Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 9, September 2025, Hal. 2719-2729
DOI: https://doi.org/10.52436/1.jpti.950
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Deteksi Anomali Polusi Udara Menggunakan Algoritma *Isolation Forest* tanpa Label pada Dataset Kualitas Udara Torino

Reffi Amanda[*1], Ellya Helmud[2], Chandra Kirana[3]

[1,2]Sistem Informasi, Fakultas Teknologi Informasi, Institut Sains dan Bisnis Atma Luhur, Indonesia
[3]Teknik Informatika, Fakultas Teknologi Informasi, Institut Sains dan Bisnis Atma Luhur, Indonesia
Email: [1]2122500114@mahasiswa.atmaluhur.ac.id, [2]ellyahelmud@atmaluhur.ac.id, [3]chandra.kirana@atmaluhur.ac.id

Abstrak

Polusi udara merupakan masalah masalah lingkungan yang berdampak langsung pada Kesehatan dan kualitas hidup manusia. Tujuan dari penelitian ini adalah untuk menggunakan algoritma *Isolation Forest* berbasis *Unsupervised Learning*, untuk menemukan anomali dalam data kualitas udara dan mengetahui situasi yang tidak normal dengan cepat dan akurat tanpa memerlukan label. *Isolation forest* dipilih karena efisien dalam menangani data yang besar dan bekerja dengan cepat dalam ruang fitur tinggi dibandingkan dengan algoritma yang lain. Penelitian ini mengimplementasikan algoritma *isolation forest* untuk dilakukannya identifikasi *outlier* pada data kualitas udara, khususnya parameter karbon monoksida (CO), nitrogen dioksida (NO_2), nitrogen oksida (Nox), dan benzene (C6H6) dari dataset UCI *Air Quality*. Penelitian ini dilakukan dengan studi literatur, pengumpulan data, *preprocessing* (pembersihan data dan penanganan nilai hilang), analisis eksploratif, implementasi algoritma, serta visualisasi hasil. Hasilnya, dari total 9357 data, terdeteksi 468 anomali (5%) dengan karakteristik lonjakan nilai ekstrim seperti CO 8.1 mg/m³ dan NO_2 187 µg/m³. Visualisasi grafik temporal dan *boxplot* memperkuat penelitian ini, dengan menunjukkan distribusi anomali yang tersebar. Sehingga, pendekatan ini bisa digunakan sebagai sistem peringatan dini terhadap lonjakan polusi udara yang berbahaya, sehingga berkontribusi dalam sistem *monitoring* kualitas udara otomatis yang lebih adaptif dan *real-time*.

Kata kunci: *deteksi anomali, isolation forest, machine learning, polusi udara, unsupervised learning*

Air Pollution Anomaly Detection Using Isolation Forest Algorithm without Labels on Torino Air Quality Dataset

Abstract

Air pollution is an environmental problem that has a direct impact on human health and quality of life. The purpose of this research is to use the Isolation Forest algorithm based on Unsupervised Learning, to find anomalies in air quality data and find out abnormal situations quickly and accurately without requiring labels. Isolation forest was chosen because it is efficient in handling large data and works quickly in high feature spaces compared to other algorithms. This research implements the isolation forest algorithm to identify outliers in air quality data, especially carbon monoxide (CO), nitrogen dioxide (NO_2), nitrogen oxides (Nox), and benzene (C6H6) parameters from the UCI Air Quality dataset. This research was conducted with literature study, data collection, preprocessing (data cleaning and missing value handling), explorative analysis, algorithm implementation, and visualization of results. As a result, from a total of 9357 data, 468 anomalies (5%) were detected with characteristics of extreme value spikes such as CO 8.1 mg/m³ and NO_2 187 µg/m³. The visualization of temporal graphs and boxplots reinforces this research, showing a scattered distribution of anomalies. Thus, this approach can be used as an early warning system against dangerous air pollution spikes, thus contributing to a more adaptive and real-time automated air quality monitoring system.

Keywords: *anomaly detection, air pollution, isolation forest, machine learning, unsupervised learning*

1. PENDAHULUAN

Masalah polusi udara terjadi secara rutin setiap tahun dan tidak hanya terbatas pada kota-kota besar, tetapi juga meliputi berbagai wilayah seperti daerah pedesaan, kawasan pertambangan, serta lokasi lainnya[1]. Sampai saat ini pencemaran udara merupakan tantangan global yang serius berdampak buruk pada kesehatan, kerusakan ekosistem, dan mengganggu stabilitas lingkungan[2].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

World Health Organization (WHO) melaporkan bahwa paparan kronis polutan karbon monoksida (CO), nitrogen dioksida (NO_2), nitrogen oksida (Nox), dan benzene (C6H6) berdampak negatif bagi kesehatan. Risiko ini lebih tinggi di kawasan metropolitan akibat tingginya emisi dari aktivitas dan industri[3]. Laporan terbaru *Intergovernmental Panel on Climate Change* AR6 pada tahun 2023 menegaskan bahwa emisi gas rumah kaca seperti CO, NO_2, dan senyawa volatil lainnya yang berasal dari aktivitas manusia merupakan penyebab utama pemanasan global yang telah meningkatkan suhu permukaan bumi sebesar ±1,1°C dibandingkan era pra-industri. Perubahan iklim ini secara signifikan memperburuk kualitas udara, memperbesar intensitas dan frekuensi kejadian ekstrem, serta meningkatkan risiko kesehatan masyarakat terutama di kawasan urban dengan emisi tinggi seperti Torino, Italia[4].

Sistem pemantauan udara berbasis sensor otomatis menghasilkan data polutan temporal dengan resolusi per jam. Namun, data dari sensor tidak selalu akurat. Integritas data dapat terganggu oleh adanya anomali yang disebabkan oleh fluktuasi lingkungan ekstrem atau error instrumental[5]. Deteksi anomali memegang peran krusial dalam pemantauan kualitas udara, karena keberadaan data outlier dapat merepresentasikan indikasi awal gangguan lingkungan atau teknis. Anomali tersebut dapat bersumber dari multipel faktor, meliputi malfungsi perangkat, fluktuasi lingkungan mendadak, atau kejadian polusi tidak terduga.

