HOAQ: JURNAL TEKNOLOGI INFORMASI, Volume 14 Nomor 2, Desember 2023, 59-64
p-ISSN: 2337-5280, e-ISSN: 26207427
DOI : https://doi.org/10.52972/hoaq.vol14no1.p59-64

Perbandingan Kinerja Model Pembelajaran Mesin dalam Prediksi Banjir menggunakan KNN, Naive Bayes, dan Random Forest

Sadam Muhammad Natzir

Program Studi Magister Informatika, Universitas AMIKOM Yogyakarta
Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta, Indonesia
Email: sadamnatzir@students.amikom.ac.id

ABSTRAK

Penelitian ini menyajikan analisis komparatif model pembelajaran mesin untuk prediksi banjir menggunakan data historis curah hujan. Tiga model, yaitu K-Nearest Neighbors (KNN), Naive Bayes, dan Random Forest, dievaluasi berdasarkan metrik kinerja mereka, yang meliputi akurasi, presisi, recall, skor F1, dan ROC AUC. Hasil evaluasi menunjukkan bahwa model Random Forest secara konsisten mengungguli KNN dan Naive Bayes. Random Forest mencapai skor sempurna (100%) pada semua indikator yang diukur. Sementara itu, KNN dan Naive Bayes juga menunjukkan kinerja yang kompetitif, meskipun terdapat beberapa trade-off antara presisi dan recall. Secara rinci, untuk akurasi, presisi, recall, skor F1, dan ROC AUC, model Random Forest memperoleh skor 100%, sedangkan KNN dan Naive Bayes berada di kisaran 90-95%. Meskipun demikian, KNN dan Naive Bayes tetap menunjukkan performa yang kompetitif dan layak dipertimbangkan sebagai alternatif model prediksi banjir. Temuan ini menyediakan wawasan berharga tentang efektivitas berbagai pendekatan pembelajaran mesin untuk prediksi banjir. Model Random Forest terbukti sebagai pendekatan yang unggul, namun KNN dan Naive Bayes juga menunjukkan potensi yang signifikan. Hasil penelitian ini berkontribusi pada pengembangan sistem prediksi banjir yang lebih andal dan akurat, dengan implikasi penting bagi manajemen bencana dan pengurangan risiko banjir.

Kata kunci: KNN, Model Machine Learning, Naïve Bayes. Prediksi Banjir, Random Forest

ABSTRACT

This study presents a comparative analysis of machine learning models for flood prediction using historical rainfall data. Three models, namely K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest, are evaluated based on their performance metrics, including accuracy, precision, recall, F1 score, and ROC AUC. The evaluation results show that the Random Forest model consistently outperforms KNN and Naive Bayes. Random Forest achieves a perfect score (100%) on all measured indicators. Meanwhile, KNN and Naive Bayes also demonstrate competitive performance, albeit with some trade-offs between precision and recall. Specifically, for accuracy, precision, recall, F1 score, and ROC AUC, the Random Forest model scores 100%, whereas KNN and Naive Bayes are in the range of 90-95%. Nevertheless, KNN and Naive Bayes still show competitive performance and are worth considering as alternative flood prediction models. These findings provide valuable insights into the effectiveness of various machine learning approaches for flood prediction. The Random Forest model proves to be the superior approach, yet KNN and Naive Bayes also show significant potential. The results of this study contribute to the development of more reliable and accurate flood prediction systems, with important implications for disaster management and flood risk reduction...

Keywords: Flood Prediction, KNN, Machine Learning Models. Naïve Bayes, Random Forest

1. PENDAHULUAN

Banjir adalah bencana paling umum di seluruh dunia dengan frekuensi tinggi, dampak besar, dan biaya kerusakan yang signifikan. Banjir merupakan kondisi sementara di mana air tiba-tiba menggenangi daratan, baik dari pasang surut atau limpasan cepat akibat hujan deras, sehingga menyebabkan genangan air yang meluas (Joy et al., 2019). Selain dampak langsungnya, banjir juga bisa menimbulkan efek jangka panjang. Banjir dapat mencemari sumber air dan menyebarkan penyakit [1].

Selama bertahun-tahun, strategi pengelolaan banjir telah dibangun berdasarkan teknik peramalan banjir konvensional yang menggunakan model hidrologi dan meteorologi, yang telah memberikan wawasan yang signifikan. Namun, dengan memperhatikan interaksi kompleks antara komponen-komponen penyebab banjir dan semakin banyaknya data yang tersedia, diperlukan metodologi yang lebih canggih, adaptif, dan berbasis data [2].

