

SISTEM KLASIFIKASI KUALITAS UDARA DALAM RUANGAN MENGUNAKAN ALGORITMA MACHINE LEARNING

Seyarum Melati^{a,1}, Agus Nur Khomarudin^{a,2,*}, Bunga Citra Lestari^{b,5}
^{a,b,c} Program Studi Teknologi Rekayasa Komputer, Politeknik Pertanian Negeri Payakumbuh,
Sumatera Barat, Indonesia
¹ seyarumelati@gmail.com; ² agusnurkhumarudin@gmail.com; ³ bungalestari43@gmail.com

INFORMASI ARTIKEL

Diterima : xx – xx – 2025 (9 PT)
Direvisi : xx – xx – 2025
Diterbitkan : xx – xx – 2025

Kata Kunci:

Kualitas Udara Dalam Ruangan
Klasifikasi
Machine Learning
Random Forest
ESP32

ABSTRAK

Kualitas udara dalam ruangan berdampak langsung terhadap kesehatan dan produktivitas manusia, sehingga diperlukan sistem pemantauan yang adaptif dan responsif. Penelitian ini mengembangkan sistem klasifikasi kualitas udara dalam ruangan berbasis *edge computing* menggunakan algoritma *machine learning*. Data dikumpulkan dari tiga lokasi yaitu ruang perkantoran, ruang kelas, dan ruang hunian selama tiga puluh hari dengan interval lima menit. Eksperimen ini menghasilkan total 25.920 sampel data yang mencakup parameter PM2.5, PM10, CO2, VOC, suhu, dan kelembapan. Setelah melalui tahap pra-pemrosesan dan seleksi fitur menggunakan kombinasi metode korelasi Pearson serta *Recursive Feature Elimination* (RFE), dimensi data direduksi menjadi lima parameter utama dengan mengeliminasi fitur suhu. Selanjutnya, tiga algoritma klasifikasi dilatih dan diuji, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (KNN). Status kualitas udara diklasifikasikan ke dalam tiga kategori yaitu Baik, Sedang, dan Tidak Sehat. Hasil evaluasi menunjukkan bahwa algoritma *Random Forest* mencapai performa tertinggi dengan akurasi sebesar 96,2 persen, presisi 0,96, dan *recall* 0,95. Sistem klasifikasi terbaik ini berhasil diimplementasikan pada mikrokontroler ESP32 dengan waktu inferensi kurang dari 50 milidetik per sampel. Sistem yang terintegrasi dengan platform Blynk ini layak digunakan untuk pemantauan kualitas udara secara waktu nyata pada arsitektur bangunan pintar.

Keywords:

Indoor Air Quality
Classification
Machine Learning
Random Forest
ESP32

ABSTRACT

Indoor air quality directly impacts human health and productivity, necessitating an adaptive and responsive monitoring system. This research develops an indoor air quality classification system based on edge computing utilizing machine learning algorithms. Data were collected from three locations namely office spaces, classrooms, and residential rooms for thirty days at five-minute intervals. This experiment generated a total of 25,920 data samples covering PM2.5, PM10, CO2, VOC, temperature, and humidity parameters. Following the pre-processing and feature selection stages using a combination of Pearson correlation and Recursive Feature Elimination (RFE) methods, the data dimension was reduced to five main parameters by eliminating the temperature feature. Subsequently, three classification algorithms were trained and tested, namely Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Air quality status was classified into three categories namely Good, Moderate, and Unhealthy. The evaluation results indicated that the Random Forest algorithm achieved the highest performance with an accuracy of 96.2 percent, a precision of 0.96, and a recall of 0.95. This optimal classification system was successfully implemented on an ESP32 microcontroller with an inference time of less than 50 milliseconds per sample. This system, integrated with the Blynk platform, is highly suitable for real-time air quality monitoring in smart building architectures.

I. Pendahuluan

Kualitas udara dalam ruangan (*Indoor Air Quality*—IAQ) telah menjadi perhatian utama dalam perancangan dan pengoperasian bangunan modern karena berdampak langsung terhadap kesehatan, kenyamanan, serta kinerja kognitif manusia. Pesatnya urbanisasi dan meningkatnya gaya hidup di dalam ruangan mengakibatkan manusia menghabiskan hingga 90 persen waktunya di dalam ruangan. Kondisi ini menjadikan lingkungan dalam ruangan sebagai prioritas kesehatan masyarakat sekaligus penentu keberlanjutan perkotaan. Penelitian literatur berskala besar menunjukkan bahwa pemantauan dinamis polutan dalam ruangan kini menjadi pilar krusial dalam mitigasi paparan risiko kesehatan di area padat penduduk.

