MODEL KLASIFIKASI RISIKO BANJIR DI WILAYAH PERKOTAAN BERBASIS XGBOOST BERDASARKAN KARAKTERISTIK DRAINASE DAN CURAH HUJAN SATELIT

TESIS

Oleh:
EVAN AFDRIANTO KOTA
NIM. 240605220004

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2026

MODEL KLASIFIKASI RISIKO BANJIR DI WILAYAH PERKOTAAN BERBASIS XGBOOST BERDASARKAN KARAKTERISTIK DRAINASE DAN CURAH HUJAN SATELIT

TESIS

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
EVAN AFDRIANTO KOTA
NIM. 240605220004

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

MODEL KLASIFIKASI RISIKO BANJIR DI WILAYAH PERKOTAAN BERBASIS XGBOOST BERDASARKAN KARAKTERISTIK DRAINASE DAN CURAH HUJAN SATELIT

TESIS

Oleh:
EVAN AFDRIANTO KOTA
NIM. 240605220004

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 11 Juni 2026

Pembimbing I,	Pembimbing II,
Dr. Ir. M. Amin Hariyadi, M.T.	Dr. Cahyo Crysdian, M.Cs.
NIP. 19670118 200501 1 001	NIP. 19740424 200901 1 008

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Muhammad Faisal, M.T.
NIP. 19740510 200501 1 007

MODEL KLASIFIKASI RISIKO BANJIR DI WILAYAH PERKOTAAN BERBASIS XGBOOST BERDASARKAN KARAKTERISTIK DRAINASE DAN CURAH HUJAN SATELIT

TESIS

Oleh:
EVAN AFDRIANTO KOTA
NIM. 240605220004

Telah Dipertahankan di Depan Dewan Penguji Tesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 11 Juni 2026

Susunan Dewan Penguji			**Tanda Tangan**
Penguji I	:	Dr. Agung Teguh Wibowo Almais, S.Kom. 19860301 2023321 1 016	( )
Penguji II	:	Dr. Ir. Fresy Nugroho, S.T., M.T., IPM., ASEAN Eng. 19710722 201101 1 001	( )
Pembimbing I	:	Dr. Ir. M. Amin Hariyadi, M.T. NIP. 19670118 200501 1 001	( )
Pembimbing II	:	Dr. Cahyo Crysdian, M.Cs. NIP. 19740424 200901 1 008	( )

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Prof. Dr. Muhammad Faisal, M.T.
NIP. 19740510 200501 1 007

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	:	Evan Afdrianto Kota
NIM	:	240605220004
Program Studi	:	Magister Informatika
Fakultas	:	Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Tesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila di kemudian hari terbukti atau dapat dibuktikan Tesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 11 Juni 2026

Yang membuat pernyataan,

Evan Afdrianto Kota

NIM. 240605220004

MOTTO

"Iman tanpa perbuatan adalah sia-sia, dan perbuatan tanpa ketekunan tidak akan membuahkan hasil."

HALAMAN PERSEMBAHAN

Tesis ini penulis persembahkan kepada:

1. Kedua orang tua dan keluarga tercinta yang senantiasa memberikan doa, kasih sayang, dan dukungan tanpa henti selama penulis menempuh pendidikan magister.
2. Segenap civitas akademika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah menjadi tempat penulis menimba ilmu, mengembangkan diri, dan menyelesaikan penelitian ini.
3. Universitas Merdeka Malang institusi tempat penulis mengabdi, yang telah memberikan kesempatan dan dukungan untuk melanjutkan pendidikan magister.

KATA PENGANTAR

Puji syukur penulis panjatkan atas terselesaikannya tesis ini sebagai salah satu syarat untuk memperoleh gelar Magister K di Universitas Islam Negeri Maulana Malik Ibrahim Malang. Penulis menyampaikan ucapan terima kasih kepada:

1. Bapak Dr. Ir. M. Amin Hariyadi, M.T. (Pembimbing I) dan Bapak Dr. Cahyo Crysdian, M.Cs. (Pembimbing II) yang telah memberikan arahan dan bimbingan selama proses penelitian dan penulisan tesis ini.
2. Bapak Dr. Agung Teguh Wibowo Almais, S.Kom., M.T. (Penguji I) dan Bapak Dr. Ir. Fresy Nugroho, S.T., M.T., IPM., ASEAN Eng. (Penguji II) yang telah memberikan kritik dan saran konstruktif demi penyempurnaan tesis ini.
3. Seluruh pimpinan, dosen, dan staf civitas akademika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang, khususnya Program Studi Magister Informatika, yang telah memfasilitasi proses perkuliahan dan penelitian.
4. Kedua orang tua dan keluarga penulis yang senantiasa memberikan doa, motivasi, dan dukungan selama menempuh pendidikan magister.
5. Rekan-rekan mahasiswa Program Studi Magister Informatika yang telah berbagi ilmu dan semangat selama masa perkuliahan.

Penulis menyadari bahwa tesis ini masih jauh dari sempurna. Kritik dan saran yang membangun sangat diharapkan untuk perbaikan di masa mendatang. Semoga tesis ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang klasifikasi risiko banjir dan pengelolaan infrastruktur drainase perkotaan.

Malang, Juni 2026

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Kota, Evan Afdrianto. 2026. **Model Klasifikasi Risiko Banjir di Wilayah Perkotaan Berbasis XGBoost Berdasarkan Karakteristik Drainase dan Curah Hujan Satelit**. Tesis. Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Ir. M. Amin Hariyadi, M.T., (II) Dr. Cahyo Crysdian, M.Cs.

Kata Kunci: Klasifikasi Risiko Banjir, XGBoost, CHIRPS, Drainase Perkotaan, Machine Learning, BorderlineSMOTE

Banjir perkotaan di Kota Malang terjadi akibat interaksi antara intensitas curah hujan dan kapasitas sistem drainase. Penelitian ini mengembangkan model klasifikasi risiko banjir tiga kelas (Rendah, Sedang, Tinggi) menggunakan algoritma XGBoost dengan 53 fitur yang terdiri dari 26 fitur drainase fisik, 19 fitur curah hujan satelit CHIRPS, dan 8 fitur interaksi. Studi ablasi dengan enam skenario (S1–S6) dilakukan untuk mengukur kontribusi seleksi fitur (MI + Pearson + RFE), BorderlineSMOTE, dan optimasi Optuna TPE. Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* (5 fold × 5 pengulangan = 25 evaluasi) dengan *preprocessing* diterapkan di dalam setiap fold pelatihan. Dataset mencakup 3.471 titik drainase dengan distribusi kelas tidak merata: Rendah 65,7%, Sedang 25,7%, dan Tinggi 8,7%. Hasil menunjukkan bahwa model *baseline* S1 mencapai *Accuracy* 0,8800 ± 0,0123, F1 Macro 0,8152 ± 0,0215, dan ROC-AUC 0,9689 ± 0,0048. S5 (Optuna) sedikit mengungguli *baseline* (+0,0019 *Accuracy*, +0,0041 F1 Macro), dan *Wilcoxon signed-rank test* mengonfirmasi bahwa peningkatan F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$) signifikan pada $\alpha = 0,05$, meskipun peningkatan *Accuracy* tidak signifikan ($p = 0,084$). Seleksi fitur mereduksi dimensi dari 53 menjadi 19 fitur dan mempertahankan 2 fitur curah hujan CHIRPS (*rainy_days* dan *rainfall_30d_max*) beserta 7 fitur interaksi, dengan penurunan F1 Macro CV sebesar 0,0174. BorderlineSMOTE meningkatkan *Recall* kelas Tinggi dari 0,5833 menjadi 0,8167–0,8833 pada *test set*, sehingga 49–53 dari 60 titik berisiko tinggi terdeteksi, dengan konsekuensi penurunan *Accuracy* CV. Optuna TPE memberikan perbaikan yang kecil secara absolut namun signifikan secara statistik pada F1 Macro dan ROC-AUC. Analisis SHAP menunjukkan bahwa fitur curah hujan (*rainy_days*, SHAP = 0,0615–0,0855) mendominasi lima dari enam skenario, termasuk skenario dengan seleksi fitur, yang membuktikan pentingnya mempertahankan fitur curah hujan dalam model. Fitur *kondisi_fisik_encoded* muncul sebagai prediktor yang konsisten di skenario S3. Pemilihan konfigurasi optimal bergantung pada prioritas aplikasi: S1 atau S5 untuk performa keseluruhan, atau S4/S6 untuk deteksi titik berisiko tinggi.

