

BAB II

TINJAUAN PUSTAKA

A. Prediksi Cuaca dengan *Machine Learning*

Penelitian oleh Syahreza dkk. (2024) [5] menggunakan algoritma *Random Forest*, *Support Vector Regression (SVR)*, dan *XGBoost* untuk melakukan prediksi suhu udara rata-rata harian berdasarkan data dari BMKG dan sensor *IoT*. Dataset yang digunakan mencakup fitur-fitur cuaca seperti curah hujan, kelembapan udara, suhu minimum dan maksimum, serta tekanan udara. Sebelum pelatihan model, data terlebih dahulu diproses melalui normalisasi min-max dan pemisahan data latih dan uji sebesar 80:20. Hasil penelitian menunjukkan bahwa *XGBoost* menghasilkan performa terbaik dengan nilai R^2 sebesar 0,8183, diikuti oleh *Random Forest* dan *SVR*, yang masing-masing memiliki akurasi lebih rendah. Hasil ini diperkuat oleh Dwiyanti dan Prianto (2023) [8] yang menggunakan algoritma *Random Forest* untuk memprediksi kelas kondisi cuaca harian di Kota Jakarta. Fitur yang digunakan meliputi kelembapan, suhu, tekanan udara, kecepatan angin, dan arah angin. Setelah dilakukan *cleaning* data, *encoding* label target, serta *split data* 70:30, model memberikan nilai *ROC-AUC* sebesar 0,92, yang menunjukkan kemampuan klasifikasi model dalam membedakan kelas cuaca dengan sangat baik.

Penelitian oleh Siregar dkk. (2020) [9] membandingkan performa beberapa algoritma klasifikasi cuaca di Australia, termasuk *Naive Bayes*, *Decision Tree*, *Random Forest*, *Deep Learning*, serta metode *ensemble learning (bagging)*. Dataset yang digunakan mencakup fitur-fitur meteorologis seperti suhu maksimum dan minimum, kelembapan, tekanan udara, curah hujan, dan arah angin. Proses prapemrosesan data mencakup penanganan nilai kosong, normalisasi data, serta *encoding* label target. Hasil evaluasi menunjukkan bahwa model *Deep Learning* memperoleh akurasi tertinggi sebesar 84,06%, sedangkan metode *ensemble learning* berbasis

bagging juga menunjukkan performa kompetitif, dengan akurasi berkisar antara 81,21% hingga 82,92%. Temuan ini menunjukkan bahwa pendekatan *ensemble* tetap menjadi pilihan yang andal dalam klasifikasi cuaca, terutama pada data berskala besar dan kompleks.

Penelitian oleh Putramulyo dan Alaa (2018) [10] menerapkan metode regresi *linier* berganda untuk memprediksi curah hujan bulanan di Kota Samarinda, dengan menggunakan suhu udara dan kelembapan udara sebagai prediktor utama. Data yang digunakan diperoleh dari stasiun cuaca setempat dan diproses melalui tahapan penghapusan *outlier* dan normalisasi sebelum dilakukan pelatihan *model*. Hasil evaluasi menunjukkan bahwa model memiliki koefisien korelasi sebesar 0,93, yang mengindikasikan adanya hubungan yang sangat kuat antara prediktor dan target. Namun demikian, model menunjukkan kecenderungan untuk mengestimasi curah hujan secara lebih rendah dari nilai sebenarnya (*underestimate*), terutama pada periode dengan intensitas hujan tinggi. Hal ini menandakan perlunya pendekatan model *non-linier* untuk menangkap pola yang lebih kompleks dalam data cuaca.

