harus dilakukan secara hati-hati dan tidak hanya bergantung pada nilai numerik semata.

Selain itu, penting untuk mempertimbangkan bahwa tidak semua anomali memiliki makna yang sama. Dalam beberapa kasus, anomali dapat dikategorikan sebagai *global outlier*, yaitu data yang secara keseluruhan berbeda jauh dari populasi utama. Namun, dalam kasus lain, anomali juga dapat berupa *contextual outlier*, yaitu data yang hanya dianggap anomali dalam kondisi tertentu. Dalam penelitian ini, metode Z-Score lebih cenderung mengidentifikasi *global outlier*, sehingga interpretasi hasil perlu disesuaikan dengan konteks data yang dianalisis.

Dari sudut pandang kualitas data, keberadaan anomali juga dapat digunakan sebagai indikator awal untuk mengevaluasi keandalan dataset. Dataset yang memiliki jumlah anomali yang terlalu banyak dapat mengindikasikan adanya masalah dalam proses pengumpulan data. Sebaliknya, jumlah anomali yang relatif sedikit, seperti yang ditemukan dalam penelitian ini, menunjukkan bahwa sebagian besar data berada dalam kondisi yang stabil dan dapat dipercaya.

Selain itu, analisis ini juga menunjukkan bahwa pendekatan statistik sederhana masih memiliki peran penting dalam pengolahan data berskala besar. Meskipun saat ini telah banyak dikembangkan metode berbasis *machine learning*, metode statistik seperti Z-Score tetap relevan karena memiliki keunggulan dalam hal interpretabilitas dan efisiensi komputasi. Hal ini menjadikannya sebagai metode yang cocok digunakan pada tahap awal analisis sebelum dilakukan pendekatan yang lebih kompleks.

Dengan demikian, hasil analisis pada bagian ini tidak hanya memberikan informasi mengenai distribusi data dan keberadaan anomali, tetapi juga memberikan pemahaman yang lebih mendalam mengenai karakteristik dataset metadata eksperimen fisika partikel. Pemahaman ini menjadi dasar penting dalam menentukan langkah analisis selanjutnya serta dalam mengembangkan metode yang lebih efektif untuk pengolahan data berskala besar.

3.1.1 Agregasi Data dan Analisis Statistik Deskriptif

Sistem komputasi menginisiasi proses pengumpulan data melalui pembacaan massal (*batch processing*) terhadap file metadata berformat JSON. Berdasarkan hasil rekaman eksekusi sistem (*console log*), proses agregasi berhasil memetakan karakteristik populasi data secara komprehensif tanpa mengalami kendala memori (*memory overflow*). Parameter statistik dari distribusi populasi data yang diperoleh dirangkum sebagai berikut:

a. Total dataset terproses: 374 file
b. Total kejadian fisika: 1.893.071.770 events
c. Rata-rata kejadian: 5.075.259,44 events
d. Standar deviasi: 18.689.769,53 events

Nilai standar deviasi yang sangat tinggi menunjukkan bahwa distribusi jumlah kejadian pada dataset memiliki tingkat penyebaran yang luas. Kondisi ini mengindikasikan bahwa data tidak terdistribusi secara merata, melainkan terdapat perbedaan yang signifikan antara sebagian besar dataset dengan sejumlah kecil dataset yang memiliki nilai jauh lebih besar. Dengan demikian, populasi metadata eksperimen menunjukkan adanya variasi yang tinggi sebelum dilakukan proses identifikasi anomali menggunakan metode statistik.

Karakteristik tersebut mencerminkan bahwa distribusi data cenderung tidak mengikuti pola distribusi normal, melainkan memiliki kecenderungan tidak simetris (*skewed distribution*). Hal ini diperkuat oleh adanya kemungkinan kemunculan nilai ekstrem yang relatif tinggi, yang dalam analisis statistik sering dikaitkan dengan distribusi *heavy-tailed*. Pola distribusi seperti ini umum ditemukan pada data berskala besar, khususnya dalam konteks eksperimen fisika partikel.

Selain itu, variasi data yang tinggi juga menunjukkan adanya heterogenitas dalam dataset metadata. Perbedaan ini dapat dipengaruhi oleh berbagai faktor, seperti variasi konfigurasi eksperimen, jenis interaksi partikel, serta kondisi operasional detektor pada saat pengambilan data. Oleh karena itu, setiap dataset tidak hanya merepresentasikan jumlah kejadian, tetapi juga mencerminkan kondisi eksperimen yang berbeda-beda.