ABSTRACT

Kota, Evan Afdrianto. 2026. **An XGBoost-Based Flood Risk Classification Model in Urban Areas Based on Drainage Characteristics and Satellite Rainfall**. Master's Thesis. Master of Informatics Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang. Supervisors: (I) Dr. Ir. M. Amin Hariyadi, M.T., (II) Dr. Cahyo Crysdian, M.Cs.

Urban flooding in Malang City results from the interaction between rainfall intensity and drainage system capacity. This study develops a three-class flood risk classification model (Low, Medium, High) using the XGBoost algorithm with 53 features comprising 26 physical drainage features, 19 CHIRPS satellite rainfall features, and 8 interaction features. An ablation study with six scenarios (S1–S6) was conducted to measure the contribution of feature selection (MI + Pearson + RFE), BorderlineSMOTE, and Optuna TPE hyperparameter optimization. The primary evaluation employed Repeated 5×5 Stratified K-Fold Cross-Validation (5 folds × 5 repeats = 25 evaluations) with preprocessing applied within each training fold. The dataset covers 3,471 drainage points with imbalanced class distribution: Low 65.7%, Medium 25.7%, and High 8.7%. Results show that the baseline model S1 achieved an Accuracy of 0.8800 ± 0.0123, F1 Macro of 0.8152 ± 0.0215, and ROC-AUC of 0.9689 ± 0.0048. S5 (Optuna) slightly outperformed the baseline (+0.0019 Accuracy, +0.0041 F1 Macro), and a Wilcoxon signed-rank test confirmed that the F1 Macro improvement ($p = 0.008$) and ROC-AUC improvement ($p = 0.034$) were statistically significant at $\alpha = 0.05$, although the Accuracy improvement was not significant ($p = 0.084$). Feature selection reduced dimensionality from 53 to 19 features while retaining 2 CHIRPS rainfall features (rainy_days and rainfall_30d_max) along with 7 interaction features, with a CV F1 Macro decrease of 0.0174. BorderlineSMOTE improved High-class Recall from 0.5833 to 0.8167–0.8833 on the test set, detecting 49–53 out of 60 high-risk points, at the cost of lower CV Accuracy. Optuna TPE provided small absolute improvements that were nonetheless statistically significant for F1 Macro and ROC-AUC. SHAP analysis revealed that the rainfall feature rainy_days (SHAP = 0.0615–0.0855) dominated five out of six scenarios, including those with feature selection, demonstrating the importance of retaining rainfall features in the model. The kondisi_fisik_encoded feature appeared as a consistent predictor in scenario S3. Optimal configuration selection depends on the application priority: S1 or S5 for overall performance, or S4/S6 for maximizing high-risk point detection.

Keywords: Flood Risk Classification, XGBoost, CHIRPS, Urban Drainage, Machine Learning, BorderlineSMOTE

الملخص

كوتا، إيفان أفدريانتو .2026** .نموذج تصنيف مخاطر الفيضانات في المناطق الحضرية القائم على استنادًا إلى خصائص الصرف وهطول الأمطار من الأقمار الصناعية **.رسالة XGBoost خوارزمية الماجستير .برنامج ماجستير المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية ،د .كاهيو كريسديان (II) ،د .المهندس .م .أمين هاريادي، الماجستير (I) :الحكومية مالانج .المشرفان .الماجستير

تصنيف مخاطر الفيضانات: الكلمات المفتاحية، XGBoost، CHIRPS، التعلم الآلي، الصرف الحضري، BorderlineSMOTE

.تحدث الفيضانات الحضرية في مدينة مالانج نتيجة التفاعل بين شدة هطول الأمطار وقدرة نظام الصرف طورت هذه الدراسة نموذج تصنيف مخاطر الفيضانات بثلاث فئات)منخفض، متوسط، مرتفع (باستخدام مع 53 سمة تتكون من 26 سمة للصرف المادي، و19 سمة لهطول الأمطار من XGBoost خوارزمية (S1–S6) و8 سمات تفاعلية .أُجريت دراسة استئصالية بستة سيناريوهات ،CHIRPS الأقمار الصناعية وتحسين ،BorderlineSMOTEو ،(MI + Pearson + RFE) لقياس مساهمة اختيار السمات (Repeated 5×5) استُخدم التحقق المتقاطع الطبقي المتكرر 5×5 .Optuna TPE المعلمات الفائقة أي 5 طيات ×5 تكرارات =25 تقييمًا (كتقييم رئيسي مع ،Stratified K-Fold Cross-Validation تطبيق المعالجة المسبقة داخل كل طية تدريبية .تغطي مجموعة البيانات 3,471 نقطة صرف بتوزيع فئات غير متوازن :منخفض 65.7%، متوسط 25.7%، ومرتفع 8.7 .%أظهرت النتائج أن النموذج الأساسي ،قدره 0.8152 ±0.0215 F1 Macroقدرها 0.8800 ±0.0123، و (Accuracy) حقق دقة S1 بشكل طفيف على النموذج الأساسي (Optuna) S5 قدره 0.9689 ±0.0048 .تفوق ROC-AUCو أن تحسن Wilcoxon signed-rank وأكد اختبار ،(F1 Macro في الدقة، +0.0041 في 0.0019+) رغم ،$\alpha = 0.05$ كان ذا دلالة إحصائية عند ($p = 0.034$) ROC-AUCو ($p = 0.008$) F1 Macro أدى اختيار السمات إلى تقليص الأبعاد من 53 إلى 19 سمة .($p = 0.084$) أن تحسن الدقة لم يكن ذا دلالة و7 (rainfall_30d_maxو rainy_days) CHIRPS مع الاحتفاظ بسمتين من سمات هطول الأمطار استدعاء BorderlineSMOTE بمقدار 0.0174 .حسّن F1 Macro CV سمات تفاعلية، مع انخفاض من 0.5833 إلى 0.8167–0.8833 في مجموعة الاختبار، مما أدى إلى (Recall) الفئة المرتفعة Optuna TPE قدم .CV اكتشاف 49–53 من أصل 60 نقطة عالية المخاطر، مقابل انخفاض في دقة .ROC-AUCو F1 Macro تحسنات صغيرة من حيث القيمة المطلقة لكنها ذات دلالة إحصائية في هيمنت (SHAP = 0.0615–0.0855) rainy_days أن سمة هطول الأمطار SHAP كشف تحليل على خمسة من ستة سيناريوهات، بما في ذلك السيناريوهات التي تتضمن اختيار السمات، مما يثبت أهمية كمؤشر تنبؤي kondisi_fisik_encoded الاحتفاظ بسمات هطول الأمطار في النموذج .ظهرت سمة للأداء العام، أو S5 أو S1 :يعتمد اختيار التكوين الأمثل على أولوية التطبيق .S3 ثابت في السيناريو لتعظيم اكتشاف النقاط عالية المخاطر S4/S6

BAB I
PENDAHULUAN

1.1 Latar Belakang

Banjir merupakan salah satu bencana alam yang paling sering terjadi di wilayah perkotaan dengan dampak terhadap kehidupan masyarakat, infrastruktur, dan aktivitas ekonomi. Pertumbuhan populasi, urbanisasi yang tidak terkendali, dan alih fungsi lahan membuat kota semakin rentan terhadap genangan yang berulang setiap tahun (Qin *et al.*, 2025; Zhu *et al.*, 2024).

Sistem prediksi banjir konvensional memiliki keterbatasan. Pendekatan deterministik dengan model hidrologi berbasis persamaan matematis yang kaku kurang adaptif terhadap kondisi dinamis perkotaan. Jang *et al.* (2025) menegaskan bahwa model numerik tradisional, meskipun akurat, memerlukan waktu komputasi yang lama sehingga kurang praktis untuk sistem peringatan dini. Sebagian besar model prediksi yang ada hanya berfokus pada satu aspek saja, baik data meteorologi maupun karakteristik fisik infrastruktur secara terpisah, sehingga menghasilkan prediksi yang kurang akurat karena mengabaikan keterkaitan antarfaktor penyebab banjir (Bersabe & Jun, 2025).

