Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Penerapan Metode *eXtreme Gradient Boosting* (*XGBoost*) pada Klasifikasi Kualitas Udara Indonesia Menggunakan Augmentasi Data *Conditional Generative Adversarial Network* (CGAN)

Application of the eXtreme Gradient Boosting (XGBoost) Method on Indonesian Air Quality Classification Using Conditional Generative Adversarial Network (CGAN) Data Augmentation

Risma Ayu Prastika[1], Indah Manfaati Nur[1], Ariska Fitriyana Ningrum[1]
[1] Universitas Muhammadiyah Semarang, Semarang
Corresponding author : indahmnur@unimus.ac.id

Abstrak

Latar belakang : Kualitas udara yang semakin menurun dari tahun ke tahun menjadi masalah penting bagi Indonesia, terutama karena dampaknya terhadap kesehatan masyarakat dan lingkungan. Salah satu contohnya adalah kasus ISPA yang semakin meningkat dan pada tahun 2023 kasusnya mencapai 1,5 hingga 1,8 juta dengan salah satu faktor penyebabnya adalah kualitas udara yang buruk. Oleh karena itu, diperlukan deteksi dan pemantauan kualitas udara yang lebih efektif. **Metode :** Metode *machine learning* seperti *eXtreme Gradient Boosting* (*XGBoost*) telah banyak digunakan untuk mengklasifikasikan kualitas udara. Namun, ketidakseimbangan data antar kelas dapat menyebabkan model klasifikasi menjadi kurang optimal, karena model cenderung bias terhadap kelas mayoritas. Oleh karena itu, penelitian ini menggunakan *Conditional Generative Adversarial Network* (CGAN) sebagai teknik augmentasi data guna meningkatkan performa klasifikasi dengan menambah jumlah sampel pada kelas minoritas. **Hasil penelitian :** Berdasarkan hasil klasifikasi, proses augmentasi data CGAN mampu meningkatkan keseimbangan prediksi antar kelas, terutama dalam mengklasifikasikan kelas minoritas. Model CGAN-X*GBoost* menunjukkan performa yang baik dengan akurasi mencapai 97.08% serta nilai presisi 97.11%, recall 97.1%, F1-Score 97.08%, dan AUC 98.82%. Hal ini menunjukkan bahwa *XGBoost* merupakan metode yang stabil dan akurat dalam memanfaatkan data sintetis CGAN. **Kesimpulan :** Dengan demikian, CGAN-*XGBoost* dinilai efektif dalam menangani ketidakseimbangan data kualitas udara di Indonesia.

Kata Kunci : CGAN, klasifikasi, kualitas udara, *XGboost*

Abstract

Background : *The declining air quality year after year has become a significant problem for Indonesia, especially due to its impact on public health and the environment. One example is the increasing number of respiratory tract infection cases, which reached 1.5 to 1.8 million in 2023, with poor air quality being one of the contributing factors. Therefore, more effective air quality detection and monitoring are needed.* ***Method :*** *Machine learning methods such as eXtreme Gradient Boosting (XGBoost) have been widely used to classify air quality. However, data imbalance between classes can cause classification models to be less than optimal, as models tend to be biased towards the majority class. Therefore, this study uses Conditional*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Generative Adversarial Network (CGAN) as a data augmentation technique to improve classification performance by increasing the number of samples in the minority class. ***Result :*** *Based on the classification results, the CGAN data augmentation process was able to improve the balance of predictions between classes, especially in classifying minority classes. The CGAN-XGBoost model showed good performance with an accuracy of 97.08%, precision of 97.11%, recall of 97.1%, F1-Score of 97.08%, and AUC of 98.82%. This shows that XGBoost is a stable and accurate method in utilizing CGAN synthetic data.* ***Conclusion :*** *Thus, CGAN-XGBoost is considered effective in handling the imbalance of air quality data in Indonesia.*