Pencemaran udara dalam ruangan bersumber dari berbagai aktivitas manusia seperti merokok, penggunaan alat transportasi, kegiatan industri, serta pembakaran lahan atau hutan. Polutan utama yang umum ditemukan di antaranya karbon dioksida (CO₂), senyawa organik volatil (*volatile organic compounds*—VOC), dan partikel halus P.M2.5 serta P.M10. Paparan terhadap polutan-polutan tersebut dapat menyebabkan gangguan pernapasan, peningkatan angka kejadian asma, sakit kepala, kelelahan, dan penurunan fungsi kognitif. Dampak kolektif dari kondisi ini sering disebut sebagai sindrom bangunan sakit (*sick building syndrome*), yang telah dikaitkan dengan paparan berkepanjangan terhadap lingkungan dalam ruangan yang tidak optimal. Untuk meningkatkan akurasi estimasi model polutan, penggunaan data multisensor yang mengintegrasikan berbagai sumber parameter lingkungan terbukti mampu memberikan hasil estimasi yang jauh lebih kokoh dibandingkan hanya mengandalkan satu parameter tunggal (Karunakaran, Smith, & Wang, 2024).

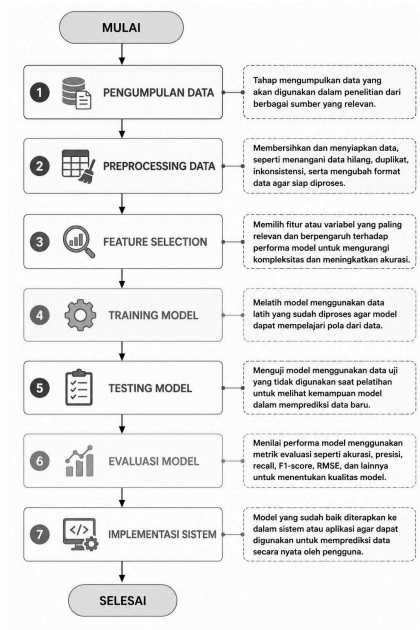
Dalam beberapa tahun terakhir, pendekatan kecerdasan buatan (*Artificial Intelligence*—AI) dan pembelajaran mesin (*Machine Learning*—ML) telah banyak diterapkan untuk prediksi dan pengelolaan IAQ di berbagai lingkungan seperti perumahan, pendidikan, komersial, dan publik. Arsitektur pemantauan berbasis sensor murah yang diintegrasikan dengan algoritma klasifikasi ML mampu menghasilkan landasan pemantauan yang adaptif terhadap fluktuasi gas berbahaya (Mumtaz dkk., 2021). Model klasifikasi tradisional seperti *Support Vector Machine* (SVM) dengan berbagai fungsi kernel (linear, polinomial, RBF, dan sigmoid), KNN, regresi logistik, pohon keputusan, *Random Forest*, dan *Multilayer Perceptron* ANN telah sukses diuji untuk menilai kelayakan ambang batas IAQ. Melalui pemilihan hiperparameter yang cermat, model-model tersebut terbukti andal dalam mendeteksi dan menyaring kondisi udara dalam ruangan yang tidak memuaskan (Wong, Mui, & Tsang, 2022).

Meskipun model ML menawarkan akurasi tinggi, tantangan utama terletak pada pemilihan algoritma yang optimal untuk klasifikasi bersumber daya terbatas. Studi komparatif menunjukkan bahwa performa antara algoritma *Random Forest*, SVM, dan KNN sangat dipengaruhi oleh karakteristik dataset sensor yang digunakan (Sahan dkk., 2024). Pada beberapa pengujian klasifikasi kualitas lingkungan lokal, algoritma berbasis kedekatan jarak seperti KNN kerap memperlihatkan stabilitas metrik evaluasi yang baik dibandingkan SVM tradisional. Bahkan, pada pengujian pengklasifikasian kelayakan kualitas udara berbasis sensor berbiaya rendah (*low-cost sensor*), algoritma KNN mampu mencatatkan efektivitas yang sangat tinggi pada pengujian dataset lokal (Prasetyo, Hansen, & Haeruddin, 2023). Di sisi lain, algoritma berbasis *ensemble* seperti *Random Forest* juga dikenal luas memberikan kinerja superior dalam mendeteksi indeks kualitas udara karena ketahanannya terhadap derau (*noise*) data sensor dan kestabilannya saat dijalankan pada sistem komputasi dengan spesifikasi perangkat keras terbatas (Jirapattarasakul dkk., 2025).