Model klasifikasi risiko banjir merupakan pendekatan yang lebih aplikatif dibandingkan model prediksi kuantitatif. Pendekatan klasifikasi memungkinkan pengelompokan wilayah ke dalam kategori risiko banjir, misalnya Rendah, Sedang, dan Tinggi, berdasarkan probabilitas kejadian. Shrestha *et al.* (2025) menunjukkan bahwa pendekatan klasifikasi mampu menghasilkan pemetaan zona risiko banjir

yang secara langsung dapat digunakan untuk perencanaan mitigasi, dengan akurasi mencapai 97,92%. Informasi berbasis kategori ini lebih mudah dipahami dan dapat digunakan oleh pengambil kebijakan dalam perencanaan mitigasi bencana dan penentuan prioritas wilayah.

Dalam konteks *machine learning*, algoritma berbasis *ensemble learning* menghasilkan performa yang lebih baik dalam pemetaan kerentanan banjir dibandingkan model tunggal. Riazi *et al.* (2023) melaporkan bahwa model hybrid ensemble mampu mencapai AUC hingga 0,997. Zhu *et al.* (2024) mengonfirmasi temuan serupa; Ensemble Model Framework (EMF) yang mengombinasikan XGBoost, SVC, MLP, dan MDL melalui Random Forest menghasilkan akurasi 0,940 dan mengungguli semua model individual. Di antara berbagai keluarga algoritma ensemble, pendekatan *boosting* menunjukkan performa yang lebih baik dalam tugas klasifikasi kerentanan banjir. Qin *et al.* (2025) mencatat bahwa XGBoost mencapai *overall accuracy* sebesar 0,96 dalam klasifikasi banjir perkotaan, sementara Goumghar *et al.* (2025) menemukan bahwa Hist Gradient Boosting mengungguli XGBoost, CatBoost, dan LightGBM dengan AUC 0,833 dan *overall accuracy* 93,1%.

Penelitian ini menggunakan algoritma Extreme Gradient Boosting (XGBoost) sebagai model utama untuk klasifikasi risiko banjir. XGBoost merupakan pengembangan teroptimasi dari Gradient Boosting yang dirancang untuk meningkatkan kinerja dan efisiensi komputasi (Qin *et al.*, 2025; Ren *et al.*, 2024). Algoritma ini dilengkapi dengan mekanisme regularisasi L1 dan L2 untuk mengendalikan *overfitting*, pemanfaatan turunan orde kedua (Hessian) untuk

optimisasi yang lebih presisi, kemampuan komputasi paralel untuk mempercepat proses pelatihan, serta penanganan *missing values* secara otomatis. Wang *et al.* (2023) dan Yuan *et al.* (2024) menemukan bahwa XGBoost, apabila dikombinasikan dengan teknik interpretabilitas seperti SHAP, menghasilkan akurasi tinggi sekaligus memberikan pemahaman tentang kontribusi masing-masing fitur terhadap prediksi banjir.

XGBoost telah digunakan dalam berbagai studi klasifikasi banjir, namun belum banyak penelitian yang menguraikan sejauh mana teknik *preprocessing* dan optimasi berkontribusi terhadap performanya. Sebagian besar penelitian sebelumnya melaporkan performa akhir tanpa memisahkan kontribusi masing-masing komponen pemrosesan data. Oleh karena itu, penelitian ini menerapkan desain studi ablasi dengan enam skenario (S1 hingga S6) yang secara sistematis menguji kontribusi tiga komponen: seleksi fitur (MI + Pearson + RFE), penanganan ketidakseimbangan kelas (BorderlineSMOTE), dan optimasi *hyperparameter* (Optuna TPE). S1 merupakan *baseline* dengan 53 fitur dan *default hyperparameter*, sementara S2 hingga S6 menambahkan komponen secara bertahap dan kombinatif. Pendekatan ablasi ini memungkinkan pengukuran kontribusi setiap komponen secara terpisah maupun gabungan terhadap performa model.

Sistem drainase yang tidak memadai merupakan salah satu faktor utama penyebab banjir perkotaan. Bersabe & Jun (2025) membuktikan bahwa integrasi data drainase, berupa *Sewer Pipe Density* (SPD) dan *Drain Station Density* (DSD), meningkatkan akurasi model sebesar 7,58% dan AUC sebesar 3,80% dibandingkan model tanpa variabel drainase. Penelitian ini menggunakan berbagai parameter fisik

drainase sebagai variabel, meliputi panjang saluran, lebar atas dan bawah, tinggi saluran, kemiringan (*slope*), luas penampang, serta keliling penampang basah. Parameter tersebut selanjutnya diolah melalui *feature engineering* untuk menghasilkan fitur turunan seperti *aspect ratio*, *taper ratio, hydraulic radius*, dan estimasi kapasitas drainase. Secara keseluruhan, data drainase menghasilkan 26 fitur.

Data curah hujan dari stasiun meteorologi konvensional sering kali memiliki keterbatasan dalam cakupan spasial dan kontinuitas data. Untuk mengatasi keterbatasan tersebut, penelitian ini memanfaatkan data curah hujan satelit CHIRPS (*Climate Hazards Group InfraRed Precipitation with Station Data*) yang menyediakan resolusi spasial tinggi (0,05 derajat atau sekitar 5,5 km) dan resolusi temporal harian. Riazi *et al.* (2023) menggunakan data CHIRPS dalam pemetaan kerentanan banjir dan menunjukkan bahwa integrasi data CHIRPS dengan citra SAR dan model *machine learning* hybrid menghasilkan AUC hingga 0,997. Dari data CHIRPS diekstraksi 19 fitur statistik yang mencakup total curah hujan, rata-rata, nilai maksimum, standar deviasi, jumlah hari hujan, dan intensitas curah hujan pada berbagai jendela waktu akumulasi. Selain itu, 8 fitur interaksi antara kapasitas drainase dan beban curah hujan juga dihitung untuk merepresentasikan hubungan antar kedua sumber data, sehingga total dataset terdiri dari 53 fitur.

Dalam penanganan ketidakseimbangan data, dataset penelitian ini memiliki distribusi kelas yang tidak merata: 2.279 sampel kelas Rendah (65,7%), 891 sampel kelas Sedang (25,7%), dan 301 sampel kelas Tinggi (8,7%). Teknik BorderlineSMOTE diterapkan pada sebagian skenario untuk menyeimbangkan

distribusi kelas pada data pelatihan. Qin *et al.* (2025) menggunakan SMOTE dalam konteks klasifikasi banjir perkotaan dan melaporkan bahwa penerapan SMOTE bersama *StratifiedKFold* menghasilkan model XGBoost dengan *overall accuracy* 0,96.

Penelitian-penelitian sebelumnya umumnya masih bergantung pada data dari satu sumber, baik yang hanya berfokus pada variabel meteorologi (Jang *et al.*, 2025) maupun hanya pada karakteristik infrastruktur (Bersabe & Jun, 2025), tanpa mengeksplorasi potensi sinergi antara keduanya. Selain itu, belum banyak penelitian yang menggunakan desain studi ablasi untuk mengukur kontribusi masing-masing komponen *preprocessing* dan optimasi terhadap performa model klasifikasi risiko banjir. Penelitian ini mengintegrasikan data drainase fisik dan curah hujan satelit CHIRPS dalam kerangka studi ablasi XGBoost, sehingga kontribusi setiap komponen dapat diukur secara terpisah dan perbandingan antarskenario dapat dilakukan.

1.2 Pernyataan Masalah

Berdasarkan latar belakang yang telah diuraikan, permasalahan yang dirumuskan dalam penelitian ini adalah sebagai berikut:

1. Bagaimana mengintegrasikan data curah hujan satelit CHIRPS yang bersifat temporal dengan data karakteristik drainase yang bersifat spasial untuk membentuk dataset klasifikasi risiko banjir tiga kategori (Rendah, Sedang, Tinggi), dan apakah integrasi kedua sumber data tersebut menghasilkan fitur yang informatif bagi model klasifikasi?