Keywords : *CGAN, classification, air quality, XGboost*

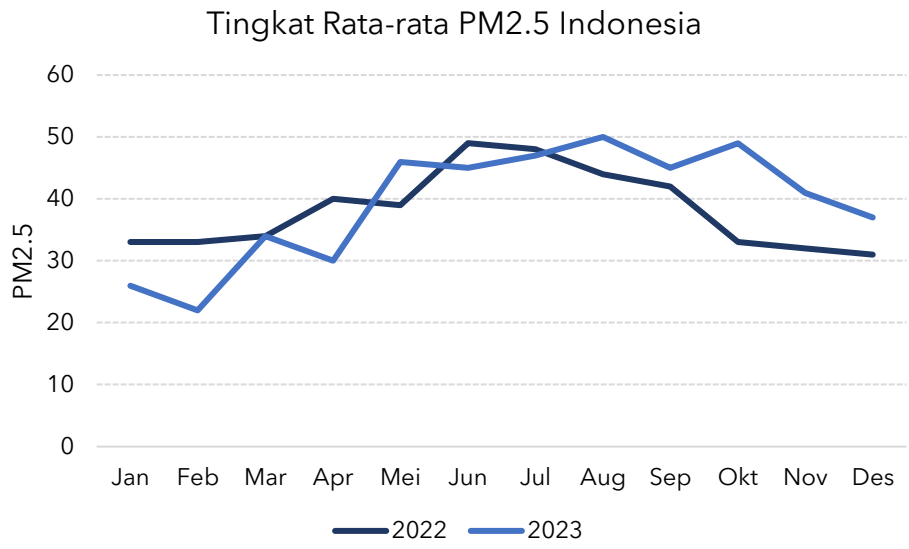
PENDAHULUAN

Kemajuan teknologi modern telah memberikan dampak ganda bagi lingkungan. Di satu sisi, perkembangan industri dan transportasi telah meningkatkan efisiensi serta produktivitas [1]. Namun di sisi lain, aktivitas ini juga menyebabkan peningkatan emisi polutan yang dapat merusak lingkungan, terutama dari sektor transportasi menyumbang sekitar 44% terhadap pencemaran udara, kemudian sektor industri 31%, manufaktur 10%, perumahan 14%, dan komersial 1% [2]. Emisi ini menghasilkan zat berbahaya antara lain, karbon monoksida (CO), sulfur dioksida (SO2), dan nitrogen dioksida (NO2), ozon (O3), dan *particulate matter* (PM) yang secara signifikan mempengaruhi kualitas udara dan berdampak kesehatan manusia dan lingkungan [3].

Menurut laporan *Health Effect Institue*, sekitar 8,1 juta kematian di seluruh dunia pada tahun 2021 disebabkan oleh kualitas udara yang buruk. Angka ini menunjukkan bahwa kualitas udara yang buruk merupakan salah satu faktor utama kematian secara global [4]. Selain itu, jutaan orang lainnya hidup dengan penyakit kronis yang secara signifikan mengurangi kualitas hidup mereka akibat paparan kualitas udara yang buruk [5]. Selama periode Januari hingga September 2023, tercatat sebanyak 1,5 hingga 1,8 juta kasus ISPA (Infeksi Saluran Pernapasan Akut) di Indonesia dengan salah satu faktor penyebabnya adalah kualitas udara yang buruk [6].

Kualitas udara yang buruk menjadi salah satu ancaman serius yang dihadapi dunia, termasuk Indonesia. Berdasarkan data AQI (*Air Quality Index*), Indonesia menempati peringkat ke-14 sebagai negara dengan kualitas udara terburuk pada tahun 2023, dengan konsentrasi rata-rata PM2.5 mencapai 7,4 kali lipat dari standar pedoman kualitas udara tahunan yang ditetapkan oleh WHO [7].

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257



Gambar 1. Tren Kualitas Udara 2022 dan 2023

Kualitas udara di Indonesia mengalami penurunan secara signifikan pada tahun 2023 dibandingkan tahun 2022. Berdasarkan Gambar 1, terdapat perbedaan tren antara kedua tahun tersebut. Pada tahun 2022, tren kualitas udara cenderung membaik menjelang akhir tahun, sedangkan pada tahun 2023 tren kualitas udara tidak mengalami perbaikan hingga akhir tahun. Penurunan kualitas udara dalam beberapa tahun terakhir menunjukkan perlunya analisis untuk membantu mendeteksi kualitas udara secara lebih akurat. Salah satu metode yang dapat digunakan adalah klasifikasi, yang berfungsi untuk mengelompokkan kondisi udara ke dalam beberapa kategori berdasarkan tingkat kualitas udara yang terdeteksi.

Secara umum, terdapat tiga tingkat kategori utama kualitas udara di Indonesia, antara lain “baik”, “sedang”, dan “tidak sehat”. Namun dalam praktiknya, distribusi data dari setiap kategori tidak selalu seimbang. Ketidakseimbangan data ini dapat menyebabkan performa model prediksi menurun dan akurasinya rendah, karena model cenderung lebih akurat dalam memprediksi kelompok mayoritas, sedangkan konstribusi kelompok minoritas seringkali diabaikan [8]. Untuk mengatasi masalah ini, salah satu teknik yang dapat digunakan adalah dengan metode *oversampling* dengan *Conditional Generative Adversarial Network* (CGAN). Metode ini dapat menghasilkan sampel sintesis dari kelas minoritas, sehingga membantu dalam menyeimbangkan dataset dan dapat meningkatkan performa model klasifikasi [9].

Pemanfaaan model klasifikasi dalam analisis kualitas udara memiliki peran penting dalam menyediakan informasi yang tepat, agar masyarakat dapat lebih mudah mengidentifikasi tingkat kualitas udara yang berpotensi berdampak terhadap kesehatan. Salah satu metode yang sering digunakan adalah metode *machine learning* yang merupakan bagian dari kecerdasan buatan dan memungkinkan komputer untuk mempelajari pola dari data tanpa memerlukan pemrograman secara eksplisit [10]. Salah satu algoritma *machine learning* yang efektif adalah *eXtreme*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Gradient Boosting (*XGBoost*) yang merupakan pengembangan dari *Gradient Boosting Decision Tree* (GBDT) dengan beberapa keunggulan antara lain regularisasi, paralelisasi, fleksibilitas, dan memiliki hasil klasifikasi yang baik. *XGBoost* juga dapat melakukan pencegahan terhadap *overfitting*, serta dapat mengoptimalkan model dengan cepat dan efisien [11].

