Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

Optimasi Hiperparameter Gradient Boosting Regressor Untuk Prediksi Curah Hujan Berdasarkan Data Cuaca Harian BMKG

Mohamad Arif Abdul Syukur[1,*], Suhartono[2], M. Imamudin[3]

Sains Dan Teknologi, Magister Informatika, UIN Maulana Malik Ibrahim, Malang, Indonesia
Email: [1,*]marifabdulsyukur@gmail.com, [2]suhartono@ti.uin-malang.ac.id, [3]imamudin@ti.uin-malang.ac.id
Email Penulis Korespondensi: marifabdulsyukur@gmail.com

Abstrak−Prediksi curah hujan harian memiliki peran krusial dalam mitigasi bencana hidrometeorologi dan perencanaan sektor pertanian. Namun, karakteristik data curah hujan yang bersifat *zero-inflated* dan non-linear menjadi tantangan utama dalam menghasilkan prediksi yang mampu menangkap pola tren. Penelitian ini bertujuan untuk meningkatkan performa prediksi curah hujan harian menggunakan algoritma *Gradient Boosting Regressor*. Dataset yang digunakan berasal dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dengan 942 observasi dan 11 variabel cuaca harian, termasuk suhu, kelembapan, durasi penyinaran, dan kecepatan angin. Tahap prapemrosesan melibatkan penanganan nilai hilang dengan *K-Nearest Neighbors Imputer* serta standardisasi fitur, sementara sifat *zero-inflated* diatasi secara komputasional melalui struktur *decision tree* pada model yang mampu mengisolasi pola curah hujan nol secara efektif. Optimasi *hyperparameter* dilakukan menggunakan *RandomizedSearchCV* dan menunjukan hasil eksperimen bahwa model teroptimasi secara signifikan mengungguli model *default* (baseline), dengan penurunan nilai RMSE dari 13.81 menjadi 12.21 (peningkatan 11.58%) dan MAE dari 8.51 menjadi 7.70 (peningkatan 9,52%). Temuan ini memberikan landasan empiris bahwa efisiensi model machine learning pada data meteorologi tidak hanya bergantung pada algoritma, tetapi juga pada kecermatan dalam mengatur parameter yang disesuaikan dengan karakteristik data lokal dan membuktikan optimasi parameter pada *gradient boosting* secara empiris terbukti efektif dalam meningkatkan kemampuan model untuk mempelajari pola variabilitas curah hujan yang kompleks, sekaligus memberikan solusi untuk menangani tantangan zero-inflated pada data meteorologi. Penelitian ini berkontribusi dalam menyajikan kerangka kerja optimasi hiperparameter berbasis *RandomizedSearchCV* pada algoritma *Gradient Boosting Regressor* untuk data meteorologi. Secara praktis, model ini menawarkan potensi sebagai instrumen pendukung keputusan awal bagi otoritas terkait dalam upaya mitigasi risiko bencana berbasis data di wilayah Kabupaten Malang.

Kata Kunci: Gradient Boosting Regressor; Prediksi Curah Hujan; Optimasi Hyperparameter; Machine Learning; Meteorologi

Abstract−Daily rainfall prediction plays a crucial role in hydrometeorological disaster mitigation and agricultural sector planning. However, the zero-inflated and non-linear nature of rainfall data poses a major challenge in generating accurate predictions. This study aims to improve the performance of daily rainfall prediction using the Gradient Boosting Regressor algorithm. The dataset used comes from the Meteorology, Climatology, and Geophysics Agency (BMKG) with 942 observations and 11 daily weather variables, including temperature, humidity, sunshine duration, and wind speed. The pre-processing stage involves handling missing values with K-Nearest Neighbors Imputer and feature standardization, while the zero-inflated nature is addressed computationally through a decision tree structure in the model that is able to effectively isolate zero rainfall patterns. Hyperparameter optimization was performed using RandomizedSearchCV and showed experimental results that the optimized model significantly outperformed the default model (baseline), with a decrease in RMSE values from 13.81 to 12.21 (an increase of 11.58%) and MAE from 8.51 to 7.70 (an increase of 9.52%). These findings provide an empirical basis that the efficiency of machine learning models on meteorological data depends not only on the algorithm, but also on the accuracy in setting parameters that are adjusted to the characteristics of local data and proves that parameter optimization in gradient boosting is empirically proven to be effective in improving the model's ability to learn complex rainfall variability patterns, while providing a solution to address the challenge of zero-inflation in meteorological data. This research contributes to the presentation of a RandomizedSearchCV based hyperparameter optimization framework for the Gradient Boosting Regressor algorithm for meteorological data. Practically, this model offers potential as an early decision-support instrument for relevant authorities in data-driven disaster risk mitigation efforts in Malang Regency.

Keywords: Gradient Boosting Regressor; Rainfall Prediction; Hyperparameter Optimization; Machine Learning; Meteorology

1. PENDAHULUAN

Prediksi curah hujan harian merupakan salah satu tantangan paling kompleks dalam bidang meteorologi karena sifatnya yang stokastik, non-linear, dan sangat dipengaruhi oleh variabel atmosfer yang saling berinteraksi[1]. Bagi Indonesia, yang merupakan negara agraris dengan kerentanan tinggi terhadap bencana hidrometeorologi seperti banjir dan tanah longsor, akurasi prediksi curah hujan menjadi krusial untuk mendukung mitigasi bencana dan ketahanan pangan. Namun, ketergantungan pada model fisik tradisional seringkali terkendala oleh keterbatasan data lokal dan tingginya biaya komputasi[2]. Pentingnya akurasi prediksi curah hujan harian ditegaskan oleh peningkatan frekuensi bencana hidrometeorologi dalam beberapa tahun terakhir. Di wilayah Sumatera Barat, Sumatera Utara dan Aceh, Badan Nasional Penanggulangan Bencana (BNPB) mencatat sejak awal tahun hingga November 2025 telah tercatat 2.726 kejadian bencana hidrometeorologi, dan banjir bandang akhir November tersebut menelan lebih dari 400 korban jiwa di tiga provinsi terdampak yang disebabkan intensitas curah hujan yang tinggi sesuai laporan BMKG yang mencatat beberapa wilayah di Sumut diguyur lebih dari 300 mm hujan per hari pada puncak kejadian. Kejadian tersebut menggarisbawahi kebutuhan mendesak akan instrumen prediktif untuk meminimalkan dampak risiko bencana bagi masyarakat.

