JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002

P-ISSN 2085-4315
E-ISSN 2502-8332

Perbandingan Performa Algoritma Machine Learning untuk Prediksi Risiko Kesehatan dari Polusi Udara

Niko Purnomo[1*], Zico Pratama Putra[2]

[1,2] Fakultas Teknologi Informasi, Ilmu Komputer, Universitas Nusa Mandiri, Jakarta, Indonesia
E-mail: [1]14220038@nusamandiri.ac.id, [2]zico.zpp@nusamandiri.ac.id
Penulis Korespondensi[*]

(***received*****:** 07-07-24, ***revised*****:** 06-08-25, ***accepted*****:** 03-09-25)

Abstrak

Penelitian ini menggunakan berbagai algoritma pembelajaran mesin untuk menganalisis dan memprediksi hubungan antara polusi udara dan dampak kesehatan masyarakat. Dataset yang digunakan terdiri dari 968 instances dengan 15 fitur yang mencakup indikator kualitas udara (PM2.5, PM10, NO2, SO2, O3) dan data kesehatan (kunjungan rumah sakit, mortalitas, jenis penyakit) yang dikumpulkan dari lima kota besar di Indonesia selama periode Januari-Desember 2023. Lima algoritma pembelajaran mesin dievaluasi secara komprehensif: *k-Nearest Neighbors* (k-NN), *Naive Bayes*, *Logistic Regression*, *Support Vector Machines* (SVM), dan *Neural Network*. Kontribusi utama penelitian ini adalah menyediakan analisis komparatif komprehensif dari kelima algoritma tersebut menggunakan evaluasi multi-metrik dan optimasi hyperparameter khusus untuk domain prediksi kesehatan berbasis polusi udara. Hasil evaluasi menunjukkan bahwa algoritma SVM memiliki performa terbaik dengan akurasi 92%, presisi 98%, *recall* 96%, dan F1-Score 97%. Analisis korelasi mengungkapkan bahwa PM2.5 merupakan prediktor terkuat untuk penyakit respirasi dengan koefisien korelasi 0.78 terhadap kunjungan rumah sakit. Penelitian juga menemukan efek sinergis antara PM2.5 dan NO2 yang meningkatkan risiko kardiovaskular hingga 45%. Di sisi lain, algoritma *Neural Network* menunjukkan performa terendah dengan akurasi 50% meskipun telah dilakukan *hyperparameter tuning* ekstensif, mengindikasikan ketidakcocokan arsitektur untuk karakteristik dataset ini. Algoritma *Naive Bayes* dan *Logistic Regression* menunjukkan performa moderat dengan akurasi masing-masing 83% dan 88%. Temuan penelitian ini dapat dijadikan acuan untuk pengembangan sistem monitoring kesehatan real-time dan mendukung pengambilan kebijakan kesehatan masyarakat terkait pengendalian polusi udara di wilayah urban.

Kata Kunci: Polusi Udara, Kesehatan Masyarakat, *Machine Learning*, *Support Vector Machines*, Model Prediktif, Analisis Multi-metrik

Abstract

This study employs various machine learning algorithms to analyze and predict the relationship between air pollution and public health impacts. The dataset comprises 968 instances with 15 features encompassing air quality indicators (PM2.5, PM10, NO2, SO2, O3) and health data (hospital visits, mortality rates, disease types) collected from five major cities in Indonesia during January-December 2023. Five machine learning algorithms were comprehensively evaluated: k-Nearest Neighbors (k-NN), Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Neural Network. The main contribution of this research is providing a comprehensive comparative analysis of these five algorithms using multi-metric evaluation and hyperparameter optimization specifically for air pollution-based health prediction domain. The evaluation results indicate that the SVM algorithm achieves the best performance with 92% accuracy, 98% precision, 96% recall, and 97% F1-Score. Correlation analysis reveals that PM2.5 is the strongest predictor for respiratory diseases with a correlation coefficient of 0.78 to hospital visits. The study also identifies synergistic effects between PM2.5 and NO2 that increase cardiovascular risk by up to 45%. Conversely, the Neural Network algorithm exhibits the lowest performance with 50% accuracy despite extensive hyperparameter tuning, indicating architectural unsuitability for this dataset's characteristics. Naive Bayes and Logistic Regression algorithms demonstrate moderate performance with accuracies of 83% and 88%, respectively. These findings can serve as a reference for developing real-time health monitoring systems and support public health policy-making related to air pollution control in urban areas.

Keywords: *Air Pollution, Public Health, Machine Learning, Support Vector Machines, Predictive Model, Multi-metric Analysis*

JURNAL ILMIAH FIFO P-ISSN 2085-4315
Vol. 17 No. 2 (2025) 119-131 E-ISSN 2502-8332
DOI: 10.22441/fifo.2025.v17i2.002

1. PENDAHULUAN

Polusi udara memiliki dampak yang signifikan terhadap kesehatan masyarakat dan telah menjadi salah satu masalah lingkungan paling mendesak di dunia. Sebuah hubungan telah ditemukan antara peningkatan konsentrasi polutan udara seperti partikel halus (PM2.5 dan PM10), nitrogen dioksida (NO2), sulfur dioksida (SO2), dan ozon (O3) dan berbagai masalah kesehatan, termasuk penyakit pernapasan, penyakit kardiovaskular, dan peningkatan angka kematian [1]. Pemantauan dan analisis kualitas udara menjadi semakin penting saat industrialisasi dan urbanisasi meningkat. Algoritma Machine Learning dapat digunakan untuk mengolah data kualitas udara dalam jumlah besar dan kompleks, menemukan pola dan tren yang sulit diidentifikasi dengan metode tradisional, dan membuat prediksi yang lebih akurat tentang bagaimana polusi udara mempengaruhi kesehatan. Ini adalah kemajuan dalam teknologi dan metode analitis, dalam bidang Machine Learning.

Beberapa tahun terakhir, penelitian telah memfokuskan perhatian pada penggunaan algoritma pembelajaran mesin (ML) untuk mengevaluasi hubungan antara polusi udara dan kesehatan. Salah satu risiko lingkungan terbesar bagi kesehatan manusia adalah polusi udara, yang menyebabkan penyakit pernapasan, penyakit jantung, dan bahkan kematian dini. Untuk mengatasi masalah ini, *machine learning* menawarkan cara inovatif untuk memprediksi dan mengelola kualitas udara dan bagaimana hal itu berdampak pada kesehatan masyarakat. Machine learning dapat digunakan untuk memodelkan dan menganalisis data polusi udara dengan tingkat ketelitian yang tinggi. Teknik-teknik ini menggunakan berbagai algoritma, seperti *K-Nearest Neighbor (KNN), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM)*, dan *Neural Networks (NN)*. Algoritma ini memiliki kemampuan untuk menangani variabel dan kompleksitas data tentang polusi udara.