Beberapa penelitian sebelumnya seperti penelitian oleh Arifianti & Salam (2024), yaitu pengklasifikasian kualitas udara menggunakan algoritma *XGBoost* dan *Random Forest*. Hasil penelitian ini menunjukkan bahwa algoritma *XGBoost* mampu memberikan performa yang baik. *XGBoost* mampu mencapai nilai akurasi sebesar 83,2% [12]. Sedangkan penelitian oleh Liu et al., (2023), yang mengklasifikasikan kesalahan peralatan kontrol kereta api menggunakan metode *XGBoost* yang disempurnakan dengan CGAN menunjukkan kekurangan metode *machine learning* konvensional dalam mengatasi masalah ketidakseimbangan data. Dengan mengombinasikan *XGBoost* dan CGAN, model prediksi dibangun menggunakan data yang tidak seimbang dan setelah dilakukan perbandingan terbukti bahwa pendekatan ini berhasil meningkatkan akurasi sebesar 2,17% dibandingkan model *XGBoost* konvensional [13]. Penelitian oleh Toharudin et al. (2023), yang berfokus pada klasifikasi tingkat konsentrasi PM2.5 di Jakarta dengan menggunakan beberapa algoritma *Boosting*, seperti *AdaBoost*, *XGBoost*, *CatBoost*, dan *LightGBM*. Hasil dari penelitian ini menunjukkan bahwa algoritma *XGBoost* menghasilkan nilai akurasi terbaik sebesar 54,2% dibanding algoritma lainnya [14]. Meskipun demikian, penelitian ini belum dapat sepenuhnya mengatasi masalah ketidakseimbangan pada data. Hal ini mengakibatkan model lebih cenderung memprediksi kelas mayoritas, sehingga dapat mengurangi akurasi pada kelas minoritas.

Berdasarkan uraian latar belakang di atas, penulis mengajukan penelitian yang berjudul "Penerapan Metode *XGBoost* pada Klasifikasi Kualitas Udara Indonesia Menggunakan Augmentasi Data CGAN". Penelitian ini bertujuan untuk mengevaluasi performa algoritma *XGBoost* dalam mengklasifikasikan kualitas udara di Indonesia dan berfokus terhadap penerapan algoritma *XGBoost* dengan mempertimbangkan berbagai aspek penting, seperti performa evaluasi dan kemampuan dalam menangani data kualitas udara yang tidak seimbang, dengan memanfaatkan CGAN sebagai metode augmentasi data untuk mengatasi masalah ketidakseimbangan tersebut.

METODE

Metode penelitian pada studi ini mencakup kumpulan teknik, prinsip, dan prosedur yang diterapkan secara sistematis untuk mendukung proses penelitian.

Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah data kualitas udara di Indonesia yang diperoleh dari website *accuweather.com*. Dataset berisi total 514 data

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

dengan 202 diantaranya kualitas udara dengan kategori baik, 215 sedang dan 97 tidak sehat, seperti Tabel 1.

Tabel 1. Distribusi Kelas pada Dataset Original

Kelas	Jumlah Kategori
Baik	202
Sedang	215
Tidak Sehat	97
Jumlah	514

Setiap kategori merepresentasikan kondisi kualitas udara di suatu wilayah berdasarkan sejumlah fitur relevan untuk mendeteksi tingkat kualitas udara. Fitur-fitur utama meliputi konsentrasi karbon monoksida (CO), sulfur dioksida (SO_2), dan nitrogen dioksida (NO_2), ozon (O_3), *particulate matter* diameter $\leq 10\ \mu g/m^3$ (PM_{10}), dan *particulate matter* diameter $\leq 2{,}5\ \mu m$ ($PM_{2.5}$).

Tabel 2. Fitur pada Dataset Kualitas Udara

Fitur	Penjelasan
Sulfur Dioksida (SO_2)	Gas hasil pembakaran batu bara, minyak, dan proses industri. Dengan kadar tinggi mampu menurunkan kualitas udara.
Karbon Monoksida (CO)	Gas tidak berwarna dan tidak berbau dari pembakaran tidak sempurna bahan bakar fosil. Kadar tinggi menurunkan kemampuan darah membawa oksigen dan menurunkan kualitas udara.
Nitrogen Dioksida (NO_2)	Gas dari emisi kendaraan dan pembangkit listrik. Konsentrasi berlebih memicu pembentukan ozon dan menurunkan kualitas udara.
Ozon (O_3)	Gas sekunder yang terbentuk dari reaksi kimia polutan di bawah sinar matahari. Kadar berlebih bersifat iritan pernapasan dan menurunkan kualitas udara.
Particulate Matter (PM)	Partikel halus di udara berdiameter $\leq 10\ \mu m$ dan $\leq 2{,}5\ \mu m$. Konsentrasi tinggi dapat masuk ke paru-paru dan menurunkan kualitas udara.