Seiring dengan kemajuan teknologi, pendekatan berbasis *machine learning* telah menunjukkan potensi besar dalam memetakan pola cuaca yang kompleks. Beberapa penelitian sebelumnya telah menerapkan model regresi untuk peramalan curah hujan dengan hasil yang bervariasi. Markuna dkk. (2023) menggunakan data dari Agromet Forest Unit (AMFU) Ranichauri dan melaporkan nilai *Root Mean Squared Error* (RMSE) sebesar 10,54 [3]. Dalam penelitian lain

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

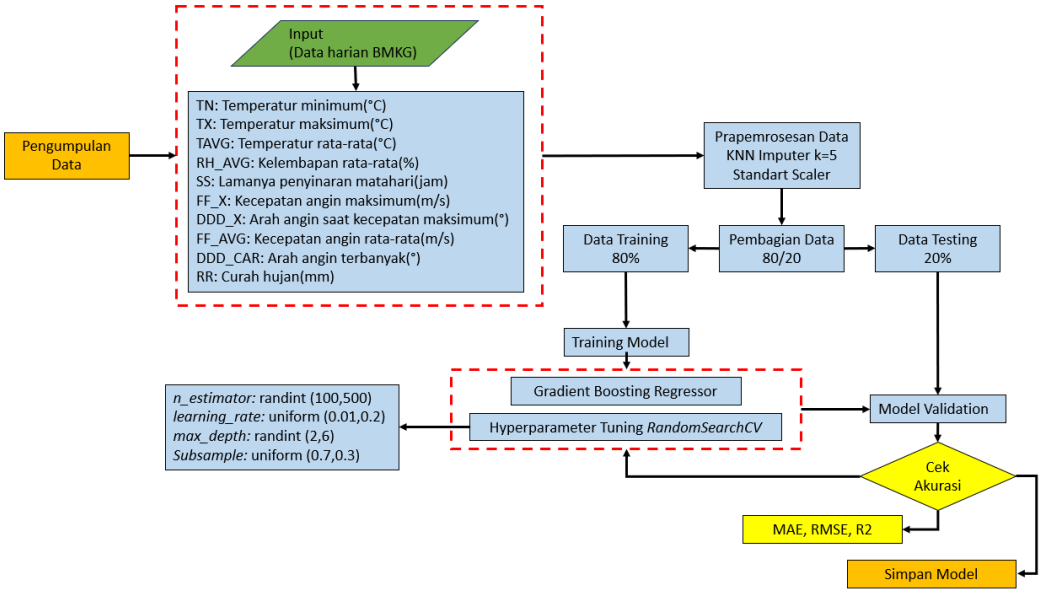
dilakukan oleh Usman dkk (2023) dengan data BMKG di Semarang periode 2018 sampai 2021 menghasilkan nilai RMSE 13,05 [4]. Sementara itu, Alvines dkk. (2025) yang menggunakan data BMKG di Sumatra Selatan tahun 2024 melaporkan RMSE sebesar 12,54 [5].

Perbedaan nilai RMSE pada studi-studi tersebut mengindikasikan bahwa performa model regresi sangat dipengaruhi oleh karakteristik topografi, resolusi spasial, dan volatilitas data meteorologi lokal . Kesenjangan utama yang diidentifikasi dalam literatur adalah keterbatasan model dalam memprediksi curah hujan dan kurangnya optimasi parameter yang spesifik untuk menangani variabilitas musiman pada data cuaca harian yang bersifat *zero-inflated*.

Penelitian ini mengusulkan penerapan algoritma *Gradient Boosting Regressor* untuk memprediksi curah hujan harian. Keunggulan algoritma ini terletak pada kemampuannya melakukan koreksi galat secara iteratif melalui pembangunan pohon keputusan sekuensial. Untuk mengatasi risiko *overfitting* dan meningkatkan generalisasi model, penelitian ini menerapkan skenario optimasi *RandomizedSearchCV* yang diintegrasikan dengan validasi silang deret waktu (*TimeSeriesSplit*) [6]. Melalui integrasi prapemrosesan data yang sistematis yang meliputi imputasi nilai hilang menggunakan *K-Nearest Neighbors* dan standardisasi fitur [7]. Penelitian ini bertujuan untuk mengevaluasi efektivitas optimasi hiperparameter dalam meningkatkan stabilitas prediksi dibandingkan model *baseline*. Pendekatan ini diharapkan dapat memberikan alternatif perangkat pendukung keputusan yang untuk memetakan pola curah hujan dengan karakteristik lokal yang kompleks. Penelitian memberikan kontribusi nyata dalam tiga aspek. Pertama, menyediakan prosedur optimasi hiperparameter yang sistematis untuk meningkatkan akurasi model *Gradient Boosting*. Kedua, mendemonstrasikan metode penanganan *missing value* berbasis KNN yang efektif untuk menjaga integritas data deret waktu BMKG. Dan ketiga, memberikan landasan empiris mengenai efektivitas penggunaan fitur historis cuaca untuk membangun sistem pendukung keputusan yang lebih responsif. Diharapkan, hasil riset ini menjadi langkah awal dalam pengembangan sistem peringatan dini yang lebih komprehensif bagi wilayah dengan karakteristik cuaca yang kompleks

2. METODOLOGI PENELITIAN

Penelitian ini mengikuti alur kerja sistematis untuk membangun model prediksi curah hujan yang mampu menangkap pola tren menggunakan algoritma *Gradient Boosting Regressor*. Tahapan penelitian terdiri dari pengumpulan data, prapemrosesan data, optimasi, pemodelan dan evaluasi. Berikut ini adalah desain sistem yang akan dilakukan untuk penelitian :



Gambar 1. Desain Sistem

Pada Gambar 1 telihat desain sistem untuk penelitian yang akan dilakukan dimulai dari tahap pengumpulan data yaitu data collection berupa data cuaca harian yang diperoleh dari BMKG, yang terdiri dari variabel input temperatur minimun, temperatur maksimum, temperatur rata-rata, kelembapan rata-rata, lamanya penyinaran matahari, kecepatan angin maksimum, arah angin saat kecepatan maksimum, kecepatan angin rata-rata, dan arah angin terbanyak serta variabel output berupa curah hujan harian. Selanjutnya dilakukan tahap prapemrosesan data yaitu data pre-processing untuk meningkatkan kualitas data melalui konversi tipe data, dan penanganan missing value. Data yang telah bersih kemudian dibagi menjadi data latih dan data uji dengan perbandingan 80/20 pemilihan nilai ini berdasarkan pada penelitian sebelumnya yang berhasil dapat meningkatkan akurasi[8]. Tahap berikutnya adalah pemodelan menggunakan algoritma Gradient Boosting Regressor dalam bidang Machine Learning untuk membangun model prediksi curah hujan. Untuk memperoleh performa optimal, dilakukan optimasi hyperparameter menggunakan metode Random Search Cross-Validation. Terakhir, model dievaluasi menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

(RMSE), dan R-Squared (R^2) untuk mengukur tingkat akurasi dan kemampuan model dalam memprediksi curah hujan secara tepat[9].

