VOL. 3. NO. 03 MEI 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

PREDIKSI KUALITAS AIR KOLAM MENGGUNAKAN ALGORITMA MACHINE LEARNING BERDASARKAN DATA IOT

Antoni Nur Yahya1, Martanto[2] , Denni Pratama[3], Umi Hayati[4], Saeful Anwar[5]

Program Studi Teknik Informatika[145]
Program Studi Manajemen Informatika[2]
Program Studi Komputerisasi Akuntansi[3]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
antoninuryahya18@gmail.com

(*) Corresponding Author : antoninuryahya18@gmail.com
Published : 30 Mei 2026

Abstract—*Pond water quality is a critical factor in the success of aquaculture, as it directly affects fish health, productivity, and ecosystem sustainability. Conventional water quality monitoring still faces limitations in measurement coverage, data continuity, and timely decision-making. This study aims to design and implement a pond water quality prediction system based on the Internet of Things (IoT) and machine learning algorithms to support early detection of water quality degradation. Water quality data were continuously collected using IoT sensors that measured key physicochemical parameters, including temperature, pH, turbidity, and dissolved oxygen. The collected data were then subjected to preprocessing stages, such as data cleaning, handling missing values, and feature engineering, before being used in supervised machine learning modeling. Several machine learning algorithms were implemented and evaluated to obtain the best-performing prediction model based on relevant accuracy metrics. The results indicate that the application of machine learning to IoT-based data is capable of producing accurate and adaptive predictions of pond water quality in response to changing environmental conditions. The developed system has the potential to provide early warnings of water quality deterioration, thereby enabling pond managers to take proactive corrective actions. This research contributes to the development of data-driven, efficient, adaptive, and sustainable pond water quality monitoring and prediction systems, and supports the implementation of intelligent aquaculture in the future.*

Keywords: *pond water quality; Internet of Things; machine learning; prediction; aquaculture*

Abstrak— *Kualitas air kolam merupakan faktor krusial dalam keberhasilan akuakultur karena berpengaruh langsung terhadap kesehatan ikan, produktivitas, dan keberlanjutan ekosistem. Pemantauan kualitas air secara konvensional masih menghadapi keterbatasan dalam cakupan pengukuran, kontinuitas data, serta keterlambatan dalam pengambilan keputusan. Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem prediksi kualitas air kolam berbasis Internet of Things (IoT) dan algoritma machine learning guna mendukung deteksi dini degradasi kualitas air. Data kualitas air dikumpulkan secara kontinu menggunakan sensor IoT yang mengukur parameter fisika-kimia utama, meliputi suhu, pH, kekeruhan, dan oksigen terlarut. Data yang diperoleh kemudian melalui tahapan pra-pemrosesan, seperti pembersihan data, penanganan nilai hilang, dan rekayasa fitur, sebelum digunakan dalam pemodelan machine learning dengan pendekatan supervised learning. Beberapa algoritma machine learning diterapkan dan dievaluasi untuk memperoleh model dengan kinerja prediksi terbaik berdasarkan metrik akurasi yang relevan. Hasil penelitian menunjukkan bahwa penerapan machine learning pada data IoT mampu menghasilkan prediksi kualitas air yang akurat dan adaptif terhadap perubahan kondisi lingkungan kolam. Sistem yang dikembangkan berpotensi memberikan peringatan dini terhadap penurunan kualitas air, sehingga memungkinkan pengelola kolam melakukan tindakan korektif secara proaktif. Penelitian ini berkontribusi dalam pengembangan sistem pemantauan dan prediksi kualitas air kolam berbasis data yang efisien, adaptif, dan berkelanjutan, serta mendukung implementasi akuakultur cerdas di masa mendatang.*
Kata kunci: kualitas air kolam; Internet of Things; machine learning; prediksi; akuakultur

Kata Kunci : *Analisis Sentimen; Naïve Bayes; RupaRupa; Google Play Store; Text Mining*