2. Bagaimana performa model XGBoost *baseline* dengan 53 fitur gabungan (drainase, curah hujan, dan interaksi), dan sejauh mana kontribusi masing-masing komponen *preprocessing* (seleksi fitur, BorderlineSMOTE) dan optimasi *hyperparameter* (Optuna TPE) terhadap performa model, yang diukur melalui studi ablasi dengan enam skenario?
3. Fitur-fitur apa yang paling berpengaruh dalam klasifikasi risiko banjir berdasarkan analisis SHAP, dan bagaimana kontribusinya berubah antarskenario?

1.3 Tujuan Penelitian

Berdasarkan permasalahan yang telah dirumuskan, tujuan penelitian ini adalah sebagai berikut:

1. Mengembangkan metode integrasi data curah hujan satelit CHIRPS dan data karakteristik drainase fisik untuk membangun dataset klasifikasi risiko banjir tiga kelas (Rendah, Sedang, Tinggi) di wilayah Kota Malang, serta membuktikan bahwa integrasi kedua sumber data menghasilkan fitur yang informatif bagi model klasifikasi.
2. Membangun dan mengevaluasi model XGBoost *baseline* dengan 53 fitur gabungan, serta mengukur kontribusi seleksi fitur (MI + Pearson + RFE), BorderlineSMOTE, dan Optuna TPE terhadap performa model melalui studi ablasi enam skenario (S1 hingga S6).
3. Mengidentifikasi fitur-fitur yang paling berpengaruh dalam klasifikasi risiko banjir menggunakan analisis SHAP dan mengevaluasi perubahan kontribusi fitur antarskenario.

1.4 Batasan Penelitian

Agar penelitian ini lebih terarah, ditetapkan beberapa batasan sebagai berikut:

1. Lokasi penelitian difokuskan pada wilayah Kota Malang, khususnya daerah rawan banjir berdasarkan data BPBD (Badan Penanggulangan Bencana Daerah) Kota Malang.
2. Hasil penelitian berupa klasifikasi risiko banjir dalam tiga kategori, yaitu Rendah, Sedang, dan Tinggi, dan tidak mencakup estimasi debit banjir atau kedalaman genangan secara kuantitatif.
3. Algoritma yang digunakan adalah XGBoost dengan *objective multi:softprob* untuk klasifikasi tiga kelas, dievaluasi melalui enam skenario ablasi (S1–S6).
4. Data curah hujan bersumber dari dataset satelit CHIRPS dengan resolusi spasial 0,05 derajat dan resolusi temporal harian.
5. Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* (5 fold × 5 pengulangan = 25 evaluasi) untuk metrik *Accuracy*, F1 Macro, dan ROC-AUC, dilengkapi satu kali *split* 80:20 untuk analisis diagnostik berupa *confusion matrix*, metrik per kelas, dan SHAP.
6. Interpretabilitas model dianalisis menggunakan SHAP (*SHapley Additive exPlanations*) dengan metode TreeExplainer.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Memberikan kerangka studi ablasi yang dapat digunakan sebagai acuan dalam mengevaluasi kontribusi komponen *preprocessing* dan optimasi pada model *machine learning* untuk klasifikasi bencana.
2. Membantu pemerintah daerah dan instansi terkait dalam mengidentifikasi wilayah rawan banjir di Kota Malang berdasarkan integrasi data drainase dan curah hujan.
3. Menyediakan informasi berbasis data untuk mendukung perencanaan perbaikan sistem drainase dan mitigasi risiko banjir.
4. Menjadi dasar dalam pengembangan sistem peringatan dini banjir dengan memanfaatkan model XGBoost yang telah dievaluasi.
5. Memberikan pemahaman tentang fitur-fitur yang paling berpengaruh terhadap risiko banjir melalui analisis SHAP, yang dapat menjadi masukan bagi pengambil kebijakan dalam pengelolaan infrastruktur drainase perkotaan.

BAB II
TINJAUAN PUSTAKA

Penelitian ini melakukan studi literatur untuk mengkaji informasi yang relevan dari berbagai jurnal ilmiah, dengan tujuan memperoleh pemahaman yang menyeluruh mengenai konteks, perkembangan, serta kerangka teoritis terkait klasifikasi risiko banjir berbasis XGBoost dengan integrasi data drainase fisik dan curah hujan satelit CHIRPS. Tinjauan difokuskan pada tiga aspek utama: (1) penggunaan *ensemble learning* dalam pemetaan kerentanan banjir, (2) metode *preprocessing* dan optimasi yang diterapkan, serta (3) posisi penelitian ini terhadap penelitian terdahulu.

2.1 Klasifikasi Risiko Banjir Berbasis Model *Ensemble Learning*

Penelitian oleh Riazi *et al.* (2023) berfokus pada peningkatan pemetaan kerentanan banjir melalui integrasi data *multi-temporal* SAR dan curah hujan satelit CHIRPS dengan model *hybrid machine learning*. Dengan menggunakan citra Sentinel-1 SAR dari Google Earth Engine serta 12 parameter input (elevasi, litologi, *drainage density*, curah hujan, NDVI, *curvature*, *slope*, SPI, TWI, *soil*, LULC, dan jarak dari sungai), penelitian ini mengevaluasi performa model Radial Basis Function (RBF) sebagai *standalone* dan tiga model *hybrid ensemble*: Bagging-RBF (BA-RBF), Random Committee-RBF (RC-RBF), dan Random Subspace-RBF (RSS-RBF) di DAS Gorganrood, Iran, dengan 309 titik banjir dan 309 non-banjir. Hasil eksperimen mengindikasikan bahwa seluruh model *hybrid* mengungguli model *standalone* RBF (AUC = 0,975), dengan RC-RBF mencapai AUC tertinggi

sebesar 0,997, BA-RBF sebesar 0,996, dan RSS-RBF sebesar 0,992. Performa ini mengonfirmasi bahwa integrasi data penginderaan jauh dengan pendekatan *ensemble* dapat meningkatkan akurasi pemetaan kerentanan banjir.

Penelitian Wang *et al.* (2023) mengkuantifikasi dampak morfologi bangunan terhadap kerentanan banjir perkotaan melalui pendekatan *Explainable AI*. Studi ini mengintegrasikan algoritma XGBoost dengan SHAP (*SHapley Additive exPlanations*) untuk menganalisis kontribusi variabel urban seperti *Mean Building Volume*, *Standard Deviation Building Volume*, *Building Height*, dan *Building Density* di Kota Shenzhen, China. Dengan menggunakan 170 titik sampel banjir dan tahapan *preprocessing* berupa analisis multikolinearitas (VIF), korelasi Spearman, serta deteksi *outlier*, hasil analisis SHAP mengidentifikasi *Mean Building Volume* sebagai prediktor utama kerentanan banjir dengan kontribusi *mean* SHAP sebesar 0,0107 m (9,70%). Temuan ini memperlihatkan bahwa pendekatan XGBoost-SHAP mampu memberikan interpretabilitas tinggi terhadap faktor-faktor yang memengaruhi kerentanan banjir perkotaan.

Penelitian Yuan *et al.* (2024) mengembangkan *framework* optimasi konfigurasi urban berbasis data untuk menyeimbangkan mitigasi banjir dan pertumbuhan ekonomi. Studi ini menerapkan algoritma XGBoost yang dikombinasikan dengan Pareto Multi-Objective Optimization untuk menganalisis variabel konfigurasi urban, termasuk *Building Congestion Degree*, vegetasi, *impervious coverage ratio*, dan *drainage network density* di Kota Shenzhen, China. Dengan *split* data 70/30,model XGBoost mencapai $R^2 = 0,82$ untuk prediksi *Pluvial Flooding Susceptibility* (PFS). Analisis *feature importance* berbasis Gini

mengidentifikasi *Building Congestion Degree* sebagai faktor dominan. Hasil ini menunjukkan bahwa XGBoost dapat digunakan untuk klasifikasi maupun optimasi berbasis kerentanan banjir.

