

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

KAJIAN DETEKSI ANOMALI INDEKS KUALITAS UDARA JAKARTA DENGAN *LONG SHORT-TERM MEMORY* DAN *EXTREME GRADIENT BOOSTING*

MUHAMMAD RIZKY NURHAMBALI



**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Kajian Deteksi Anomali Indeks Kualitas Udara Jakarta dengan *Long Short-Term Memory* dan *Extreme Gradient Boosting*” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2025

Muhammad Rizky Nurhambali
G1501231089



RINGKASAN

MUHAMMAD RIZKY NURHAMBALI. Kajian Deteksi Anomali Indeks Kualitas Udara Jakarta dengan *Long Short-Term Memory* dan *Extreme Gradient Boosting*. Dibimbing oleh YENNI ANGRAINI dan ANWAR FITRIANTO.

Salah satu kasus lingkungan yang memungkinkan adanya anomali adalah indeks kualitas udara (IKU). IKU dapat berubah secara drastis seperti saat pandemi. Pembatasan kegiatan manusia saat pandemi maupun faktor meteorologi, seperti hujan dan angin menyebabkan langit Jakarta menjadi cerah dan menyebabkan masyarakat lebih peduli tentang kasus pencemaran udara. IKU yang dicatat dalam rentang waktu menjadikan data IKU sebagai data deret waktu. Oleh karena itu, untuk melakukan deteksi anomali harus dilakukan dengan pendekatan deret waktu.

Pendeteksian anomali pada deret waktu berkembang pesat dari metode konvensional hingga *machine learning*. Namun, metode *machine learning* dianggap lebih unggul karena kemampuannya dalam menangani data kompleks, dan nonlinear, serta dapat mengabaikan asumsi yang ada pada metode konvensional. Metode *machine learning* yang banyak berkembang antara lain *long short-term memory* (LSTM) dan *extreme gradient boosting* (XGBoost). Kedua metode terkenal akan kemampuannya dalam menangani berbagai kondisi data deret waktu. Akan tetapi, faktor-faktor yang memengaruhi keakuratan metode belum banyak dikaji dalam penelitian. Oleh karena itu, penelitian ini bertujuan untuk mengkaji kinerja LSTM dan XGBoost melalui kajian empiris dan simulasi, mengidentifikasi faktor meteorologi yang memengaruhi keberadaan anomali, dan mengidentifikasi faktor-faktor yang memengaruhi keakuratan kedua metode pada data IKU Jakarta melalui beragam skenario.

Penelitian ini menggunakan data IKU Jakarta yang bersumber pada AirNow (www.airnow.gov) sebagai peubah respons dan data meteorologi berupa kelembapan, angin, dan suhu yang bersumber pada NASA Power (<https://power.larc.nasa.gov/data-access-viewer/>) sebagai peubah penjelas. Data tersebut merupakan data per jam menggunakan zona waktu Indonesia bagian barat (WIB / GMT+7) dengan periode 1 Januari 2018 pukul 00:00 s.d. 31 Desember 2023 pukul 23:00. Data IKU Jakarta dilabelkan dengan *moving range* (MR) dan dikombinasikan dengan aturan sigma. IKU Jakarta divalidasi dengan LSTM dan XGBoost lalu dilakukan klasifikasi anomali dengan ambang batas sisaan hasil validasi.

Kajian data empiris menunjukkan LSTM sebagai metode yang lebih baik dan sesuai dibandingkan dengan XGBoost. Meskipun kajian data empiris menambahkan *feature engineering* pada kedua metode, hasil yang diperoleh masih menunjukkan metode LSTM sebagai metode terbaik meskipun ada penurunan nilai metrik evaluasi yang digunakan. Oleh karena itu, kombinasi MR (2), 4-sigma, dan tanpa *feature engineering* pada LSTM merupakan kombinasi terbaik untuk deteksi anomali IKU Jakarta pada data empiris dengan rata-rata MAPE 10,3840% dan RMSE 10,5913 pada hasil validasi, serta *balanced accuracy* 0,9424 pada hasil klasifikasi. Hasil deteksi menunjukkan LSTM lebih baik dalam menangkap perubahan nilai ekstrem pada data, sementara XGBoost lebih baik dalam menangkap nilai ekstrem data. Anomali IKU Jakarta banyak terdeteksi pada pukul 21:00 s.d. 09:00 dan pada musim hujan (November-Maret). Aktivitas manusia dan



faktor meteorologi curah hujan, kelembapan, dan angin ikut memengaruhi keberadaan anomali.