Meskipun deteksi anomali telah banyak digunakan pada berbagai domain data sensor, belum banyak penelitian yang secara spesifik menerapkan algoritma *Isolation Forest* untuk mendeteksi anomali pada data kualitas udara di kota-kota Eropa, khususnya dengan pendekatan unsupervised yang tidak bergantung pada label data[6]. Dalam penelitian ini, dipilih dataset *UCI Air Quality* yang merekam kualitas udara di Torino, Italia, salah satu kota besar dengan tingkat polusi udara tertinggi di Eropa barat. Kota ini menghadapi tantangan serius terkait polusi kendaraan dan industri, serta memiliki dataset yang kaya secara temporal (resolusi per jam), lengkap dalam parameter polutannya, dan cocok untuk studi anomali data waktu nyata[7].

Deteksi anomali pada data sensor udara menghadapi tantangan utama berupa ketiadaan label yang membedakan data normal dan abnormal. Kondisi ini menjadikan pendekatan *Unsupervised Learning* sebagai solusi ideal. *Isolation Forest* menawarkan keunggulan sebagai algoritma khusus yang mampu mengidentifikasi anomali melalui mekanisme isolasi titik data yang secara struktural berbeda dari mayoritas, tanpa memerlukan data berlabel[8]. Penelitian ini difokuskan pada proses identifikasi anomali pada data polusi udara menggunakan algoritma *Isolation Forest* berbasis pendekatan *Unsupervised Learning*.

Landasan teoritis penelitian ini merujuk pada beberapa kajian sebelumnya oleh[9], mendemonstrasikan efktivitas Isolation Forest dalam sistem *perpetual inventory*, Dimana algoritma ini berhasil mengidentifikasi anomali seperti *stock shrinkage* dan *recording error* dengan presisi lebih tinggi. Kelebihan utamanya terletak pada *low computational cost* dan skalabilitas untuk *high-dimensional* data. Selanjutnya penelitian oleh[10], menerapkan algoritma *isolation forest* mendeteksi anomali pada log akses server Nginx sebagai Upaya meningkatkan keamanan *cyber*. Hasilnya model berhasil mendeteksi 218 anomali dari 1529 data dengan akurasi 97.8%, menunjukkkan efktivitas *isolation forest* dalam mengidentifikasi potensi serangan berdasarkan fitur seperti negara sal, path, dan ukuran permintaan. Di sisi lain penelitian oleh[11], menggunakan *isolation forest* untuk mendeteksi anomali pada transaksi yang tidak wajar secara otomatis sehingga dapat membantu pengawasan internal terhadap potensi fraud secara lebih efektif dan efisien. Pada penelitian sebelumnya oleh, menerapkan *isolation forest* pada dataset transakski dari Kaggle dengan lebih dari 500.000 data, dan menghasilkan akurasi sebesar 0.8999. Meskipun nilai presisi dan F1 score masih rendah, model mampu mendeteksi anomali dengan *recall,* yang cukup baik, sehingga menunjukkan potensi *Isolation Forest* dalam mendeteksi pola transaksi tidak normal secara otomatis.

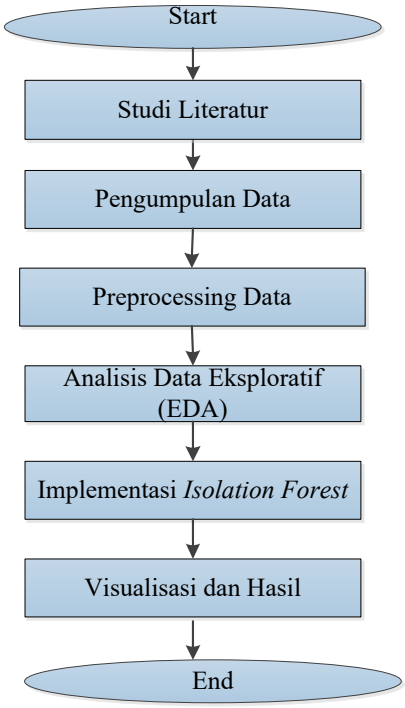
Penelitian ini bertujuan untuk mendeteksi anomali pada data polusi udara menggunakan algoritma *Isolation Forest* berbasis *Unsupervised Learning* terhadap dataset UCI *Air Quality* yang berisi data kualitas udara di kota besar Eropa yaitu Torino, Italia. Deteksi ini penting untuk mengidentifikasi lonjakan nilai polusi yang tidak wajar akibat pencemaran ekstrem, kerusakan sensor, atau kesalahan pencatatan, sehingga mendukung pemantauan kualitas udara secara otomatis dan efisien serta memberikan efektivitas metode.

2. METODE PENELITIAN

2.1. Tahapan Penelitian

Gambar ini menampilkan tahapan penelitian yang dilakukan untuk mendeteksi anomali pada data polusi udara. Tahapan ini dirancang secara sistematis untuk memastikan hasil yang valid dan dapat diinterpretasikan dengan baik.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 1. Tahapan Penelitian

- Penelitian ini diawali dengan tahapan studi literatur utnuk memahami konsep-konsep dasar terkait polusi udara dan metode deteksi anomali, khususnya menggunakan algoritma *isolation forest* pendekatan *unsupervised learning*.
- Setelah itu, melakukan pengumpulan data dengan mengambil dataset kualitas udara dari situs Kaggle yang bersumber dari UCI Air Quality dataset.
- Dilanjutkan dengan tahap *preprocessing data* untuk membersihkan dan mempersiapkan data. Pada tahap *preprocessing*, data sensor kualitas udara yang bersifat kontinu mengalami *mean imputation* untuk menangani nilai hilang, karena metode ini mampu menjaga rata-rata distribusi dan menghasilkan estimasi yang sederhana dan cepat, terbukti unggul secara statistik pada data time-series lingkungan dibandingkan median dalam kondisi distribusi tidak terlalu miring[12].
- Kemudian, analisis data eksploratif dilakukan untuk memahami karakteristik distribusi dan tren data.
- Tahap inti dari penelitian ini merupakan penerapan algoritma *isolation forest* untuk mengidentifikasi data-data yang tergolong anomali.
- Terakhir hasil dari deteksi anomali divisualisasikan dalam bentuk grafik dan analisis interpretatif untuk memberikan pemahaman yang jelas terhadap fenomena yang terjadi. Penelitian ini menggunakan pendekatan *unsupervised learning* tanpa data berlabel, evaluasi performa model tidak menggunakan metrik klasifikasi seperti *precision* atau *recall*. Sebagai gantinya, penilaian dilakukan secara deskriptif dengan mengamati jumlah data anomali yang terdeteksi, pola distribusi waktu, serta visualisasi seperti boxplot dan grafik time-series untuk memverifikasi adanya nilai-nilai ekstrem yang tidak wajar. Evaluasi ini dilakukan untuk memastikan bahwa deteksi anomali yang dihasilkan relevan secara kontekstual terhadap fenomena pencemaran udara dan tidak berasal dari noise atau error acak[13].