INTRODUCTION

Kualitas air kolam merupakan faktor penentu utama dalam produktivitas akuakultur, keseimbangan ekologi, dan keberlanjutan sumber daya perairan tawar. Dalam beberapa tahun terakhir, integrasi teknologi Internet of Things (IoT) dengan model machine learning (ML) berkembang pesat sebagai pendekatan yang menjanjikan untuk pemantauan dan prediksi parameter kualitas air secara berkelanjutan. Meskipun demikian, praktik pemantauan dan pemeliharaan kualitas air kolam hingga saat ini masih menghadapi berbagai tantangan. Tantangan tersebut meliputi cakupan sensor yang jarang dan heterogen, keterbatasan ketersediaan sensor berbiaya rendah yang andal untuk beberapa parameter biologis, serta permasalahan kualitas data dan gangguan transmisi yang umum terjadi pada implementasi IoT. Selain itu, tingginya variabilitas temporal dan spasial pada lingkungan kolam menyulitkan integrasi data sensor ke dalam model prediksi, sementara ketidakpastian dan rendahnya interpretabilitas hasil prediksi berbasis ML mengurangi efektivitasnya sebagai alat pendukung pengambilan keputusan(Abbas et al., 2024)(Singh & Walingo, 2024)(Yan et al., 2024)(Hussein et al., 2023)(Islam, 2023)

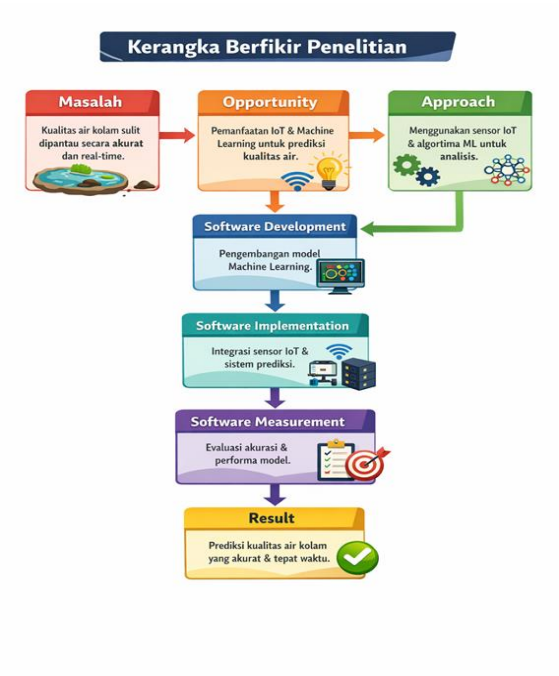
Berbagai artikel tinjauan dan studi empiris terkini secara konsisten melaporkan adanya hambatan teknis yang masih bertahan dalam pengembangan sistem prediksi kualitas air kolam pada skala operasional. Permasalahan utama mencakup pemilihan dan kalibrasi sensor yang tepat, penerapan teknik imputasi data hilang yang efektif, serta rekayasa fitur yang sesuai untuk dataset kolam yang berukuran kecil atau tidak seimbang. Di samping itu, keterbatasan analitik waktu nyata (real-time) pada sisi edge seperti kapasitas komputasi yang terbatas, konsumsi energi, dan latensi—membatasi penerapan model ML yang kompleks dalam sistem IoT. Isu penting lainnya adalah minimnya kuantifikasi ketidakpastian pada prediksi ML, yang berdampak pada keandalan dan tingkat kepercayaan pengguna terhadap keluaran model dalam pengelolaan kolam (Abbas et al., 2024; Yan et al., 2024). Penelitian yang berfokus pada IoT juga menekankan adanya keterbatasan konektivitas, tantangan pasokan daya, serta beban pemeliharaan perangkat, khususnya pada kolam yang berada di wilayah pedesaan. Sementara itu, publikasi berbasis data dan studi kasus akuakultur menyoroti kebutuhan mendesak akan ketersediaan dataset kualitas air kolam yang terbuka, berlabel, dan berkualitas tinggi untuk mendukung proses pelatihan, validasi, serta pembandingan pendekatan ML secara objektif (Islam, 2023; Singh & Walingo, 2024; Hussein et al., 2023).