Kajian simulasi menunjukkan hasil yang sejalan dengan kajian empiris, yaitu metode LSTM superior dibandingkan dengan XGBoost. LSTM memiliki nilai MAPE (14,7024%) dan RMSE (13,9909) yang lebih rendah, serta *balanced accuracy* (0,9935) lebih tinggi. Bahkan, nilai metrik terburuk dari LSTM tidak pernah lebih rendah dibandingkan nilai metrik terburuk pada XGBoost. Kajian simulasi menunjukkan faktor yang memengaruhi keakuratan metode berupa panjang periode data, cara pelabelan, dan jumlah anomali. Hasil tersebut diperkuat dengan uji non-parametrik Mann-Whitney dan Kruskal-Wallis yang signifikan pada kombinasi metode dengan faktor yang dicobakan. LSTM memiliki kinerja lebih baik dengan periode data yang panjang, sementara XGBoost untuk periode pendek. Hal ini diakibatkan perbedaan arsitektur di antara kedua metode, di mana LSTM memiliki sel memori yang mampu mengingat informasi berurutan layaknya jaringan syaraf. Namun, kedua metode menunjukkan hasil sejalan pada penambahan anomali pada data, yaitu mengalami penurunan kinerja seiring peningkatan jumlah anomali.

Kata kunci: deret waktu, deteksi anomali, indeks kualitas udara, LSTM, XGBoost



SUMMARY

MUHAMMAD RIZKY NURHAMBALI. Anomaly Detection Study of Jakarta Air Quality Index with Long Short-Term Memory and Extreme Gradient Boosting. Supervised by YENNI ANGRAINI and ANWAR FITRIANTO.

One environmental case that allows for anomalies is the air quality index (AQI). The AQI can change drastically, such as during a pandemic. Restrictions on human activities during a pandemic and meteorological factors, such as rain and wind, cause Jakarta's skies to clear, and cause people to be more concerned about air pollution. AQIs that are recorded over a span of time make AQI data a time series. Therefore, anomaly detection must be performed with a time series approach.

Anomaly detection in time series rapidly evolves from conventional methods to machine learning. However, machine learning methods are considered superior due to their ability to handle complex, nonlinear data and can ignore assumptions in conventional methods. Machine learning methods that have been widely developed include long short-term memory (LSTM) and extreme gradient boosting (XGBoost). Both methods are well-known for their ability to handle various conditions of time series data. However, the factors that affect the accuracy of the methods have not been widely studied. Therefore, this study aims to assess the performance of LSTM and XGBoost, identify factors that influence the presence of anomalies, and identify factors that affect the accuracy of both methods on Jakarta AQI data through various scenarios.

This study uses Jakarta's AQIs data sourced from AirNow (www.airnow.gov) as response variables and meteorological data in the form of humidity, wind, and temperature sourced from NASA Power (<https://power.larc.nasa.gov/data-access-viewer/>) as explanatory variables. The data is hourly data using the western Indonesian time zone (WIB / GMT+7) with a period from January 1, 2018, at 00:00 to December 31, 2023, at 23:00. The Jakarta AQIs data is labeled with a moving range (MR) and combined with the sigma rule. IKU Jakarta is validated with LSTM and XGBoost, and then anomaly classification is carried out with the threshold of the validation results.