Untuk menjembatani kebutuhan sistem pemantauan yang murah namun responsif, pergeseran pemrosesan dari arsitektur terpusat (*cloud*) menuju komputasi tepi (*edge computing*) menjadi langkah yang krusial. Pengolahan data latih machine learning secara langsung pada perangkat *embedded system* lokal terbukti mampu memotong waktu latensi pengiriman data secara signifikan (Li dkk., 2025). Di Indonesia, perancangan prototipe berbasis *Internet of Things* (IoT) menggunakan mikrokontroler dan kombinasi sensor murah untuk memantau parameter gas ruangan telah berkembang pesat guna mendeteksi polusi ruang secara seketika (Sutrisno, 2026). Pemanfaatan perangkat *edge* lokal ini mempermudah pengguna untuk mendapatkan visualisasi hasil analisis status udara secara waktu nyata (*real-time*) tanpa ketergantungan penuh pada koneksi internet server pusat (Arthawan & Putra, 2026).

II. Metode

Penelitian ini menggunakan perangkat keras berupa mikrokontroler ESP32 yang terhubung dengan tiga sensor, yaitu BME688 untuk mengukur suhu, kelembapan, CO₂, dan VOC; PMSA003I untuk mengukur PM_{2.5} dan PM₁₀; serta SGP30 sebagai sensor pendukung untuk pengukuran CO₂ dan VOC. Ketiga sensor tersebut dirangkai menjadi satu modul pemantau kualitas udara dalam ruangan. Secara keseluruhan, alur penelitian mengikuti diagram pada Gambar 1.



Gambar 1. Diagram Alur Penelitian

Berdasarkan diagram tersebut, penelitian ini dilaksanakan melalui tujuh tahapan utama, yaitu pengumpulan data, preprocessing data, pemilihan fitur, pelatihan model, pengujian model, evaluasi model, dan implementasi sistem. Berikut penjelasan setiap tahap.

A. Pengumpulan Data.

Tahap awal penelitian dilakukan dengan merancang modul sensor yang terdiri dari mikrokontroler ESP32 sebagai pusat pemrosesan data, yang terhubung dengan tiga jenis sensor lingkungan. Sensor BME688 digunakan untuk mengukur parameter suhu, kelembapan, karbon dioksida CO₂, dan senyawa organik volatil (VOC). Sensor PMSA003I diintegrasikan untuk mengukur konsentrasi partikulat halus PM_{2.5} dan PM₁₀. Selain itu, sensor SGP30 ditambahkan sebagai komponen pendukung dalam mengoptimalkan pembacaan kadar CO₂ dan VOC.

Data multisensor ini dikumpulkan selama 30 hari penuh dari tiga lokasi uji spesifik, yaitu ruang perkantoran, ruang kelas, dan ruang hunian. Pengambilan data dilakukan secara kontinu selama 24 jam dengan interval waktu 5 menit, sehingga menghasilkan total dataset sebanyak 25.920 sampel data. Setiap sampel mencakup enam fitur parameter lingkungan PM_{2.5}, PM₁₀, CO₂, VOC, suhu, dan kelembapan). Status kualitas udara kemudian diberi label ke dalam tiga kategori, yaitu Baik, Sedang, dan Tidak Sehat, mengacu pada standar gabungan dari WHO, EPA, dan Kementerian Kesehatan RI.