2.1 Pengumpulan Data

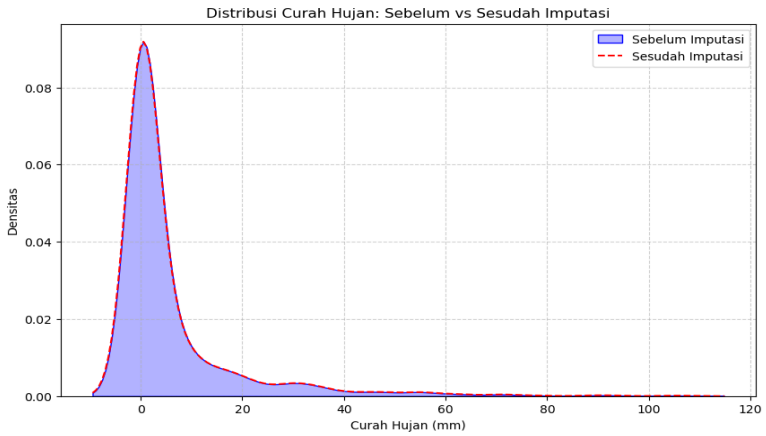
Penelitian ini menggunakan data cuaca harian yang diperoleh dari hasil pengamatan stasiun klimatologi BMKG Kabupaten Malang selama periode 1 Januari 2023 hingga 30 Juli 2025. Data tersebut mencakup berbagai parameter meteorologi yang memengaruhi kondisi cuaca harian, terutama curah hujan. Terdapat 942 baris dan 11 kolom fitur yang setiap baris data mewakili satu hari pengamatan dengan sebelas atribut utama, yaitu Tanggal, Temperatur Minimum (TN), Temperatur Maksimum (TX), Temperatur Rata-rata (TAVG), Kelembapan Rata-rata (RH_AVG), Curah Hujan (RR), Lamanya Penyinaran Matahari (SS), Kecepatan Angin Maksimum (FF_X), Kecepatan Angin Rata-rata (FF_AVG), Arah Angin saat Kecepatan Maksimum (DDD_X), Arah Angin Terbanyak (DDD_CAR). Variabel FF_X (kecepatan angin maksimum) dan DDD_X (arah angin) digunakan untuk menangkap dinamika atmosferik lokal yang mempengaruhi potensi curah hujan harian. Tabel 1 dibawah ini adalah fitur dan keterangan dataset yang digunakan :

Tabel 1. Dataset

Fitur	Keterangan
TANGGAL	Tanggal pada saat nilai diukur
TN	Temperatur minimum (°C)
TX	Temperatur maksimum (°C)
TAVG	Temperatur rata-rata (°C)
RH_AVG	Kelembapan rata-rata (%)
RR	Curah hujan (mm)
SS	Lamanya penyinaran matahari (jam)
FF_X	Kecepatan angin maksimum (m/s)
DDD_X	Arah angin saat kecepatan maksimum (°)
FF_AVG	Kecepatan angin rata-rata (m/s)
DDD_CAR	Arah angin terbanyak (°)

2.2 Prapemrosesan Data

Data awal dikumpulkan dalam format CSV (Comma Separated Values) dan melalui beberapa tahap prapemrosesan untuk menjamin kualitas data[10]. Tahapan tersebut meliputi pembersihan data yaitu data cleaning untuk mengatasi nilai hilang yaitu missing value dan data tidak valid, seperti kode 8888 atau 9999 yang menandakan data tidak tercatat. Selanjutnya dilakukan konversi tipe data, khususnya pada kolom numerik yang masih bertipe objek akibat perbedaan format desimal seperti tanda koma. Kolom tanggal juga diubah menjadi format datetime agar dapat digunakan dalam analisis berbasis waktu. Setelah itu, dilakukan standarisasi satuan pengukuran menggunakan *StandardScaler* untuk menyamakan skala fitur agar model memiliki bobot yang adil bagi setiap variabel serta imputasi nilai hilang menggunakan *K-Nearest Neighbors (KNN) Imputer* dengan nilai k=5 untuk mengisi data yang hilang berdasarkan kemiripan pola fitur cuaca agar distribusi data tetap representatif[11]. Nilai k=5 dipilih karena menurut penelitain sebelumnya dari Alvines (2025) yang melaporkan bahwa nilai k=5 ini nilai yang ideal dalam imputasi nilai pada data cuaca harian [5]. Hasil akhir dari proses prapemrosesan ini adalah dataset yang bersih, konsisten, dan siap digunakan untuk analisis eksploratif maupun pembangunan model prediksi curah hujan berbasis data cuaca harian.



Gambar 2. Sebelum dan sesudah penanganan missing value

Pada Gambar 2 menampilkan grafik hasil penanganan nilai yang hilang dengan imputasi. Grafik menunjukkan bahwa proses imputasi missing value tidak mengubah distribusi utama curah hujan secara signifikan yang ditunjukan dengan garis merah putus-putus menempel degan garis biru. Hal ini mengindikasikan bahwa metode imputasi yang

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

digunakan mampu mempertahankan karakteristik statistik data sehingga dataset tetap valid untuk proses pemodelan prediksi.