B. Prediksi Banjir dan Penanganan Data Tidak Seimbang

Banyak penelitian menyoroti bahwa salah satu tantangan utama dalam prediksi banjir adalah ketidakseimbangan data (*class imbalance*) antara kejadian banjir (kelas minoritas) dan tidak banjir (kelas mayoritas). Akbar dan Sanjaya (2023) [4] meneliti penerapan teknik *class weight* pada algoritma *Random Forest* dalam tugas klasifikasi intensitas curah hujan. Dataset yang digunakan mengandung dominasi data dengan intensitas rendah, sehingga diperlukan strategi khusus agar model tetap mampu mengenali pola pada data kelas minoritas. Dalam penelitian tersebut, penggunaan parameter `class_weight=balanced` terbukti mampu meningkatkan sensitivitas model terhadap kelas minoritas. Namun demikian, hasil evaluasi menggunakan *area under the precision-recall curve (AUC-*

PR) menunjukkan bahwa meskipun performa membaik, penanganan terhadap *imbalance* data belum sepenuhnya optimal, terutama dalam konteks generalisasi terhadap data baru. Hal ini menegaskan perlunya pendekatan tambahan seperti *oversampling*, *undersampling*, atau *synthetic data generation* dalam pengembangan sistem prediksi banjir yang andal.

Sementara itu, Mursianto dkk. (2021) [11] membandingkan kinerja algoritma *Random Forest* dan *XGBoost* dalam tugas klasifikasi curah hujan. Penelitian ini secara khusus menyoroti masalah ketidakseimbangan kelas, dan menerapkan *Synthetic Minority Over-sampling Technique (SMOTE)* sebagai metode *resampling* untuk meningkatkan representasi kelas minoritas. Hasil evaluasi menunjukkan bahwa setelah penerapan *SMOTE*, nilai *recall* dari kedua model meningkat secara signifikan, bahkan mencapai lebih dari 90%, yang menandakan kemampuan model dalam mendeteksi kejadian hujan meningkat tajam. Temuan ini menunjukkan bahwa kombinasi antara model *ensemble* dan teknik *resampling* mampu secara efektif mengurangi bias terhadap kelas mayoritas, serta meningkatkan kemampuan model dalam menangani distribusi data yang tidak seimbang.

Nugroho dan Harini (2024) [12] juga menegaskan bahwa penggunaan teknik *Synthetic Minority Over-sampling Technique (SMOTE)* dapat secara signifikan meningkatkan akurasi dan presisi terhadap kelas minoritas dalam dataset yang tidak seimbang. Dalam studi mereka, *SMOTE* berhasil memperbaiki distribusi data dan mendorong performa model menjadi lebih adil terhadap kelas minoritas. Namun demikian, penelitian ini juga menekankan pentingnya eksplorasi algoritma *ensemble* lainnya di luar *Random Forest* dan *XGBoost*. Mereka merekomendasikan untuk menguji algoritma seperti *AdaBoost* dan *LightGBM*, yang memiliki potensi lebih baik dalam hal efisiensi pemrosesan dan kemampuan menangani data numerik dalam jumlah besar. Rekomendasi ini sejalan dengan tren terkini di bidang prediksi cuaca dan bencana, di mana pemilihan algoritma yang adaptif terhadap struktur data menjadi kunci keberhasilan model prediktif.

C. Prediksi dan Implementasi Sistem Banjir

Implementasi sistem prediksi banjir dilakukan oleh beberapa peneliti melalui *web* dan perangkat keras. Sasoko dkk. (2024) [13] menggabungkan algoritma *K-Means* dan *Random Forest* untuk memprediksi risiko banjir di DKI Jakarta. Model yang dikembangkan diintegrasikan ke dalam aplikasi *web*, dan menghasilkan performa yang sangat tinggi dengan nilai *recall* dan *F1-score* mendekati 1.00, yang menunjukkan kemampuan deteksi banjir yang sangat baik.