Bidang kecerdasan buatan berfokus pada pembuatan mesin yang mampu memproses data, mempelajarinya, dan membuat penilaian. Penggunaan pembelajaran mesin adalah pendekatan menarik untuk prakiraan banjir karena dapat mengungkap korelasi kompleks dalam kumpulan data besar. Kemampuannya untuk menggabungkan data dari berbagai sumber, termasuk citra satelit, data pengukur sungai, dan model iklim, memberikan peluang untuk meningkatkan presisi, prediktabilitas, dan waktu respons terhadap banjir [3].

Penelitian terdahulu sudah mengusulkan untuk menerapkan beberapa metode untuk melakukan prediksi banjir dengan menggunakan metode Deep learning hybrid ConvLSTM blishes [4]. pada penelitian ini dilakukan penerapan algoritma deep learning hybrid (ConvLSTM) dengan menggabungkan kelebihan prediktif Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM) Network untuk merancang dan mengevaluasi model peramalan banjir. Hasil penelitian menunjukkan bahwa model peramalan banjir berbasis algoritma deep learning ConvLSTM yang diusulkan memiliki akurasi yang unggul dibandingkan metode acuan.

Pada penelitian sebelumnya yang berjudul "Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques" [5], pendekatan machine learning terkini digunakan, termasuk decision tree, random forest, naive Bayes, artificial neural network (terutama Multilayer Perceptron), support vector machine, dan fuzzy logic. Hasil penelitian menunjukkan bahwa sistem peramalan banjir yang dikembangkan dapat meramalkan kejadian banjir, termasuk lokasi yang terdampak dan tingkat keparahannya. Sistem dengan konfigurasi Multilayer Perceptron (MLP) Artificial Neural Network memberikan performa peramalan terbaik, dengan persentase akurasi 97,93%, indeks Kappa 0,89, Mean Absolute Error (MAE) 0,01, dan Root Mean Squared Error (RMSE) 0,10. Hasil ini menunjukkan bahwa sistem peramalan banjir yang dibangun dengan mengintegrasikan berbagai data besar dan menggunakan teknik machine learning terkini dapat memberikan hasil peramalan yang akurat dan efektif.

Penelitian berjudul "Flood Forecasting Based on Machine Learning Pattern Recognition and Dynamic Migration of Parameters" [6], berfokus pada peramalan banjir menggunakan pendekatan machine learning, pengenalan pola, dan dinamika migrasi parameter. Penelitian ini dilakukan di sub-DAS Jingle, anak sungai Sungai Kuning, China, dengan menganalisis data hujan dan aliran dari 98 kejadian banjir antara tahun 1971-2014 menggunakan teknik clustering dinamis dan random forest. Hasilnya menunjukkan bahwa sistem yang dikembangkan mampu meramalkan kejadian banjir, termasuk lokasi yang terdampak dan tingkat keparahannya, dengan Multilayer Perceptron (MLP) Artificial Neural Network memberikan performa terbaik, mencapai akurasi 97,93%, indeks Kappa 0,89, Mean Absolute Error (MAE) 0,01, dan Root Mean Squared Error (RMSE) 0,10. Penelitian ini membuktikan bahwa integrasi data besar dan teknik machine learning terkini dapat menghasilkan peramalan banjir yang akurat dan efektif.

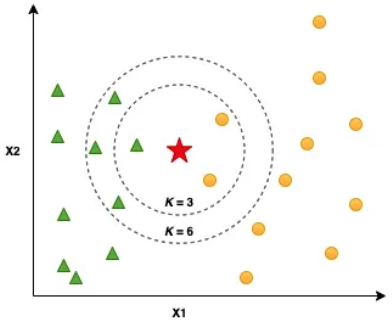
Penelitian sebelumnya yang berjudul "Penerapan Machine Learning untuk Prediksi Bencana Banjir" [7] peneliti mengusulkan sebuah model baru yang dinamakan algoritma Deep Neural Investigation Network (DNIN). Algoritma ini mengombinasikan Convolutional Neural Network (CNN) dan Bidirectional Long Short-Term Memory (BiLSTM). Proses dari metode yang diusulkan terdiri dari tiga bagian utama: ekstraksi fitur spasial menggunakan CNN, penangkapan pola temporal menggunakan BiLSTM, serta penggabungan hasil dari kedua metode tersebut untuk memprediksi tingkat bahaya banjir. Hasil penelitian menunjukkan bahwa model DNIN yang diusulkan lebih unggul dibandingkan dengan model-model sebelumnya dalam melakukan prediksi bencana banjir.