Terlepas dari berbagai keterbatasan tersebut, prediksi dini terhadap degradasi kualitas air kolam diakui sebagai kebutuhan yang sangat penting dalam pengelolaan kolam yang efektif dan berkelanjutan. Deteksi awal terhadap tren negative seperti eutrofikasi, hipoksia, atau fluktuasi mendadak parameter fisika-kimia memungkinkan dilakukannya intervensi sebelum kondisi air melampaui ambang batas kritis. Kemampuan prediktif ini berperan dalam mencegah kematian ikan, menekan kerugian ekonomi, serta menjaga fungsi layanan ekosistem dalam sistem akuakultur (Hussein et al., 2023)(Islam, 2023) Prakiraan cepat yang dihasilkan oleh model ML berbasis data IoT memungkinkan pengelola kolam untuk menjadwalkan aerasi, mengoptimalkan strategi pemberian pakan, dan menerapkan tindakan korektif secara proaktif, sehingga efisiensi produksi meningkat dan penggunaan bahan kimia yang tidak perlu dapat dikurangi (Jayaraman et al., 2024)

Lebih lanjut, sistem peringatan dini yang didukung oleh penginderaan IoT secara kontinu dan prediksi berbasis ML membantu alokasi sumber daya pemeliharaan yang terbatas secara lebih efisien, meningkatkan biosekuriti dengan menurunkan risiko wabah penyakit, serta menyediakan aliran data berfrekuensi tinggi yang diperlukan untuk melatih model ML yang adaptif dan tangguh. Dengan mengubah data sensor mentah menjadi informasi prediktif berupa waktu antisipasi (lead-time) yang dapat ditindaklanjuti, sistem prediksi mendorong pergeseran praktik pengelolaan kolam dari respons reaktif terhadap krisis menuju pendekatan pencegahan yang bersifat proaktif dan berbasis data. Pergeseran paradigma ini menjadi semakin penting bagi sistem akuakultur modern yang berupaya menyeimbangkan produktivitas, perlindungan lingkungan, dan keberlanjutan jangka panjang (Hussein et al., 2023)(Islam, 2023)(Jayaraman et al., 2024)(Z. Deng, 2024)(Jaywant, 2024)

MATERIALS AND METHODS

Tahapan alur penelitian ini menjelaskan langkah-langkah sistematis yang dilakukan dalam proses penelitian, mulai dari pengumpulan data hingga evaluasi hasil model pada Gambar 1 Tahapan Alur Penelitian.



Gambar 1 Kerangka Berfikir

Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan memanfaatkan sistem Internet of Things (IoT) yang dirancang untuk memperoleh data kualitas air kolam secara kontinu dan waktu nyata. Data utama diperoleh dari sensor in-situ yang terdistribusi pada kolam penelitian, terdiri atas sensor suhu, pH, dissolved oxygen (DO), kekeruhan, dan konduktivitas atau total dissolved solids (TDS). Sensor-sensor tersebut terhubung dengan mikrokontroler dan node edge yang berfungsi sebagai titik akuisisi awal data. Pada tahap ini, data dikumpulkan dalam frekuensi tinggi (high-frequency sampling) guna menangkap dinamika temporal dan kejadian transien yang umum terjadi pada ekosistem kolam.

Data sensor yang diperoleh selanjutnya diproses pada lapisan edge melalui mekanisme pra-pemrosesan awal, meliputi penyaringan noise, pemeriksaan rentang nilai, deteksi anomali awal, serta pencatatan metadata penting seperti waktu pengukuran, identitas sensor, dan status perangkat. Data yang telah melalui pra-pemrosesan kemudian ditransmisikan ke sistem penyimpanan terpusat menggunakan protokol komunikasi berdaya rendah dan andal, seperti LoRa, NB-IoT, atau Wi-Fi, sesuai dengan kondisi infrastruktur lokasi penelitian. Pendekatan edge–cloud hybrid ini memungkinkan pengurangan latensi dan beban transmisi sekaligus menjaga kontinuitas aliran data.

Pada sisi cloud, data IoT diintegrasikan ke dalam pipeline pengelolaan data yang menerapkan prosedur quality assurance dan quality control (QA/QC) secara sistematis. Tahapan ini mencakup koreksi drift sensor, imputasi nilai hilang secara streaming, validasi konsistensi antar-parameter, serta kurasi data untuk memastikan data siap digunakan dalam pemodelan machine learning. Seluruh proses pengumpulan dan pengolahan data dilengkapi dengan pencatatan provenance dan metadata, sehingga mendukung reprodusibilitas penelitian, audit data, serta pelacakan sumber kesalahan apabila terjadi anomali pada hasil analisis.