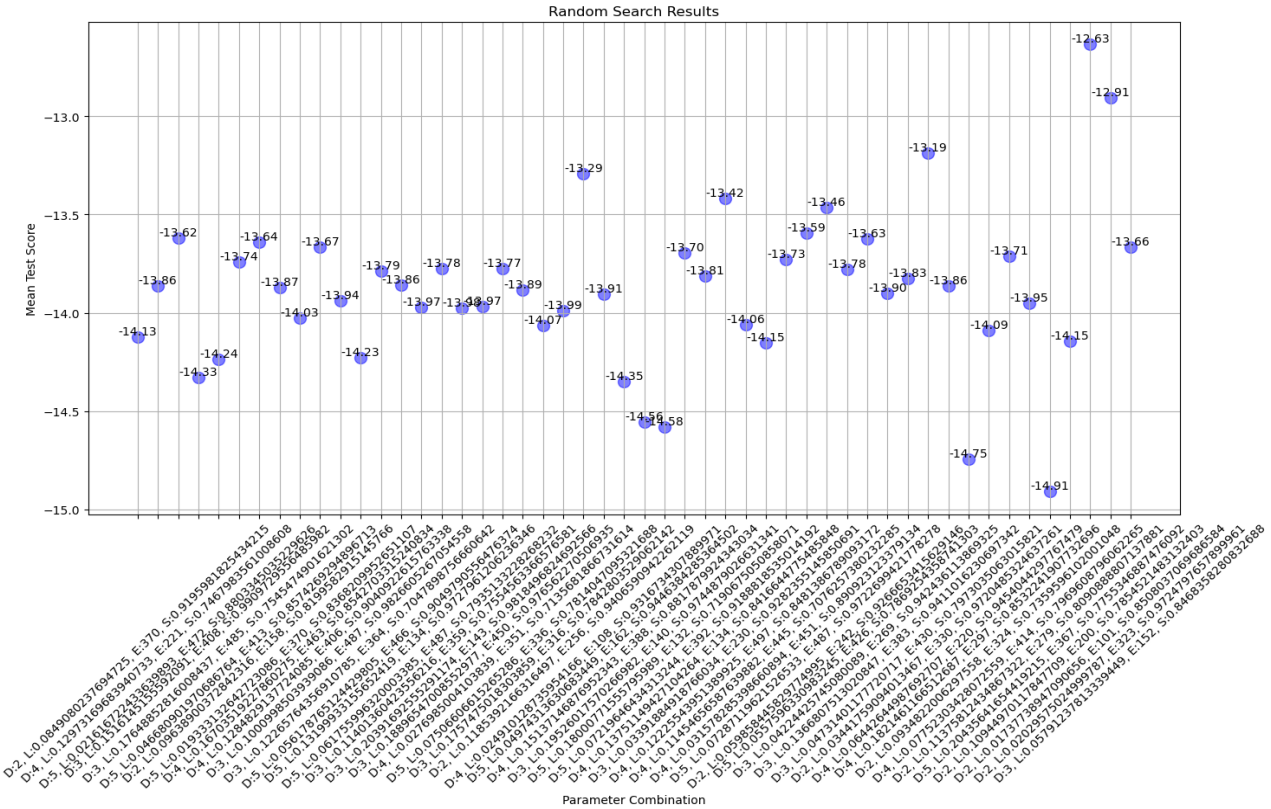
2.3 Optimasi

Setelah melalui tahap prapemrosesan, dilakukan proses optimasi hyperparameter untuk mendapatkan konfigurasi model Gradient Boosting Regressor yang paling optimal dalam memprediksi curah hujan[12]. Proses ini dilakukan menggunakan metode RandomizedSearchCV. Pencarian dilakukan dengan menguji 50 kombinasi parameter secara acak dari ruang pencarian yang telah ditentukan. Evaluasi setiap kombinasi dilakukan berdasarkan nilai Root Mean Squared Error RMSE terkecil. Hal ini memastikan bahwa setiap iterasi pengujian tetap menghormati urutan kronologis data, sehingga model yang dihasilkan memiliki kemampuan generalisasi yang baik terhadap data masa depan [13]. Berdasarkan hasil eksperimen, diperoleh kombinasi parameter terbaik yang memberikan nilai error minimum. Parameter hasil optimasi disajikan pada Tabel 2 berikut :

Tabel 2. Parameter optimal

No.	Hyperparameter	Deskripsi	Nilai Terpilih
1.	n_estimator	Jumlah pohon keputusan yang dibangun	101
2.	learning_rate	Kecepatan model dalam mengoreksi residu	0,017
3.	max_depth	Kedalaman maksimum setiap pohon	2
4.	subample	Fraksi data yang digunakan untuk tiap pohon	0,850
5.	**RMSE score**	**Rata-rata kesalahan prediksi**	**12,63**

Berikut pada Gambar 3 merupakan Hasil RandomsearchCV.



Gambar 3. Hasil RandomsearchCV

2.4 Pemodelan

Tahap pemodelan bertujuan untuk memetakan hubungan non-linear antara fitur meteorologi dengan curah hujan harian. Pemilihan *Gradient Boosting Regressor* didasari oleh karakteristik data curah hujan harian yang bersifat *zero-inflated* di mana nilai nol yaitu hari tidak hujan muncul dengan frekuensi sangat tinggi[14]. Algoritma ini menangani karakteristik tersebut secara implisit melalui struktur *decision tree* yang berurutan[15]. Pada setiap iterasi, algoritma akan mempartisi ruang fitur secara langkah per langkah dengan mencari poin pembagian optimal yang meminimalkan fungsi kerugian[16]. Dalam konteks data *zero-inflated*, *decision tree* pada model ini mampu membentuk percabangan yang secara efektif memisahkan observasi bernilai nol dari observasi bernilai positif. Dengan terus membagi ruang data ke dalam sub-wilayah yang lebih homogen, model mampu mengisolasi subset data bercurah hujan nol dan mengestimasi nilai residual yang mendekati nol pada subset tersebut[17]. Kemampuan iteratif ini memungkinkan *Gradient Boosting* untuk secara adaptif

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

membangun aturan keputusan yang membedakan kondisi hari tidak hujan dan hari hujan tanpa memerlukan teknik *oversampling* atau *undersampling* eksternal yang kompleks[18]. Algoritma *Gradient Boosting Regressor* dipilih juga karena kemampuannya dalam meminimalisir *residual error* secara iteratif melalui pembangunan *ensemble* pohon keputusan[19]. Model dibangun menggunakan pendekatan *Gradient Boosting* yang mengoptimalkan fungsi kerugian (*loss function*) berupa *Least Squares*. Setiap pohon baru dibangun untuk memprediksi residu dari pohon sebelumnya, sehingga model secara progresif memperbaiki akurasi prediksinya[13]. Persamaan umum dari proses *boosting* yang diterapkan dalam penelitian ini adalah:

$$F_m(x) = F_{m-1}(x) + \eta . \gamma_m h_m(x) \quad (1)$$

Dalam persamaan tersebut, $F_m(x)$ merepresentasikan model prediksi pada iterasi ke-m, di mana m menunjukkan indeks iterasi yang sedang berlangsung. Selanjutnya, γ_m melambangkan bobot optimal yang dihitung untuk setiap iterasi, sedangkan $h_m(x)$ merupakan fungsi yang dirancang untuk meminimalisir residu prediksi. Terakhir, η didefinisikan sebagai learning rate yang berfungsi untuk mengontrol laju pembaruan model guna mencapai konvergensi yang optimal.

Pemodelan dilakukan dengan menggunakan dua skenario, skenario pertama adalah menggunakan *gradient boosting regresor* yang default atau baseline dengan model dijalankan menggunakan konfigurasi parameter standar dari pustaka scikit-learn tanpa modifikasi tambahan. Skenario kedua adalah menggunakan *gradient boosting regressor* optimized yang sudah di tuning dengan *RandomSearchCV* yang mana model dijalankan menggunakan kombinasi hyperparameter terbaik yang ditemukan melalui tuning dan ruang pencarian mencakup parameter *n_estimators*, *learning_rate*, *max_depth*, dan *subsample*.