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Prepocessing Data

Tahapan ini meliputi pembersihan data dengan menghapus/ memperbaiki data yang tidak konsisten dan *noise*, menghapus data *duplicate*, memperbaiki kesalahan pada data, dan memperkaya data dengan data eksternal yang relevan [15]. Selain itu, pada tahap ini juga fitur numerik diubah menjadi *one-hot-encoding* untuk menghindari asumsi hubungan ordinal antar kategori. Sementara itu, fitur numerik dinormalisasikan agar memiliki skala yang seragam, yang penting dilakukan sebagai bagian untuk input pada tahap *oversampling* CGAN.

Augmentasi Data dengan CGAN

CGAN adalah salah satu teknik *oversampling* varian dari GAN yang menghasilkan data sintesis dari kelas minoritas untuk menyeimbangkan dataset [9]. CGAN memiliki keunggulan dalam memperkirakan distribusi data yang lebih akurat dan mampu menghasilkan data untuk kelas minoritas pada kumpulan data yang tidak seimbang. CGAN memiliki batasan pada jaringan yang digunakan untuk membatasi tingkat kebebasan dalam konvergensi. Pendekatan ini dapat membantu mengurangi risiko konvergensi yang lambat atau kegagalan konvergensi yang sering terjadi akibat tingkat kebebasan yang tinggi pada GAN [16].

CGAN terdiri dari dua jaringan saraf, yaitu generator dan diskriminator dengan menerima informasi tambahan berupa *condition label* yang digunakan untuk mengontrol proses pembangkitan data sintetis. Generator menerima input berupa vektor acak dan *condition label*, kemudian mengolahnya untuk menghasilkan data sintetis. Hasil dari data sintetis tersebut akan di evaluasi oleh diskriminator. Apabila data sintetis terdeteksi sebagai data palsu, maka akan dikembalikan ke generator untuk diolah kembali hingga generator dapat menghasilkan data sintetis yang memiliki karakteristik yang mirip dengan data asli. Sedangkan data sintetis yang terdeteksi sebagai data asli, akan digabungkan dengan data original. Data gabungan ini yang kemudian akan digunakan untuk proses analisis klasifikasi.

Pelatihan Model

Proses pelatihan dan pengujian model dilakukan dengan 2 skenario yaitu, menggunakan data asli yang tidak seimbang (*imbalance*) dan data hasil augmentasi yang sudah seimbang (*balance*). Hal ini bertujuan untuk mengevaluasi kinerja model pada data yang telah melalui proses augmentasi data menggunakan CGAN.

Pembagian dataset dilakukan menggunakan teknik *stratified* 5-*fold cross valiadtion*, di mana data dibagi menjadi lima *fold* dengan proporsi kelas yang seimbang pada setiap *fold*. Setiap *fold* secara bergantian berperan sebagai data validasi, sedangkan *fold* lainnya digunakan sebagai data latih, sehingga seluruh data dapat dimanfaatkan secara optimal sebagai data latih dan pengujian [17].

Model yang digunakan adalah XGBoost, algoritma berbasis pohon keputusan yang merupakan implementasi lanjutan dari algoritma *Gradient Boosting* yang

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

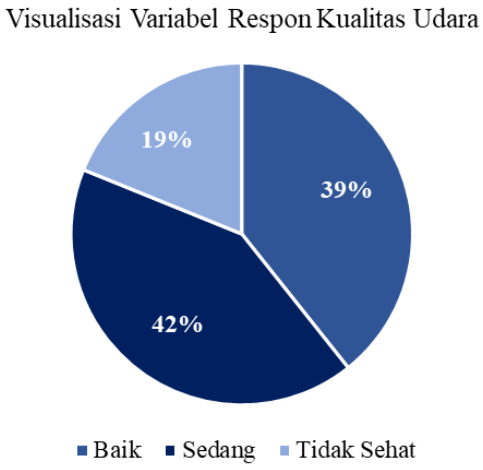
dioptimalkan dan dirancang untuk menjadi sangat efisien, fleksibel, dan portabel [14]. Pelatihan *XGBoost* dilakukan dengan *random_state* = 42 untuk memastikan bahwa model menghasilkan hasil yang konsisten ketika proses pelatihan model. Parameter inisialisasi meliputi *eval_metric*= [*'mlogloss', 'merror'*] yang digunakan untuk memantau performa setiap iterasi karena sesuai dengan kebutuhan klasifikasi multikelas, serta *objective= 'multi:softprob'*. *XGBooost* bekerja secara iteratif dengan dengan membangun pohon keputusan baru yang berfokus pada area kelemahan model sebelumnya. Pada setiap iterasi, *XGBoost* menghitung gradien dari fungsi *loss* untuk mengidentifikasi area yang perlu diperbaiki, lalu proses ini berlanjut hingga jumlah pohon yang ditentukan atau kriteria penghentian lain tercapai [9].

Hyperparameter tuning yang digunakan dalam penelitian ini ada teknik *GridSearch*, dimana teknik ini secara sistematis akan mencari kombinasi nilai parameter yang paling optimal dengan mecoba seluruh kombinasi nilai parameter yang sudah ditentukan dalam grid [18]. Kemudian kinerja model akan di evaluasi menggunakan metrik evaluasi seperti akurasi, presisi, *recall*, *F1-Score*, dan AUC.