Meskipun berbagai penelitian sebelumnya telah dilakukan untuk memprediksi kejadian banjir, sebagian besar penelitian tersebut hanya mengandalkan metode tradisional seperti analisis regresi, tanpa mengeksplorasi potensi teknik pembelajaran mesin yang lebih canggih. Dalam penelitian ini, peneliti akan menerapkan tiga metode pembelajaran mesin yang berbeda, yaitu K-Nearest Neighbors (KNN), Naive Bayes, dan Random Forest, untuk membandingkan kinerja masing-masing pendekatan dalam memprediksi kejadian banjir. Dengan memanfaatkan keunggulan dari beragam algoritma pembelajaran mesin, peneliti berharap dapat menghasilkan model prediksi banjir yang lebih akurat dan andal dibandingkan dengan penelitian-penelitian sebelumnya.

2. METODE PENELITIAN

K-Nearest Neighbors (KNN)

Algoritma K-Nearest Neighbor (K-NN) adalah a algoritma klasifikasi yang mengklasifikasikan objek baru berdasarkan mayoritas kategori dari K-tetangga terdekatnya, yaitu data pembelajaran yang jaraknya paling dekat dengan objek tersebut[8]. Tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel latih. Algoritma K-NN tidak menggunakan model apapun untuk dicocokkan, melainkan hanya berdasarkan pada memori. Ketika diberikan titik uji, algoritma ini akan menemukan sejumlah K objek (titik pelatihan) yang paling dekat dengan titik uji tersebut [9] Gambar 1 mengilustrasikan model K-Nearest Neighbors (KNN).



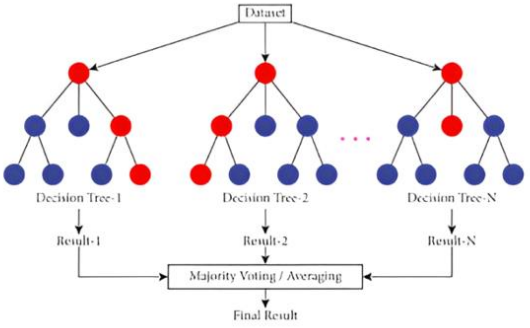
Gambar 1. K-Nearest Neighbors (KNN)

Naïve Bayes

Algoritma Naïve Bayes merupakan salah satu metode untuk mengklasifikasikan fenomena berdasarkan probabilitas terjadinya atau tidak terjadinya fenomena tersebut. Metode ini didasarkan pada Teorema Bayes, yang memungkinkan pembaruan probabilitas a posteriori berdasarkan bukti baru. Dengan pendekatan ini, kita dapat menggunakan informasi sebelumnya (prior) dan bukti baru (likelihood) untuk menghasilkan perkiraan probabilitas yang lebih akurat [10].

Random Forest

Teknik ensemble learning yang disebut algoritma Random Forest menggabungkan beberapa pohon keputusan seperti yang ditunjukan pada gambar 2. Random Forest adalah sekumpulan metode pembelajaran yang terdiri dari sejumlah pohon pilihan yang dibangun secara acak. Setiap pohon keputusan terdiri dari subset acak dari data pelatihan dan subset acak dari fitur yang tersedia melalui teknik bagging [11]. Random Forest secara acak memilih subset fitur untuk dipertimbangkan dalam splitting node selama pembangunan setiap pohon keputusan [12]. Prediksi data baru dibuat melalui mayoritas suara dari semua cabang keputusan di hutan. Random Forest, yang dapat di-paralelkan untuk meningkatkan kecepatan pelatihan model pada dataset besar, efektif dalam mengatasi overfitting berkat variasi yang dihasilkan oleh bagging dan pemilihan fitur acak [13]. Metode ini terkenal karena dapat membuat prediksi yang akurat dan memiliki toleransi yang baik terhadap noise dalam data [14].