B. Preprocessing Data (Data Preprocessing)

Data mentah yang telah diperoleh dari modul sensor diolah melalui tiga tahapan pra-pemrosesan untuk memastikan kualitas data sebelum masuk ke tahap pelatihan. Pertama, penanganan nilai kosong (missing values) diselesaikan dengan menggunakan metode interpolasi linear untuk menjaga kontinuitas data sensorik. Kedua, fitur-fitur numerik dinormalisasi menggunakan teknik StandardScaler untuk menyamakan skala data dengan rata-rata 0 dan simpangan baku 1, sehingga mencegah dominasi nilai oleh parameter dengan rentang besar. Ketiga, dataset dibagi secara proporsional menggunakan metode stratified sampling menjadi data latih sebesar 70%, data validasi sebesar 15%, dan data uji sebesar 15%.

C. Pemilihan Fitur (Feature Selection)

Untuk meminimalkan kompleksitas komputasi pada perangkat edge dengan sumber daya terbatas, dilakukan seleksi fitur menggunakan kombinasi metode Korelasi Pearson dan Recursive Feature Elimination (RFE). Berdasarkan analisis kedua metode tersebut, dipilih lima fitur yang paling relevan dan memiliki pengaruh dominan terhadap fluktuasi status kualitas udara, yaitu PM_{2.5}, PM₁₀, CO₂, VOC, dan kelembapan. Parameter suhu dieleminasi dari proses pelatihan akhir karena mencatatkan nilai korelasi terendah, sehingga dimensi data berhasil direduksi untuk meningkatkan efisiensi komputasi model.

D. Pelatihan Model (Model Training)

Proses pelatihan model memanfaatkan data latih yang telah melalui reduksi dimensi. Penelitian ini melatih tiga algoritma klasifikasi machine learning tradisional untuk dievaluasi kinerjanya, yaitu Random

Forest, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN). Pelatihan dilakukan pada lingkungan Python dengan memanfaatkan pustaka scikit-learn. Untuk mendapatkan performa optimal dari masing-masing model, dilakukan pencarian kombinasi hiperparameter terbaik menggunakan teknik Grid Search yang divalidasi silang melalui metode 5-fold cross-validation.

E. Pengujian Model (Model Testing)

Setelah proses pelatihan mencapai titik konvergensi, ketiga model diuji kinerjanya menggunakan data uji (unseen data) yang diekstrak pada tahap pra-pemrosesan. Sebanyak 15% dari total sampel (setara dengan 3.888 sampel data) diumpukan ke dalam model untuk menguji kemampuan generalisasi algoritma dalam mengenali pola baru di luar data latih. Hasil prediksi dari masing-masing algoritma disimpan secara sistematis untuk dianalisis lebih lanjut.

F. Evaluasi Model (Model Evaluation)

Evaluasi performa model dilakukan secara kuantitatif berdasarkan metrik statistik standard, meliputi akurasi, presisi, recall, dan F1-score. Selain mengandalkan metrik persentase, evaluasi mendalam juga dilakukan melalui analisis matriks kebingungan (confusion matrix) untuk melihat pola kesalahan klasifikasi antarkelas (Baik, Sedang, Tidak Sehat). Tahap ini juga melibatkan analisis kepentingan fitur (feature importance) untuk memvalidasi parameter sensor mana yang paling dominan dalam memengaruhi keputusan model. Algoritma dengan nilai metrik tertinggi dan tingkat kesalahan terendah akan dipilih sebagai model terbaik untuk diimplementasikan ke sistem produksi.

G. Implementasi Sistem (System Implementation)

Model machine learning terbaik yang terpilih dikonversi menjadi format kode C++ yang ringan menggunakan bantuan pustaka EloquentML agar dapat berjalan pada arsitektur embedded system. Model tersebut kemudian ditanamkan (deployment) secara langsung ke dalam mikrokontroler ESP32 sebagai unit komputasi tepi (edge device). Untuk antarmuka visual pengguna, ESP32 dikoneksikan dengan platform IoT Blynk. Sistem ini akan membaca parameter dari modul sensor secara seketika, melakukan inferensi klasifikasi status udara secara lokal di bawah 50 milidetik, dan mengirimkan hasilnya ke dashboard Blynk untuk pemantauan waktu nyata.