HASIL DAN PEMBAHASAN

Statistika Deskriptif

Analisis deskriptif dilakukan untuk memberikan gambaran umum mengenai karakteristik data yang digunakan. Berikut adalah hasil analisis deskriptif pada kualitas udara di Indonesia.



Gambar 2. Visualisasi Variabel Respon Kualitas Udara

Gambar 2 menujukkan bahwa dari 514 wilayah Kabupaten/ Kota yang diamati, sebanyak 42% atau 215 wilayah termasuk dalam kategori kualitas udara "sedang". Sementara itu, 39% atau 202 wilayah dikategorikan memiliki kualitas udara "baik", dan 19% sisanya atau sebanyak 97 wilayah tergolong dalam kualitas udara "tidak sehat". Perbedaan proporsi ini menandakan adanya ketidakseimbangan kelas (*class*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

imbalance) pada data, karena jumlah sampel kategori tidak sehat jauh lebih sedikit dibandingkan dua kategori lainnya. Kondisi tersebut dapat menyebabkan model klasifikasi cenderung bias dalam memprediksi kelas mayoritas (baik dan sedang) dengan lebih akurat, namun mengabaikan kelas minoritas (tidak sehat). Oleh karena itu perlu dilakukan penanganan seperti augmentasi data menggunakan CGAN untuk menjaga performa model dan meningkatkan keadilan prediksi pada seluruh kelas.

Augmentasi Data CGAN

Teknik *oversampling* CGAN digunakan untuk mengatasi ketidakseimbangan pada data dengan cara menambahkan data sintesis baru pada kelas minoritas agar distribusi kelas menjadi lebih seimbang. Dari total 514 data kualitas udara yang sebelumnya mengalami ketidakseimbangan distribusi kelas data, setelah dilakukan augmentasi data CGAN menjadi memiliki distribusi data dengan kelas yang lebih seimbang. Seperti pada tabel berikut.

Tabel 3. Jumlah Kelas Hasil *Oversampling* CGAN

Kelas	Sebelum	Setelah
Baik	202	202
Sedang	215	215
Tidak Sehat	97	200
Jumlah	514	617

Tabel 3 menunjukkan distribusi data pada kelas target sebelum dan setelah penerapan CGAN. Visualisasi perbandingan kelas target sebelum penerapan CGAN menunjukkan adanya ketidakseimbangan yang signifikan, terlihat dari perbedaan jumlah frekuensi data yang cukup jauh antar kelasnya. Dengan penambahan data sebanyak 13 data baru pada kelas baik dan 118 data baru pada kelas tidak sehat, maka diperoleh distribusi data yang seimbang antara kelas baik, kelas sedang, dan kelas tidak sehat dengan masing-masing kelas memiliki jumlah data sebanyak 215 data. Jumlah data yang sama pada setiap kelas mengindikasikan bahwa masalah ketidakseimbangan data telah teratasi dengan menggunakan teknik CGAN.

Pembagian Data Penelitian

Pada langkah ini dilakukan pembagian data menggunakan teknik *stratified 5-fold cross validation*. Teknik ini membagi dataset menjadi lima *fold* dengan ukuran sama besar, yaitu masing-masing sebesar 20% dari keseluruhan data. Pembagian dilakukan dengan mempertahankan proporsi kelas pada setiap fold agar sama atau mendekati proporsi kelas pada keseluruhan dataset (*stratified*).

Tabel 4. Pembagian Data Menggunakan *Stratified 5-Fold Cross Validation*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Iterasi	*Fold* 1	*Fold* 2	*Fold* 3	*Fold* 4	*Fold* 5	Total
1	**20%**	20%	20%	20%	20%	100%
2	20%	**20%**	20%	20%	20%	100%
3	20%	20%	**20%**	20%	20%	100%
4	20%	20%	20%	**20%**	20%	100%
5	20%	20%	20%	20%	**20%**	100%

Pada setiap iterasi, satu *fold* digunakan sebagai data validasi (*testing*) dan empat *fold* lainnya digunakan sebagai data pelatihan (*training*). Proses ini di ulang sebanyak lima kali, sehingga setiap *fold* berperan sebagai data validasi satu kali. Dengan pendekatan ini, model dapat di evaluasi secara adil karena setiap data memiliki kesempatan yang sama untuk menjadi data pelatihan maupun validasi, sekaligus meminimalkan bias akibat pembagian data yang tidak proporsional.

Hasil *Hyperparameter Tuning Grid Search*

Proses ini merupakan langkah awal dalam analisis berbasis *machine learning*, terutama klasifikasi *XGBoost*. Tahap ini bertujuan untuk mencari kombinasi parameter terbaik melalui teknik *Grid Search*. Proses *tuning* dilakukan dengan menggunakan beberapa parameter seperti berikut.