Selain data sensor IoT utama, penelitian ini juga memanfaatkan data pendukung (complementary data) untuk memperkaya konteks dan meningkatkan kualitas label dalam pemodelan supervised learning. Data pendukung tersebut meliputi catatan pengelolaan kolam (misalnya jadwal pemberian pakan dan kepadatan tebar), data meteorologi lokal, serta hasil pengukuran laboratorium berkala seperti BOD/COD atau parameter biologis tertentu apabila tersedia. Integrasi data multisumber ini bertujuan untuk meningkatkan kemampuan model dalam merepresentasikan kondisi ekosistem kolam secara lebih komprehensif.

Penggunaan data IoT waktu nyata sebagai sumber utama pengumpulan data memberikan sejumlah keunggulan dibandingkan pemanfaatan dataset sekunder atau data historis yang teragregasi. Data waktu nyata memungkinkan pemantauan berkelanjutan dengan resolusi temporal tinggi, mendukung deteksi dini anomali dan kejadian kritis, serta menyediakan aliran data kontekstual dan berlabel yang lebih kaya untuk rekayasa fitur dan pembelajaran model secara daring (online learning). Selain itu, integrasi data IoT secara langsung memfasilitasi analitik edge dan cloud, memungkinkan pengambilan keputusan dengan latensi rendah, serta mendukung pelatihan ulang model secara berkelanjutan sesuai dengan perubahan kondisi lingkungan kolam.

Dengan metode dan teknik pengumpulan data tersebut, penelitian ini diharapkan mampu menghasilkan dataset kualitas air kolam yang andal, terdokumentasi dengan baik, dan relevan secara operasional. Dataset ini menjadi fondasi utama dalam pengembangan, evaluasi, dan validasi sistem prediksi kualitas air kolam berbasis IoT dan machine learning yang adaptif serta berkelanjutan.

RESULTS AND DISCUSSION

Bab ini menyajikan hasil penelitian yang diperoleh dari seluruh tahapan eksperimen yang telah dilakukan, mulai dari pemuatan dan praproses data, perancangan arsitektur model, proses pelatihan model Random Forest, hingga evaluasi

kinerja dan analisis feature importance. Penyajian hasil difokuskan pada keluaran objektif dari setiap tahap tanpa disertai interpretasi mendalam, yang selanjutnya akan dibahas pada subbab pembahasan.Untuk lebih Detail dapat dilihat pada gambar 4.1 berikut ini :

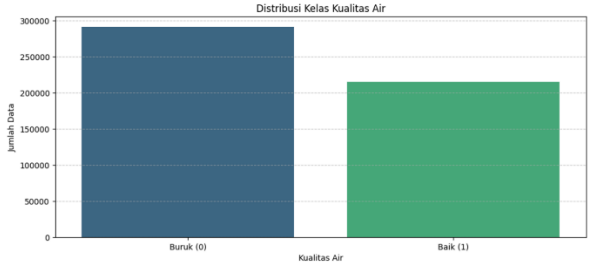
[Tabel 1: Lima Baris Data Awal]

id	created_date	water_pH	TDS	water_temp
181775	1/26/2023 11:24	7.6	250	24
181777	1/26/2023 11:33	7.6	247	24.06
181778	1/26/2023 12:02	7.6	249	24.19
181780	1/26/2023 12:05	6.92	247	24.19
181782	1/26/2023 12:10	7.3	245	24.25

```
[Ringkasan Informasi Data]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 624016 entries, 0 to 624015
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   id            624016 non-null  int64
 1   created_date  624016 non-null  object
 2   water_pH      624016 non-null  float64
 3   TDS           624016 non-null  int64
 4   water_temp    624016 non-null  float64
dtypes: float64(2), int64(2), object(1)
memory usage: 23.8+ MB
```

Gambar 1 Hasil Pembacaan Datasete & Processing

Pada tahap awal, ditampilkan lima baris data pertama yang menunjukkan bahwa dataset terdiri dari atribut id, created_date, water_pH, TDS, dan water_temp. Informasi ringkasan data menunjukkan bahwa dataset memiliki total 624.016 entri dengan lima variabel dan tidak terdapat nilai hilang (missing values), sehingga kualitas data awal tergolong baik. Selain itu, teridentifikasi sebanyak 118.286 data duplikat yang kemudian dihapus untuk menjaga validitas analisis. Selanjutnya, dibentuk variabel target berupa kelas kualitas air berdasarkan standar kualitas air yang mengacu pada nilai pH air dan TDS, sehingga data berhasil diklasifikasikan ke dalam dua kelas, yaitu kualitas air buruk (0) dan baik (1).