Dari perspektif analisis data, kondisi ini memiliki implikasi penting terhadap proses pemodelan dan interpretasi. Dataset dengan tingkat variasi yang tinggi cenderung lebih kompleks untuk dianalisis menggunakan pendekatan statistik sederhana, karena keberadaan nilai ekstrem dapat memengaruhi parameter distribusi seperti rata-rata dan standar deviasi. Oleh sebab itu, diperlukan metode analisis yang mampu mengidentifikasi dan menangani nilai-nilai tersebut secara tepat.

Lebih lanjut, analisis statistik deskriptif pada tahap ini juga berfungsi sebagai langkah awal dalam proses validasi data. Dengan memahami pola distribusi dan tingkat penyebaran data, peneliti dapat mendeteksi potensi

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

ketidakwajaran sebelum dilakukan analisis lanjutan. Dalam penelitian ini, tingginya nilai standar deviasi menjadi indikator awal adanya variasi yang signifikan, sehingga mendukung perlunya penerapan metode deteksi anomali pada tahap berikutnya.

Dengan demikian, analisis statistik deskriptif tidak hanya memberikan gambaran umum mengenai kondisi dataset, tetapi juga menjadi dasar dalam menentukan pendekatan analisis yang tepat, khususnya dalam proses identifikasi anomali menggunakan metode Z-Score.

3.1.2 Identifikasi Anomali Berbasis Z-Score

Untuk mengidentifikasi keberadaan penyimpangan data secara lebih presisi, sistem menerapkan metode statistik Z-Score terhadap setiap dataset yang dianalisis. Metode ini digunakan untuk mengukur jarak suatu nilai terhadap rata-rata distribusi data dalam satuan standar deviasi.

Dengan menetapkan ambang batas absolut $|Z| > 3$, sistem menghitung bahwa batas atas normalitas kejadian berada pada kisaran 61,14 juta kejadian ($\mu + 3\sigma$). Dataset yang memiliki nilai di atas batas tersebut dikategorikan sebagai anomali atau pencilan (outlier).

Hasil komputasi menunjukkan bahwa terdapat 6 dataset yang teridentifikasi sebagai anomali. Rincian dataset tersebut disajikan pada Tabel 2

Tabel 2. Dataset Anomali Berdasarkan Nilai Z-Score

ID Dataset	Jumlah Kejadian (nEvent)	Skor Standard (Z-Score)	Klasifikasi Algoritma
cern_batch_700322	85403200	4.297963	Anomali
cern_batch_700343	158012000	8.182912	Anomali Ekstrem
cern_batch_700340	177913000	9.247719	Anomali Ekstrem
cern_batch_700325	85443000	4.300093	Anomali
cern_batch_700339	133882000	6.891831	Anomali Ekstrem
cern_batch_700342	124010000	6.363628	Anomali Ekstrem

Berdasarkan hasil pada Tabel 2, dataset dengan ID *cern_batch_700340* memiliki nilai penyimpangan tertinggi dengan Z-Score sebesar 9,247719. Nilai ini menunjukkan bahwa jumlah kejadian pada dataset tersebut (177.913.000 events) berada jauh di atas rata-rata populasi, dengan deviasi lebih dari sembilan kali standar deviasi. Kondisi ini mengindikasikan bahwa dataset tersebut merupakan anomali yang sangat signifikan dalam distribusi data.

Secara keseluruhan, identifikasi enam dataset sebagai anomali menunjukkan bahwa sebagian kecil data memiliki nilai yang sangat dominan dibandingkan mayoritas populasi. Hal ini memperkuat temuan sebelumnya bahwa distribusi data bersifat tidak homogen dan memiliki tingkat variasi yang tinggi. Dengan demikian, keberadaan anomali tidak hanya mencerminkan penyimpangan individu, tetapi juga memberikan gambaran mengenai karakteristik distribusi data secara keseluruhan.

Secara metodologis, penggunaan ambang batas $|Z| > 3$ mengacu pada prinsip distribusi normal, di mana sebagian besar data berada dalam rentang ±3 standar deviasi dari rata-rata. Oleh karena itu, data yang melampaui batas tersebut dapat dikategorikan sebagai kejadian yang jarang terjadi (*rare events*) dan memerlukan perhatian khusus dalam analisis.