Tabel 5. Hasil *Hyperparameter Tuning Grid Search* CGAN-*XGBoost*

Parameter	Nilai Parameter	Parameter Terbaik				
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
learning_rate	[0.03, 0.05, 0.07]	0.03	0.05	0.03	0.03	0.05
n_estimator	[100, 200, 300]	200	100	100	100	100
max_depth	[4, 5]	5	5	5	5	4
min_child_weight	[3, 4]	3	3	3	3	3
gamma	[1, 2, 3]	1	2	3	2	2
reg_lambda	[3, 4, 5]	3	3	3	3	5
sub_sample	[0.6]	0.6	0.6	0.6	0.6	0.6
colsample_bytree	[0.6]	0.6	0.6	0.6	0.6	0.6

Tabel 5 menunjukkan kecenderungan pemilihan *hyperparameter* pada CGAN-*XGBoost* relatif konsisten meskipun terdapat sedikit variasi antar-*fold*. Parameter

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

learning_rate optimal bervariasi antara 0.03 hingga 0.05, dengan nilai 0.03 muncul pada tiga *fold* dan 0.05 pada dua *fold*. Nilai yang rendah ini menunjukkan bahwa kecepatan pembelajaran yang konservatif lebih sesuai untuk menjaga proses konvergensi yang stabil setelah penambahan data sintetis CGAN, sehingga risiko *overfitting* dapat diminimalkan. Jumlah *n_estimators* optimal berkisar antara 100 dan 200, dengan dominasi nilai 100 pada empat *fold*, menandakan bahwa model mampu mencapai performa terbaik dengan jumlah *boosting round* yang moderat.

Parameter *max_depth* sebagian besar memilih kedalaman 5, kecuali satu *fold* yang memilih 4, mengindikasikan bahwa pohon dengan kedalaman menengah sudah cukup menangkap pola penting pada data hasil augmentasi. Untuk *min_child_weight*, seluruh *fold* konsisten pada nilai 3, menunjukkan bahwa model lebih efektif ketika node daun diizinkan terbentuk dengan jumlah sampel minimum yang relatif kecil. Parameter *gamma* bervariasi antara 1 hingga 3, menandakan adanya kebutuhan regularisasi pemisahan *node* yang menyesuaikan karakter distribusi data di tiap *fold*.

Nilai *reg_lambda* yang mengontrol regularisasi L2 umumnya berada pada nilai 3, dengan satu *fold* memilih 5, menegaskan pentingnya regularisasi moderat untuk menjaga kestabilan bobot. Sementara itu, *subsample* dan *colsample_bytree* konsisten di 0.6 pada seluruh *fold*, sesuai rancangan awal untuk menjaga keberagaman pohon dan meningkatkan kemampuan generalisasi. Secara keseluruhan, hasil *tuning* ini menegaskan bahwa konfigurasi parameter yang konservatif dengan *learning_rate* rendah, kedalaman pohon menengah, dan regularisasi moderat memberikan keseimbangan terbaik bagi model CGAN-*XGBoost*. Pendekatan ini terbukti efektif dalam mengendalikan kompleksitas model setelah augmentasi data, sehingga performa klasifikasi tetap stabil dan mampu menangani variasi distribusi data pada setiap *fold*.

Hasil Klasifikasi

Langkah selanjutnya adalah klasifikasi *XGBoost* menggunakan data hasil *oversampling* CGAN dengan kombinasi parameter terbaik. Hasil klasifikasi tersebut dapat dilihat melalui *confussion matrix* berikut.

Tabel 6. Hasil Confussion Matrix CGAN-*XGBoost*

Nilai Aktual	Nilai Observasi			
	Baik	Sedang	Tidak sehat	Jumlah
Baik	201	0	1	202
Sedang	6	206	3	215
Tidak Sehat	5	3	192	200
Jumlah	212	209	196	617

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Tabel 6 menujukkan bahwa model klasifikasi CGAN-*XGBoost* menghasilkan ketepatan prediksi klasifikasi pada kelas kualitas udara dari total 212 wilayah yang diprediksi memiliki kualitas udara yang baik, 201 wilayah tepat terprediksi memiliki kualitas udara baik (*True Positive*) dan terdapat kesalahan prediksi 6 wilayah yang masuk ke dalam kelas sedang serta 5 wilayah masuk ke dalam kelas tidak sehat (*False Positive*). Pada kelas sedang, dari total 209 wilayah yang diprediksi memiliki kulitas udara sedang, 206 wilayah tepat terprediksi memiliki kualitas udara yang sedang (*True Positive*) dan terdapat kesalahan prediksi 3 wilayah masuk ke dalam kelas tidak sehat (*False Positive*). Pada kelas tidak sehat, dari total 196 wilayah yang diprediksi memiliki kulitas udara tidak sehat, 192 wilayah terprediksi dengan benar memiliki kualitas udara tidak sehat (*True Positive*) dan terdapat kesalahan prediksi 1 wilayah masuk ke dalam kelas baik serta 3 wilayah lainnya masuk ke dalam kelas sedang (*False Positive*).