Selanjutnya, Asnawi dkk. (2024) [14] membangun model prediksi banjir berbasis *XGBoost* menggunakan dataset dari *Kaggle* yang mencakup berbagai fitur lingkungan dan manajemen banjir, seperti curah hujan, ketinggian tanah, dan kapasitas drainase. Hasil evaluasi menunjukkan bahwa model *XGBoost* memberikan performa terbaik dengan nilai *Mean Absolute Percentage Error (MAPE)* hanya sebesar 0,16%, jauh lebih unggul dibandingkan model lainnya dalam studi tersebut. Sementara itu, Pamungkas (2025) [10] membandingkan algoritma *XGBoost* dan *Random Forest* untuk prediksi bencana di wilayah Kota Semarang. Berdasarkan nilai *Mean Squared Error (MSE)*, *XGBoost* terbukti lebih unggul secara signifikan dibandingkan *Random Forest*, yang menegaskan konsistensi keunggulan *XGBoost* dalam menangani data prediksi bencana dengan kompleksitas tinggi.

D. Analisis Faktor Penyebab dan Kombinasi Model

Penelitian Suaif dan Rahayu (2025) [16] melakukan analisis pola kejadian banjir di wilayah Bandar Lampung dengan membandingkan tiga pendekatan yaitu *ARIMA*, *Random Forest*, dan *XGBoost*. Hasil analisis menunjukkan bahwa model *Random Forest* dan *XGBoost* mampu mengidentifikasi curah hujan harian sebagai faktor dominan penyebab banjir, sehingga lebih unggul dibandingkan pendekatan statistik

konvensional seperti *ARIMA* yang kurang adaptif terhadap data non-linier dan fluktuatif

Sementara itu, Fitriyaningsih dan Basani (2019) [17] memanfaatkan algoritma *Backpropagation Neural Network (BP-NN)* dan *Support Vector Machine (SVM)* untuk memprediksi risiko banjir berdasarkan data historis curah hujan dan debit air sungai. Penelitian ini menekankan bahwa ketersediaan data historis yang memadai serta pemilihan model yang tepat merupakan kunci utama dalam prediksi banjir jangka pendek maupun menengah, terutama pada wilayah dengan pola hidrologi yang dinamis.

Adapun Rachmawardani dkk. (2022) [18] dalam studi literturnya menyimpulkan bahwa peningkatan kualitas prediksi banjir dapat dicapai melalui pendekatan *hybrid machine learning*, *model ensemble*, teknik dekomposisi data, serta optimalisasi parameter algoritma. Pendekatan-pendekatan ini diyakini mampu mengatasi berbagai tantangan dalam pemodelan bencana hidrometeorologi, seperti kompleksitas data spasial-temporal dan ketidakseimbangan kelas.

E. Implementasi Sistem Prediksi Berbasis Web dan Hardware

Kurniawan dkk. (2016) [19] merancang sebuah sistem pemantauan cuaca berbasis mikrokontroler sebagai bagian dari sistem peringatan dini banjir. Sistem ini terhubung secara langsung dengan layanan *OpenWeatherMap API* untuk memperoleh data cuaca *real-time*, seperti suhu, kelembapan, dan curah hujan. Dalam pengujian, sistem menunjukkan tingkat akurasi yang baik, dengan *margin error* pengukuran suhu di bawah 5%. Penelitian ini menegaskan bahwa integrasi antara perangkat keras (mikrokontroler) dan data daring (*API*) dapat meningkatkan kecepatan dan keandalan respons sistem dalam mendeteksi potensi bencana, serta dapat dijadikan komponen pendukung dalam sistem prediksi atau peringatan banjir berbasis teknologi.

F. GAP Penelitian

Berdasarkan berbagai penelitian di atas, belum banyak yang mengembangkan aplikasi prediksi cuaca dan banjir *real-time* secara terintegrasi berbasis *web* dengan *model ensemble weighted voting*, khususnya di wilayah DIY dan Jawa Barat, serta implementasi prediksi cuaca (*real-time*, mingguan). Selain itu, kebanyakan penelitian masih mengandalkan data historis saja tanpa integrasi data *API real-time* dan visualisasi berbasis peta interaktif.