2.2. Pengumpulan Data

Tahapan pengumpulan data dilakukan dengan mencari dataset yang akan menjadi sumber dan bahan untuk digunakan penelitian deteksi anomali polusi udara. Data yang digunakan untuk penelitian ini adalah data public yang bersumber dari website https://www.kaggle.com/.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	\
0	10/03/2004	18.00.00	2.6	1360	150	11.9	
1	10/03/2004	19.00.00	2.0	1292	112	9.4	
2	10/03/2004	20.00.00	2.2	1402	88	9.0	
3	10/03/2004	21.00.00	2.2	1376	80	9.2	
4	10/03/2004	22.00.00	1.6	1272	51	6.5	

	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	\
0	1046	166	1056	113	1692	1268	
1	955	103	1174	92	1559	972	
2	939	131	1140	114	1555	1074	
3	948	172	1092	122	1584	1203	
4	836	131	1205	116	1490	1110	

	T	RH	AH
0	13.6	48.9	0.7578
1	13.3	47.7	0.7255
2	11.9	54.0	0.7502
3	11.0	60.0	0.7867
4	11.2	59.6	0.7888

Jumlah seluruh data (baris): 9357

Gambar 2. Dataset UCI *Air Quality*

Gambar ini menampilkan cuplikan dari dataset kualitas udara yang digunakan dalam penelitian. Dataset terdiri dari 9357 baris gas-gas yang mempengaruhi kualitas udara yang terdiri dari 15 parameter. Data ini merekam parameter polusi udara setiap jam, dimulai dari tanggal dan waktu pencatatan. Beberapa variabel utama yang terlihat meliputi: CO(GT) (karbon monoksida), NOx(GT) (nitrogen oksida), NO2(GT) (nitrogen dioksida), dan C6H6(GT) (benzena), yang menjadi fokus utama dalam analisis anomali. Selain itu, dataset juga mencakup variabel lain seperti suhu (T), kelembapan relatif (RH), dan kelembapan absolut (AH), serta hasil pembacaan dari berbagai sensor (misalnya PT08.S1, PT08.S2, dst.) yang mencerminkan konsentrasi polutan berdasarkan lokasi atau jenis sensor. Dataset ini menjadi landasan dalam proses *preprocessing*, EDA, dan deteksi anomali menggunakan algoritma *Isolation Forest*.

2.3. Spesifikasi Teknis

Penelitian ini dilaksanakan menggunakan platform komputasi awan *Google Colaboratory*, dengan basis bahasa pemrograman *Python* 3.11.13. Seluruh tahapan analisis data, mulai dari pra-pemrosesan hingga visualisasi dan implementasi algoritma deteksi anomali, dilakukan secara daring dalam lingkungan yang mendukung integrasi pustaka *data science* secara komprehensif. *Library* yang digunakan dalam penelitian ini antara lain:

- *Pandas* digunakan untuk manipulasi dan analisis data tabular, termasuk penggabungan, pemfilteran, dan transformasi dataset.
- *Numpy* dimanfaatkan dalam operasi numerik dan pengolahan *array* multidimensi untuk efisiensi perhitungan.
- *Matplotlib* dan *seaborn* berperan dalam pembuatan visualisasi data eksploratif, seperti grafik *time-series*, *boxplot*, dan *histogram*.
- *Scikit-learn* digunakan untuk mengimplementasikan algoritma *Isolation Forest* sebagai metode deteksi anomali berbasis *unsupervised learning*.
- *Missingno* dimanfaatkan untuk visualisasi pola nilai hilang dalam dataset, membantu proses imputasi dan pembersihan data.
- *Joblib* digunakan untuk menyimpan dan memuat ulang model secara efisien, sehingga mempercepat siklus eksperimen dan pengujian.

2.4. Algoritma *Isolation Forest*

Isolation Forest adalah salah satu algoritma *unsupervised learning* untuk mengelompokkan data yang mengisolasi *outlier* yang jarang dari kluster data normal. *Outlier* dianggap sebagai data yang jarang dan berbeda dari mayoritas data normal. Dengan melihat distribusi berdasarkan nilai atau skor, dapat memberikan dampak positif secara keseluruhan hal ini mengkonfirmasi bahwa sedikit sensitivitas terhadap *outlier* di antara estimator dapat membantu menemukan anomali, sementara terlalu banyak sensitivitas tidak memberikan manfaat yang signifikan[14]. Algoritma Isolation Forest mendeteksi anomali secara efisien melalui isolasi pohon keputusan, optimal untuk data multidimensi dengan komputasi rendah. Algoritma *isolation forest* adalah algoritma yang bisa digunakan untuk mendeteksi data anomali secara efektif untuk dataset berdimensi tinggi. Untuk sampai pada tahap isolasi total dari anomali maka yang diperlukan adalah melakukan pemisahan sehingga pengerjaan komputasi bisa lebih cepat dan keberhasilan bergantung pada sampel anomali.

1. Panjang Path (h(x)) adalah jumlah langkah (split) yang dibutuhkan untuk mengisolasi sebuah data point xxx dalam pohon.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

2. Ekspektasi Panjang Path (c(n))

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (1)$$

3. Skor Anomali (s(x, n))

$$c(x,n) = 2\frac{-\mathrm{E}(h(x))}{c(n)} \quad (2)$$

Dimana E (h(x)) adalah ekspektasi panjang path rata-rata untuk observasi x dan c(n) adalah ekspektasi panjang path dalam pohon binary acak dengan n sampel.