Tabel 7. Hasil evaluasi model CGAN-*XGBoost*

Kelas	Presisi	*Recall*	*F1-Score*	Akurasi	AUC
Baik	94.81%	99.5%	97.1%	97.08%	98.82%
Sedang	98.56%	95.81%	97.17%		
Tidak Sehat	97.96%	96%	96.97%		

Berdasarkan tabel 7, hasil evaluasi model CGAN-*XGBoost* menunjukkan nilai akurasi sebesar 97.08% yang menunjukkan bahwa secara umum model mampu memprediksi dengan baik dan AUC sebesar 98.82% menunjukkan bahwa model mampu membedakan antar kelas, sehingga distribusi data dapat dipisahkan dengan jelas. Pada kelas "baik", nilai presisi sebesar 94.81% yang menunjukkan bahwa hampir seluruh data yang diprediksi sebagai kelas "baik", benar-benar bersala dari kelas tersbeut. Sedangkan nilai *recall* mencapai 99.5%, yang berarti hampir seluruh data kelas ini berhasil dikenali dengan benar. Kombinasi ini menghasilkan nilai F1-Score sebesar 97.1%, yang merepresentasikan keseimbangan antara presisi dan *recall*. Kelas Sedang menunjukkan hasil yang baik, dengan nilai presisi mencapai 98.56% dan *recall* 95.81%, sehingga menghasilkan nilai *F1-Score* sebesar 97.17%, yang menandakan kemampuan prediksi yang konsisten antara data yang benar dikenali dan prediksi yang tepat sasaran. Pada kelas "tidak sehat" memiliki presisi 97.96% dan *recall* 96% yang menunjukkan bahwa model mampu mengidentifikasi kelas ini dengan cukup baik, sehingga menghasilkan niali *F1-Score* sebesar 96.97%.

Untuk memperkuat justifikasi keunggulan performa model klasifikasi, akan dilakukan perbandingan hasil evaluasi model CGAN-*XGBoost* dengan model *XGBoost* tanpa CGAN. Hasil perbandingan dapat dilihat pada Tabel 7 berikut.

Tabel 8. Hasil evaluasi model *XGBoost* dan CGAN-*XGBoost*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

	XGBoost	CGAN-XGBoost
Akurasi	98.05%	97.08%
Presisi	97.56%	97.11%
Recall	98.45%	97.1%
F1-Score	97.95%	97.08%
AUC	98.54%	98.82%

Hasil perbandingan antara model CGAN-*XGBoost* dan model *XGBoost* biasa menunjukkan perbedaan kinerja yang cukup signifikan dalam berbagai metrik evaluasi. Model CGAN-*XGBoost* berhasil mencapai akurasi sebesar 97.08%. Meskipun sedikir menurun jika dibandingkan dengan *XGBoost* tanpa CGAN (98.05% menjadi 97.08%), tetapi nilai AUC meningkat (98.54% menjadi 98.82%). Hal ini menunjukkan bahwa model *XGBoost* dengan CGAN memiliki kemampuan diskriminasi antar kelas yang lebih baik dan stabil.

Secara keseluruhan, meskipun akurasi CGAN-*XGBoost* sedikit lebih rendah, nilai AUC yang lebih tinggi serta distribusi presisi, *recall*, dan *F1-Score* yang lebih seimbang antar kelas menunjukkan bahwa penggunaan CGAN berhasil memperbaiki kelemahan *XGBoost* tanpa CGAN. Dengan kata lain, CGAN membuat model lebih adil dalam mengenali ketiga kelas, sehingga hasil klasifikasi lebih representatif terhadap kondisi nyata. Penelitian terdahulu oleh Sutou & Wang (2024) yang melakukan penelitian mengenai *Influence-Balanced XGBoost* dengan menggunakan beberapa metode balancing data menunjukkan bahwa kinerja yang signifikan dalam penanganan data tidak seimbang dapat mengurangi kecenderungan terhadap kelas mayoritas, meskipun hasil yang diperoleh belum sepenuhnya optimal [19]. Dengan demikian, meskipun performa keseluruhan CGAN-*XGBoost* tidak sebaik model pada dataset tidak seimbang dalam hal akurasi, pendekatan ini terbukti mampu meningkatkan keadilan model dalam mengenali kelas minoritas dan menghasilkan distribusi prediksi yang lebih seimbang.

KESIMPULAN

Model *XGBoost* pada dataset yang tidak seimbang menghasilkan akurasi 98,05% dengan nilai AUC 98,54%. Meskipun kinerjanya tinggi, masih terdapat indikasi ketidakseimbangan prediksi antar kelas, sehingga model belum sepenuhnya adil dalam mengklasifikasikan seluruh kategori kualitas udara. Penerapan augmentasi data menggunakan CGAN mampu memperbaiki hal tersebut, terbukti dari peningkatan kualitas performa model. Kombinasi CGAN–*XGBoost* mencatat akurasi 97,08% dan nilai AUC 98,82%, yang menunjukkan distribusi prediksi antar kelas menjadi lebih seimbang meskipun terjadi sedikit penurunan akurasi.

DAFTAR PUSTAKA

[1] A. D. Izzati *et al.*, *Analisis Dampak Teknologi Modern Terhadap Masalah Lingkungan*. Semarang: CV. Alinea Media Dipantara, 2022.