The empirical data study shows LSTM as a more suitable method than XGBoost. Although the empirical data study added feature engineering to both methods, the results still showed the LSTM method as the best method despite the decrease in the evaluation metrics' value. Therefore, the combination of MR (2), 4-sigma, and no feature engineering on LSTM is the best combination for anomaly detection of AQIs Jakarta in empirical data with an average MAPE of 10.3840% and RMSE of 10.5913 in validation results, and balanced accuracy of 0.9424 in classification results. The detection results show that LSTM is better at capturing extreme value changes in the data, while XGBoost is better at capturing extreme data values. Anomalies in Jakarta's AQI are mostly detected at 21:00 to 09:00 and during the rainy season (November-March). Human activities and meteorological factors of rainfall, humidity, and wind also influence the presence of anomalies.

The simulation study shows results that align with the empirical study, namely that the LSTM method is superior to XGBoost. LSTM has lower MAPE (14.7024%) and RMSE (13.9909) values, and higher balanced accuracy (0.9935). The worst metric value of LSTM is never lower than the worst metric value of

XGBoost. Simulation studies show that the factors that affect the method's accuracy are the length of the data period, the labeling method, and the number of anomalies. These results are reinforced by the non-parametric Mann-Whitney and Kruskal-Wallis tests, which are significant in the combination of the method and the factors tested. LSTM performs better with long data periods, while XGBoost performs better with short periods. This is due to the difference in architecture between the two methods, where LSTM has memory cells that can remember sequential information like a neural network. However, both methods showed similar results when anomalies were added to the data, which decreased performance as the number of anomalies increased.

Keywords: air quality index, anomaly detection, LSTM, time series, XGBoost.

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

KAJIAN DETEKSI ANOMALI INDEKS KUALITAS UDARA JAKARTA DENGAN *LONG SHORT-TERM MEMORY* DAN *EXTREME GRADIENT BOOSTING*

MUHAMMAD RIZKY NURHAMBALI

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Tesis:
Sachnaz Desta Oktarina, S.Stat., M.Agr., Ph.D.

Judul Tesis : Kajian Deteksi Anomali Indeks Kualitas Udara Jakarta Dengan
Long Short-Term Memory dan *Extreme Gradient Boosting*
Nama : Muhammad Rizky Nurhambali
NIM : G1501231089

Disetujui oleh

Pembimbing 1:

Dr. Yenni Angraini, S.Si., M.Si.

Pembimbing 2:

Dr. Anwar Fitrianto, S.Si., M.Sc.

Diketahui oleh

Ketua Program Studi:

Dr. Agus Mohammad Soleh, S.Si., M.T

NIP 197503151999031004

Dekan Sekolah Sains Data, Matematika, dan Informatika:

Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.

NIP 196607021993021001

Tanggal Ujian:
14 Juli 2025

Tanggal Lulus:



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Desember 2023 sampai bulan Juli 2025 ini ialah deteksi anomali deret waktu, dengan judul “Kajian Deteksi Anomali Indeks Kualitas Udara Jakarta dengan *Long Short-Term Memory* dan *Extreme Gradient Boosting*”.

Penulisan karya ilmiah ini tidak luput atas bantuan dan dukungan berbagai pihak sehingga karya ilmiah ini dapat terselesaikan dengan baik. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Ibu Dr. Yenni Angraini, S.Si., M.Si. dan Bapak Dr. Anwar Fitrianto, S.Si., M.Sc. selaku komisi pembimbing yang telah membantu, membimbing, mengarahkan, dan mendukung selama penyusunan karya ilmiah ini;
2. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi yang sudah memberikan hibah dana penelitian tesis magister (PTM) melalui program Penelitian dan Pengabdian kepada Masyarakat Tahun Anggaran 2024;
3. Ibu Sachnaz Desta Oktarina, S.Stat., M.Agr., Ph.D. selaku dosen penguji luar yang telah memberikan saran dan masukan dalam penulisan karya ilmiah ini;
4. seluruh dosen dan tenaga pendidik Program Studi Statistika dan Sains Data yang telah memberi ilmu yang bermanfaat dan menunjang segala kebutuhan penulis selama perkuliahan dan penyusunan karya ilmiah ini;
5. Babe, Mama, adik Rizka, adik Riyan, dan keluarga besar penulis yang selalu memberikan doa, dukungan, dan kasih sayang kepada penulis;
6. teman-teman Program Studi Magister Statistika dan Sains Data Angkatan 2022 dan 2023, serta sahabat-sahabat penulis yang senantiasa mendukung dan kebersamaan penulis dalam penyusunan karya ilmiah ini; dan
7. seluruh pihak yang telah membantu penulis dalam penyusunan karya ilmiah ini yang tidak dapat penulis sebutkan satu persatu.