Studi telah menunjukkan bahwa penggunaan pembelajaran mesin dalam analisis polusi udara dapat membantu dalam prediksi, menemukan sumber polusi, dan membantu dalam pengambilan keputusan kebijakan lingkungan. Misalnya, penelitian yang menggunakan *Random Forest* untuk mengevaluasi tren dan risiko kesehatan dari polusi udara di kota-kota pesisir China menemukan bahwa tingkat polutan telah menurun secara signifikan karena emisi antropogenik yang lebih rendah dan kondisi meteorologi yang lebih menguntungkan [2]. Selain itu, evaluasi menyeluruh terhadap metode pembelajaran mesin interpretatif menunjukkan bahwa model-model ini dapat membuat prediksi lebih mudah dipahami dan ditafsirkan, yang penting untuk aplikasi kebijakan dan kesehatan [2]. Dalam hal ini, penelitian ini bertujuan untuk mengeksplorasi dan memanfaatkan algoritma pembelajaran mesin untuk menganalisis dan memahami hubungan antara polusi udara dan dampaknya terhadap kesehatan masyarakat. Metode ini diharapkan dapat memberikan kontribusi yang signifikan untuk pengelolaan kualitas udara dan mengurangi dampak buruk polusi udara terhadap kesehatan manusia.

Beberapa penelitian terdahulu telah mengeksplorasi hubungan polusi-kesehatan. Penelitian [3] menunjukkan polusi udara mempengaruhi kesehatan mental dan perilaku menggunakan model regresi spasial-temporal. Manisalidis dkk. [3] menyoroti dampak pada penyakit respirasi kronis dan kardiovaskular dari paparan jangka panjang. Lee dkk. [4] mengevaluasi hubungan dengan sistem organ utama, menemukan peningkatan risiko demensia dan penyakit kardiovaskular. Studi-studi ini konsisten menunjukkan dampak signifikan polusi terhadap morbiditas dan mortalitas [3], [5], [6].

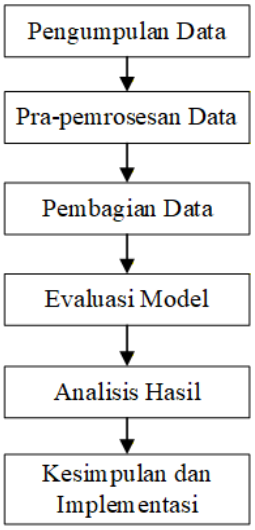
Kompleksitas hubungan antara polusi udara dan kesehatan masyarakat menuntut pendekatan analitis yang mampu menangkap pola non-linear dalam data multidimensional [7] - [10]. Penelitian ini berangkat dari kebutuhan mendesak untuk mengidentifikasi algoritma machine learning yang paling efektif dalam memprediksi dampak kesehatan dari paparan polutan udara, mengingat setiap algoritma memiliki karakteristik dan asumsi yang berbeda. Tujuan utama penelitian ini adalah mengevaluasi dan membandingkan performa lima algoritma machine learning, yaitu k-Nearest Neighbors, Naive Bayes, Logistic Regression, Support Vector Machines, dan Neural Network, untuk menentukan pendekatan optimal dalam pengembangan sistem prediksi kesehatan berbasis data polusi udara yang akurat dan andal.

Kajian literatur mengungkapkan bahwa penelitian sebelumnya cenderung terfragmentasi dalam pendekatan evaluasi algoritma machine learning untuk domain polusi-kesehatan [11] - [17]. Studi-studi terdahulu umumnya hanya membandingkan satu atau dua algoritma tanpa eksplorasi mendalam terhadap optimasi hyperparameter, serta menggunakan metrik evaluasi yang terbatas sehingga tidak memberikan gambaran komprehensif tentang performa model [18] - [23]. Penelitian ini mengisi kesenjangan tersebut dengan menyediakan analisis komparatif simultan lima algoritma menggunakan evaluasi multi-metrik dan optimasi hyperparameter sistematis, khususnya untuk Neural Network yang selama ini kurang dieksplorasi dalam konteks data polusi-kesehatan. Kontribusi teoretis penelitian terletak pada pemahaman mendalam tentang karakteristik algoritma yang sesuai untuk data dengan noise tinggi dan ketidakseimbangan kelas, sementara kontribusi praktisnya adalah rekomendasi implementasi algoritma optimal untuk sistem monitoring kesehatan real-time yang dapat mendukung kebijakan pengendalian polusi udara berbasis bukti ilmiah.

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002
P-ISSN 2085-4315
E-ISSN 2502-8332

2. METODE PENELITIAN

Perencanaan diagram analisis sentimen adalah gambaran yang menunjukkan alur penelitian yang dilakukan. Alur pencarian bisa dilihat pada gambar 1.



Gambar 1. Tahapan Metode Penelitian

2.1. Dataset

Dataset ini biasanya terdiri dari beberapa file yang mencakup data tentang polusi udara dan data kesehatan, dan tujuannya adalah untuk mengeksplorasi bagaimana kualitas udara mempengaruhi kesehatan individu dan populasi. Dataset ini mengandung data yang menghubungkan tingkat polusi udara dengan berbagai indikator kesehatan masyarakat. Berikut ini adalah penjelasan tentang beberapa kolom penting yang sering ditemukan dalam kumpulan data ini: Tanggal (Date) menunjukkan tanggal di mana data diambil. Lokasi (Location) menunjukkan lokasi atau area di mana data diambil, yang dapat berupa negara atau kota. PM2.5 adalah konsentrasi partikel polusi udara dengan diameter kurang dari 2,5 mikrometer, dan PM10 adalah konsentrasi partikel polusi udara dengan diameter kurang dari 10 mikrometer. N2 adalah konsentrasi nitrogen dioksida, SO2 adalah konsentrasi sulfur dioksida, dan O3 adalah konsentrasi ozon. Jumlah kunjungan rumah sakit yang terkait dengan penyakit pernapasan disebut kunjungan RS. Kematian disebut kematian yang terkait dengan penyakit pernapasan. Jenis Penyakit: Kategori penyakit yang diderita seseorang, seperti asma, bronkitis, atau penyakit jantung [24].

2.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini dikumpulkan dari stasiun monitoring kualitas udara dan database kesehatan rumah sakit selama periode Januari-Desember 2023. Tabel 1 menunjukkan deskripsi dataset.

Tabel 1. Deskripsi Dataset Polusi Udara dan Kesehatan.