Penelitian Shrestha *et al.* (2025) melakukan penilaian risiko banjir di daerah aliran sungai perkotaan Briar Creek, Charlotte, North Carolina, yang secara historis sering terdampak banjir. Studi ini membandingkan tiga algoritma *machine learning* yaitu Logistic Regression, XGBoost, dan Random Forest (Bagging), menggunakan 750 titik data (375 banjir + 375 non-banjir) dengan variabel topografi (*slope*, *aspect*, *curvature*), hidrologi (*flow velocity*, *flow concentration*, *discharge*), dan 8 tahun data curah hujan. Analisis korelasi digunakan untuk mengevaluasi hubungan antarvariabel, *balanced sampling* diterapkan melalui stratified random sampling, dan optimasi *hyperparameter* dilakukan menggunakan Grid SearchCV dengan 5-fold *cross-validation*. Hasil eksperimen memperlihatkan bahwa Logistic Regression mencapai performa terbaik dengan akurasi 97,92%, F1 = 0,9787, dan AUC = 0,9861, sementara XGBoost mencapai akurasi 95,83% dan Bagging mencapai 93,75%. Klasifikasi menghasilkan lima zona risiko: sangat tinggi (5,55%), tinggi (8,66%), sedang (12,04%), rendah (21,56%), dan sangat rendah (52,20%).

Penelitian Zhu *et al.* (2024) mengusulkan integrasi *remote sensing* dan *social sensing* melalui Ensemble Model Framework (EMF) untuk meningkatkan akurasi pemetaan kerentanan banjir perkotaan. Studi ini secara unik mengombinasikan data Sentinel-1 SAR dengan data *social sensing* dari media sosial Weibo untuk mengidentifikasi 1.500 titik sampel banjir, yang kemudian

digunakan melatih empat model konstituen, yaitu XGBoost, SVC, MLP, dan MDL yang diensemble melalui Random Forest sebagai *meta-learner*. Hasil pengujian memperlihatkan bahwa EMF mencapai akurasi tertinggi sebesar 0,940,mengungguli seluruh model individual: MDL, XGBoost, SVC, dan MLP. Temuan ini mengindikasikan bahwa pendekatan *stacking ensemble* mampu menangani heterogenitas penggunaan lahan perkotaan lebih baik daripada model individual.

Penelitian Goumghar *et al.* (2025) membandingkan empat algoritma *boosting* (XGBoost, CatBoost, LightGBM, dan Hist Gradient Boosting/HGB) untuk prediksi kerentanan banjir di wilayah *semi-arid* Upper Drâa Basin, Maroko. Dengan memanfaatkan data GIS dan *remote sensing* serta tahapan seleksi fitur berupa *Mutual Information*, RFE, dan analisis VIF, hasil eksperimen memperlihatkan bahwa HGB memiliki performa terbaik dengan OA = 93,1%, AUC = 0,833, dan F1 = 0,80,sementara XGBoost mencapai AUC = 0,727. Perbedaan performa antar varian *boosting* ini mengindikasikan bahwa pemilihan algoritma bergantung pada karakteristik dataset dan konteks masalah, sehingga perbandingan langsung antara varian *boosting* pada dataset yang sama menjadi penting.

Berbagai studi di atas mengonfirmasi bahwa metode *ensemble learning* seperti Random Forest, XGBoost, dan berbagai model *hybrid* secara konsisten memberikan performa lebih stabil dan akurat dibandingkan model *standalone*. XGBoost menjadi algoritma yang paling dominan, diadopsi oleh 7 dari 10 studi yang ditinjau (Wang *et al.*, 2023; Yuan *et al.*, 2024; Qin *et al.*, 2025; Shrestha *et al.*, 2025; Zhu *et al.*, 2024; Ren *et al.*, 2024; Goumghar *et al.*, 2025), sedangkan

Random Forest menempati posisi kedua dengan penggunaan di 6 studi (Riazi *et al.*, 2023; Shrestha *et al.*, 2025; Zhu *et al.*, 2024; Bersabe & Jun, 2025; Ren *et al.*, 2024; Jang *et al.*, 2025). Dominasi XGBoost di literatur menjadi landasan pemilihan algoritma ini sebagai model utama dalam penelitian ini. Namun demikian, studi-studi tersebut umumnya menerapkan XGBoost dengan satu konfigurasi *preprocessing* tunggal tanpa melakukan studi ablasi yang sistematis untuk mengevaluasi kontribusi individual maupun gabungan dari komponen *feature selection*, penanganan ketidakseimbangan kelas, dan optimasi *hyperparameter*.

2.2 Metode *Preprocessing* dan Optimasi pada Klasifikasi Risiko Banjir

Penelitian oleh Qin *et al.* (2025) mengusulkan identifikasi faktor kunci dan analisis heterogenitas spasial banjir perkotaan melalui *framework* "*model-factor-space*" yang menggabungkan XGBoost dengan SHAP. Dengan menggunakan 300 titik sampel banjir di area Kangyi, Kota Ordos, China, penelitian ini menerapkan SMOTE (*Synthetic Minority Oversampling Technique*) dalam kombinasi dengan *StratifiedKFold* untuk menangani ketidakseimbangan kelas, serta RFE (*Recursive Feature Elimination*) dan *one-hot encoding* untuk seleksi dan transformasi fitur. Hasil eksperimen menunjukkan bahwa XGBoost dengan integrasi SMOTE mencapai *overall accuracy* sebesar 0,96. Kelebihan penelitian ini terletak pada kemampuannya mengatasi masalah ketidakseimbangan data yang umum pada dataset banjir sekaligus memberikan analisis heterogenitas spasial melalui SHAP.

Penelitian Bersabe & Jun (2025) berfokus pada integrasi data drainase sebagai faktor pengkondisi banjir yang sering diabaikan dalam studi terdahulu. Studi ini menambahkan dua variabel baru berupa *Sewer Pipe Density* (SPD) dan

Drain Station Density (DSD) ke dalam 16 faktor pengkondisi untuk pemetaan kerentanan banjir di Seoul, Korea Selatan, dengan data banjir periode 2010-2022. Penelitian ini membandingkan tiga algoritma, yaitu Logistic Regression, Random Forest, dan SVM, dengan tahapan *preprocessing* berupa analisis multikolinearitas (VIF, korelasi Pearson), deteksi *outlier*, dan *random sampling*. Hasil menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 0,837 dan AUC 0,902. Studi ini menemukan bahwa penambahan variabel drainase meningkatkan akurasi model sebesar 7,58% dan AUC sebesar 3,80%, yang membuktikan bahwa data infrastruktur drainase berkontribusi terhadap klasifikasi kerentanan banjir perkotaan.

Penelitian Ren *et al.* (2024) mengusulkan strategi *random negative sampling* untuk mengatasi bias pemilihan sampel negatif yang memengaruhi kinerja model prediksi kerentanan banjir. Menggunakan 340 titik banjir dari inventaris kejadian 2018-2022 di Kunming, China, penelitian ini membandingkan empat algoritma, yaitu Random Forest, XGBoost, ANN, dan SVM, dengan seleksi fitur berbasis RFE dan *5-fold cross-validation*. Hasil pengujian menunjukkan bahwa Random Forest mencapai performa terbaik dengan AUC = 0,87 dan Kappa = 0,89, diikuti XGBoost (AUC = 0,84), ANN (AUC = 0,83), dan SVM (AUC = 0,82). Strategi *random negative sampling* terbukti efektif dalam meningkatkan representativitas sampel pelatihan dan mereduksi bias spasial dalam pemodelan kerentanan banjir.