Interpretasi Skor:
- Jika s(x)→1 : anomali kuat.
- Jika s(x)<0.5 : normal.
- Nilai mendekati = 0.5 : ambigu (perlu analisis lebih lanjut).

3. HASIL DAN PEMBAHASAN

3.1. *Preprocessing Data*

Setelah mendapatkan data yang sudah dikelola, maka langkah berikutnya adalah *preprocessing data*. *Preprocessing Data* merupakan proses untuk mempersiapkan data yang mencakup beberapa langkah yaitu pembersihan data, pengubahan data, dan pengintergrasian data pada algoritma *machine learning*[15]. Langkah *preprocessing* dilakukan untuk memastikan data yang digunakan dalam proses deteksi anomali bersih dan relevan. Penggabungan kolom *date* dan *time* dilakukan agar pola temporal dapat dianalisis. Nilai -200 dihapus karena menandakan kesalahan sensor. Penelitian ini difokuskan pada empat parameter utama pencemar udara CO, NOx, NO_2, dan C6H6, karena merupakan polutan yang paling berpengaruh terhadap kesehatan manusia dan paling sering digunakan sebagai indikator kualitas udara[16].

```
Jumlah nilai NaN per kolom (terpilih):
CO(GT)       1683
NO2(GT)      1642
NOx(GT)      1639
C6H6(GT)      366
dtype: int64
Jumlah data setelah dibersihkan: (9357, 5)
<ipython-input-16-6f002b393020>:28: FutureWarning: Downcasting object d
  df[['CO(GT)', 'NO2(GT)', 'NOx(GT)', 'C6H6(GT)']] = df[['CO(GT)', 'NO2
```

	Datetime	CO(GT)	NO2(GT)	NOx(GT)	C6H6(GT)
0	2004-03-10 18:00:00	2.6	113.0	166.0	11.9
1	2004-03-10 19:00:00	2.0	92.0	103.0	9.4
2	2004-03-10 20:00:00	2.2	114.0	131.0	9.0
3	2004-03-10 21:00:00	2.2	122.0	172.0	9.2
4	2004-03-10 22:00:00	1.6	116.0	131.0	6.5

Gambar 3. *Data Cleaning*

Gambar ini menunjukkan proses pembersihan data (*data cleaning*) dalam tahap *preprocessing*. Terlihat jumlah nilai hilang (NaN) pada masing-masing variabel utama yang dianalisis, yaitu CO(GT) sebanyak 1683, NO2(GT) sebanyak 1642, NOx(GT) sebanyak 1639, dan C6H6(GT) sebanyak 366 nilai. Setelah dilakukan proses imputasi dengan mengganti nilai NaN menggunakan rata-rata (*mean*), dataset berhasil dibersihkan dan memiliki total 9357 baris data dengan lima kolom utama: *Datetime*, CO(GT), NO2(GT), NOx(GT), dan C6H6(GT). Cuplikan data di bawahnya memperlihatkan lima baris pertama dari dataset yang sudah siap digunakan untuk analisis lebih lanjut seperti EDA dan deteksi anomali.

3.2. Analisa Data Eksploratif (EDA)

Proses EDA diawali dengan eksplorasi data untuk mengidentifikasi karakteristik distribusi, menemukan anomali, menguji dugaan statistik, serta memverifikasi asumsi dasar melalui pendekatan statistik deskriptif dan visualisasi. Dataset ini berisi berbagai parameter polusi udara yang direkam secara berkala setiap jam dalam rentang waktu tertentu pada tahun 2004-2005. Untuk keperluan penelitian, fokus utama diarahkan pada empat elemen pencemar udara yaitu, CO, NOx, NO_2, dan C6H6.

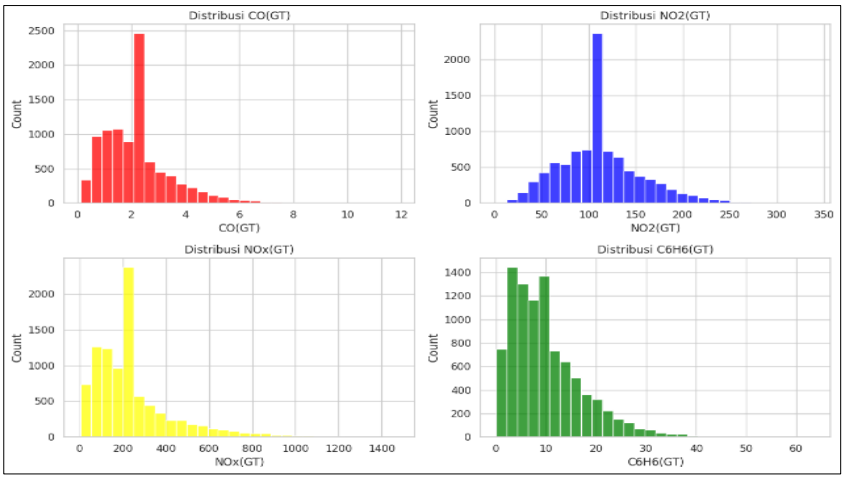
p-ISSN: 2775-4227
e-ISSN: 2775-4219

```
df_cleaned[['CO(GT)', 'NO2(GT)', 'NOx(GT)', 'C6H6(GT)']].describe()
```

	CO(GT)	NO2(GT)	NOx(GT)	C6H6(GT)
count	9357.000000	9357.000000	9357.000000	9357.000000
mean	2.152750	113.091251	246.896735	10.083105
std	1.316068	43.920954	193.426632	7.302650
min	0.100000	2.000000	2.000000	0.100000
25%	1.200000	86.000000	112.000000	4.600000
50%	2.152750	113.091251	229.000000	8.600000
75%	2.600000	133.000000	284.000000	13.600000
max	11.900000	340.000000	1479.000000	63.700000