Klasifikasi “anomali ekstrem” pada beberapa dataset dengan nilai Z-Score tinggi menunjukkan adanya tingkat penyimpangan yang jauh lebih besar dibandingkan anomali biasa. Hal ini mengindikasikan bahwa dataset tersebut memiliki karakteristik yang berbeda secara signifikan dari pola umum distribusi, yang dapat dipengaruhi oleh volume kejadian yang sangat besar atau kondisi eksperimen tertentu.

Dalam konteks analisis ilmiah, keberadaan anomali memiliki dua kemungkinan interpretasi, yaitu sebagai representasi fenomena langka yang bernilai tinggi atau sebagai indikasi kesalahan dalam proses pengumpulan data. Oleh karena itu, dataset yang teridentifikasi sebagai anomali perlu ditindaklanjuti melalui proses validasi dan analisis lanjutan untuk memastikan penyebabnya.

Selain itu, hasil deteksi ini juga memiliki implikasi terhadap pengolahan data selanjutnya. Dataset anomali tidak seharusnya langsung dihapus, melainkan perlu dianalisis secara terpisah agar informasi penting yang terkandung di dalamnya tidak hilang. Pendekatan ini penting terutama dalam konteks data eksperimen, di mana nilai ekstrem dapat memiliki signifikansi ilmiah yang tinggi.

Dari sisi komputasi, keberhasilan sistem dalam mengidentifikasi anomali secara otomatis menunjukkan bahwa metode Z-Score efektif diterapkan pada data berskala besar. Proses ini mendukung prinsip otomasi dalam kerangka kerja Knowledge Discovery in Database (KDD), khususnya dalam tahap analisis dan evaluasi data.

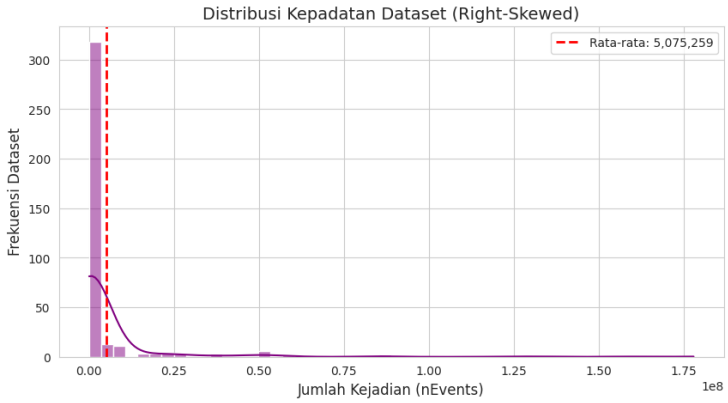
Dengan demikian, hasil identifikasi anomali tidak hanya menghasilkan klasifikasi data, tetapi juga memberikan pemahaman yang lebih mendalam mengenai struktur distribusi serta membuka peluang untuk eksplorasi lebih lanjut terhadap fenomena yang tidak umum dalam dataset.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

3.2 Analisis Distribusi Populasi (Skewness)

Selain menerapkan pemisahan kelas menggunakan metrik Z-Score, sistem juga melakukan evaluasi terhadap bentuk sebaran data untuk memahami karakteristik *Variety* pada himpunan metadata. Berdasarkan hasil komputasi statistik dan visualisasi kepadatan data (*KDE/Histogram*), ditemukan bahwa distribusi populasi memiliki tingkat kemiringan (*skewness*) positif yang sangat ekstrem.

Kurva distribusi secara visual bersifat asimetris dan condong ke arah kanan (*right-skewed*). Mayoritas absolut dari 374 dataset terkonsentrasi pada rentang frekuensi yang sangat padat di dekat area *baseline*. Namun, keberadaan 6 dataset anomali masif (hingga menyentuh 177 juta kejadian) menciptakan efek "ekor panjang" (*long-tail*) ke arah kanan yang menarik nilai rata-rata menjauhi nilai median populasi. Fenomena asimetris ini memberikan konfirmasi empiris mengenai tingginya tingkat keberagaman data operasional detektor ATLAS.