Gambar 2. Random Forest

Dataset

Dalam penelitian ini, digunakan dataset publik yang diambil dari situs web https://www.kaggle.com/code/mukulthakur177/flood-prediction-model/data, yang berisi data curah hujan yang berjumlah 117 data dari tahun 1901 hingga 2018. Dataset yang diperoleh mencakup 16 atribut, yaitu: *subdivision, year, jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec, annual rainfall, dan floods*. Untuk mengaplikasikan dataset ini menggunakan Python, langkah pertama yang dilakukan adalah proses impor dataset. Gambar 3 menunjukkan contoh dataset yang digunakan.

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	YES
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	YES
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	YES
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	YES
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	NO

Gambar 3. Contoh Dataset

Performance Evaluation

Dalam rangka mengevaluasi performa model prediksi, penelitian ini menjalankan analisis dengan mempertimbangkan beragam metrik kunci guna memastikan keakuratan serta reliabilitas prediksi yang terkait dengan kemungkinan kejadian banjir. Evaluasi tersebut meliputi presisi, recall, skor F1, akurasi, kurva ROC (Receiver Operating Characteristic), dan AUC (Area Under the Curve). Akurasi, yang mengukur proporsi prediksi yang tepat dibandingkan dengan total prediksi, dihitung dengan menggunakan rumus:

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

Presisi mengukur relevansi dari sampel yang dipilih oleh model, yang bertujuan untuk mengurangi jumlah positif palsu.

$$Presisi = \frac{TP}{TP+FP} \qquad (2)$$

Sensitivitas (Recall) adalah metrik evaluasi yang menggambarkan seberapa baik suatu model dalam mengidentifikasi kelas positif dengan benar. Rumus menentukan recall:

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

F1 Score merupakan metrik evaluasi yang mencerminkan keseimbangan antara Presisi (*Precision*) dan Sensitivitas (*Recall*)

$$Recall = \frac{2\ x\ Presisi\ x\ Recall}{Presisi\ +Recall} \qquad (4)$$

Dalam konteks evaluasi model klasifikasi, parameter ROC (Receiver Operating Characteristic) memberikan representasi visual tentang dampak variasi ambang batas keputusan pada kinerja model. ROC mengilustrasikan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) seiring dengan perubahan ambang batas keputusan. TPR mengukur kemampuan model untuk mengidentifikasi kelas positif, sementara FPR mengindikasikan tingkat kesalahan dalam memprediksi kelas negatif. Secara numerik, Area Under the Curve (AUC) pada kurva ROC memberikan ukuran kualitas keseluruhan dari model. Model yang efektif akan memiliki AUC yang lebih tinggi, sementara nilai AUC mendekati 0 menunjukkan kapasitas pemisahan kelas yang buruk. Misalnya, jika AUC adalah 0,5, menandakan bahwa model tersebut tidak memiliki kemampuan dalam memisahkan kelas secara efektif [15].

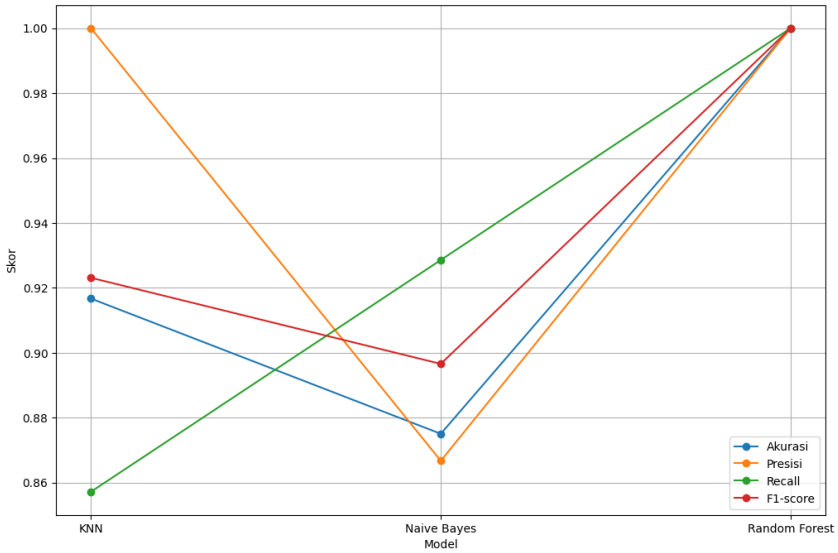
3. HASIL DAN PEMBAHASAN

Berikut adalah hasil dari evaluasi tiga model klasifikasi yang berbeda, yaitu K-Nearest Neighbors (KNN), Naive Bayes, dan Random Forest, yang ditampilkan pada tabel 1, disertai dengan grafik untuk memperjelas perbandingan kinerja masing-masing model.