Penelitian Jang *et al.* (2025) mengembangkan pendekatan *machine learning* berbasis karakteristik statistik curah hujan untuk prediksi banjir perkotaan yang

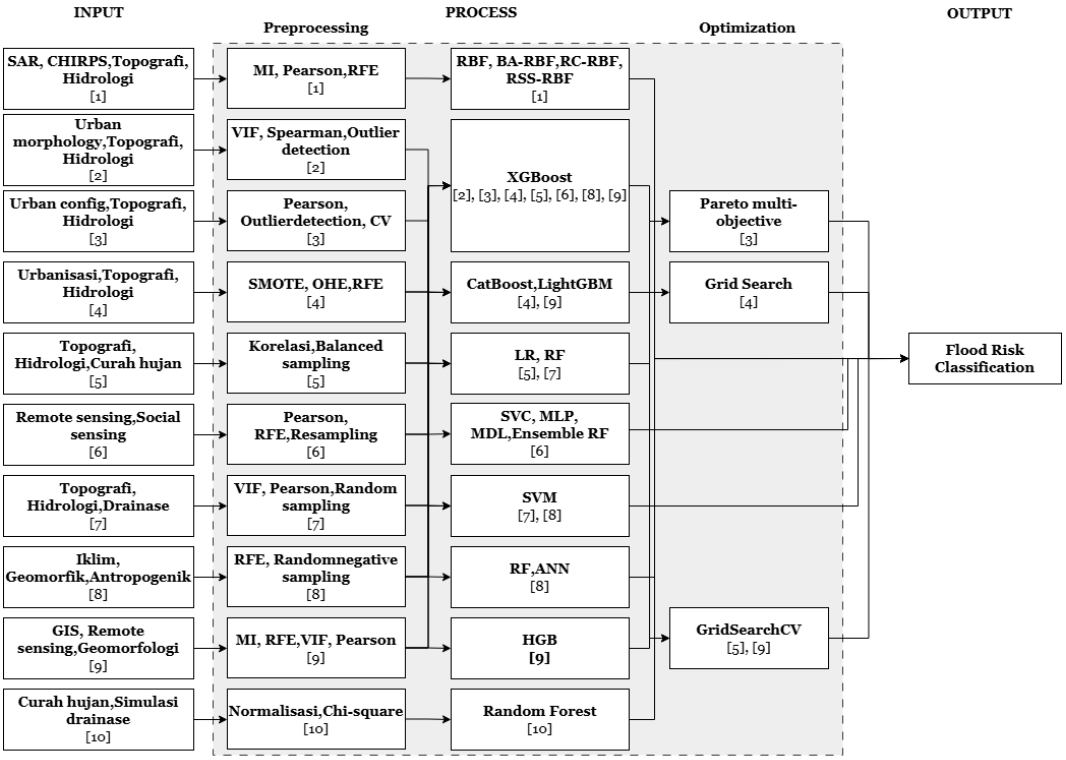
lebih cepat dibandingkan model numerik tradisional. Studi ini menggunakan output simulasi drainase 1D (EPA SWMM) dan banjir 2D dari 99 kejadian hujan di Republik Korea sebagai data pelatihan, dengan algoritma Random Forest sebagai model klasifikasi utama. *Preprocessing* meliputi normalisasi, *cross-validation* (k-fold), dan uji Chi-square untuk seleksi fitur. Hasil penelitian ini menunjukkan bahwa penggunaan fitur statistik curah hujan, meliputi *mean*, standar deviasi, *skewness*, dan *kurtosis* menghasilkan performa prediksi yang lebih baik (R^2 = 0,9761; RMSE = 2,7354) dibandingkan penggunaan total curah hujan saja (R^2 = 0,9682; RMSE = 3,1573). Hasil ini memiliki implikasi langsung bagi penelitian yang menggunakan data curah hujan satelit CHIRPS, karena fitur statistik curah hujan dapat diekstraksi dari data CHIRPS untuk meningkatkan akurasi model klasifikasi.

Permasalahan yang dihadapi dalam klasifikasi risiko banjir adalah ketidakseimbangan distribusi kelas dan tingginya dimensi fitur yang berpotensi menyebabkan *overfitting* serta menurunkan akurasi klasifikasi. Untuk penanganan ketidakseimbangan kelas, berbagai penelitian menerapkan teknik SMOTE (Qin *et al.*, 2025), *balanced sampling* manual (Shrestha *et al.*, 2025), dan *random negative sampling* (Ren *et al.*, 2024). Namun demikian, penerapan varian SMOTE yang lebih adaptif seperti BorderlineSMOTE, yang secara selektif menghasilkan sampel sintetis hanya pada titik-titik di dekat batas keputusan antarkelas, belum dieksplorasi dalam konteks klasifikasi kerentanan banjir. Di sisi seleksi fitur, *Recursive Feature Elimination* (RFE) merupakan teknik yang paling banyak diadopsi (Riazi *et al.*, 2023; Qin *et al.*, 2025; Ren *et al.*, 2024; Goumghar *et al.*,

2025), diikuti oleh *Mutual Information* (Riazi *et al.*, 2023; Goumghar *et al.*, 2025), analisis korelasi (Shrestha *et al.*, 2025), dan uji Chi-square (Jang *et al.*, 2025). Penanganan multikolinearitas melalui VIF dan korelasi Pearson juga diterapkan secara luas (Wang *et al.*, 2023; Bersabe & Jun, 2025; Goumghar *et al.*, 2025), namun belum ada studi yang mengombinasikan *Mutual Information*, korelasi Pearson, dan RFE secara berurutan dalam satu *pipeline* untuk mereduksi dimensi sambil mempertahankan fitur yang informatif terhadap target. Untuk optimasi *hyperparameter*, mayoritas studi menggunakan *Grid Search* (Qin *et al.*, 2025; Shrestha *et al.*, 2025; Goumghar *et al.*, 2025) atau *cross-validation* standar, namun belum ada yang mengadopsi pendekatan *Bayesian optimization* seperti Optuna TPE yang secara teoritis lebih efisien dalam menavigasi ruang *hyperparameter* berdimensi tinggi.

2.3 Kerangka Teori

Bagian kerangka teori membahas teori-teori yang mendukung penelitian terkait klasifikasi risiko banjir berbasis XGBoost dengan integrasi data drainase fisik dan curah hujan satelit CHIRPS. Kerangka teori pada Gambar 2.1 memuat berbagai pendekatan yang telah banyak digunakan serta *gap* yang teridentifikasi dari penelitian terdahulu, untuk memberikan landasan konseptual sekaligus menjadi acuan dalam pemilihan teknik yang relevan untuk penelitian ini.



Gambar 2. 1 Kerangka Teori

Tabel 2. 1 Literatur Penelitian Sebelumnya

Peneliti	**Kategori Data**	**Model / Algoritma**	***Pre-processing***	**Optimasi**	**Performa**
Riazi *et al.* (2023)	SAR, CHIRPS, Topografi, Hidrologi	RBF, BA-RBF, RC-RBF, RSS-RBF	MI, Pearson, RFE	Tidak disebutkan	RC-RBF AUC=0,997
Wang *et al.* (2023)	*Urban morphology*, Topografi, Hidrologi	XGBoost	VIF, Spearman, *Outlier detection*	Tidak disebutkan	SHAP mean=0,0107
Yuan *et al.* (2024)	*Urban config*, Topografi, Hidrologi	XGBoost	Pearson, *Outlier detection*	Pareto *multi-objective*	R^2=0,82

Peneliti	**Kategori Data**	**Model / Algoritma**	***Pre-processing***	**Optimasi**	**Performa**
Qin *et al.* (2025)	Urbanisasi, Topografi, Hidrologi	XGBoost, CatBoost, LightGBM	SMOTE, *One-hot encoding*, RFE	*Grid Search*	XGBoost OA=0,96; CatBoost AUC=0,85
Shrestha *et al.* (2025)	Topografi, Hidrologi, Curah hujan	LR, XGBoost, RF	Korelasi, *Balanced sampling*	*Grid SearchCV*	LR Acc=97,92%, AUC=0,9861
Zhu *et al.* (2024)	*Remote sensing, Social sensing*	XGBoost, SVC, MLP, MDL, *Ensemble* RF	Pearson, RFE, *Resampling*	Tidak disebutkan	EMF Acc=0,940
Bersabe & Jun (2025)	Topografi, Hidrologi, Drainase	LR, RF, SVM	VIF, Pearson, *Random sampling*	Tidak disebutkan	RF Acc=0,837, AUC=0,902
Ren *et al.* (2024)	Iklim, Geomorfik, Antropogenik	RF, XGBoost, ANN, SVM	RFE, *Random negative sampling*	Tidak disebutkan	RF AUC=0,87
Goumghar *et al.* (2025)	GIS, *Remote sensing*, Geomorfologi	XGBoost, CatBoost, LightGBM, HGB	MI, RFE, VIF, Pearson	*GridSearchCV*	HGB OA=93,1%, AUC=0,833
Jang *et al.* (2025)	Curah hujan, Simulasi drainase	Random Forest	Normalisasi, *Chi-square*	Tidak disebutkan	R^2=0,976, RMSE=2,74

Studi-studi pada Tabel 2.1 menunjukkan bahwa pendekatan berbasis *ensemble learning*, khususnya XGBoost, memberikan akurasi yang tinggi dalam

klasifikasi kerentanan banjir. Riazi *et al.* (2023) memperkuat temuan ini secara empiris: seluruh model *hybrid ensemble* (AUC = 0,992-0,997) mengungguli model *standalone* RBF (AUC = 0,975), menunjukkan bahwa mekanisme agregasi prediksi mampu mereduksi varians dan bias model individual. Goumghar *et al.* (2025) membandingkan empat varian *boosting* di Upper Drâa Basin, dengan Hist Gradient Boosting mencapai AUC = 0,833 dan OA = 93,1%, sementara XGBoost mencapai AUC = 0,727 , yang menunjukkan bahwa performa relatif antar algoritma bergantung pada karakteristik dataset.