Gambar 2 Distribusi Kelas Kualitas Air

Distribusi kelas kualitas air divisualisasikan dalam bentuk diagram batang yang menunjukkan bahwa jumlah data dengan kualitas air buruk lebih banyak dibandingkan dengan kualitas air baik. Hal ini mengindikasikan adanya ketidakseimbangan kelas (class imbalance) yang perlu diperhatikan pada tahap pemodelan selanjutnya. Setelah proses imputasi dan standardisasi, dilakukan analisis statistik deskriptif terhadap variabel numerik utama. Hasil statistik menunjukkan bahwa suhu air memiliki rata-rata sekitar 24,19°C dengan variasi yang relatif kecil, sedangkan nilai TDS memiliki rata-rata sekitar 388,98 dengan rentang yang cukup lebar, mencerminkan variasi tingkat zat terlarut dalam air. Nilai pH air memiliki rata-rata mendekati netral, yaitu sekitar 7,90, namun dengan rentang yang cukup besar hingga nilai ekstrem. Secara keseluruhan, gambar ini menunjukkan bahwa dataset telah melalui proses pembersihan dan transformasi yang memadai serta siap digunakan untuk tahap analisis lanjutan atau pemodelan machine learning dalam penentuan kualitas air.

Tabel 1 Statistik Imputasi & Standarisasi

	count	mean	std	min
Water_temp	505730	24.1872	0.973159	21.63
TDS	505730	388.979	137.266	200
Water_ pH	505730	7.90038	1.38645	5.51

Tahap awal penelitian menghasilkan data yang telah berhasil dimuat dari sumber yang telah ditentukan dan dikonversi ke dalam format numerik yang sesuai untuk proses pemodelan. Seluruh atribut yang bersifat kategorikal telah ditransformasikan menjadi representasi numerik, sedangkan data numerik telah melalui proses normalisasi dan pembersihan. Pada tahap ini juga dilakukan penanganan terhadap data hilang (missing values) sehingga dataset akhir berada dalam kondisi siap digunakan untuk proses pembelajaran mesin. Hasil praproses menunjukkan bahwa data telah terstruktur dengan baik dalam bentuk tabel numerik dan tidak ditemukan anomali signifikan yang dapat mengganggu proses pelatihan model

```
Ukuran Data Latih (Train): 303438 (60.0%)
Ukuran Data Validasi (Validation): 101146 (20.0%)
Ukuran Data Uji (Test): 101146 (20.0%)
Data latih mengandung 2 kelas unik. Pelatihan dilanjutkan.

[Arsitektur Model]
Model 1: Random Forest Classifier (Ensemble Learning)
Model 2: Multi-Layer Perceptron (Neural Network Sederhana)
```

Gambar 3 Hasil Pemuatan dan Prepocessing Data

Dataset yang telah melalui tahap prapemrosesan selanjutnya dibagi menjadi tiga bagian, yaitu data latih (training) sebesar 60% dengan jumlah 303.438 data, data validasi (validation) sebesar 20% sebanyak 101.146 data, dan data uji (testing) sebesar 20% dengan jumlah yang sama, yaitu 101.146 data. Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih secara optimal, divalidasi untuk pemilihan parameter terbaik, serta diuji secara objektif terhadap data yang belum pernah dilihat sebelumnya. Informasi pada gambar juga menunjukkan bahwa data latih mengandung dua kelas unik, sehingga proses

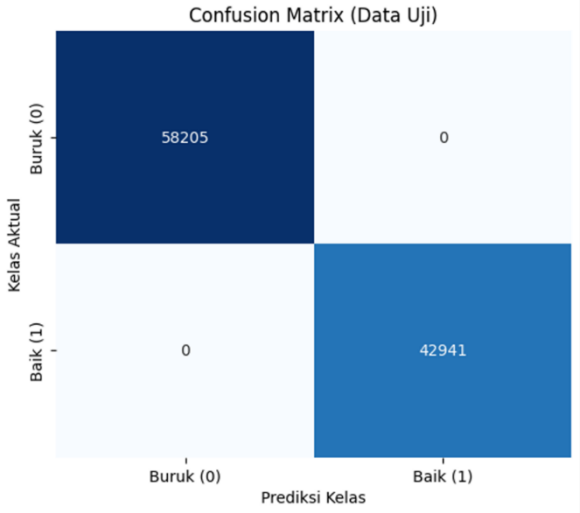
pelatihan dapat dilanjutkan tanpa kendala terkait kelengkapan kelas target.