TABEL 1. Evaluasi Kinerja Model Klasifikasi

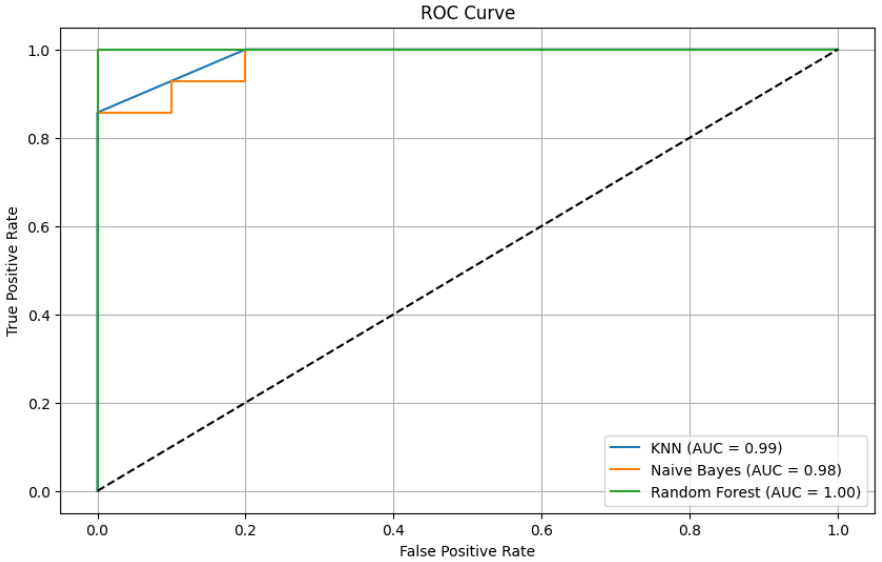
Model	Akurasi	Presisi	Recall	F1-Score	AUC
KNN	0.9167	1.0000	0.8571	0.9231	0.9857
Naïve Bayes	0.8750	0.8667	0.9286	0.8966	0.9786
Random Forest	1.0000	1.0000	1.0000	1. 0000	1. 0000

Pada gambar 4 memvisualisasikan perbandingan kinerja ketiga model dalam memprediksi kejadian banjir berdasarkan data curah hujan historis.



Gambar 4. Perbandingan Metrik Kinerja Model Pembelajaran Mesin

Pada gambar 5 menunjukkan perbandingan kurva ROC untuk ketiga model, yang memvisualisasikan perbandingan antara False Positive Rate (FPR) dan True Positive Rate (TPR).



Gambar 5. Perbandingan Kurva ROC Model Pembelajaran Mesin

Dari hasil evaluasi, dapat dilihat bahwa model Random Forest memberikan kinerja terbaik dengan akurasi, presisi, recall, dan F1-score mencapai 100%, serta Area Under Curve (AUC) dari kurva ROC mencapai nilai maksimal 1.0000. Model KNN dan Naive Bayes juga menunjukkan kinerja yang baik, namun dengan trade-off yang berbeda antara presisi dan recall.

Kinerja model Random Forest yang sangat baik menandakan bahwa model tersebut memiliki kemampuan yang tinggi dalam memprediksi kejadian banjir berdasarkan data curah hujan historis. Namun, keputusan akhir untuk memilih model tergantung pada prioritas spesifik dari akurasi, presisi, atau recall dalam aplikasi praktis.

4. SIMPULAN

Dalam penelitian ini, peneliti mengevaluasi tiga model klasifikasi yang berbeda, yakni K-Nearest Neighbors (KNN), Naive Bayes, dan Random Forest, untuk memproyeksikan kemungkinan kejadian banjir berdasarkan data historis curah hujan. Hasil evaluasi menunjukkan bahwa model Random Forest memiliki kinerja yang sangat baik dengan mencapai nilai akurasi, presisi, recall, F1-score, dan AUC dari kurva ROC sebesar 100%. Ini menandakan kemampuan model Random Forest dalam membedakan antara kelas positif dan negatif dengan sangat baik.