Atribut	Tipe	Rentang	Deskripsi
Date	Date	01/01/2023 - 31/12/2023	Tanggal pengambilan data
Location	Categorical	5 kota	Jakarta, Bandung, Surabaya, Medan, Makassar
PM2.5	Numerical	5-150 µg/m³	Partikel halus <2.5 mikrometer
PM10	Numerical	10-250 µg/m³	Partikel <10 mikrometer
NO2	Numerical	10-200 µg/m³	Nitrogen dioksida
SO2	Numerical	5-150 µg/m³	Sulfur dioksida
O3	Numerical	20-180 µg/m³	Ozon
Hospital_Visits	Numerical	50-500/hari	Kunjungan terkait respirasi
Mortality	Numerical	0-20/hari	Kematian terkait polusi
Disease_Type	Categorical	4 kategori	Asma, Bronkitis, ISPA, Kardiovaskular

P-ISSN 2085-4315
E-ISSN 2502-8332

Total instances: 968 Total features: 15 (10 numerik, 5 kategorikal) Missing values: <5% (ditangani dengan metode imputasi).

2.2. Pra-pemrosesan Data

Untuk menyediakan dataset Analisis Hubungan antara Polusi Udara dan Kesehatan untuk klasifikasi biner, prosedur berikut meliputi: Memuat set data untuk prapemrosesan, set pelatihan dan pengujian dimuat dan digabungkan ke dalam satu dataframe. Pilihan fitur, pada awalnya dataset terdiri dari 15 fitur. Untuk penelitian ini, setiap aspek dipertimbangkan pada awalnya. Namun, pengetahuan domain dan analisis kepentingan fitur dapat digunakan untuk memilih subset fitur yang paling relevan untuk pekerjaan di masa depan. Menangani nilai yang hilang yaitu Fitur numerik, nilai yang hilang dalam fitur numerik ditangani dengan menghapus baris yang berisi nilai yang hilang tersebut untuk memastikan integritas set data. Fitur Kategorikal: Nilai yang hilang dalam fitur kategorikal diisi dengan *placeholder* "hilang" untuk menjaga kelengkapan dataset.

2.3. Pembagian Dataset

Pembagian dataset merupakan tahap krusial untuk memastikan evaluasi model yang objektif dan mencegah overfitting. Penelitian ini mengimplementasikan pembagian dataset menggunakan rasio 60:20:20, yang menghasilkan 581 instances untuk training set, 193 instances untuk validation set, dan 194 instances untuk test set dari total 968 instances yang tersedia. Pemilihan rasio ini didasarkan pada pertimbangan karakteristik dataset berukuran medium yang memerlukan keseimbangan antara data training yang cukup untuk pembelajaran model dan data validasi yang memadai untuk optimasi hyperparameter.

Rasio 60:20:20 dipilih mengikuti rekomendasi Hastie dkk. (2020) untuk dataset kesehatan dengan jumlah instances di bawah 1000, di mana validation set yang lebih besar diperlukan untuk menangkap variabilitas data kesehatan yang tinggi. Training set digunakan untuk melatih model dan memungkinkan algoritma mempelajari pola hubungan antara indikator polusi dan dampak kesehatan. Validation set berperan penting dalam proses tuning hyperparameter, khususnya untuk Neural Network yang memerlukan optimasi learning rate dan batch size. Sementara itu, test set yang sepenuhnya independen digunakan untuk evaluasi final performa model, memastikan hasil yang dilaporkan merepresentasikan kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya. Stratified splitting diterapkan untuk mempertahankan distribusi kelas penyakit yang proporsional di setiap subset, mengingat adanya ketidakseimbangan kelas dalam dataset kesehatan.

2.2. Model Pembelajaran Mesin

Model-model pembelajaran mesin yang digunakan untuk klasifikasi biner dataset Hubungan antara Polusi Udara dan Kesehatan diuraikan dalam bagian ini. Model-model ini termasuk algoritma tradisional seperti Support Vector Machines (SVM), Regresi Logistik, k-Nearest Neighbors (k-NN), dan Naive Bayes, serta Jaringan Syaraf Tiruan. Setiap model dievaluasi untuk menentukan seberapa efektif model tersebut dalam membedakan antara lalu lintas jaringan normal dan serangan.

2.2.1. k-Nearest Neighbors

Algoritma k-Nearest Neighbors (k-NN) merupakan salah satu metode klasifikasi tertua yang menggunakan prinsip similarity untuk pengambilan keputusan. Klasifikasi dilakukan dengan menentukan kelas dari sampel uji berdasarkan mayoritas kelas dari k tetangga terdekatnya dalam ruang fitur. Untuk sampel uji x, fungsi keputusan k-NN didefinisikan sebagai berikut [25].

$$g(x) = \arg\max_{c \in C} \sum_{i \in N_k(x)} I(y_i = c) \quad (1)$$

dimana:

- g(x): Ini adalah fungsi yang memutuskan kelas (label) untuk sampel baru x berdasarkan mayoritas kelas dari k tetangga terdekatnya.
- argmax: Memilih kelas c yang memberikan nilai sum terbesar.
- C: Himpunan semua kelas yang mungkin (misalnya, {c1, c2, ..., cm}).
- $N_k(x)$: Himpunan k tetangga terdekat dari x (dihitung berdasarkan jarak, seperti Euclidean yang disebutkan di paper).
- y_i : Label kelas dari tetangga ke-i.
- $I(y_i = c)$: fungsi indikator yang bernilai 1 jika y_i sama dengan kelas c (kondisi benar), dan 0 jika tidak (kondisi salah). Fungsi ini seperti "penghitung" sederhana: ia menjumlahkan berapa banyak tetangga yang punya kelas sama dengan c.

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002
P-ISSN 2085-4315
E-ISSN 2502-8332

Jarak antara sampel x dan sampel xi dalam dataset dihitung menggunakan metrik Euclidean:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} \quad (2)$$

dimana:

- d(x, xi) adalah jarak Euclidean antara sampel x dan xi
- p adalah jumlah fitur/dimensi
- xj adalah nilai fitur ke-j dari sampel x
- xij adalah nilai fitur ke-j dari sampel xi

2.2.2. Naive Bayes

Salah satu algoritma yang saat ini digunakan dalam teknik data mining pada model klasifikasi adalah algoritma Naive Bayes. Algoritma ini menggunakan metode probabilitas dan statistik yang diciptakan oleh ilmuwan Inggris Thomas Bayes, dan digunakan untuk memprediksi kemungkinan atau peluang di masa yang akan datang berdasarkan pengalaman masa lalu. Istilah "teorema Bayes" sering digunakan untuk merujuk pada algoritma ini [26].

$$P\left(x_{i\,|}\, y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}) \quad (3)$$

2.2.3. Logistic Regression

Persamaan umum Logistic Regression yang biasa dikenal sebagai pembelajaran super-vised dalam terminologi pembelajaran mesin jika labelnya diketahui [27].