Penulis menyadari bahwa dalam penulisan karya ilmiah ini masih banyak kekurangan. Penulis berharap karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2025

Muhammad Rizky Nurhambali



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR ISI

DAFTAR TABEL	iv
DAFTAR GAMBAR	iv
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
II TINJAUAN PUSTAKA	4
2.1 Anomali	4
2.2 Kualitas Udara	4
2.3 Long Short-Term Memory (LSTM)	6
2.4 Extreme Gradient Boosting (XGBoost)	10
2.5 Ukuran Evaluasi Model	11
III PENDEKATAN ALTERNATIF UNTUK DETEKSI ANOMALI DATA INDEKS KUALITAS UDARA JAKARTA	12
3.1 Pendahuluan	12
3.2 Data dan Metode	13
3.3 Hasil dan Pembahasan	16
3.4 Simpulan	24
IV IDENTIFIKASI FAKTOR BERPENGARUH PADA KINERJA LSTM DAN XGBOOST UNTUK DETEKSI ANOMALI DATA DERET WAKTU DILABELKAN	25
4.1 Pendahuluan	25
4.2 Data dan Metode	26
4.3 Hasil dan Pembahasan	28
4.4 Simpulan	34
V SIMPULAN DAN SARAN	35
5.1 Simpulan	35
5.2 Saran	35
DAFTAR PUSTAKA	36
RIWAYAT HIDUP	44



DAFTAR TABEL

1	Tingkatan IKU menurut US EPA (2023) dan US EPA (2024)	5
2	Daftar peubah yang digunakan	13
3	Metode imputasi berdasarkan kriteria data hilang	14
4	Daftar <i>hyperparameter</i> optimal yang digunakan setiap metode	15
5	Jumlah data hilang per bulan untuk setiap tahun	17
6	Evaluasi metrik klasifikasi untuk setiap kombinasi	20
7	Persentase perlakuan anomali	26
8	Daftar kombinasi antarperlakuan	27
9	Perbandingan median nilai metrik evaluasi antarmetode	29
10	Hasil uji non-parametrik nilai metrik evaluasi terhadap faktor perlakuan	30

DAFTAR GAMBAR

1	IKU Jakarta tanggal 21 Juni tahun 2021, 2022, 2023	6
2	Arsitektur LSTM tanpa <i>forget gate</i> (Yu <i>et al.</i> 2019)	7
3	Arsitektur LSTM dengan <i>forget gate</i> (Yu <i>et al.</i> 2019)	8
4	Arsitektur umum dalam XGBoost (Wang <i>et al.</i> 2019)	10
5	Rata-rata IKU selama 24 jam per tahun (A) dan sebaran IKU selama 24 jam (B)	17
6	Korelasi antarpeubah yang digunakan	18
7	Akurasi validasi model LSTM dan XGBoost	19
8	Jumlah anomali terdeteksi dengan LSTM (A) dan XGBoost (B)	22
9	Jumlah anomali terdeteksi oleh metode terbaik berdasarkan (A) jam dan (B) bulan	23
10	Perbandingan nilai metrik evaluasi antarmetode dengan <i>boxplot</i> dan <i>violin plot</i>	29
11	Perbandingan nilai metrik evaluasi metode berdasarkan panjang data	31
12	Perbandingan nilai metrik klasifikasi metode berdasarkan <i>moving range</i>	32
13	Perbandingan nilai metrik evaluasi metode berdasarkan jumlah anomali	33
14	Contoh hasil deteksi anomali dan sisaan LSTM (A, C) dan XGBoost (B, D) untuk data tahun 2020 dengan perlakuan jumlah anomali 0,5%	34