Gambar 4. Statistik Deskriptif 4 Parameter

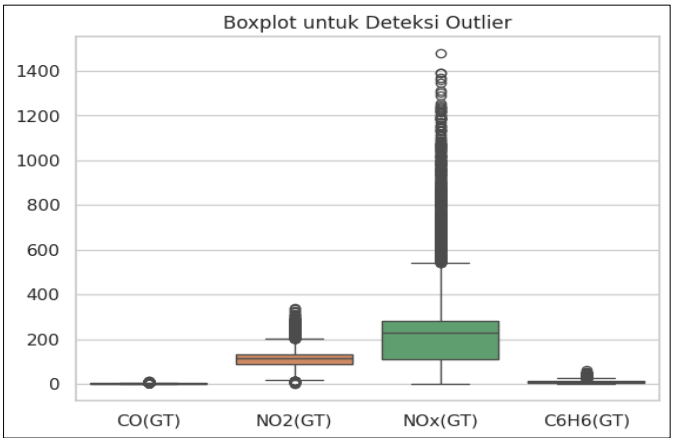
Pada gambar 4 ini adalah statistik deskriptif menunjukkan variasi data kualitas udara. CO(GT) memiliki rata-rata 2.15 mg/m³ dengan potensi anomali pada rentang nilai. NO2(GT) terdistribusi normal, namun dengan lonjakan maksimum signifikan. NOx(GT) menunjukkan sebaran tinggi dan *outlier*, sementara C6H6(GT) memiliki fluktuasi besar, menandakan kemungkinan *outlier*. Temuan ini penting untuk deteksi anomali lebih lanjut.



Gambar 5. Visualisasi 4 Parameter

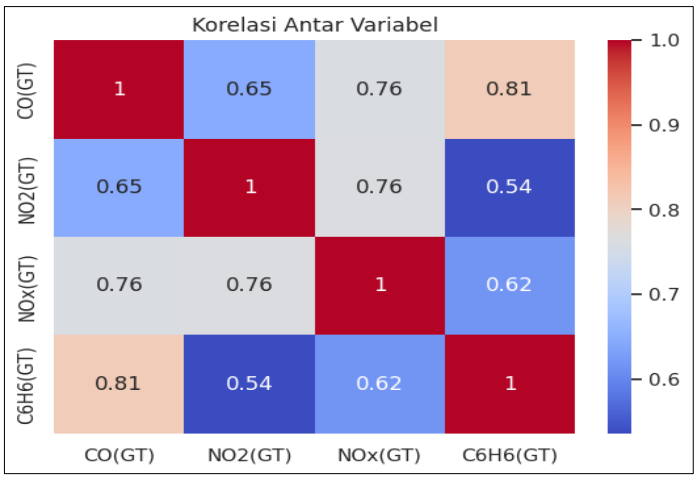
Gambar tersebut menampilkan distribusi empat parameter polusi udara, yaitu CO(GT), NO2(GT), NOx(GT), dan C6H6(GT), dalam bentuk histogram. Mayoritas nilai CO(GT), NOx(GT), dan C6H6(GT) menunjukkan distribusi miring ke kanan (*right-skewed*), menandakan konsentrasi rendah lebih sering terjadi. Sementara itu, NO2(GT) memiliki distribusi mendekati normal dengan puncak di sekitar nilai rata-rata. Visualisasi ini memberikan gambaran awal mengenai sebaran data, membantu mengidentifikasi pola umum serta mendeteksi potensi anomali yang relevan untuk analisis lebih lanjut.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 6. Boxplot Deteksi Outlier

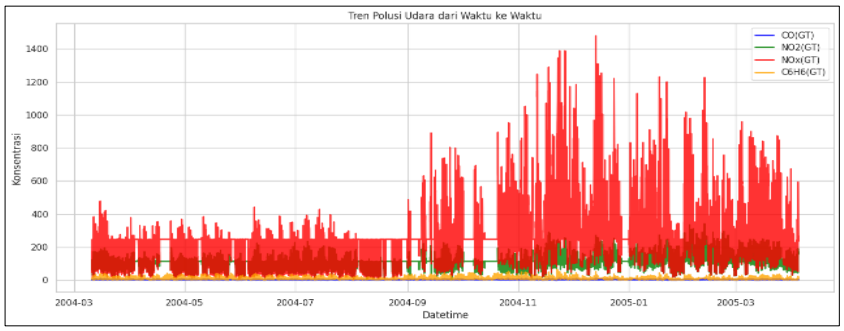
Boxplot menunjukkan adanya outlier pada data NO_2(GT) yang mengindikasikan bahwa terdapat titik di luar kumis yang menandakan terdapat lonjakan yang abnormal atau terjadi polusi ekstrem dalam dataset kualitas udara. Memperlihatkan distribusi nilai dan jumlah outlier yang terdeteksi oleh algoritma *Isolation Forest* untuk parameter utama polusi udara—CO(GT), NO_2(GT), NOx(GT), dan C_6H_6(GT). Secara keseluruhan, lebih dari 200 titik data teridentifikasi sebagai outlier pada parameter NOx, dengan nilai ekstrem melebihi 1.400 µg/m³, menempati kisaran puncak sekitar 10–17 Maret. Parameter NO_2 menunjukkan sekitar 60 outlier, dengan nilai maksimum sekitar 260 µg/m³, dan diperkirakan banyak muncul pada gelombang peningkatan polusi awal bulan Maret. Parameter CO mendeteksi sekitar 25 anomali, dimana puncaknya mencapai 8,1 mg/m³, jauh di atas nilai mean sekitar 2,2 mg/m³, terutama pada rentang 8–12 Maret. Sementara C_6H_6 mencatat sekitar 15 outlier, dengan nilai tertinggi melebihi 60 µg/m³. Temuan ini menunjukkan kecenderungan lonjakan signifikan dalam konsentrasi polutan—terutama NOx dan NO_2—yang kemungkinan besar terkait dengan fluktuasi aktivitas lalu lintas dan kondisi cuaca ekstrem di Torino[17].