Selain itu, gambar tersebut menampilkan arsitektur model yang digunakan dalam penelitian, yaitu dua pendekatan pembelajaran mesin yang berbeda. Model pertama adalah Random Forest Classifier, yang merepresentasikan metode ensemble learning dengan menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan stabilitas prediksi. Model kedua adalah Multi-Layer Perceptron (MLP), yaitu jaringan saraf tiruan sederhana yang mampu mempelajari pola nonlinier dalam data. Pemilihan dua model dengan karakteristik yang berbeda ini bertujuan untuk membandingkan kinerja metode berbasis ensemble dan jaringan saraf dalam mengklasifikasikan kualitas air. Secara keseluruhan, gambar ini menggambarkan kesiapan data dan strategi pemodelan yang sistematis sebelum memasuki tahap pelatihan, evaluasi, dan analisis hasil.

4.1.2 Hasil Perancangan Arsitektur Model dan Pembagian Data

Pada tahap ini, arsitektur model Random Forest berhasil dirancang sesuai dengan parameter penelitian yang telah ditetapkan. Dataset yang telah dipraproses kemudian dibagi ke dalam data latih (training data) dan data uji (testing data) dengan proporsi yang telah ditentukan. Pembagian data ini bertujuan untuk memastikan bahwa proses evaluasi model dilakukan secara objektif terhadap data yang tidak dilibatkan dalam proses pelatihan. Hasil pembagian menunjukkan bahwa distribusi kelas pada data latih dan data

Tabel 2 Evaluasi Confusion Matrix



Gambar 4 Hasil Data Uji Klasifikasi air

Datatersebut menunjukkan ringkasan performa utama model Naïve Bayes, di mana *accuracy* mencapai 88.34%, menandakan bahwa sebagian besar prediksi model sesuai dengan label aktual. *Precision* sebesar 90.09% menunjukkan bahwa model sangat tepat dalam mengidentifikasi ulasan yang diprediksi sebagai positif, dengan tingkat kesalahan yang rendah. Sementara itu, *recall* sebesar 88.34% mengindikasikan bahwa model mampu menemukan sebagian besar data yang benar-benar relevan pada masing-masing kelas. Nilai *F1-score* sebesar 88.27% mencerminkan keseimbangan antara *precision* dan *recall*, sehingga menggambarkan performa model yang stabil dan konsisten. Secara keseluruhan, keempat metrik ini menegaskan bahwa model *Naïve Bayes* memiliki kemampuan klasifikasi yang kuat terhadap data ulasan yang telah dipraproses dan diseimbangkan.

CONCLUSION

Berasarkan hasil dan pembahasan yang telah dijabarkan, maka dapat ditarik kesimpulan terkait Penelitian ini membuktikan bahwa data kualitas air kolam yang dikumpulkan melalui sistem IoT, setelah melalui tahapan prapemrosesan yang meliputi pembersihan data, normalisasi, dan transformasi ke bentuk numerik, dapat digunakan secara efektif sebagai input bagi model machine learning.

Model Random Forest yang dikembangkan mampu memberikan performa prediksi yang sangat tinggi serta menunjukkan konsistensi kinerja antara data latih dan data uji. Hal ini menandakan bahwa model memiliki kemampuan generalisasi yang sangat baik dan tidak mengalami overfitting maupun underfitting. Analisis learning curve pada model Multi-Layer Perceptron (MLP) menunjukkan bahwa proses pembelajaran berlangsung secara stabil seiring dengan bertambahnya jumlah data latih. Konvergensi antara skor pelatihan dan validasi mengindikasikan bahwa pendekatan pembelajaran yang digunakan bersifat robust dan andal.

Hasil analisis feature importance menunjukkan bahwa parameter pH air merupakan faktor yang paling dominan dalam menentukan kualitas air kolam, diikuti oleh Total Dissolved Solids (TDS), sedangkan suhu air memiliki pengaruh yang relatif lebih kecil. Temuan ini sejalan dengan prinsip ilmiah dalam pengelolaan kualitas air dan memperkuat validitas model yang dibangun. Secara keseluruhan, integrasi IoT dan machine learning dalam penelitian ini berhasil menghasilkan sistem prediksi kualitas air kolam yang akurat, stabil, dan dapat diinterpretasikan, sehingga berpotensi besar untuk dimanfaatkan sebagai sistem pendukung pengambilan keputusan dalam pengelolaan kolam secara efektif dan berkelanjutan.