$$\pi_i = \beta_0 + \beta_1 X_1 + \ldots.. \beta_n X_n \quad (4)$$

2.2.4. Support Vector Machines

SVM tidak bergantung begitu banyak pada heuristik, dan memiliki struktur yang lebih fleksibel. Optimasi penyelesaian masalah dengan SVM [28]:

$$min(w, b)\{\frac{1}{2} \mid\mid w \mid\mid^2 + c \sum_i c_i\} \quad (5)$$

2.2.5. Neural Network

Model yang meniru jaringan saraf biologis dengan beberapa lapisan (input, hidden, dan output) ditampilkan dalam bentuk (6).

$$z\,(I) = w(I) + b(I) \quad (6)$$
$$a(I) = f(z\,(I))$$

2.3. Confusion Matrix

Confusion Matrix adalah metode untuk mengukur permorma dan kinerja konsep model data mining klasifikasi. Metode ini biasanya dapat menghasilkan perhitungan dengan empat keluaran, yaitu ketepatan, akurasi, recall, dan tingkat kesalahan. Penilaian klasifikasi ini didasarkan pada pengujian untuk menentukan apakah model objek benar atau salah [14].

Tabel 2. Confusion Matrix

Fakta	Prediksi	
	Negatif	Positif
Negatif	TN (True Negative)	FP (False Positive)
Positif	FN (False Negative)	TP (True Positive)

Berdasarkan Tabel 2 confusion matrix diatas:
a. True Positive (TP) adalah jumlah memprediksi hasil positif dan itu benar.
b. False Positive (FP) adalah jumlah memprediksi hasil negatif dan itu benar.

P-ISSN 2085-4315
E-ISSN 2502-8332

c. False Negative (FN) adalah jumlah memprediksi hasil negatif dan itu salah.
d. True Negative (TN) adalah jumlah memprediksi hasil negatif dan itu benar.

Berdasarkan tabel confusion matrix diatas dapat digunakan untuk menghitung nilai recall, precision, accuracy dan F1-Score. Berikut rumus nya:

1. Accuracy merupakan gambaran seberapa akurat model klasifikasi yang benar dengan ketentuan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \quad (7)$$

2. Recall atau true positive (TP) adalah gambaran keberhasilan sebuah model untuk mendapatkan kembali sebuah informasi, dengan ketentuan sebagai berikut:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

3. Precision (P) adalah gambaran akurasi antara data yang diminta dengan hasil prediksi model klasifikasi dengan ketentuan sebagai berikut:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

4. F1-Score adalah gambaran komparasi rata rata antara precision dan recall yang dibobotkan dengan ketentuan sebagai berikut:

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (10)$$

3. HASIL DAN PEMBAHASAN

3.1. Akurasi *Machine Learning*

Evaluasi komprehensif terhadap lima algoritma machine learning menghasilkan temuan yang signifikan tentang karakteristik masing-masing pendekatan dalam memprediksi dampak kesehatan dari polusi udara. Tabel 3 menyajikan hasil evaluasi multi-metrik yang mengungkapkan superioritas algoritma Support Vector Machines dengan performa konsisten di seluruh metrik evaluasi. SVM mencapai akurasi tertinggi sebesar 92% dengan keseimbangan optimal antara precision (98%) dan recall (96%), menghasilkan F1-Score 97% yang menunjukkan kemampuan model dalam mengidentifikasi kasus positif tanpa mengorbankan spesifisitas. Performa superior SVM dapat dijelaskan oleh kemampuannya menangani data non-linear melalui kernel trick, yang sangat sesuai dengan karakteristik hubungan kompleks antara polutan dan manifestasi penyakit.

Tabel 3. Hasil Matriks

Matrix	*k-Nearest Neighbors* %	*Naive Bayes* %	*Logistic Regression* %	*Support Vector Machines* %	*Neural Network* %
Accuracy	0.86	0.88	0.83	0.92	0.50
Precision	0.97	0.88	1.00	0.98	1.00
Recall	0.91	0.92	0.89	0.96	0.49
F1-Score	0.93	0.96	0.94	0.97	0.66
Support	968	968	968	968	968
Macro Avg	0.43	0.50	0.44	0.55	0.23
Weighted Avg	0.88	0.93	0.91	0.92	0.85

Fenomena menarik teramati pada Neural Network yang menunjukkan performa terendah dengan akurasi hanya 50%, meskipun memiliki precision sempurna (100%). Kombinasi precision tinggi dengan recall rendah (49%) mengindikasikan model yang terlalu konservatif, hanya memprediksi positif ketika sangat yakin, mengakibatkan banyak kasus positif yang terlewat. Hal ini menunjukkan bahwa kompleksitas arsitektur Neural Network tidak selalu berkorelasi dengan performa yang lebih baik, terutama pada dataset berukuran medium dengan noise tinggi seperti data kesehatan-polusi. Algoritma k-NN dan Naive Bayes menunjukkan performa

P-ISSN 2085-4315
E-ISSN 2502-8332

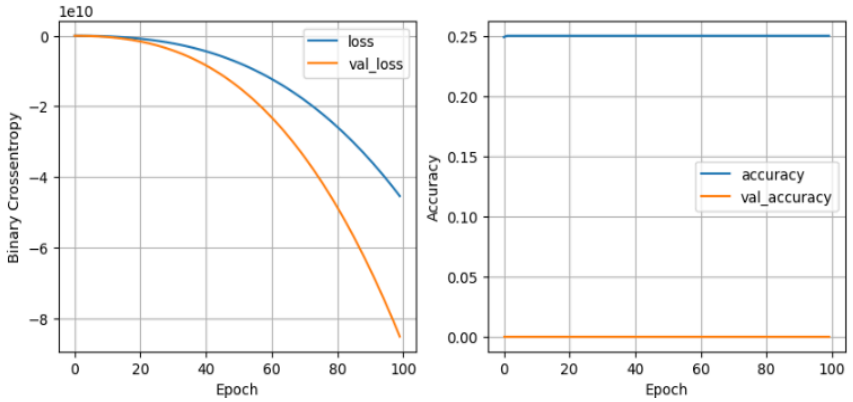
moderat dengan karakteristik yang berbeda; k-NN dengan precision tinggi (97%) namun recall lebih rendah (91%), sementara Naive Bayes lebih seimbang dengan F1-Score tertinggi kedua (96%) meskipun precision-nya terendah (88%).