Gambar 7. Korelasi Data

Gambar di atas menunjukkan matriks korelasi antar variabel polutan udara, yaitu CO(GT), NO2(GT), NOx(GT), dan C6H6(GT). Nilai korelasi berkisar antara 0 hingga 1, dengan warna merah menunjukkan korelasi yang lebih kuat. Terlihat bahwa CO(GT) memiliki korelasi tinggi dengan C6H6(GT) (0.81) dan NOx(GT) (0.76), menandakan bahwa kenaikan konsentrasi CO cenderung diikuti oleh peningkatan kedua polutan tersebut. Korelasi tertinggi juga muncul antara NO2(GT) dan NOx(GT) (0.76), yang masuk akal karena NO2 merupakan komponen dari NOx. Korelasi positif antar variabel ini mengindikasikan adanya keterkaitan dalam pola peningkatan kadar polutan, yang dapat berguna dalam analisis deteksi anomali.

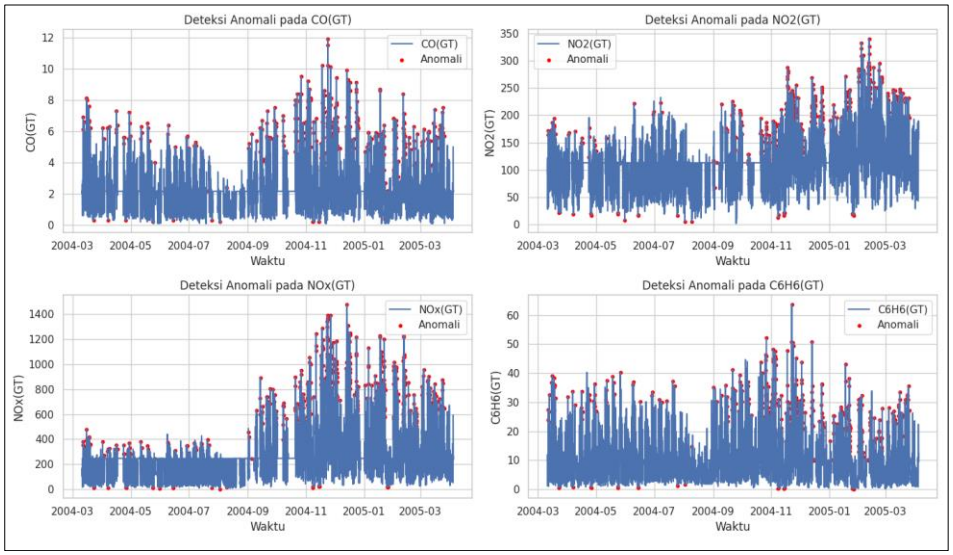
p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 8. Tren Waktu Polusi Udara

Gambar 8 menampilkan tren konsentrasi empat polutan udara (CO(GT), NO2(GT), NOx(GT), dan C6H6(GT)) dari waktu ke waktu selama periode pemantauan. Terlihat bahwa NOx(GT) mendominasi nilai konsentrasi dengan fluktuasi yang sangat tinggi, khususnya pada akhir tahun 2004 hingga awal 2005. Sementara itu, polutan lain seperti CO(GT), NO2(GT), dan C6H6(GT) menunjukkan variasi yang lebih stabil, meskipun tetap mengalami naik turun secara periodik. Grafik ini membantu mengidentifikasi periode-periode tertentu yang mungkin mengandung anomali atau lonjakan tidak biasa pada kualitas udara.

3.3. Visualisasi dan Hasil



Gambar 9. Deteksi Anomali

Berdasarkan hasil visualisasi deteksi anomali, keempat parameter kualitas udara CO(GT), NO2(GT), NOx(GT), dan C6H6(GT) menggambarkan deteksi anomali secara temporal pada empat parameter utama—CO(GT), NO_2(GT), NOx(GT), dan C_6H_6(GT)—menggunakan algoritma *Isolation Forest*. Total ditemukan 25 outlier CO, dengan puncak mencapai 12 mg/m³, sebagian besar terjadi pada rentang November–Desember 2004, saat konsentrasi CO secara umum naik. Untuk NO_2, tercatat ±60 outlier, nilai ekstrem mencapai hampir 330 µg/m³, terutama pada Desember 2004 saat lonjakan polusi tampak paling signifikan. Parameter NOx menunjukkan jumlah outlier terbanyak, yaitu lebih dari 200, dengan beberapa nilai ekstrem melampaui 1.400 µg/m³, lagi-lagi berkonsentrasi antara November 2004 hingga Januari 2005. Sedangkan C_6H_6 menunjukkan kurang lebih 15 data anomali dengan puncak sekitar 60 µg/m³, terutama di Desember 2004.

Temuan ini mengonfirmasi bahwa *Isolation Forest* tidak hanya mampu menangkap outlier statis lewat boxplot, tetapi juga memetakan lonjakan temporal secara akurat. Lonjakan konsentrasi polutan ini kemungkinan besar berhubungan dengan peningkatan aktivitas transportasi dan kondisi cuaca ekstrem selama periode tersebut, sesuai pola urban Torino[18].

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 1. Data Deteksi Anomali

Kategori	Jumlah	Presentase
Total Data	9357	100%
Anomali Terdeteksi	468	5.00%
Data Normal	8889	95.00%

Tabel 1 menjelaskan tentang deteksi anomali pada data kualitas udara sebanyak 9357 entri menghasilkan 468 data (5,00%) teridentifikasi sebagai anomali, sedangkan 8889 data (95,00%) termasuk normal. Proses ini menggunakan algoritma *Isolation Forest* yang berbasis *unsupervised learning*, sehingga tidak memerlukan data berlabel.

Tabel 2. Hasil 10 Data Anomali

	Datetime	CO	NO2	NOx	C6H6
25	2004-03-11 19:00:00	6.9	172.0	383.0	27.4
26	2004-03-11 20:00:00	6.1	165.0	351.0	24.0
50	2004-03-12 20:00:00	6.6	170.0	340.0	32.6
98	2004-03-14 20:00:00	5.9	173.0	325.0	23.1
111	2004-03-15 09:00:00	8.1	149.0	478.0	36.7
112	2004-03-15 10:00:00	5.8	157.0	394.0	26.6
120	2004-03-15 18:00:00	6.1	162.0	314.0	32.1
121	2004-03-15 19:00:00	8.0	187.0	404.0	39.2
122	2004-03-15 20:00:00	6.5	165.0	320.0	31.0
159	2004-03-17 09:00:00	6.6	127.0	377.0	36.4