REFERENCE

Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Arifullah, Alrefaei, A. F., & Albeshr, M. F. (2024). Machine learning models for water quality prediction: A comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan. Water, 16(7), 941. https://doi.org/10.3390/w16070941

Agbo, B., Al-Aqrabi, H., Hill, R., & Alsboui, T. (2022). Missing data imputation in the Internet of Things sensor networks. Future Internet, 14(5), 143. https://doi.org/10.3390/fi14050143

Alam, S., Yakopcic, C., Wu, Q., Barnell, M., Khan, S., & Taha, T. M. (2024). Survey of deep learning accelerators for edge and emerging computing. Electronics, 13(15), 2988. https://doi.org/10.3390/electronics13152988

Bhatt, N., Bhatt, N., Prajapati, P., Sorathiya, V., Alshathri, S., & El-Shafai, W. (2024). A data-centric approach to improve performance of deep learning models. Scientific Reports, 14, 22329. https://doi.org/10.1038/s41598-024-73643-x

Bzai, J., Alam, F., Dhafer, A., Bojović, M., Altowaijri, S. M., Niazi, I. K., & Mehmood, R. (2022). Machine learning-enabled Internet of Things (IoT): Data, applications, and industry perspective. Electronics, 11(17), 2676. https://doi.org/10.3390/electronics11172676

Chui, K. T., Gupta, B. B., Liu, J., Arya, V., Nedjah, N., Almomani, A., & Chaurasia, P. (2023). A survey of Internet of Things and cyber-physical systems: Standards, algorithms, applications, security, challenges, and future directions. Information, 14(7), 388. https://doi.org/10.3390/info14070388

de Haro-Olmo, F. J., Valencia-Parra, A., Varela-Vaca, Á. J., Álvarez-Bermejo, J. A., & Gómez-López, M. T. (2023). ELI: An IoT-aware big data pipeline with data curation and data quality. PeerJ Computer Science, 9, e1605. https://doi.org/10.7717/peerj-cs.1605

De Moor, B. J., Gijsbrechts, J., & Boute, R. N. (2022). Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. European Journal of Operational Research, 301(2), 535–545. https://doi.org/10.1016/j.ejor.2021.10.045

Decorte, T., Mortier, S., Lembrechts, J. J., Meysman, F. J. R., Latré, S., Mannens, E., & Verdonck, T. (2024). Missing value imputation of wireless sensor data for environmental monitoring. Sensors, 24(8), 2416. https://doi.org/10.3390/s24082416

Deng, Y., Zhang, Y., Pan, D., Yang, S. X., & Gharabaghi, B. (2024). Review of recent advances in remote sensing and machine learning methods for lake water quality management. Remote Sensing, 16(22), 4196. https://doi.org/10.3390/rs16224196

Deng, Z. (2024). Reward shaping via expectation maximization method. Neurocomputing, 609, 128471. https://doi.org/10.1016/j.neucom.2024.128471

El-Shafeiy, E., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-time anomaly detection for water quality sensor monitoring based on multivariate deep-learning technique. Sensors, 23(20), 8613. https://doi.org/10.3390/s23208613

Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Advances in machine learning and IoT for water quality monitoring: A comprehensive review. Heliyon, 10, e27920. https://doi.org/10.1016/j.heliyon.2024.e27920

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. Data Mining and Knowledge Discovery, 37, 788–832. https://doi.org/10.1007/s10618-022-00894-5

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 15, 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hussein, E. E., Jat Baloch, M. Y., Nigar, A., Abualkhair, H. F., Aldawood, F. K., & Tageldin, E. (2023). Machine learning algorithms for predicting the water quality index. Water, 15(20), 3540. https://doi.org/10.3390/w15203540

Islam, M. M. (2023). Real-time dataset of pond water for fish farming using IoT devices. Data in Brief, 51, 109761. https://doi.org/10.1016/j.dib.2023.109761