Meskipun demikian, model KNN dan Naive Bayes juga menunjukkan kinerja yang baik, namun terdapat beberapa trade-off antara presisi dan recall. Rekomendasi berdasarkan hasil evaluasi adalah penggunaan model Random Forest sebagai pilihan utama dalam sistem prediksi banjir, terutama untuk aplikasi yang memprioritaskan akurasi tinggi. Namun, jika presisi yang tinggi lebih diutamakan, model KNN dapat menjadi alternatif yang baik,

sementara Naive Bayes cocok untuk situasi di mana recall yang tinggi lebih penting. Untuk penelitian mendatang, peneliti merekomendasikan penggunaan fitur tambahan seperti data topografi dan penggunaan tanah untuk meningkatkan kinerja model dalam memprediksi kejadian banjir yang jarang terjadi

DAFTAR PUSTAKA

[1] C. J. Talbot *et al.*, "The impact of flooding on aquatic ecosystem services," *Biogeochemistry*, vol. 141, no. 3, pp. 439–461, Dec. 2018, doi: 10.1007/s10533-018-0449-7.

[2] Wang Guangsheng, Yang Jianqing, Hu Yuzhong, Li Jingbing, and Yin Zhijie, "Application of a novel artificial neural network model in flood forecasting," *Environ Monit Assess*, Jan. 2022.

[3] A. Rajab *et al.*, "Flood Forecasting by Using Machine Learning: A Study Leveraging Historic Climatic Records of Bangladesh," *Water (Switzerland)*, vol. 15, no. 22, Nov. 2023, doi: 10.3390/w15223970.

[4] M. Moishin, R. C. Deo, R. Prasad, N. Raj, and S. Abdulla, "Designing deep-based learning flood forecast model with ConvLSTM hybrid algorithm," *IEEE Access*, vol. 9, pp. 50982–50993, 2021, doi: 10.1109/ACCESS.2021.3065939.

[5] S. Puttinaovarat and P. Horkaew, "Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 5885–5905, 2020, doi: 10.1109/ACCESS.2019.2963819.

[6] Y. Tang *et al.*, "flood forecasting based on machine learning pattern recognition and dynamic migration of parameters," *J Hydrol Reg Stud*, vol. 47, Jun. 2023, doi: 10.1016/j.ejrh.2023.101406.

[7] Vidya S, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Rainfall Based Flood Prediction in Kerala Using Machine Learning," 2024. [Online]. Available: www.ijisae.org

[8] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water (Switzerland)*, vol. 11, no. 11, 2019, doi: 10.3390/w11112210.

[9] R. A. Asmara *et al.*, "Prediksi Banjir Lahar Dingin Pada Lereng Merapi Menggunakan Data Curah Hujan Dari Satelit," *JIP (Jurnal Informatika Polinema)* , 2021.

[10] A. Habibi, M. R. Delavar, M. S. Sadeghian, and B. Nazari, "Flood Susceptibility Mapping And Assessment Using Regularized Random Forest And Naïve Bayes Algorithms," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus Publications, Jan. 2023, pp. 241–248. doi: 10.5194/isprs-annals-X-4-W1-2022-241-2023.

[11] S. Vijay and K. Kamaraj, "Ground Water Quality Prediction using Machine Learning Algorithms in R," *International Journal of Research and Analytical Reviews*, vol. 6, no. 1, 2019.

[12] "Smart Prediction of Water Quality System for Aquaculture Using Machine Learning Algorithms," *Journal of Current Trends in Computer Science Research*, vol. 2, no. 3, 2023, doi: 10.33140/jctcsr.02.03.01.

[13] M. YURTSEVER and M. EMEÇ, "Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability," *Ege Akademik Bakis (Ege Academic Review)*, 2023, doi: 10.21121/eab.1252167.

[14] Md. M. Hassan *et al.*, "Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms," *Human-Centric Intelligent Systems*, vol. 1, no. 3–4, 2021, doi: 10.2991/hcis.k.211203.001.

[15] A. Tikaningsih, P. Lestari, A. Nurhopipah, I. Tahyudin, E. Winarto, and N. Hassa, "Telematika Optuna Based Hyperparameter Tuning for Improving the Performance Prediction Mortality and Hospital Length of Stay for Stroke Patients," vol. 17, no. 1, pp. 1–16, 2024, doi: 10.35671/telematika.v17i1.2816.