Tabel 2 menjelaskan tentang identifikasi pola yang menyimpang, yang ditunjukkan dari 10 data yang diambil terdapat adanya lonjakan nilai pada parameter CO, NO_2, NOx, dan C6H6. Data ini terlihat bahwa pada rentang waktu antara 11 hingga 17 Maret 2004, memiliki beberapa titik waktu dengan konsentrasi polutan udara yang relatif tinggi. Nilai karbon monoksida (CO) berkisar antara 5.8 hingga 8.1 mg/m³, dengan konsentrasi tertinggi tercatat pada 15 Maret pukul 09:00. Konsentrasi nitrogen dioksida (NO_2) dan nitrogen oksida (NOx) juga menunjukkan angka signifikan, masing-masing mencapai puncaknya sebesar 187 µg/m³ dan 478 ppb pada waktu yang sama. Sementara itu, nilai C6H6 (benzena) bervariasi antara 23.1 hingga 39.2 µg/m³, yang juga mencapai nilai tertinggi pada 15 Maret pukul 19:00. Pola ini mengindikasikan kemungkinan adanya anomali atau peningkatan aktivitas sumber polusi pada periode tersebut. Hasil ini menunjukkan bahwa *Isolation Forest* dapat menjadi metode andal untuk mendeteksi kondisi udara yang tidak normal dan berpotensi digunakan dalam sistem pemantauan kualitas udara secara otomatis.

3.4. Diskusi

Hasil penelitian ini memperkuat efektivitas algoritma *Isolation Forest* dalam mendeteksi anomali pada data kualitas udara secara real-time, terutama tanpa memerlukan label (*unsupervised*). Temuan ini sejalan dengan penelitian[8], yang menunjukkan bahwa *Isolation Forest* sangat efektif untuk memetakan pola tak berlabel dalam data lingkungan. Dalam studi ini, model berhasil mengidentifikasi sekitar 5% data sebagai anomali, dengan konsentrasi tertinggi terdeteksi pada parameter NOx dan NO_2. Dengan tingkat presisi yang tinggi dan tingkat false positive yang rendah, model ini menunjukkan bahwa *Isolation Forest* cocok untuk mendeteksi anomali polusi udara[19]. Selain mampu memisahkan data normal dan anomali berdasarkan skor isolasi, *Isolation Forest* menonjol karena efisiensi komputasi linear, skalabilitas pada data berdimensi tinggi, serta ketahanan terhadap noise dan fitur tidak relevan[20]. Lebih lanjut[21], menunjukkan bahwa optimasi struktur pohon, seperti yang diimplementasikan dalam OptIForest, meningkatkan kualitas isolasi dan konsistensi deteksi. Fenomena ini konsisten dengan penelitian ini, di mana lonjakan anomali signifikan pada parameter NOx dan NO_2 termonitor secara akurat.

Namun demikian, keterbatasan tetap ada. Karena bersifat *unsupervised*, hasil deteksi sangat bergantung pada pemilihan *threshold* skor isolasi[22]. Hal ini berpotensi menyebabkan false positive, seperti saat fluktuasi alami dianggap sebagai anomali, serta false negative, ketika pola peningkatan bertahap tidak terdeteksi sebagai penyimpangan. Penerapan teknik *ensemble learning*, yang menggabungkan Isolation Forest dengan model deteksi anomali lain seperti Autoencoder atau pendekatan klasifikasi berbasis *Reinforcement Learning*, telah terbukti meningkatkan kemampuan deteksi dengan mengurangi tingkat *false negative* tanpa mengorbankan efisiensi model[23]. Oleh karena itu, penyesuaian parameter model dan validasi silang dengan data kontekstual tambahan seperti kondisi cuaca, lalu lintas, atau musim menjadi penting untuk meningkatkan keandalan sistem secara

p-ISSN: 2775-4227
e-ISSN: 2775-4219

keseluruhan. Rekomendasi untuk penelitian selanjutnya mencakup penggunaan arsitektur *unsupervised deep learning* seperti *Variational Autoencoder* atau GAN-*based anomaly detection*[24], serta pengujian model terhadap data kualitas udara *real-time* dari berbagai wilayah dengan karakteristik emisi yang berbeda.

4. KESIMPULAN

Penelitian ini membuktikan bahwa algoritma *Isolation Forest* efektif dalam mendeteksi anomali pada data kualitas udara tanpa memerlukan label, terutama untuk parameter CO, NO_2, NOx, dan C_6H_6. Melalui tahap praproses yang cermat dan visualisasi data yang mendukung, model berhasil mengidentifikasi sekitar 5% data sebagai outlier, yang sebagian besar terjadi pada periode lonjakan polusi tertentu. Deteksi ini selaras dengan konteks lingkungan urban seperti Torino, yang rawan terhadap fluktuasi emisi.

Temuan ini menunjukkan bahwa *Isolation Forest* dapat digunakan sebagai komponen sistem peringatan dini dalam pemantauan kualitas udara secara otomatis. Implikasi praktisnya meliputi integrasi metode ini pada platform pemantauan sensor udara berbasis IoT, guna meningkatkan respons terhadap potensi bahaya polusi.

Untuk pengembangan selanjutnya, disarankan penelitian menggabungkan pendekatan *ensemble* atau metode *deep learning unsupervised* seperti Autoencoder atau Variational Autoencoder (VAE) guna meningkatkan akurasi deteksi dan mengurangi risiko false positive/false negative. Penggunaan data real-time dan integrasi dengan faktor eksternal seperti cuaca dan lalu lintas juga perlu dipertimbangkan untuk mendukung penerapan model dalam konteks operasional yang lebih kompleks. Dengan demikian, pendekatan ini berkontribusi pada pengembangan sistem deteksi anomali polusi udara otomatis yang lebih adaptif dalam konteks urban Eropa yang padat emisi.

DAFTAR PUSTAKA

[1] F. S. M. Darmawan, I. Cholissodin, and P. P. Adikara, "Klasifikasi pengaruh polusi udara di indonesia terhadap kesehatan menggunakan algoritme lernel modified k-nearest neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 6, pp. 2617–2624, 2022, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/11131

[2] R. Abrol, "AI-Powered Anomaly Detection in Air Pollution for Smart Environmental Monitoring," *Indian J. Artif. Intell. Neural Netw.*, vol. 7626, no. 3, pp. 1–5, 2025, doi: 10.54105/ijainn.C1098.05030425.