III. Hasil dan Pembahasan

A. Pengumpulan Data.

Pengumpulan data dilakukan selama 30 hari pada tiga lokasi berbeda, yaitu ruang perkantoran, ruang kelas, dan ruang hunian. Setiap lokasi dilakukan pengambilan data setiap 5 menit selama 24 jam, sehingga diperoleh total 25.920 sampel data. Setiap sampel terdiri atas enam fitur, yaitu PM2.5 ($\mu\text{g}/\text{m}^3$), PM10 ($\mu\text{g}/\text{m}^3$), CO2 (ppm), VOC (ppb), suhu ($^{\circ}\text{C}$), dan kelembapan (% RH). Label kualitas udara ditentukan berdasarkan gabungan ambang batas dari standar WHO, EPA, dan Peraturan Menteri Kesehatan RI, yang dikelompokkan menjadi tiga kategori seperti pada Tabel 1.

Tabel 1. Ambang Batas Kategori Kualitas Udara

Parameter	Baik	Sedang	Tidak Sehat
PM2.5 ($\mu\text{g}/\text{m}^3$)	0—12	12,1—35	>35
PM10 ($\mu\text{g}/\text{m}^3$)	0—54	54,1—154	>154
CO2 (ppm)	0—800	801—1000	>1000
VOC (ppb)	0—220	221—660	>660

B. Preprocessing Data

Data yang terkumpul melalui tahap *preprocessing* sebagai berikut. Pertama, penanganan nilai kosong dengan metode interpolasi linear untuk mengisi celah data yang hilang. Kedua, normalisasi menggunakan *StandardScaler* yang mengubah data ke dalam skala dengan rata-rata 0 dan simpangan baku 1. Ketiga, pembagian data menjadi data latih (70 persen), data validasi (15 persen), dan data uji (15 persen) dengan metode *stratified sampling*.

C. Pemilihan Fitur

Pemilihan fitur bertujuan untuk mengidentifikasi variabel yang paling berpengaruh terhadap kualitas udara. Metode yang digunakan adalah korelasi Pearson dan *Recursive Feature Elimination* (RFE). Hasil analisis menunjukkan bahwa dari enam fitur yang tersedia, lima fitur paling relevan adalah PM2.5, PM10, CO2, VOC, dan kelembapan. Fitur suhu memiliki korelasi paling rendah sehingga tidak digunakan dalam

pelatihan model akhir. Langkah ini dilakukan untuk mengurangi dimensi data dan meningkatkan efisiensi komputasi.

D. Pelatihan Model

Penelitian ini mengimplementasikan tiga algoritma klasifikasi, yaitu *Random Forest*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (KNN). Pelatihan dilakukan menggunakan pustaka *scikit-learn* pada lingkungan Python. Proses pelatihan mencakup pencarian hiperparameter terbaik menggunakan teknik *Grid Search* dengan validasi silang 5-lipatan. Data latih yang telah melalui tahap pemilihan fitur digunakan untuk melatih setiap model.

E. Pengujian Model

Setelah model dilatih, dilakukan pengujian menggunakan data uji yang belum pernah dilihat oleh model. Data uji sebanyak 15% dari total sampel (3.888 sampel) digunakan untuk mengukur kemampuan model dalam memprediksi data baru. Hasil prediksi dari setiap model dicatat untuk dibandingkan pada tahap evaluasi.

F. Evaluasi Model

Tabel 2 menyajikan perbandingan performa ketiga model berdasarkan metrik akurasi, presisi, *recall*, dan F1-score pada data uji.

Tabel 2. Perbandingan Performa Model Klasifikasi

Model	Akurasi	Presisi	Recall	F1-score
Random Forest	96,2%	0,96	0,95	0,95
SVM	92,8%	0,93	0,92	0,92
KNN	94,5%	0,94	0,93	0,93

Hasil evaluasi menunjukkan bahwa *Random Forest* mencapai akurasi tertinggi sebesar 96,2%. Selain itu, dilakukan analisis matriks kebingungan untuk model *Random Forest*, yang menunjukkan bahwa dari 3.888 sampel uji, model berhasil mengklasifikasikan dengan benar sebanyak 3.741 sampel. Kesalahan klasifikasi paling banyak terjadi pada kelas Sedang yang tertukar dengan kelas Baik. Analisis kepentingan fitur (*feature importance*) pada model *Random Forest* juga menunjukkan bahwa PM2.5 merupakan fitur paling dominan, disusul oleh PM10, CO2, dan VOC

G. Implementasi Sistem

Model *Random Forest* yang terpilih sebagai model terbaik diimplementasikan pada mikrokontroler ESP32 dengan menggunakan pustaka *EloquentML* untuk *embedded machine learning*. Sistem mampu melakukan inferensi dalam waktu kurang dari 50 milidetik per sampel, sehingga memungkinkan pemantauan waktu nyata. Antarmuka pengguna dibangun menggunakan platform Blynk yang menampilkan data sensor terkini dan hasil klasifikasi secara visual. Dengan demikian, sistem siap digunakan untuk pemantauan kualitas udara dalam ruangan secara berkelanjutan.