Gambar 2. Visualisasi histogram dan KDE yang menunjukkan distribusi populasi bersifat condong ke kanan (*Right-Skewed*).

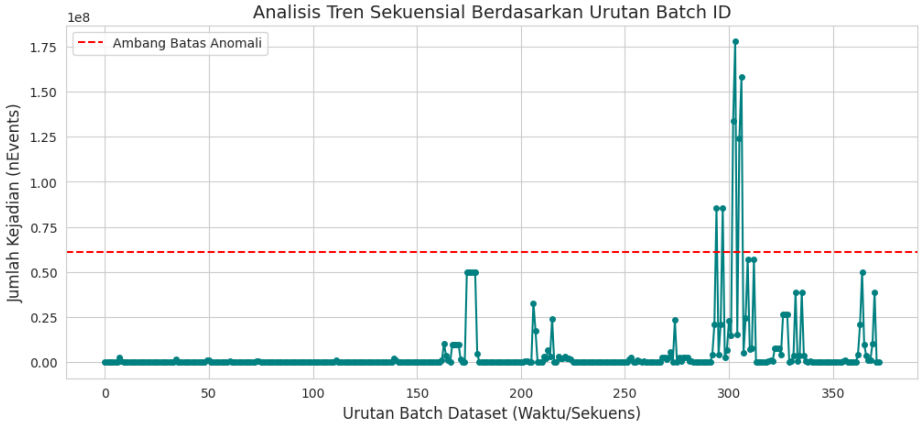
Dari visualisasi pada Gambar 2, dapat diamati secara jelas bahwa sebaran data sangat tidak simetris (asimetris). Karakteristik utama yang menonjol adalah adanya kemiringan positif (*positive skewness*) atau condong ke kanan (*right-skewed*). Puncak histogram (modus) berada sangat dekat dengan area kiri (nilai rendah), yang merepresentasikan mayoritas absolut dataset memiliki jumlah kejadian yang relatif kecil (di bawah 10 juta). Di sisi lain, terdapat ekor panjang (*long tail*) kurva KDE yang menjuntai jauh ke arah kanan. Keberadaan ekor panjang ini disebabkan oleh segelintir dataset anomali masif seperti pencilan ekstrem berlabel *Anomali: 302521* yang memiliki lebih dari 177 juta kejadian yang nilainya berada jauh di atas kerumunan populasi standar.

Kemiringan positif ini memiliki dampak statistik yang signifikan: pencilan ekstrem di sisi kanan bertindak sebagai penarik vertikal yang kuat, yang menyeret nilai rata-rata populasi (μ=5.075.259,44) menjauhi nilai median populasi. Gambar 2 secara spesifik menyoroti anomali tersebut serta garis ambang batas manual (garis merah), yang secara visual mempertegas keberadaan deviasi masif pada segelintir observasi. Fenomena asimetris ini memberikan konfirmasi empiris mengenai tingginya tingkat keberagaman (*Variety*) ukuran data eksperimen sains Big Data di laboratorium CERN, dan membuktikan bahwa dalam manajemen data berskala besar, pemahaman terhadap bentuk distribusi adalah krusial karena penggunaan rata-rata populasi saja dapat menyesatkan akibat pengaruh pencilan ekstrem yang sangat kuat.

3.2 Analisis Tren Sekuensial (Pattern Recognition)

Untuk menginvestigasi lebih jauh mengenai sifat dan karakteristik anomali yang telah teridentifikasi, sistem tidak hanya mengevaluasi besaran nilai penyimpangan, tetapi juga melakukan analisis pengenalan pola (*pattern recognition*). Analisis ini dilakukan dengan memetakan jumlah kejadian (*nEvents*) secara sekuensial berdasarkan urutan identitas *batch* dataset ke dalam bentuk grafik garis (*line chart*). Pemetaan tren waktu atau sekuens ini bertujuan secara spesifik untuk menguji hipotesis apakah kemunculan anomali dalam populasi data bersifat acak (*random error*) atau memiliki kecenderungan sistemik pada periode observasi tertentu.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939



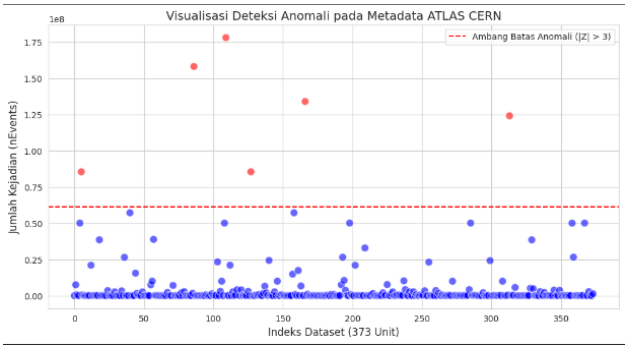
Gambar 3. Pemetaan tren sekuensial jumlah kejadian berdasarkan urutan identitas *batch* dataset.