Integrasi data yang tepat terbukti memberikan kontribusi yang berarti terhadap akurasi model. Bersabe & Jun (2025) membuktikan bahwa penambahan variabel drainase (*Sewer Pipe Density* dan *Drain Station Density*) meningkatkan akurasi model sebesar 7,58% dan AUC sebesar 3,80%, menunjukkan bahwa data infrastruktur drainase berkontribusi terhadap akurasi model. Riazi *et al.* (2023) mendemonstrasikan efektivitas integrasi citra SAR dan data curah hujan satelit CHIRPS, sementara Jang *et al.* (2025) menunjukkan bahwa fitur statistik curah hujan (*mean*, standar deviasi, *skewness*, *kurtosis*) menghasilkan prediksi yang lebih baik dibandingkan total curah hujan. Interpretabilitas model juga menjadi aspek penting: Wang *et al.* (2023) dan Qin *et al.* (2025) menerapkan SHAP untuk mengkuantifikasi kontribusi variabel, sedangkan Yuan *et al.* (2024) menggunakan *feature importance* berbasis Gini.

Selain pemilihan dan integrasi data, rekayasa fitur (*feature engineering*) merupakan tahapan kritis yang menentukan kualitas representasi data untuk model *machine learning*. Zheng & Casari (2018) menegaskan bahwa fitur mentah sering

kali tidak cukup untuk menangkap hubungan kompleks antarvariabel, sehingga diperlukan konstruksi fitur turunan berupa rasio, perkalian, atau transformasi yang merepresentasikan interaksi antarvariabel berdasarkan pengetahuan domain. Dalam bidang hidrologi, praktik konstruksi fitur turunan berbasis pengetahuan domain memiliki preseden yang kuat dan telah diadopsi secara luas. *Topographic Wetness Index* (TWI = ln(A/s / tan β)) yang diperkenalkan oleh Beven & Kirkby (1979) dan *Stream Power Index* (SPI = $A/s \times \tan \beta$) yang diformulasikan oleh Moore *et al.* (1991) merupakan contoh klasik fitur interaksi yang menggabungkan luas tangkapan (*contributing area*) dan kemiringan (*slope*) menjadi satu indeks bermakna secara hidrologis. Kedua indeks ini bukan variabel yang diukur secara langsung di lapangan, melainkan diturunkan secara matematis dari variabel dasar untuk merepresentasikan proses hidrologi yang tidak dapat ditangkap oleh variabel individual. TWI mengukur kecenderungan akumulasi air pada suatu titik berdasarkan topografi, sementara SPI mengukur daya erosif aliran permukaan. Keduanya secara konsisten digunakan sebagai fitur masukan (*conditioning factors*) dalam studi pemetaan kerentanan banjir berbasis *machine learning* (Tehrany *et al.*, 2014; Riazi *et al.*, 2023; Pourzangbar *et al.*, 2025), dan telah terbukti menjadi prediktor yang signifikan terhadap kerentanan banjir.

Preseden TWI dan SPI menunjukkan bahwa konstruksi fitur interaksi dari variabel dasar merupakan praktik ilmiah yang mapan dalam hidrologi, bukan pendekatan *ad hoc*. Prinsip yang sama dapat diterapkan pada konteks integrasi data drainase fisik dan curah hujan: variabel drainase (kapasitas, dimensi, kemiringan) dan variabel curah hujan (intensitas, frekuensi, akumulasi) masing-masing

menangkap satu aspek dari sistem hidrologi, namun interaksi antara keduanya yaitu seberapa besar beban curah hujan relatif terhadap kapasitas drainase merupakan informasi kritis yang tidak terwakili oleh fitur individual. Konsep ini sejalan dengan prinsip dasar hidraulika saluran terbuka (Chow, 1959; Chow *et al.*, 1988) dan kerangka risiko bencana (UNDRR, 2017) yang memodelkan risiko sebagai fungsi dari *hazard*, *exposure*, dan *vulnerability*.

Terkait penanganan ketidakseimbangan kelas, Qin *et al.* (2025) menerapkan SMOTE dengan *StratifiedKFold* untuk menyeimbangkan distribusi kelas, menghasilkan XGBoost dengan *overall accuracy* 0,96. Shrestha *et al.* (2025) menggunakan *balanced sampling* manual, sementara Ren *et al.* (2024) mengembangkan strategi *random negative sampling* untuk mengurangi bias sampel negatif. Untuk seleksi fitur, RFE menjadi teknik yang paling banyak diadopsi (Riazi *et al.*, 2023; Qin *et al.*, 2025; Ren *et al.*, 2024; Goumghar *et al.*, 2025), dilengkapi analisis multikolinearitas melalui VIF dan korelasi Pearson (Wang *et al.*, 2023; Bersabe & Jun, 2025; Goumghar *et al.*, 2025).

Berdasarkan tinjauan terhadap sepuluh studi tersebut, teridentifikasi beberapa *gap* yang menjadi landasan penelitian ini. Pertama, dari sisi data, mayoritas studi menggunakan data topografi dan *remote sensing*, sementara integrasi karakteristik drainase fisik yang menyeluruh (26 fitur) dengan fitur statistik curah hujan satelit CHIRPS (19 fitur) belum dieksplorasi; hanya Bersabe & Jun (2025) dan Jang *et al.* (2025) yang menyertakan variabel terkait drainase secara terbatas. Kedua, dari sisi metodologi, seluruh studi menerapkan satu konfigurasi *preprocessing* tunggal tanpa melakukan studi ablasi sistematis untuk

mengevaluasi kontribusi individual masing-masing komponen (*feature selection*, penanganan *imbalance*, optimasi *hyperparameter*). Ketiga, dari sisi optimasi, belum ada studi yang menggunakan *Bayesian optimization* (Optuna TPE) untuk *hyperparameter tuning* XGBoost dalam konteks klasifikasi kerentanan banjir.