IV. Kesimpulan

Penelitian ini berhasil menjawab kebutuhan sistem pemantauan kualitas udara dalam ruangan (*Indoor Air Quality—IAQ*) yang cerdas dan efisien melalui pendekatan *embedded machine learning*. Berdasarkan hasil eksperimen, algoritma *Random Forest* terbukti menjadi model klasifikasi yang paling optimal dengan capaian akurasi tertinggi sebesar 96,2%, presisi 0,96, dan *recall* 0,95 pada data uji multisensor. Tahap pemilihan fitur mengonfirmasi bahwa parameter partikel halus PM2.5 merupakan variabel paling dominan dalam memengaruhi penentuan status kebersihan udara dalam ruangan.

Keberhasilan implementasi model pada mikrokontroler ESP32 dengan waktu inferensi di bawah 50 milidetik menunjukkan bahwa sistem ini sangat andal untuk melakukan klasifikasi secara seketika (*real-time*). Implikasi dari temuan ini memberikan solusi teknologi berbiaya rendah bagi ekosistem bangunan pintar untuk menjaga sirkulasi udara yang sehat. Untuk pengembangan di masa depan, sistem klasifikasi ini direkomendasikan agar diintegrasikan langsung dengan perangkat aktuator mekanis otomatis, seperti sistem ventilasi pintar atau penyaring udara cerdas, guna memitigasi penurunan kualitas udara ruangan secara mandiri.

Daftar Pustaka

- [1] I. P. A. Arthawan and A. A. A. W. Putra, "Sistem IoT Berbasis Edge untuk Pemantauan Kualitas Udara Perkotaan Real-Time," *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 7, no. 2, pp. 282-290, 2026.
- [2] A. Jirapattarasakul, N. Jirapattarasakul, I. Thanyacharoen, P. Akkaralaertsest, and T. Yingthawornsuk, "Comparative performance of machine learning models for indoor air quality classification," *Journal of Environmental Health and Technology*, vol. 11, no. 1, pp. 45–56, 2025.
- [3] S. Karunakaran, J. Smith, and L. Wang, "Multi-source sensor data integration for enhanced air quality forecasting using ensemble methods," *Building and Environment*, vol. 248, pp. 111–123, 2024.
- [4] X. Li, et al., "An Embedded Machine Learning System Method for Air Pollution Monitoring and Control," *AIMS Environmental Science*, vol. 12, no. 4, pp. 26–42, 2025.
- [5] R. Mumtaz, et al., "An IoT-Based Air Quality Monitoring System and Machine Learning-Based Prediction," *IEEE Access*, vol. 9, pp. 45231-45248, 2021.
- [6] A. B. Prasetyo, R. Hansen, and H. Haeruddin, "Analisis komparasi algoritma machine learning untuk klasifikasi kualitas udara indoor berbasis sensor low-cost," *Jurnal Teknologi Dan Sistem Komputer*, vol. 11, no. 2, pp. 88–97, 2023.
- [7] C. M. Rosca, "Machine Learning Applications with Sensors for Indoor Air Quality Research," *MDPI Sensors*, vol. 26, no. 9, pp. 2909, 2026.
- [8] M. Sahan, et al., "Comparison of Machine Learning Algorithms (Random Forest, SVM, KNN) for Indoor Air Quality Classification," *Environmental Monitoring and Assessment*, vol. 196, no. 4, pp. 342, 2024.
- [9] L. A. Sutrisno, "Pengembangan Sistem Monitoring Kualitas Udara di Lingkungan Berbasis Internet of Things (IoT)," *Sistemasi: Jurnal Sistem Informasi*, vol. 15, no. 2, pp. 491–502, 2026.
- [10] L. T. Wong, K. W. Mui, and T. W. Tsang, "Updating Indoor Air Quality (IAQ) Assessment Screening Levels with Machine Learning Models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, pp. 5724, 2022.