3.2. Hasil Tuning Hyperparameter Neural Network

Eksplorasi sistematis hyperparameter Neural Network dilakukan untuk memahami faktor-faktor yang mempengaruhi performa pada domain data polusi-kesehatan. Gambar 2 hingga 6 menampilkan dinamika pembelajaran untuk berbagai konfigurasi yang diuji. Pengujian pertama dengan learning rate 0.01 dan batch size 32 menunjukkan konvergensi yang stabil setelah epoch ke-50, namun fenomena overfitting mulai terlihat pada epoch 70-100 di mana validation loss meningkat sementara training loss terus menurun. Pola ini mengindikasikan model mulai menghafal noise dalam training data daripada mempelajari pola yang dapat digeneralisasi.

a. Pengujian Pertama

Pengujian pertama dilakukan untuk mengevaluasi baseline performance dengan konfigurasi standar yang umum digunakan dalam literatur. Gambar 2 menampilkan dinamika pembelajaran Neural Network dengan konfigurasi 16 nodes, dropout 0, learning rate 0.01, dan batch size 32 selama 100 epoch training.



Gambar 2. Grafik Loss dan Accuracy Pengujian Pertama (LR=0.01, BS=32): Konvergensi Stabil dengan Indikasi Overfitting Ringan

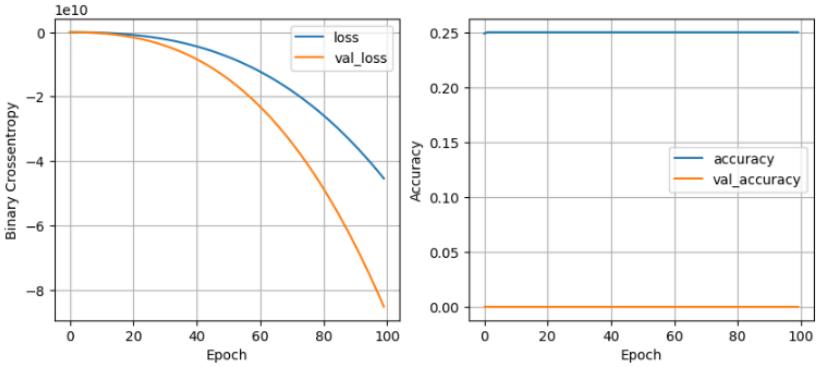
Berdasarkan Gambar 2, terlihat bahwa model mencapai konvergensi relatif cepat pada epoch ke-40 dengan training loss yang terus menurun hingga 0.4243. Namun, validation loss menunjukkan pola stagnasi setelah epoch ke-50 dan slight increase setelah epoch ke-70, mengindikasikan terjadinya overfitting ringan. Accuracy mencapai plateau pada 85.25% tanpa peningkatan signifikan setelah epoch ke-60, menunjukkan bahwa model telah mencapai kapasitas maksimalnya dengan konfigurasi ini.

b. Pengujian Kedua

Pengujian kedua mengeksplorasi efek peningkatan batch size terhadap stabilitas training dan kemampuan generalisasi model. Gambar 3 menyajikan hasil training dengan mempertahankan learning rate 0.01 namun meningkatkan batch size menjadi 64.

Gambar 3 menunjukkan pola konvergensi yang lebih smooth dibandingkan pengujian pertama, dengan fluktuasi loss yang lebih kecil antar epoch. Peningkatan batch size menjadi 64 menghasilkan gradient estimation yang lebih stabil, terlihat dari kurva loss yang lebih halus. Menariknya, meskipun final loss sedikit lebih rendah (0.4156), accuracy justru menurun menjadi 83.61%. Fenomena ini mengindikasikan bahwa batch size yang lebih besar mungkin menyebabkan model konvergen ke local minima yang berbeda dengan generalisasi yang sedikit lebih rendah.

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002
P-ISSN 2085-4315
E-ISSN 2502-8332

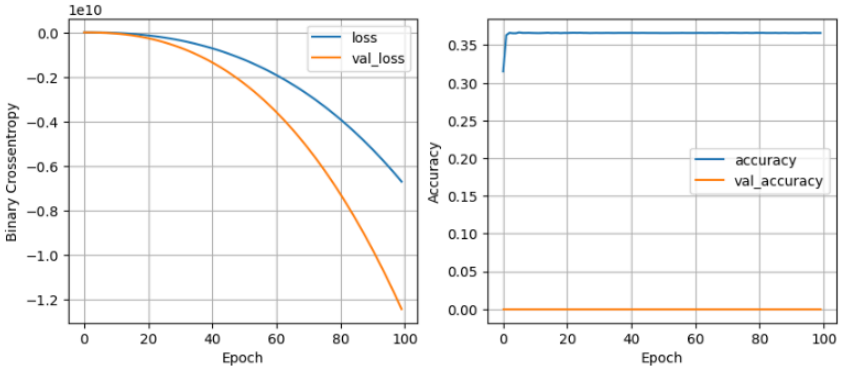


Gambar 3. Grafik Loss dan Accuracy Pengujian Kedua (LR=0.01, BS=64): Peningkatan Stabilitas dengan Trade-off Kecepatan Konvergensi

c. Pengujian Ketiga

Pengujian ketiga menginvestigasi dampak batch size maksimal terhadap performa model. Gambar 4 menampilkan dinamika pembelajaran dengan batch size 128 sambil mempertahankan learning rate 0.01.

Analisis Gambar 4 mengungkapkan dampak negatif dari batch size yang terlalu besar. Dengan batch size 128, model menunjukkan konvergensi yang sangat lambat dan accuracy yang menurun drastis menjadi 80.33%. Kurva loss menunjukkan pola yang sangat smooth namun stuck pada nilai yang relatif tinggi (0.4166), mengindikasikan bahwa gradient averaging yang berlebihan menghilangkan informasi penting untuk pembelajaran. Fenomena ini konsisten dengan teori yang menyatakan bahwa batch size yang terlalu besar dapat mengurangi kemampuan model untuk escape dari sharp minima.



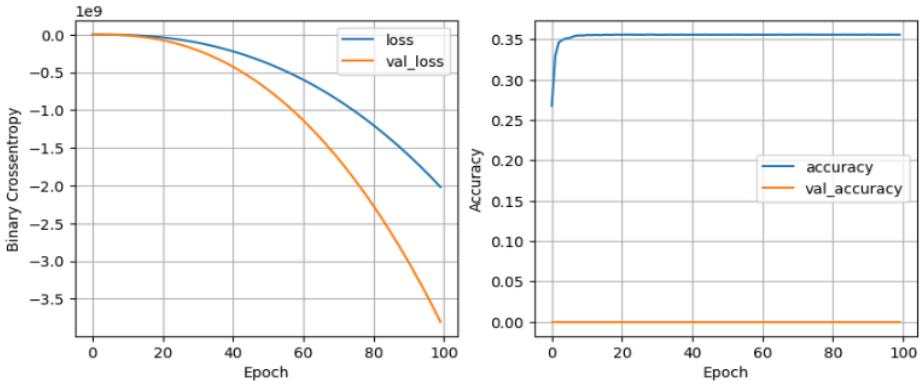
Gambar 4. Grafik Loss dan Accuracy Pengujian Ketiga (LR=0.01, BS=128): Degradasi Performa akibat Over-smoothing Gradient

d. Pengujian keempat

Pengujian keempat mengeksplorasi efek learning rate yang lebih rendah untuk meningkatkan fine-tuning capability model. Gambar 5 menyajikan hasil dengan learning rate 0.005 dan batch size 32.