Prosedur pengujian dilakukan dengan menguji model pada data uji (*testing data*) yang telah disisihkan sejak awal (20% dari total dataset). Hal ini dilakukan untuk menghindari *data leakage* dan memastikan bahwa model memiliki kemampuan generalisasi yang baik pada data yang belum pernah diproses selama tahap pelatihan. Model dengan nilai RMSE terkecil dan R2 positif tertinggi pada data uji ditetapkan sebagai model terbaik.

2.5 Evaluasi

Pada penelitian ini evaluasi yang dilakukan menggunakan MAE untuk mengukur rata-rata selisih absolut antara nilai aktual dan prediksi yang mana jika semakin kecil nilai MAE maka model semakin baik, RMSE untuk mengukur rata-rata eror yang kembali ke satuan asli data yang lebih sensitif terhadap outlier dibanding MAE, dan juga R^2 untuk mengukur seberapa baik model menjelaskan variansi data diaman jika hasilnya mendekati nilai 1 maka model sangat baik dan jika haslnya mendekati nilai 0 maka model sangat buruk[20]. Untuk perbandingan skenario dilakukan perbandingan pada pembagian data training dan juga data testing tanpa adanya kandidat algoritma model lain. Sehingga fokus kepada model gradient boosting yang dikembangkan. Berikut ini adalah rumus persamaan dari MAE, RMSE dan juga R^2 :

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (4)$$

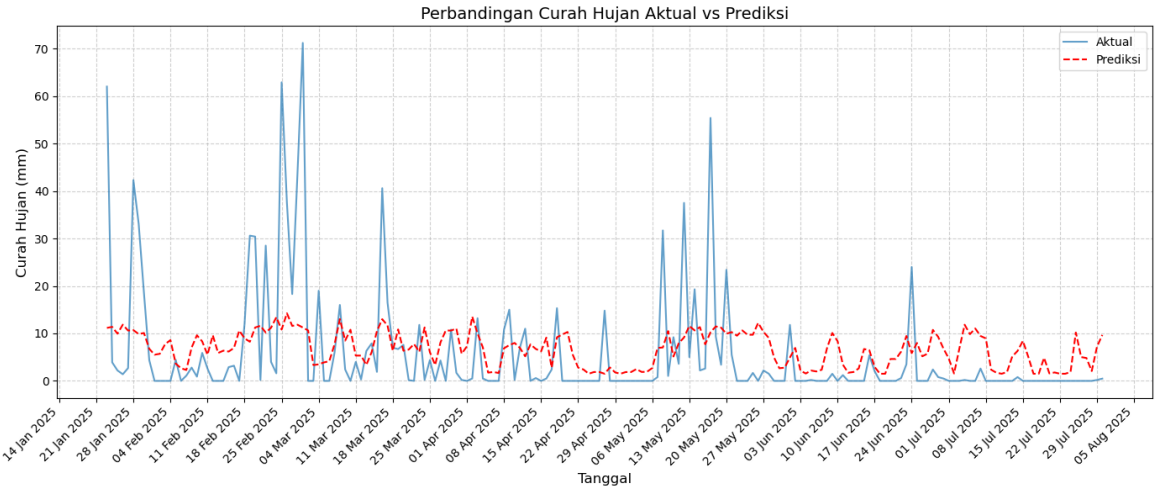
Variabel y_i merepresentasikan nilai aktual curah hujan hasil observasi, sedangkan $\hat{y}_i$ merupakan nilai curah hujan hasil prediksi model. Selanjutnya, $\bar{y}_i$ melambangkan rata-rata dari seluruh nilai aktual dalam dataset, dan n menyatakan total jumlah data observasi yang digunakan dalam proses evaluasi model.

3. HASIL DAN PEMBAHASAN

3.1 Skenario 1 Baseline

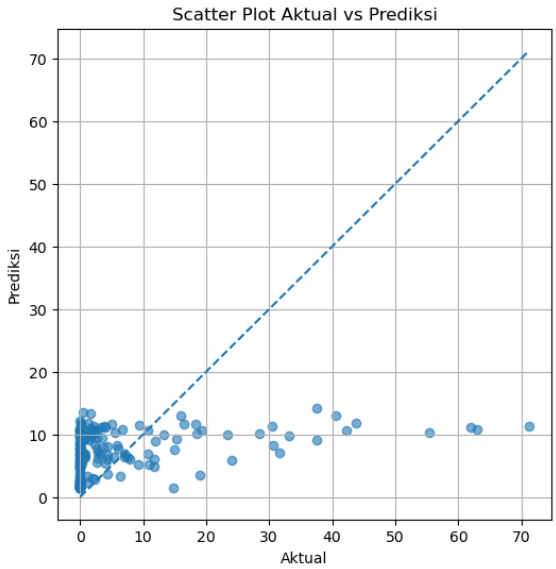
Model *baseline* yang dibangun dengan konfigurasi parameter standar *scikit-learn* menghasilkan performa yang mencerminkan kapabilitas dasar algoritma *Gradient Boosting* sebelum adanya penyesuaian spesifik terhadap karakteristik data meteorologi. Hasil model baseline untuk RMSE sebesar 13,81 dan hasil MAE sebesar 8,51 sedangkan hasil R2 -0,03. Meskipun model ini mampu memberikan tren prediksi secara umum, nilai *RMSE* dan *MAE* yang relatif tinggi menunjukkan bahwa model kurang sensitif dalam menangkap volatilitas data curah hujan yang bersifat stokastik. Kurangnya pengaturan pada *hyperparameter* seperti *learning rate* dan *depth* menyebabkan model cenderung menangkap pola makro namun sering melewatkan detail perubahan cuaca harian yang dinamis, sehingga menghasilkan akurasi yang moderat dan *R-squared* yang belum optimal dalam merepresentasikan variabilitas data secara keseluruhan. Untuk lebih jelasnya pada Gambar 4 berikut ini adalah grafik perbandingan antara data aktual dan prediksi:

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119–127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 4. Perbandingan Curah Hujan Aktual dan Prediksi Baseline

Pada Gambar 5 *scatterplot* model *baseline*, titik-titik sebaran data terlihat lebih menyebar secara lebar menjauhi garis diagonal ideal y = x. Hal ini mengindikasikan bahwa model memiliki tingkat variabilitas kesalahan yang tinggi, di mana prediksi sering kali meleset cukup jauh dari nilai curah hujan aktual. Terlihat adanya konsentrasi titik yang menumpuk di dekat sumbu horizontal, yang menandakan model kesulitan menangkap intensitas hujan yang lebih tinggi dan cenderung memberikan estimasi yang "tumpul" atau kurang responsif. Sebaran yang tidak simetris ini membuktikan bahwa konfigurasi parameter *default* belum mampu memetakan hubungan non-linear antar variabel cuaca dengan presisi, sehingga model masih sangat rentan terhadap residu yang besar.