Jayaraman, P., Nagarajan, K. K., Partheeban, P., & Krishnamurthy, V. (2024). Critical

review on water quality analysis using IoT and machine learning models. International Journal of Information Management Data Insights, 4(1), 100210. https://doi.org/10.1016/j.jjimei.2023.100210

Jaywant, S. A. (2024). Remote sensing techniques for water quality monitoring: A review. Sensors, 24(24), 8041. https://doi.org/10.3390/s24248041

Jierula, A., Wang, S., Oh, T.-M., & Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. Applied Sciences, 11(5), 2314. https://doi.org/10.3390/app11052314

Jouini, O., Sethom, K., Namoun, A., Aljohani, N., Alanazi, M. H., & Alanazi, M. N. (2024). A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions. Technologies, 12(6), 81. https://doi.org/10.3390/technologies12060081

Kaliappan, J. (2021). Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. Frontiers in Public Health, 9, 729795. https://doi.org/10.3389/fpubh.2021.729795

Kolltveit, A. B., & Li, J. (2022). Operationalizing machine learning models: A systematic literature review BT - Proceedings of the 1st Workshop on Software Engineering for Responsible AI (SE4RAI '22). 1–8. https://doi.org/10.1145/3526073.3527584

Malerba, D. (2024). Data-centric AI. Journal of Intelligent Information Systems. https://doi.org/10.1007/s10844-024-00901-9

Miller, C., Portlock, T., Nyaga, D. M., & O'Sullivan, J. M. (2024). A review of model evaluation metrics for machine learning in genetics and genomics. Frontiers in Bioinformatics, 4, 1457619. https://doi.org/10.3389/fbinf.2024.1457619

Miller, M., Kisiel, A., Cembrowska-Lech, D., Durlik, I., & Miller, T. (2023). IoT in water quality monitoring—Are we really here? Sensors, 23(2), 960. https://doi.org/10.3390/s23020960

Monios, N., Peladarinos, N., Cheimaras, V., Papageorgas, P., & Piromalis, D. D. (2024). A thorough review and comparison of commercial and open-source IoT platforms for smart city applications. Electronics, 13(8), 1465. https://doi.org/10.3390/electronics13081465

Popescu, S. M., Mansoor, S., Wani, O. A., Kumar, S. S., Sharma, V., Sharma, A., Arya, V. M., Kirkham, M. B., Hou, D., Bolan, N., & Chung, Y. S. (2024). Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management. Frontiers in Environmental Science, 12, 1336088. https://doi.org/10.3389/fenvs.2024.1336088

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics, 2, 927312. https://doi.org/10.3389/fbinf.2022.927312

Rosero-Montalvo, P. D., López-Batista, V. F., & Peluffo-Ordóñez, D. H. (2022). A new data-preprocessing-related taxonomy of sensors for IoT applications. Information, 13(5), 241. https://doi.org/10.3390/info13050241

Sahoo, G. R., Freed, J. H., & Srivastava, M. (2024). Optimal wavelet selection for signal denoising. IEEE Access, 12, 45369–45380. https://doi.org/10.1109/ACCESS.2024.3377664

Schackart, K. E., Imker, H. J., & Cook, C. E. (2024). Detailed implementation of a reproducible machine-learning-enabled workflow. Data Science Journal, 23(1). https://doi.org/10.5334/dsj-2024-023

Singh, Y., & Walingo, T. (2024). Smart water quality monitoring with IoT wireless sensor networks. Sensors, 24(9), 2871. https://doi.org/10.3390/s24092871

Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., & Engel, T. (2024). Time-series data preprocessing: A survey and an empirical analysis. Journal of Engineering Research, 13(2). https://doi.org/10.1016/j.jer.2024.02.018

Wiryasaputra, R., Huang, C.-Y., Lin, Y.-J., & Yang, C.-T. (2024). An IoT real-time potable water quality monitoring and prediction model based on cloud computing architecture. Sensors, 24(4), 1180.

https://doi.org/10.3390/s24041180
Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. (2024). A comprehensive review of machine learning for water quality prediction over the past five years. Journal of Marine Science and Engineering, 12(1), 159. https://doi.org/10.3390/jmse12010159
Yu, R., Wan, S., Wang, Y., Gao, C.-X., Gan, L., Zhang, Z., & Zhan, D.-C. (2025). Reward models in deep reinforcement learning: A survey. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). https://doi.org/10.24963/ijcai.2025/1199