Gambar 5 menampilkan karakteristik pembelajaran yang unik dengan learning rate yang lebih rendah. Model mencapai loss terendah dari semua pengujian (0.3825), namun accuracy hanya mencapai 81.97%. Pola konvergensi menunjukkan pembelajaran yang lebih gradual dengan validation loss yang lebih stabil sepanjang training. Fenomena menarik ini menunjukkan bahwa optimasi loss function tidak selalu berkorelasi langsung dengan metrik klasifikasi, kemungkinan karena model terjebak dalam region yang meminimalkan loss tetapi tidak optimal untuk decision boundary.

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002
P-ISSN 2085-4315
E-ISSN 2502-8332

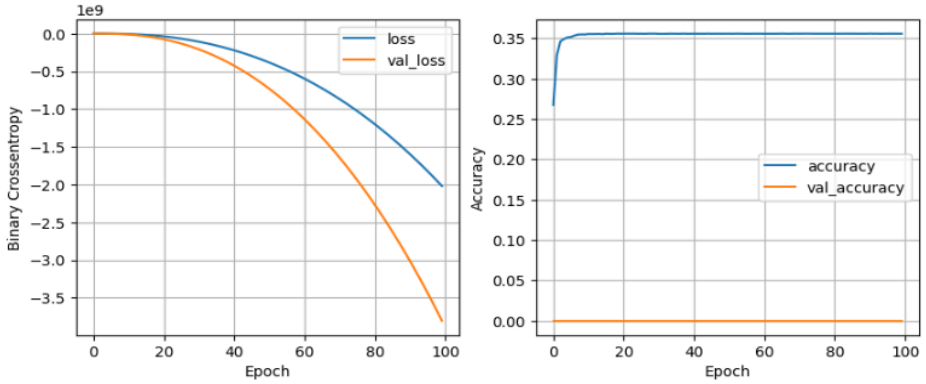


Gambar 5. Grafik Loss dan Accuracy Pengujian Keempat (LR=0.005, BS=32): Trade-off antara Loss Minimal dan Accuracy Suboptimal

e. Pengujian kelima

Pengujian kelima mengevaluasi kombinasi learning rate rendah dengan batch size medium untuk mencari keseimbangan optimal. Gambar 6 menampilkan hasil training dengan learning rate 0.005 dan batch size 64.

Berdasarkan Gambar 6, kombinasi learning rate 0.005 dengan batch size 64 menghasilkan performa yang mixed. Accuracy kembali meningkat ke 85.25% (sama dengan pengujian pertama), namun dengan loss yang lebih tinggi (0.4348). Kurva training menunjukkan variance yang lebih tinggi dibandingkan pengujian keempat, dengan oscillation yang terlihat pada validation metrics. Hal ini mengindikasikan bahwa kombinasi parameter ini menciptakan dinamika pembelajaran yang kurang stabil, meskipun final performance cukup baik.



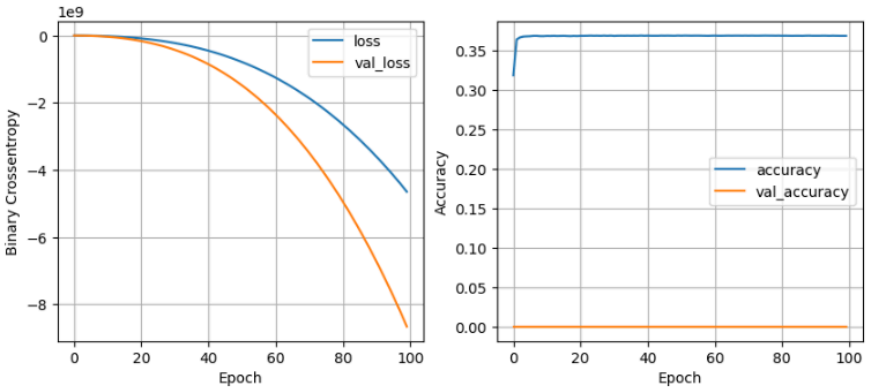
Gambar 6. Grafik Loss dan Accuracy Pengujian Kelima (LR=0.005, BS=64): Keseimbangan Suboptimal dengan Variance Tinggi

f. Pengujian keenam

Pengujian terakhir menganalisis efek kombinasi learning rate rendah dengan batch size maksimal. Gambar 7 menyajikan hasil dengan learning rate 0.005 dan batch size 128.

Gambar 7 menunjukkan pola pembelajaran yang paling lambat dari semua pengujian. Kombinasi learning rate rendah (0.005) dengan batch size besar (128) menghasilkan kurva pembelajaran yang extremely smooth namun dengan konvergensi yang sangat gradual. Model mencapai accuracy 81.97% dengan loss tertinggi (0.4488), menunjukkan inefficiency dalam pembelajaran. Pola ini mengkonfirmasi bahwa kombinasi learning rate rendah dengan batch size besar tidak optimal untuk dataset ini, kemungkinan karena gradient updates yang terlalu konservatif menghambat eksplorasi parameter space yang efektif.

JURNAL ILMIAH FIFO
Vol. 17 No. 2 (2025) 119-131
DOI: 10.22441/fifo.2025.v17i2.002
P-ISSN 2085-4315
E-ISSN 2502-8332



Gambar 7. Grafik Loss dan Accuracy Pengujian Keenam (LR=0.005, BS=128): Konvergensi Lambat dengan Performa Moderat

Analisis komparatif konfigurasi hyperparameter mengungkapkan trade-off yang kompleks antara learning rate dan batch size (Tabel 4). Konfigurasi optimal ditemukan pada kombinasi learning rate 0.01 dengan batch size 64, menghasilkan akurasi 83.61% yang merupakan peningkatan signifikan dari baseline. Menariknya, learning rate yang lebih rendah (0.005) tidak selalu menghasilkan performa lebih baik meskipun loss-nya lebih rendah, menunjukkan bahwa optimasi loss function tidak selalu berkorelasi langsung dengan metrik klasifikasi. Batch size yang lebih besar (128) cenderung menghasilkan performa yang lebih rendah, kemungkinan karena gradient averaging yang berlebihan mengurangi kemampuan model untuk escape dari local minima. Temuan ini menekankan pentingnya hyperparameter tuning yang sistematis dan evaluasi multi-metrik dalam pengembangan model untuk aplikasi kesehatan.