Interpretasi spasial pada Gambar 3 mengungkapkan sebuah pola (*pattern*) sekuensial yang sangat terstruktur. Garis putus-putus berwarna merah merepresentasikan ambang batas anomali absolut, sementara garis hijau tosca menunjukkan pergerakan volume data dari 374 dataset yang diurutkan. Dapat diamati bahwa fluktuasi data pada sebagian besar indeks observasi (dari awal hingga mendekati indeks ke-300) cenderung datar dan stabil mendekati angka nol. Namun, seluruh titik data yang melampaui garis ambang batas merah tidak tersebar secara acak atau sporadis di sepanjang garis waktu, melainkan terklasterisasi secara sangat rapat pada satu jendela observasi yang spesifik. Klaster lonjakan ekstrem ini secara presisi berpusat pada rentang indeks observasi yang merujuk pada identitas *batch* 700322 hingga 700343.

Dalam kerangka kerja *Knowledge Discovery in Database* (KDD), penemuan pola ini menolak hipotesis awal bahwa anomali disebabkan oleh galat sensor yang terisolasi secara kebetulan. Sebaliknya, klasterisasi titik anomali yang beruntun ini membuktikan terjadinya fenomena yang bersifat sistemik dan persisten selama rentang waktu pemrosesan *batch* tersebut. Kondisi ini memberikan indikasi operasional yang kuat bahwa akselerator di laboratorium CERN kemungkinan besar sedang dioperasikan pada mode luminositas ekstrem atau sedang difokuskan untuk menangkap peristiwa interaksi partikel yang sangat intensif secara kontinu pada periode tersebut. Dengan demikian, penerapan teknik pengenalan pola ini terbukti tidak hanya mampu menyoroti anomali data, tetapi juga berhasil merekonstruksi informasi kontekstual yang esensial mengenai jalannya eksperimen fisika itu sendiri.

3.4 Implementasi

Untuk memvalidasi hasil kalkulasi matematis secara visual, sistem menghasilkan representasi spasial dari distribusi 374 dataset ke dalam bentuk grafik *Scatter Plot*. Visualisasi ini bertindak sebagai alat bantu penafsiran (*interpretation tool*) dalam metodologi KDD.



Gambar 4. Visualisasi Distribusi Dataset dan Batas Anomali

Pada Gambar 4, garis putus-putus berwarna merah merepresentasikan batas keputusan (*decision boundary*) dengan ambang |Z| = 3. Berdasarkan visualisasi tersebut, sebaran titik observasi terbagi menjadi dua klaster utama, yaitu data normal dan data anomali. Klaster pertama terdiri dari 368 titik berwarna biru yang terkonsentrasi padat di bawah batas ambang, menunjukkan bahwa mayoritas dataset berada dalam rentang nilai yang relatif stabil. Sebagian besar titik pada klaster ini terletak pada kisaran nol hingga sepuluh juta kejadian, yang mengindikasikan konsistensi dalam proses perekaman data eksperimen.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

Sebaliknya, klaster kedua terdiri dari enam titik berwarna merah yang terletak jauh di atas garis batas. Posisi titik-titik ini menunjukkan penyimpangan yang signifikan dibandingkan distribusi umum data. Jarak yang mencolok antara data anomali dan kelompok data normal mencerminkan tingkat deviasi yang tinggi, sekaligus mengonfirmasi hasil perhitungan Z-Score pada Tabel 2. Pola persebaran yang tidak merata ini juga menunjukkan bahwa dataset memiliki distribusi yang cenderung tidak simetris (*skewed*) dan mengandung nilai ekstrem.