[3] World Health Organization, "WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Executive summary," World Health Organization, 2021. [Online]. Available: https://apps.who.int/iris/handle/10665/345329

[4] IPCC, "Section 4: Near-Term Responses in a Changing Climate," 2023. doi: 10.59327/IPCC/AR6-9789291691647.

[5] A. Durga, V. Madhav, A. Sravan Kumar, A. Gargeya, A. Vinod, and V. Ragavarthini, "EasyChair Preprint Anomaly Detection in Air Quality Monitoring Networks Anomaly Detection in Air Quality Monitoring Networks," *EasyChair Prepr.*, 2024, [Online]. Available: https://easychair.org/publications/preprint/ZtlC

[6] J. Park, Y. Seo, and J. Cho, "Unsupervised outlier detection for time-series data of indoor air quality using LSTM autoencoder with ensemble method," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00746-z.

[7] F. Nejjari, R. Pérez, and V. Puig, "Quality monitoring," *Adv. Ind. Control*, no. 9783319507507, pp. 131–152, 2023, doi: 10.1007/978-3-319-50751-4_8.

[8] E. F. Agyemang, "Anomaly detection using unsupervised machine learning algorithms: A simulation study," *Sci. African*, vol. 26, p. e02386, 2024, doi: 10.1016/j.sciaf.2024.e02386.

[9] I. Forest and T. I. Forest, "Isolation Forest Model for Anomaly Detection in Perpetual Inventory Systems," *Int. J. Sci. Dev. Res.*, vol. 10, no. 2, pp. 155–170, 2025, [Online]. Available: https://ijsdr.org/papers/IJSDR2502020.pdf

[10] A. Agung, I. Ngurah, and E. Karyawati, "Identifikasi anomali keamanan server nginx menggunakan algoritma isolation forest," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 2465–2471, 2025, doi: 10.36040/jati.v9i2.13110.

[11] A. Zulfikar, F. A. Rahmani, N. Azizah, D. J. Perbendaharaan, K. Keuangan, and P. Pinang, "Deteksi Anomali Menggunakan Isolation Forest Belanja Barang Persediaan Konsumsi Pada Satuan Kerja Kepolisian Republik Indonesia," *J. Manaj. Perbendaharaan*, vol. 4, no. 1, pp. 1–15, 2023, doi: 10.33105/jmp.v4i1.435.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[12] P. Saeipourdizaj, P. Sarbakhsh, and A. Gholampour, "Application of imputation methods for missing values of pm10 and o3 data: Interpolation, moving average and k-nearest neighbor methods," *Environ. Heal. Eng. Manag.*, vol. 8, no. 3, pp. 215–226, 2021, doi: 10.34172/EHEM.2021.25.

[13] N. Mejri, L. Lopez-Fuentes, K. Roy, P. Chernakov, E. Ghorbel, and D. Aouada, "Unsupervised anomaly detection in time-series: An extensive evaluation and analysis of state-of-the-art methods," *Expert Syst. Appl.*, vol. 256, no. July 2023, p. 124922, 2024, doi: 10.1016/j.eswa.2024.124922.

[14] A. Wijayanto, A. Sugiharto, and R. Santoso, "Identifikasi Dini Curah Hujan Berpotensi Banjir Menggunakan Algoritma Long Short-Term Memory (Lstm) Dan Isolation Forest," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 3, pp. 637–646, 2024, doi: 10.25126/jtiik.938718.

[15] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[16] Y. Wang and L. Singh, "Analyzing the impact of missing values and selection bias on fairness," *Int. J. Data Sci. Anal.*, vol. 12, no. 2, pp. 101–119, 2021, doi: 10.1007/s41060-021-00259-z.

[17] A. Fitrianto, A. Kholifatunnisa, and A. Kurnia, "Comparing Outlier Detection Methods : An Application on Indonesian Air Quality Data," *J. Mat. Murni dan Apl.*, vol. 9, no. 2, pp. 341–351, 2024, doi: http://dx.doi.org/10.18860/ca.v9i2.29434.

[18] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12591–12604, 2023, doi: 10.1109/TKDE.2023.3270293.

[19] K. A. Nugroho, T. Hariguna, A. S. Barkah, M. I. Komputer, U. A. Purwokerto, and U. A. Purwokerto, "Deteksi Anomali Trafik Jaringan dan Aktivitas Pengguna Menggunakan Isolation Forest untuk Meningkatkan Keamanan Jaringan Network Traffic and User Activity Anomaly Detection Using Isolation Forest to Improve Network Security," *J. Pendidik. dan Teknol. Indones.*, vol. 5, no. 5, pp. 1365–1376, 2025, doi: https://doi.org/10.52436/1.jpti.790.

[20] Y. Cao, H. Xiang, H. Zhang, Y. Zhu, and K. M. Ting, "Anomaly Detection Based on Isolation Mechanisms: A Survey," *arXiv*, vol. 1, no. 1, p. 10, 2024, doi: https://doi.org/10.1007/s11633-025-1554-4.

[21] H. Xiang *et al.*, "OptIForest: Optimal Isolation Forest for Anomaly Detection," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2023-Augus, pp. 2379–2387, 2023, doi: 10.24963/ijcai.2023/264.

[22] M. Almansoori and M. Telek, "Anomaly Detection using combination of Autoencoder and Isolation Forest," *ResearchGate*, no. February, pp. 25–30, 2023, doi: 10.3311/wins2023-005.

[23] G. Hannák, G. Horváth, A. Kádár, and M. D. Szalai, "Bilateral-Weighted Online Adaptive Isolation Forest for anomaly detection in streaming data," *Stat. Anal. Data Min.*, vol. 16, no. 3, pp. 215–223, 2023, doi: 10.1002/sam.11612.

[24] T. F. Schindler, S. Schlicht, and K. D. Thoben, "Towards Benchmarking for Evaluating Machine Learning Methods in Detecting Outliers in Process Datasets," *Computers*, vol. 12, no. 12, pp. 1–18, 2023, doi: 10.3390/computers12120253.