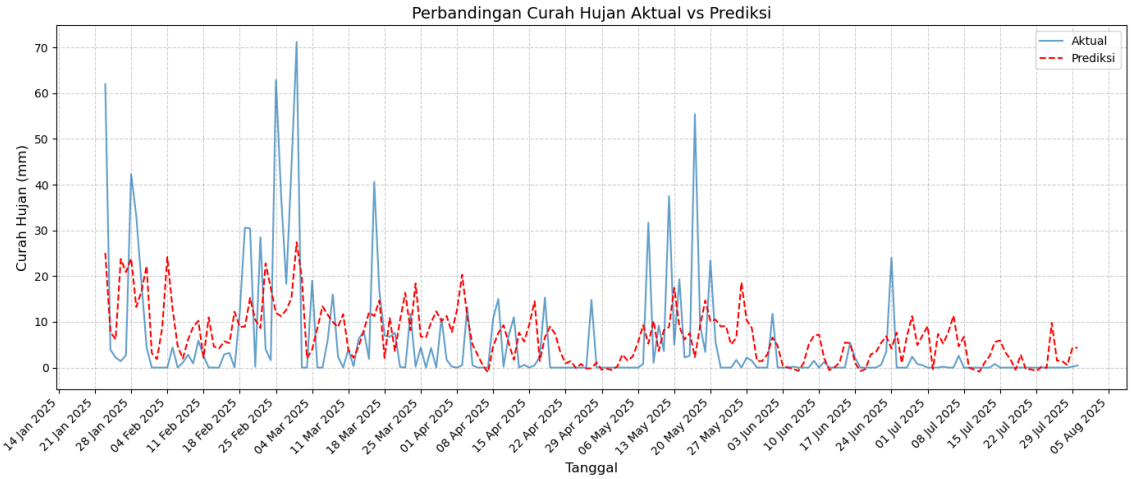


Gambar 5. Scater Plot Aktual dan Prediksi Baseline

3.2 Skenario 2 Optimmized

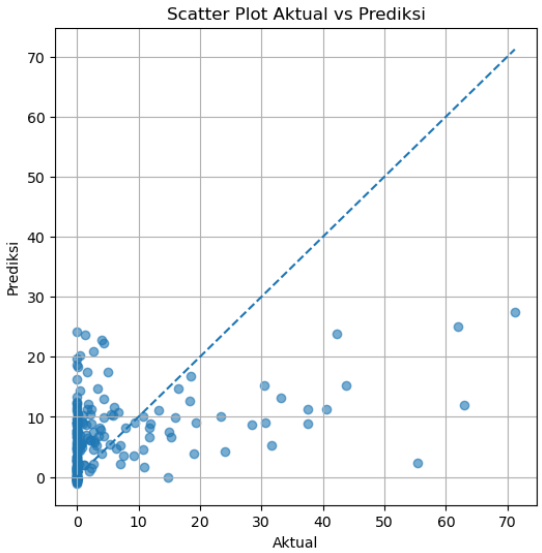
Skenario *optimized* menunjukkan peningkatan performa yang signifikan setelah penerapan teknik pencarian parameter melalui *RandomizedSearchCV* yang terintegrasi dengan validasi silang deret waktu. Hasil model Optimized untuk RMSE sebesar 12,21 dan hasil MAE sebesar 7,70 sedangkan hasil R2 0,20. Dengan penyesuaian yang presisi pada *n_estimators*, *max_depth*, dan *subsample*, model mampu melakukan koreksi kesalahan prediksi secara jauh lebih iteratif dan stabil, yang terbukti dari penurunan nilai *RMSE* yang signifikan dibandingkan model *baseline*. Optimasi ini secara efektif menyeimbangkan kompleksitas model guna meminimalisir *error* tanpa terjebak dalam *overfitting*, sehingga model tidak hanya lebih mampu menangkap pola tren dalam mengestimasi intensitas hujan harian, tetapi juga memiliki kemampuan generalisasi yang lebih tangguh saat dihadapkan pada pola cuaca yang bervariasi selama periode uji. Untuk lebih jelasnya pada Gambar 6 berikut ini adalah grafik perbandingan antara data aktual dan prediksi:

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 6. Perbandingan Curah Hujan Aktual Dan Prediksi Optimized

Pada Gambar 7 *scatterplot* pada model *optimized* menunjukkan peningkatan kepadatan titik yang jauh lebih rapat di sepanjang garis diagonal ideal, yang menjadi indikator kuat peningkatan akurasi prediksi. Sebaran data tampak lebih terpusat dan memiliki *spread* (jarak sebar) yang lebih sempit dibandingkan skenario *baseline*, yang berarti penyimpangan residu telah berhasil ditekan melalui optimasi *hyperparameter*. Meskipun pada bagian curah hujan ekstrem (titik-titik di area kanan atas) masih terdapat sebaran yang melebar, model *optimized* secara keseluruhan berhasil mengurangi jumlah *outlier* prediksi secara drastis. Kerapatan titik yang meningkat di area tengah ini membuktikan bahwa optimasi telah membantu model dalam "belajar" membedakan kondisi cuaca dengan lebih baik, menghasilkan estimasi yang jauh lebih konsisten dan mendekati angka observasi BMKG yang sesungguhnya.