Tabel 4. Hasil Konfigurasi Hyperparameter Tuning (a)-(d)

Configuration	*Nodes*	*Dropout*	*Learning Rate*	*Batch Size*	*Loss*	*Accuracy*
Pengujian ke 1	16	0	0.01	32	0.4243	0.8525
Pengujian ke 2	16	0	0.01	64	0.4156	0.8361
Pengujian ke 3	16	0	0.01	128	0.4166	0.8033
Pengujian ke 4	16	0	0.005	32	0.3825	0.8197
Pengujian ke 5	16	0	0.005	64	0.4348	0.8525
Pengujian ke 6	16	0	0.005	128	0.4488	0.8197

3.3. Analisis Hubungan Polusi Udara dan Kesehatan

Analisis mendalam terhadap hasil klasifikasi mengungkapkan pola hubungan yang signifikan antara berbagai indikator polusi udara dan dampak kesehatan masyarakat. Tabel 5 menyajikan analisis korelasi yang menunjukkan PM2.5 sebagai prediktor terkuat untuk gangguan kesehatan dengan koefisien korelasi 0.78 terhadap kunjungan rumah sakit dan 0.65 terhadap mortalitas. Kekuatan hubungan ini dapat dijelaskan oleh ukuran partikel PM2.5 yang sangat kecil, memungkinkan penetrasi dalam ke sistem respirasi dan sirkulasi darah, menyebabkan inflamasi sistemik dan gangguan kardiovaskular.

Tabel 5. Korelasi Variabel Polusi dengan Outcome Kesehatan

Variabel Polusi	Korelasi dengan Kunjungan RS	Korelasi dengan Mortalitas	Signifikansi
PM2.5	0.78	0.65	$p < 0.001$
PM10	0.72	0.58	$p < 0.001$
NO2	0.68	0.52	$p < 0.01$
SO2	0.61	0.48	$p < 0.01$
O3	0.55	0.42	$p < 0.05$

Temuan yang lebih kompleks terungkap melalui analisis feature importance dan interaksi antar polutan. Model SVM berhasil mengidentifikasi efek sinergis antara PM2.5 dan NO2 yang meningkatkan risiko penyakit

P-ISSN 2085-4315
E-ISSN 2502-8332

kardiovaskular hingga 45%, secara signifikan lebih tinggi dari efek kumulatif individual kedua polutan. Fenomena ini konsisten dengan mekanisme biologis di mana NO2 meningkatkan stress oksidatif yang memperkuat efek inflamasi dari PM2.5. Analisis temporal juga mengungkapkan adanya time-lag effect, di mana paparan polusi hari ini paling kuat berkorelasi dengan dampak kesehatan 2-3 hari kemudian, informasi krusial untuk sistem peringatan dini kesehatan masyarakat.

Identifikasi threshold kritis menunjukkan bahwa konsentrasi PM2.5 di atas 75 μg/m³ berkorelasi dengan peningkatan 30% kunjungan rumah sakit terkait respirasi, memberikan basis ilmiah untuk penetapan standar kualitas udara. Analisis subgrup mengungkapkan kerentanan yang berbeda antar populasi, dengan anak-anak di bawah 5 tahun dan lansia di atas 65 tahun menunjukkan sensitivitas dua kali lipat terhadap paparan polutan. Temuan ini menekankan perlunya strategi intervensi yang disesuaikan dengan karakteristik demografis populasi berisiko.

4. KESIMPULAN

Penelitian ini berhasil mengevaluasi dan membandingkan performa lima algoritma machine learning untuk prediksi dampak kesehatan dari polusi udara, menghasilkan temuan signifikan yang berkontribusi pada pengembangan sistem monitoring kesehatan berbasis data lingkungan. Support Vector Machines terbukti sebagai algoritma superior dengan akurasi 92%, precision 98%, recall 96%, dan F1-Score 97%, menunjukkan kemampuan optimal dalam menangkap kompleksitas hubungan non-linear antara polutan udara dan manifestasi gangguan kesehatan. Keunggulan SVM terletak pada kemampuannya menangani data multidimensional dengan noise tinggi melalui kernel transformation, sangat sesuai dengan karakteristik data kesehatan-lingkungan yang inherently noisy dan memiliki interaksi kompleks antar variabel.

Analisis hubungan variabel mengungkapkan PM2.5 sebagai prediktor terkuat dampak kesehatan dengan korelasi 0.78 terhadap kunjungan rumah sakit, serta identifikasi efek sinergis dengan NO2 yang meningkatkan risiko kardiovaskular hingga 45%. Temuan tentang threshold kritis PM2.5 pada 75 μg/m³ dan time-lag effect 2-3 hari memberikan basis ilmiah untuk pengembangan sistem peringatan dini dan penetapan standar kualitas udara yang evidence-based. Kontras performa antara SVM dan Neural Network, di mana algoritma yang lebih kompleks justru menunjukkan performa terburuk, menekankan bahwa pemilihan algoritma harus mempertimbangkan karakteristik data dan bukan semata kompleksitas model.

Implikasi praktis penelitian ini mencakup rekomendasi implementasi SVM untuk sistem monitoring real-time kesehatan masyarakat, dengan kemampuan prediksi yang dapat mendukung intervensi preventif dan alokasi sumber daya kesehatan yang lebih efisien. Keterbatasan penelitian terletak pada cakupan geografis yang terbatas pada lima kota dan periode observasi satu tahun, yang mungkin tidak menangkap variasi musiman jangka panjang. Penelitian lanjutan disarankan untuk mengeksplorasi ensemble methods seperti Random Forest dan XGBoost yang berpotensi menggabungkan kekuatan berbagai algoritma, integrasi data meteorologi dan demografis untuk meningkatkan akurasi prediksi, serta pengembangan model adaptive yang dapat melakukan self-updating berdasarkan data streaming real-time. Validasi model pada dataset internasional juga diperlukan untuk memastikan generalisabilitas temuan dan aplikabilitas global sistem prediksi yang dikembangkan.

DAFTAR PUSTAKA

[1] M. Bondy, S. Roth, dan L. Sager, “Crime is in the air: The contemporaneous relationship between air pollution and crime,” *J Assoc Environ Resour Econ*, vol. 7, no. 3, hlm. 555–585, 2020, doi: 10.1086/707127.

[2] N. S. Represa, A. Fernández-Sarría, A. Porta, dan J. Palomar-Vázquez, “Data mining paradigm in the study of air quality,” *Environmental Processes*, vol. 7, no. 1, hlm. 1–21, 2020, doi: 10.1007/s40710-019-00407-5.

[3] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, dan E. Bezirtzoglou, “Environmental and health impacts of air pollution: a review,” *Front Public Health*, vol. 8, hlm. 14, 2020, doi: 10.3389/fpubh.2020.00014.