[2] KLHK, "Uji Emisi dan Kendaraan Listrik Jadi Solusi Tekan Polusi," Kementerian Lingkungan Hidup dan Kehutanan. Accessed: Feb. 21, 2025. [Online]. Available: https://ppid.menlhk.go.id/berita/siaran-pers/7311/uji-emisi-dan-kendaraan-listrik-jadi-solusi-tekan-polusi

[3] A. K. Gorai, F. Tuluri, P. B. Tchounwou, and S. Ambinakudige, "Influence of local meteorology and NO2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA," *Air Qual Atmos Health*, vol. 8, no. 1, pp. 81–96, Feb. 2015, doi: 10.1007/s11869-014-0276-5.

[4] HEI, "New State of Global Air Report Finds Air Pollution is Second Leading Risk Factor for Death Worldwide," Health Effects Institute.

[5] UNICEF, "Air pollution accounted for 8.1 million deaths globally in 2021, becoming the second leading risk factor for death, including for children under five years," *UNICEF for every child*, Jun. 18, 2024. Accessed: Nov. 16, 2024. [Online]. Available: https://www.unicef.org/press-releases/air-pollution-accounted-81-million-deaths-globally-2021-becoming-second-leading-risk

[6] Rokom, "Polusi Ancam Saluran Pernapasan," Kementrian Kesehatan RI. Accessed: Nov. 24, 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/blog/20240108/5644635/polusi-ancam-saluran-pernapasan/

[7] IQAir, "Air quality in Indonesia," IQAir. Accessed: Nov. 16, 2024. [Online]. Available: https://www.iqair.com/indonesia

[8] I. M. Nur, D. Rosadi, and Abdurakhman, "Multi-Class Imbalance Classification of Diabetes Cases Using Light Gradient Boosting Machine," *ITM Web of Conferences*, vol. 67, p. 01012, 2024, doi: 10.1051/itmconf/20246701012.

[9] Sarmini, Sunardi, and A. Fadlil, "Performa Random Forest dan XGBoost pada Deteksi Penipuan E-Commerce Menggunakan Augmentasi Data CGAN," *Technology and Science (BITS)*, vol. 6, no. 3, 2024, doi: 10.47065/bits.v6i3.6430.

[10] A. F. B. Sajiwo, B. Rahmat, and A. Junaidi, "KLASIFIKASI INDEKS STANDAR PENCEMARAN UDARAN (ISPU) MENGGUNAKAN ALGORITMA XGBOOST DENGAN TEKNIK IMBALANCED DATA (SMOTE)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4699.

[11] A. M. Sapari, A. Id Hadiana, and F. R. Umbara, "Air Quality Classification Using Extreme Gradient Boosting (XGBOOST) Algorithm ARTICLE INFORMATION ABSTRACT," 2023. [Online]. Available: http://innovatics.unsil.ac.id

[12] F. P. Arifianti and A. Salam, "XGBoost and Random Forest Optimization using SMOTE to Classify Air Quality," *Advance Sustainable Science, Engineering and Technology (ASSET)*, vol. 6, no. 1, Nov. 2024, doi: 10.26877/asset.v6i1.17951.

[13] J. Liu, K. Xu, B. Cai, and Z. Guo, "Fault Prediction of On-Board Train Control Equipment Using a CGAN-Enhanced XGBoost Method with Unbalanced Samples," *Machines*, vol. 11, no. 1, Jan. 2023, doi: 10.3390/machines11010114.

[14] T. Toharudin *et al.*, "Boosting Algorithm to Handle Unbalanced Classification of PM2.5Concentration Levels by Observing Meteorological Parameters in Jakarta-Indonesia Using AdaBoost, XGBoost, CatBoost, and LightGBM," *IEEE Access*, vol. 11, pp. 35680–35696, 2023, doi: 10.1109/ACCESS.2023.3265019.

[15] K. Erwansyah, B. Andika, and R. Gunawan, "J-SISKO TECH Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD Implementasi Data Mining Menggunakan Asosiasi Dengan Algoritma Apriori Untuk Mendapatkan Pola Rekomendasi Belanja Produk Pada Toko Avis Mobile," ν, vol. 148, no. 1, pp. 148–161, 2021.

[16] R. Ahsan, W. Shi, X. Ma, and W. Lee Croft, "A comparative analysis of CGAN-based oversampling for anomaly detection," *IET Cyber-Physical Systems: Theory and Applications*, vol. 7, no. 1, pp. 40–50, Mar. 2022, doi: 10.1049/cps2.12019.

[17] J. Kaliappan, A. R. Bagepalli, S. Almal, R. Mishra, Y. C. Hu, and K. Srinivasan, "Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise," *Diagnostics*, vol. 13, no. 10, May 2023, doi: 10.3390/diagnostics13101692.

[18] S. George and B. Sumathi, "Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction," 2020. [Online]. Available: www.ijacsa.thesai.org

[19] A. Sutou and J. Wang, "Influence-Balanced XGBoost: Improving XGBoost for Imbalanced Data Using Influence Functions," *IEEE Access*, vol. 12, pp. 193473–193486, 2024, doi: 10.1109/ACCESS.2024.3520159.