Penelitian ini hadir untuk menjawab *gap* tersebut dengan membangun aplikasi prediksi banjir dan cuaca berbasis *machine learning ensemble* (*Random Forest, XGBoost, LightGBM*) menggunakan data historis dan *API* cuaca *real-time*, serta implementasi *web* interaktif yang dapat diakses masyarakat luas.

G. Tabel Perbandingan Penelitian

Tabel perbandingan di bawah ini menunjukkan bahwa penelitian ini tidak hanya mengadopsi algoritma *ensemble* seperti penelitian terdahulu, tetapi juga menggabungkan prediksi banjir dan prediksi cuaca berbasis data historis di Indonesia, terintegrasi API *real-time*, serta ditampilkan dalam aplikasi *web*. Penanganan *imbalance* juga dilakukan tanpa *oversampling* (seperti *SMOTE*), melainkan dengan *threshold tuning* dan *voting ensemble*, sehingga lebih relevan dan aplikatif untuk kondisi data lokal.

Tabel 1. 1 Tabel Perbandingan Penelitian

No	Peneliti / Tahun	Lokasi / Kasus	Algoritma / Model	Dataset & Fitur	Penanganan Imbalanced	Output / Implementasi	Evaluasi	Gap & Relevansi dengan Penelitian Ini
1	Syahreza et al. [1]	Prediksi Suhu (BMKG)	RF, SVR, XGBoost	Data suhu, kelembapan, curah hujan, 1650 baris	-	Prediksi suhu rata-rata	MAE, MSE, R ²	Tidak membahas prediksi banjir, belum ada integrasi web dan API real-time

2	Mursianto et al. [6]	Prediksi Hujan	RF, XGBoost + SMOTE	WeatherAUS Kaggle, 145.460 baris, 23 fitur	SMOTE	Prediksi hujan	Akurasi, Recall	SMOTE meningkatkan recall, tapi belum pada data banjir di Indonesia
3	Akbar & Sanjaya [5]	Klasifikasi Hujan	Random Forest (class_weight)	5 kelas intensitas curah hujan, AWS BMKG	Class Weight	Prediksi curah hujan	Akurasi, AUC-PR	Kelas minoritas lebih terdeteksi, belum fokus prediksi banjir
4	Sasoko et al. [8]	Banjir Jakarta	K-Means, Random Forest	Data tinggi air sungai, real-time Open Data	Clusterisasi + RF	Aplikasi web prediksi	Recall, F1, Akurasi	Fokus tinggi air & status siaga, fitur banjir sederhana
5	Asnawi et al. [9]	Prediksi Banjir	XGBoost	Faktor lingkungan, 50.000 data (Kaggle)	-	Web prediksi risiko banjir	RMSE, MAE, MAPE	XGBoost kuat, tapi dataset bukan lokal Indonesia, fitur berbeda
6	Pamungkas [10]	Prediksi Bencana (Semarang)	RF, XGBoost	Data BPBD, jumlah kejadian tahunan	-	Estimasi jumlah kejadian	MSE, RMSE	Belum integrasi data real-time & prakiraan cuaca multi-skala

7	Siregar et al. [3]	Prediksi Cuaca Australia	Ensemble, DL, RF, DT, NB	142.193 record, 24 atribut (Kaggle)	-	Klasifikasi cuaca	Akurasi, MSE	Belum integrasi banjir & cuaca sekaligus, tidak berbasis Indonesia
8	Penelitian Ini	DIY & Jabar, 2018-2023	RF, XGBoost, LightGBM (ensemble voting berbobot)	BMKG, BNPB, OWM, data real-time, >6000 baris	Threshold tuning, tanpa oversampling	Prediksi banjir (binary), prediksi cuaca real-time, weekly, aplikasi web Flask	Akurasi, Precision, Recall, F1, Confusion Matrix, R ² , MAE, RMSE, validasi temporal	Integrasi prediksi banjir & cuaca multi-skala, berbasis data historis, integrasi API real-time, peta interaktif, penanganan imbalance dengan threshold & voting, tanpa SMOTE