Visualisasi scatter plot dalam penelitian ini berperan sebagai alat interpretasi yang memperjelas pola distribusi data secara intuitif. Kepadatan titik pada area tertentu merepresentasikan dominasi data normal, sedangkan titik yang terpisah jauh mengindikasikan keberadaan anomali. Dengan demikian, posisi relatif setiap titik terhadap pusat distribusi secara langsung menggambarkan tingkat penyimpangannya terhadap rata-rata.

Keberadaan batas anomali berbasis $|Z| = 3$ memberikan dasar keputusan yang objektif dalam membedakan data normal dan anomali. Hal ini memastikan bahwa interpretasi visual tetap konsisten dengan pendekatan statistik yang digunakan, sehingga hasil analisis tidak bersifat subjektif, melainkan terukur dan dapat dipertanggungjawabkan secara ilmiah.

Dari sisi implementasi, penggunaan Google Colab memungkinkan seluruh proses analisis dilakukan secara terintegrasi, mulai dari pengolahan data hingga visualisasi. Lingkungan ini mendukung efisiensi komputasi serta reproduktibilitas hasil, sehingga analisis dapat dijalankan kembali dengan kondisi yang sama tanpa perubahan hasil yang signifikan.

Selain itu, visualisasi ini juga mendukung tahap *interpretation and evaluation* dalam metodologi Knowledge Discovery in Database (KDD). Dengan adanya representasi visual, peneliti dapat lebih mudah memahami pola distribusi data serta mengevaluasi efektivitas metode deteksi anomali yang digunakan.

Dengan demikian, visualisasi yang dihasilkan tidak hanya berfungsi sebagai representasi grafis, tetapi juga sebagai sarana validasi yang memperkuat hasil analisis statistik serta memberikan pemahaman yang lebih komprehensif terhadap struktur dataset.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa implementasi algoritma *Z-Score Outlier Detection* berbasis komputasi awan (*Cloud Computing*) melalui lingkungan Google Colab terbukti sangat efektif dalam mengeksekusi kerangka kerja *Knowledge Discovery in Database* (KDD) berskala masif. Sistem berhasil memproses 374 dataset metadata eksperimen ATLAS yang mengakumulasi 1,89 miliar kejadian fisika secara stabil. Analisis statistik deskriptif mengungkap tingginya variansi data dengan nilai standar deviasi melebihi 18,6 juta *events*, yang secara empiris menunjukkan adanya ketimpangan signifikan antara mayoritas dataset eksperimen standar dengan segelintir observasi bernilai ekstrem. Melalui penetapan ambang batas $|Z| > 3$, penelitian ini berhasil mengidentifikasi 6 dataset (sekitar 1,6%) yang tergolong sebagai pencilan ekstrem, dengan dataset cern_batch_700340 mencatat penyimpangan tertinggi sebesar $Z = 9{,}24$. Analisis lanjutan terhadap kepadatan data membuktikan bahwa distribusi populasi memiliki karakteristik yang sangat tidak homogen dan bersifat *Right-Skewed* (condong ke kanan) secara ekstrem. Hal ini membuktikan bahwa metode Z-Score tidak hanya tangguh dalam memfilter data mentah, tetapi juga esensial untuk memahami bentuk distribusi *Big Data* yang asimetris pada eksperimen fisika energi tinggi. Lebih lanjut, integrasi visualisasi spasial (*scatter plot*) dan analisis pengenalan pola (*pattern recognition*) memberikan dimensi interpretasi yang jauh lebih komprehensif. Pemetaan tren sekuensial secara meyakinkan membuktikan bahwa keenam anomali tersebut tidak terjadi secara sporadis atau akibat galat acak (*random error*), melainkan membentuk sebuah klaster sistemik yang terpusat pada rentang *batch* 700322 hingga 700343. Dengan demikian, penelitian ini membuktikan bahwa pendekatan terotomasi ini sangat efisien tidak hanya dalam memvalidasi kualitas observasi (*Veracity*), tetapi juga mampu mengekstraksi informasi operasional berharga mengenai dinamika sistem eksperimen di CERN guna mendukung riset sains fisika yang lebih akurat dan terukur.

REFERENCES

[1] I. K. Mahardika, S. Handono, E. Ernasari, H. A. Rofida, F. Zahro, dan M. A. Seftiyani, "Hakikat fisika sebagai pilar kehidupan," *Jurnal Pendidikan Ilmiah Transformatif*, vol. 7, no. 12, pp. 30–34, Des. 2023. [Online].