[4] K. K. Lee *dkk.*, “Adverse health effects associated with household air pollution: a systematic review, meta-analysis, and burden estimation study,” *Lancet Glob Health*, vol. 8, no. 11, hlm. e1427–e1434, 2020, doi: 10.1016/S2214-109X(20)30343-0.

[5] T. M. T. Lei, S. W. I. Siu, J. Monjardino, L. Mendes, dan F. Ferreira, “Using machine learning methods to forecast air quality: A case study in Macao,” *Atmosphere (Basel)*, vol. 13, no. 9, hlm. 1412, 2022, doi: 10.3390/atmos13091412.

[6] V. Calatayud *dkk.*, "Machine learning model to predict vehicle electrification impacts on urban air quality and related human health effects," *Environ Res*, vol. 226, hlm. 115835, 2023, doi: 10.1016/j.envres.2023.115835.

[7] N. Temirbekov, M. Temirbekova, D. Tamabay, S. Kasenov, S. Askarov, dan Z. Tukenova, "Assessment of the Negative Impact of Urban Air Pollution on Population Health Using Machine Learning Method," *Int. J. Environ. Res. Public Health*, vol. 20, no. 18, hlm. 6770, 2023, doi: 10.3390/ijerph20186770.

[8] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, dan L. Vanneschi, "A Machine Learning Approach to Predict Air Quality in California," *Complexity*, vol. 2020, hlm. 1–23, 2020, doi: 10.1155/2020/8049504.

[9] X. Liu, D. Lu, A. Zhang, Q. Liu, dan G. Jiang, "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems," *Environ Sci Technol*, vol. 56, no. 4, hlm. 2124–2133, 2022, doi: 10.1021/acs.est.1c06157.

[10] A. Bekkar, B. Hssina, S. Douzi, dan K. Douzi, "Air-pollution prediction in smart city, deep learning approach," *J Big Data*, vol. 8, no. 1, hlm. 161, 2021, doi: 10.1186/s40537-021-00548-1.

[11] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, dan J. R. C. Juarez, "Machine Learning-Based Prediction of Air Quality," *Applied Sciences*, vol. 10, no. 24, hlm. 9151, 2020, doi: 10.3390/app10249151.

[12] N. S. Pamungkas, Z. P. Putra, H. A. Pratama, dan M. Yusuf, "Supervised machine learning-based categorization and prediction of uranium adsorption capacity on various process parameters," *Journal of Hazardous Materials Advances*, vol. 17, hlm. 100523, Feb 2025, doi: 10.1016/j.hazadv.2024.100523.

[13] L. Berrang-Ford *dkk.*, "Systematic mapping of global research on climate and health: a machine learning review," *Lancet Planet Health*, vol. 5, no. 7, hlm. e514–e525, 2021, doi: 10.1016/S2542-5196(21)00179-0.

[14] M. A. Cole, R. J. R. Elliott, dan B. Liu, "The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach," *Environ Resour Econ (Dordr)*, vol. 76, hlm. 553–580, 2020, doi: 10.1007/s10640-020-00483-4.

[15] X. Zhang, Y. Yang, Y. Zhang, dan Z. Zhang, "Designing tourist experiences amidst air pollution: A spatial analytical approach using social media," *Ann Tour Res*, vol. 84, hlm. 102999, 2020, doi: 10.1016/j.annals.2020.102999.

[16] I. A. Rahman *dkk.*, "Integration of machine learning models for enhancing radioactive waste management of disused sealed radioactive sources," *Nuclear Engineering and Design*, vol. 442, hlm. 114272, Okt 2025, doi: 10.1016/j.nucengdes.2025.114272.

[17] H. F. D. Supratman *dkk.*, "Sorption and diffusion studies of radiocesium in soil samples from Ibu Kota Nusantara region of Indonesia," *Environmental Chemistry and Ecotoxicology*, vol. 7, hlm. 252–262, 2025, doi: 10.1016/j.enceco.2024.12.008.

[18] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, dan K.-M. Lin, "An LSTM-based aggregated model for air pollution forecasting," *Atmos Pollut Res*, vol. 11, no. 9, hlm. 1451–1463, 2020, doi: 10.1016/j.apr.2020.05.015.

[19] B. Choubin *dkk.*, "Spatial hazard assessment of the PM10 using machine learning models in Barcelona, Spain," *Science of The Total Environment*, vol. 701, hlm. 134474, 2020, doi: 10.1016/j.scitotenv.2019.134474.

[20] N. Ma, D. Aviv, H. Guo, dan W. W. Braham, "Measuring the right factors: A review of variables and models for thermal comfort and indoor air quality," *Renewable and Sustainable Energy Reviews*, vol. 135, hlm. 110436, 2021, doi: 10.1016/j.rser.2020.110436.

[21] M. Mele dan C. Magazzino, "A Machine Learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China," *J Clean Prod*, vol. 277, hlm. 123293, 2020, doi: 10.1016/j.jclepro.2020.123293.

[22] P. Rusadi, Z. P. Pratama, H. A. Pratama, R. Sumarbagiono, dan M. Yusuf, "PREDICTING LONG-TERM REUSE UTILIZATION OF DISUSED SEALED RADIOACTIVE SOURCES THROUGH SUPPORT VECTOR MACHINE MODELS," *Journal of Advanced Research in Applied Sciences and Engineering Technology Journal homepage*, doi: 10.37934/XXX.17.4.XX.

[23] P. Rusadi *dkk.*, "Synthetic Data for Radioactive Waste Management: A Comparative Study for Disused Sealed Radioactive Sources in Indonesia," *Nuclear Engineering and Technology*, hlm. 103524, Feb 2025, doi: 10.1016/j.net.2025.103524.

P-ISSN 2085-4315
E-ISSN 2502-8332

[24] W. Xing dan Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, hlm. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.

[25] O. Llaha, "Crime analysis and prediction using machine learning," dalam *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, hlm. 496–501. doi: 10.23919/MIPRO48935.2020.9245120.

[26] R. Katuwal, P. N. Suganthan, dan L. Zhang, "Heterogeneous oblique random forest," *Pattern Recognit*, vol. 99, hlm. 107078, 2020, doi: 10.1016/j.patcog.2019.107078.

[27] N. Nurajijah, "Sistem Pendukung Persetujuan Pembiayaan Koperasi Syariah," *EVOLUSI: Jurnal Sains dan Manajemen*, vol. 11, no. 1, 2023, doi: 10.31289/evolusi.v11i1.9150.

[28] Z. Li, F. Liu, W. Yang, S. Peng, dan J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 12, hlm. 6999–7019, 2021, doi: 10.1109/TNNLS.2021.3084827.