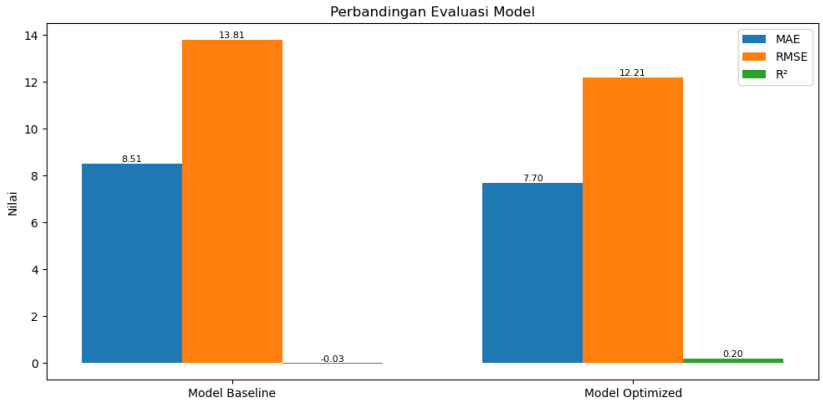


Gambar 7. Scater Plot Aktual dan Prediksi Optimized

3.3 Evaluasi

Hasil evaluasi model dalam penelitian ini memberikan gambaran komprehensif mengenai efektivitas algoritma Gradient Boosting Regressor dalam memetakan kompleksitas curah hujan harian. Secara kuantitatif, metrik RMSE dan MAE berfungsi sebagai indikator utama besaran kesalahan prediksi, sementara R2 Score merepresentasikan sejauh mana model mampu menjelaskan variabilitas data cuaca yang bersifat stokastik. Penurunan nilai error yang signifikan dari skenario baseline ke skenario optimized membuktikan bahwa proses penyetelan hyperparameter melalui RandomizedSearchCV bukan sekadar teknis komputasi, melainkan langkah krusial untuk meningkatkan ketajaman prediksi model terhadap pola meteorologi lokal. Gambar 8 berikut ini adalah perbandingan evaluasi model baseline dengan model optimasi:

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119–127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds



Gambar 8. Perbandingan Evaluasi Model

Dalam penelitian ini, optimasi *Gradient Boosting Regressor* melalui *RandomizedSearchCV* berhasil meningkatkan performa model secara terukur, dengan penurunan nilai RMSE dari 13,81 menjadi 12,21 dan peningkatan skor R2 dari -0,03 menjadi 0,20. Meskipun peningkatan ini menunjukkan responsivitas model yang lebih baik dalam menangkap tren dibandingkan skenario *baseline*, nilai R2 sebesar 0,20 secara objektif mengindikasikan bahwa model hanya mampu menjelaskan 20% variabilitas curah hujan harian. Kami mengakui bahwa angka ini jauh dari kategori ideal untuk sebuah instrumen prediktif dalam pengambilan keputusan yang bersifat krusial dan memiliki risiko tinggi, seperti mitigasi bencana banjir atau tanah longsor. Terdapat batasan fundamental yang mendasari hasil tersebut yaitu curah hujan di wilayah tropis seperti Kabupaten Malang memiliki karakteristik stokastik yang sangat tinggi dan dipengaruhi oleh fenomena atmosferik skala besar yang tidak tertangkap oleh variabel meteorologi permukaan yang digunakan dalam penelitian ini. Penting untuk ditegaskan bahwa penelitian ini tidak mengklaim model yang dikembangkan sebagai solusi final atau instrumen yang sudah sempurna. Sebaliknya, posisi penelitian ini adalah sebagai langkah awal dalam pengembangan sistem pendukung keputusan yang bersifat iteratif. Hasil ini berfungsi sebagai bukti empiris bahwa pendekatan *machine learning* yang dioptimasi secara terstruktur mampu memberikan pola dasar yang lebih stabil dibandingkan estimasi manual atau model tanpa optimasi.

3.4 Pembahasan

Hasil eksperimen menunjukkan bahwa optimasi Gradient Boosting Regressor memberikan peningkatan performa yang terukur dibandingkan model baseline. Namun, nilai R2 sebesar 0,20 mengindikasikan bahwa model saat ini belum mampu menangkap seluruh kompleksitas dinamika curah hujan harian secara utuh. Algoritma *Gradient Boosting Regressor* dengan fungsi kerugian *Least Squares* standar cenderung memberikan *penalty* yang lebih besar pada kesalahan prediksi di hari-hari kering, sehingga model secara tidak langsung mengabaikan fluktuasi intensitas hujan yang tinggi untuk meminimalkan *error* total. Hal ini menegaskan bahwa penggunaan variabel meteorologi permukaan saja belum mencukupi untuk menjelaskan transisi stokastik dari kondisi kering ke hujan ekstrem. Hal ini merefleksikan batasan fundamental dari penggunaan variabel meteorologi permukaan saja, tanpa menyertakan dinamika atmosferik skala besar yang menjadi karakteristik khas wilayah tropis seperti Kabupaten Malang. Secara operasional, model ini tidak diklaim sebagai sistem prediksi mutlak, melainkan sebagai instrumen pendukung keputusan awal yang bersifat iteratif. Temuan ini memberikan kontribusi pada pemahaman mengenai batasan algoritma tree-based dalam menangani data zero-inflated yang sering ditemui pada pencatatan BMKG. Oleh karena itu, langkah strategis ke depan perlu difokuskan pada pengayaan variabel prediktor dan penerapan fungsi kerugian yang lebih robust terhadap outlier. Dengan demikian, model yang diusulkan saat ini dipandang sebagai landasan teknis yang perlu terus dievaluasi dan disempurnakan sebelum diintegrasikan lebih dalam ke dalam protokol mitigasi bencana yang lebih luas.

4. KESIMPULAN

Penelitian ini mengevaluasi efektivitas algoritma Gradient Boosting Regressor dengan optimasi RandomizedSearchCV untuk memprediksi curah hujan harian di Kabupaten Malang. Hasil eksperimen menunjukkan bahwa optimasi hyperparameter *learning_rate*, *max_depth*, *n_estimators*, dan *subsample* memberikan peningkatan performa yang moderat dibandingkan model baseline. Secara kuantitatif, model teroptimasi berhasil menekan nilai RMSE dari 13,81 menjadi 12,21 dan meningkatkan skor R2 dari -0,03 menjadi 0,20. Peningkatan ini mengonfirmasi bahwa penyesuaian parameter secara sistematis membantu algoritma dalam mempelajari pola non-linear curah hujan serta memitigasi overfitting pada data deret waktu. Meskipun menunjukkan peningkatan performa, nilai R2 sebesar 0,20 mengindikasikan bahwa model saat ini belum mampu menjelaskan 80% variabilitas data curah hujan yang bersifat stokastik. Evaluasi lebih lanjut menunjukkan bahwa model masih memiliki keterbatasan dalam memprediksi intensitas curah hujan ekstrem, yang sebagian besar disebabkan oleh distribusi data yang tidak seimbang dan keterbatasan fungsi kerugian *Least Squares* yang cenderung sensitif terhadap outlier cuaca. Dengan demikian, model ini belum direkomendasikan sebagai instrumen tunggal untuk pengambilan keputusan mitigasi bencana yang krusial pada tahap saat ini. Hasil penelitian ini memberikan

ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 119−127
DOI: 10.47065/bulletinds.v5i3.10071
https://ejurnal.seminar-id.com/index.php/bulletinds

landasan empiris bahwa efisiensi model *machine learning* pada data meteorologi tidak hanya bergantung pada algoritma, tetapi juga pada kecermatan dalam mengatur parameter yang disesuaikan dengan karakteristik data lokal. Untuk pengembangan di masa depan, disarankan melakukan eksplorasi pada fungsi kerugian yang lebih robust seperti *Huber Loss* atau *Quantile Regression*, penerapan pendekatan hybrid untuk menangani data zero-inflated, serta penambahan variabel prediktor berbasis remote sensing guna memperkaya informasi atmosferik yang belum tertangkap oleh observasi stasiun cuaca saat ini. Penelitian ini memberikan kontribusi penting dalam mengidentifikasi batasan model pada data *zero-inflated* serta menyediakan *roadmap* bagi pengembangan model meteorologi di masa depan. Model yang diusulkan tidak diposisikan sebagai solusi mutlak, melainkan sebagai instrumen pendukung keputusan iteratif yang perlu disempurnakan lebih lanjut dengan penambahan variabel atmosferik guna mendukung ketahanan masyarakat terhadap bencana hidrometeorologi.

REFERENCES

[1] A. A. Sinaga, D. Silvia, D. Arianti, J. A. Sinaga, and O. S. R. Purba, "Curah Hujan Di Indonesia," *J. Intelek dan Cendikiawan Nusant.*, vol. 2, no. 6, pp. 11497–11504, 2025.
[2] R. Meenal, K. Kailash, P. A. Michael, J. J. Joseph, F. T. Josh, and E. Rajasekaran, "Machine learning based smart weather prediction," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 508–515, 2022, doi: 10.11591/ijeecs.v28.i1.pp508-515.
[3] S. Markuna, P. Kumar, R. Ali, D. I. K. U. V Ishwkarma, K. U. S. I. K. Ushwaha, and V. I. K. U. S. Ingh, "Application of Innovative Machine Learning Techniques for Long-Term Rainfall Prediction," *Pure Appl. Geophys.*, vol. 180, p. 3189, 2023.
[4] C. D. Usman, A. P. Widodo, K. Adi, and R. Gernowo, "Rainfall prediction model in Semarang City using machine learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, pp. 1224–1231, 2023, doi: 10.11591/ijeecs.v30.i2.pp1224-1231.
[5] M. Alvines *et al.*, "Komparasi Ridge Regression, Random Forest, Dan Gradient Boosting Untuk Prediksi Curah Hujan Harian Di Sumatra Selatan Berbasis Time Series Cross-Validation," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 4, pp. 5742–5748, 2025, doi: 10.36040/jati.v9i4.13915.
[6] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.
[7] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Mach. Learn. with Appl.*, vol. 7, no. August 2021, p. 100204, 2022, doi: 10.1016/j.mlwa.2021.100204.
[8] M. S. Pathan, P. Nadella, and Y. U. L. Haq, "A Systematic Analysis of Meteorological Parameters in Predicting Rainfall Events," *IEEE Access*, vol. 13, no. July, pp. 111529–111541, 2025, doi: 10.1109/ACCESS.2025.3573091.
[9] M. T. Anwar, E. Winarno, W. Hadikurniawati, and M. Novita, "Rainfall prediction using Extreme Gradient Boosting," *J. Phys. Conf. Ser.*, vol. 1869, no. 1, 2021, doi: 10.1088/1742-6596/1869/1/012078.
[10] Hendra Di Kesuma, D. Apriadi, H. Juliansa, and E. Etriyanti, "Implementasi Data Mining Prediksi Mahasiswa Baru Menggunakan Algoritma Regresi Linear Berganda," *J. Ilm. Bin. STMIK Bina Nusant. Jaya Lubuklinggau*, vol. 4, no. 2, pp. 62–66, 2022, doi: 10.52303/jb.v4i2.74.
[11] H. Li, Q. Guo, T. Zhang, S. Zhou, and C. Guo, "Interpretable Machine Learning for Predicting Anterior Uveitis in Axial Spondyloarthritis," *JCR J. Clin. Rheumatol.*, vol. 31, no. 5, 2025.
[12] E. M. Z. Darmawan and A. Fauzan Dianta, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM," *Teknol. J. Ilm. Sist. Inf.*, vol. 13, no. 1, pp. 8–15, 2023.
[13] M. S. Islam *et al.*, "Explainable deep learning for rainfall prediction: A CNN-XGBoost hybrid approach in the northern region of Bangladesh," *Neural Comput. Appl.*, vol. 37, no. 33, pp. 28125–28160, 2025, doi: 10.1007/s00521-025-11646-z.
[14] W. Sun *et al.*, "Improved Prediction of Extreme Rainfall Using a Machine Learning Approach," *Adv. Atmos. Sci.*, vol. 42, no. 8, pp. 1661–1674, 2025, doi: 10.1007/s00376-024-4269-5.
[15] W. Zhou *et al.*, "Machine learning enhanced framework for rainfall-induced debris flows spatio-temporal prediction: Integrating mobilized materials and hydrological processes," *J. Hydrol. Reg. Stud.*, vol. 63, no. July 2025, 2026, doi: 10.1016/j.ejrh.2025.103046.
[16] T. Talan, "Machine learning-based rainfall prediction across temporal scales: model benchmarking and explainability analysis," *Stoch. Environ. Res. Risk Assess.*, vol. 40, no. 5, pp. 1–17, 2026, doi: 10.1007/s00477-026-03245-8.
[17] H. Farman, M. A. Hussain, S. Hassan, S. Shaikh, and K. Ali, "Activation function impact on rainfall prediction: comparative insights across ML and DL architectures," *Model. Earth Syst. Environ.*, vol. 11, no. 6, 2025, doi: 10.1007/s40808-025-02630-6.
[18] S. Ghosh, M. K. Gourisaria, B. Sahoo, and H. Das, "A pragmatic ensemble learning approach for rainfall prediction," *Discov. Internet Things*, vol. 3, no. 1, 2023, doi: 10.1007/s43926-023-00044-3.
[19] Z. H. Haq Doost, A. Alsuwaiyan, A. Abdulraheem, N. M. Al-Areeq, and Z. M. Yaseen, "Rainfall Prediction Using Integrated Machine Learning Models With K-Means Clustering: A Representative Case Study of Harirud Murghab Basin-Afghanistan," *IEEE Access*, vol. 13, no. April, pp. 111628–111646, 2025, doi: 10.1109/ACCESS.2025.3581921.
[20] S. D. Latif *et al.*, "Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches," *Alexandria Eng. J.*, vol. 82, no. July, pp. 16–25, 2023, doi: 10.1016/j.aej.2023.09.060.