[2] D. D. Ramadhani, F. Fuad, dan D. Nurmalitasari, "Pengaruh hambatan ontogenik terhadap keefektifan pembelajaran statistika dalam Kurikulum Merdeka," *Jurnal Media Akademik (JMA)*, vol. 3, no. 1, Jan. 2025, doi: 10.62281/v3i1.1505.

[3] J. Pebralia, "Orientasi gerak jatuh bebas pada sistem tiga partikel," *Jurnal Ilmu Fisika dan Pembelajarannya (JIFP)*, vol. 1, no. 2, pp. 55–58, Des. 2017, doi: 10.19109/jifp.v1i2.2835.

[4] M. Ramdhan dan R. Oktova, "Pengembangan media pembelajaran tentang fisika partikel elementer menggunakan Android Studio Bumblebee untuk mahasiswa S-1 Pendidikan Fisika," *Jurnal Inovasi dan Pembelajaran Fisika*, vol. 11, no. 1, pp. 96–104, 2024, doi: 10.3670/jipf.v11i1.313.

[5] G. Guerrieri, on behalf of the ATLAS Collaboration, "Open Data at ATLAS: Bringing TeV collisions to the World," *EPJ Web of Conferences*, vol. 337, 01293, 2025, doi: 10.1051/epjconf/202533701293.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1790-1799
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.939

[6] Z. Marshall, L. A. Heinrich, M. Lassnig, M. I. Vivas Albornoz, dan S. Willocq, "The first release of ATLAS Open Data for Research," *EPJ Web of Conferences*, vol. 337, 01345, 2025, doi: 10.1051/epjconf/202533701345.

[7] P. V. Karamoy, D. A. Tulandi, dan P. M. Silangen, "Efektivitas penggunaan File Manager pada pembelajaran dinamika partikel," *Charm Sains: Jurnal Pendidikan Fisika*, vol. 4, no. 2, pp. 98–104, 2023, doi: 10.53682/charmsains.v4i2.251.

[8] R. Ajie dan T. Sutabri, "Optimasi kinerja basis data terdistribusi menggunakan algoritma replication dan partitioning," *Jurnal Sains Student Research*, vol. 3, no. 2, pp. 371–378, 2025, doi: 10.61722/jssr.v3i2.4318.

[9] B. W. Al Aluf, A. Hernawan, dan G. S. Nugraha, "Teknik distributed Naive Bayes untuk analisis sentimen ulasan pelanggan Amazon," *Universitas Mataram*, 2025.

[10] P. Hendikawati, N. A. S. Harwanti, Wardono, A. Prabowo, M. D. Zahra, dan W. R. Saefurrochman, "Pembelajaran Big Data di perguruan tinggi: Potensi masa depan, faktor pendukung dan penghambat," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 3, pp. 2232–2241, 2025, doi: 10.29100/jipi.v10i3.7711.

[11] A. Wibowo, *Pengantar AI, Big Data dan Ilmu Data*. Semarang: Yayasan Prima Agus Teknik bekerja sama dengan Universitas STEKOM, 2023.

[12] M. R. Aditya dan C. Dewi, "Optimisasi pengecekan anomali pada proses job: Analisis waktu dan data untuk identifikasi anomali yang efisien," *Jurnal Indonesia: Manajemen Informatika dan Komunikasi (JIMIK)*, vol. 5, no. 2, pp. 1819–1832, 2024, doi: 10.35870/jimik.v5i2.737.

[13] I. M. Karo Karo dan Hendriyana, "Klasifikasi penderita diabetes menggunakan algoritma machine learning dan Z-score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, 2022, doi: 10.54914/jtt.v8i2.564.

[14] V. Bansal et al., "ATLAS sensitivity to leptoquarks, W_R and heavy Majorana neutrinos in final states with high pt dileptons and jets with early LHC data at 14 TeV proton proton collisions," *ATLAS Collaboration preprint*, arXiv:0910.2215, 2009.

[15] F. D. Utami dan F. D. Astuti, "Comparison of Hadoop MapReduce and Apache Spark in data processing with Hgrid247-DE," *Journal of Applied Informatics and Computing (JAIC)*, vol. 8, no. 2, pp. 390–399, 2024, doi: 10.30871/jaic.v8i2.8557.