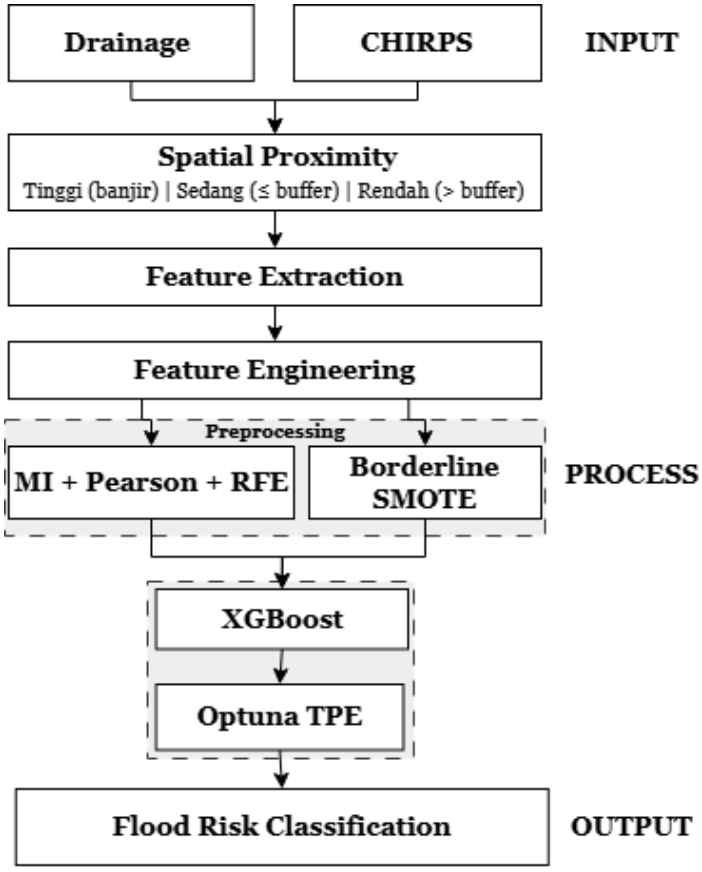
Dengan mengacu pada *gap* tersebut, penelitian ini menggunakan XGBoost sebagai model klasifikasi utama dan merancang enam skenario ablasi (S1–S6) untuk mengevaluasi kontribusi setiap komponen secara sistematis. *Pipeline* yang digunakan terdiri dari seleksi fitur MI + Pearson + RFE untuk mereduksi dimensi sambil mempertahankan fitur yang informatif terhadap target, BorderlineSMOTE untuk penanganan ketidakseimbangan kelas, dan Optuna TPE untuk optimasi *hyperparameter*. Keenam skenario mengikuti pola: S1 (*Baseline*), S2 (*Feature Selection Only*), S3 (*Imbalance Handling Only*), S4 (*Feature Selection + Imbalance Handling*), S5 (*Optuna Only*), dan S6 (*Full Pipeline*). Eksperimen didukung oleh integrasi 53 fitur dari data drainase fisik (sejalan dengan Bersabe & Jun, 2025), curah hujan satelit CHIRPS (sejalan dengan Riazi *et al.*, 2023 dan Jang *et al.*, 2025), serta fitur interaksi antara drainase dan curah hujan. Desain ablasi ini memungkinkan evaluasi kuantitatif terhadap efektivitas masing-masing komponen *preprocessing* dan optimasi, sehingga menghasilkan model yang akurat dan informatif dalam mengklasifikasikan risiko banjir di wilayah Kota Malang.

BAB III
METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi klasifikasi risiko banjir berbasis XGBoost dengan integrasi data drainase fisik dan curah hujan satelit CHIRPS di Kota Malang.

3.1 Kerangka Konseptual

Kerangka konseptual penelitian ini dirancang untuk menjawab tiga *research gap* yang teridentifikasi dari tinjauan pustaka pada Bab II. Gambar 3.1 menunjukkan alur konseptual dari data masukan hingga keluaran berupa klasifikasi risiko banjir.



Gambar 3. 1 Kerangka Konseptual

Kerangka konseptual pada Gambar 3.1 terdiri dari tujuh komponen utama yang membentuk *pipeline* sekuensial. Komponen pertama adalah *input* berupa dua sumber data: data drainase fisik Kota Malang dari survei lapangan dan data curah hujan satelit CHIRPS dalam format NetCDF harian. Komponen kedua adalah *risk labeling*, yaitu pelabelan kelas risiko banjir pada setiap titik drainase berdasarkan kedekatan spasial terhadap titik banjir historis menggunakan jarak Haversine; titik banjir diberi label Tinggi, titik non-banjir dalam zona buffer diberi label Sedang, dan titik non-banjir di luar zona buffer diberi label Rendah (detail pada sub-bagian 3.2.3c). Komponen ketiga adalah *feature extraction*, yaitu konversi data mentah menjadi representasi numerik: 26 fitur diekstraksi dari data drainase (Tabel 3.1) dan 19 fitur statistik dari *time series* CHIRPS (Tabel 3.2), menghasilkan 45 fitur dasar. Komponen keempat adalah *feature engineering*, yaitu konstruksi 8 fitur interaksi drainase × curah hujan berdasarkan pengetahuan domain hidrologi (Tabel 3.3), sehingga total menjadi 53 fitur. Komponen kelima adalah *preprocessing* yang terdiri dari *feature selection* (MI + Pearson + RFE) dan penanganan ketidakseimbangan kelas (BorderlineSMOTE). Komponen keenam adalah model klasifikasi XGBoost dengan optimasi *hyperparameter* Optuna TPE. Komponen ketujuh adalah *output* berupa klasifikasi risiko banjir tiga kelas (Rendah, Sedang, Tinggi) pada setiap titik drainase.

Tabel 3. 1 Daftar 26 Fitur Drainase Fisik

No	**Nama Fitur**	**Subkategori**	**Keterangan**
1	*panjang_saluran_m*	Geometri	Panjang saluran (meter)
2	*tinggi_saluran_m*	Geometri	Tinggi saluran (meter)
3	*lebar_atas_saluran_m*	Geometri	Lebar atas saluran (meter)

No	Nama Fitur	Subkategori	Keterangan
4	*lebar_bawah_saluran_m*	Geometri	Lebar bawah saluran (meter)
5	*luas_penampung_m2*	Geometri	Luas penampung (m^2)
6	*keliling_penampung_m*	Geometri	Keliling penampung (meter)
7	*elevasi_hulu_m*	Elevasi	Elevasi titik hulu (mdpl)
8	*elevasi_hilir_m*	Elevasi	Elevasi titik hilir (mdpl)
9	*selisih_elevasi_m*	Elevasi	Selisih elevasi hulu-hilir (meter)
10	*slope_asli*	Elevasi	Kemiringan asli saluran
11	*catchment_area_m2*	Kapasitas	Luas area tangkapan (m^2)
12	*catchment_area_km2*	Kapasitas	Luas area tangkapan (km^2)
13	*luas_penampang_m2*	Kapasitas	Luas penampang saluran (m^2)
14	keliling_penampang_m	Kapasitas	Keliling penampang saluran (meter)
15	*catchment_density_m2perm*	Kapasitas	Kepadatan area tangkapan per meter
16	*beda_elevasi_m*	Turunan Elevasi	Beda elevasi terhitung (meter)
17	*slope_hitung*	Turunan Elevasi	Kemiringan terhitung
18	*tipe_saluran_encoded*	Kategorikal	Tertutup=2, Terbuka=1, Alami=0
19	*kondisi_fisik_encoded*	Kategorikal	Baik=3, Sedang=2, Buruk=1, Rusak=0
20	*sisi_saluran_encoded*	Kategorikal	Kiri=0,Kanan=1, Kiri&Kanan=2, Tengah=3
21	*rasio_lebar*	Turunan	*lebar_atas / lebar_bawah*
22	*kapasitas_per_meter*	Turunan	*luas_penampang / panjang_saluran*
23	*rasio_catchment_kapasitas*	Turunan	*catchment_area / luas_penampang*
24	*hydraulic_radius*	Turunan	*luas_penampang / keliling_penampang*
25	*volume_tampung_est*	Turunan	*luas_penampang* × *panjang_saluran*

No	Nama Fitur	Subkategori	Keterangan
26	*rasio_volume_catchment*	Turunan	*volume_tampung / catchment_area*

Tabel 3. 2 Daftar 19 Fitur Curah Hujan CHIRPS

No	Nama Fitur	Subkategori	Keterangan
1	*rainfall_mean*	Statistik Dasar	Rata-rata curah hujan harian (mm)
2	*rainfall_max*	Statistik Dasar	Curah hujan harian maksimum (mm)
3	*rainfall_total*	Statistik Dasar	Total akumulasi curah hujan (mm)
4	*rainfall_std*	Statistik Dasar	Standar deviasi curah hujan harian
5	*rainy_days*	Frekuensi	Jumlah hari hujan (≥1 mm)
6	*heavy_rain_days*	Frekuensi	Jumlah hari hujan deras (≥20 mm)
7	*very_heavy_rain_days*	Frekuensi	Jumlah hari hujan sangat deras (≥50 mm)
8	*max_consecutive_rain_days*	Frekuensi	Hari hujan berturut-turut maksimum
9	*dry_spell_max*	Frekuensi	Hari kering berturut-turut maksimum
10	*rainfall_intensity*	Intensitas	Rata-rata curah hujan pada hari hujan
11	*rainfall_monthly_max*	Intensitas	Curah hujan bulanan maksimum (mm)
12	*rainfall_7d_max*	Multi-temporal	Curah hujan akumulasi 7 hari maks (mm)
13	*rainfall_14d_max*	Multi-temporal	Curah hujan akumulasi 14 hari maks (mm)
14	*rainfall_30d_max*	Multi-temporal	Curah hujan akumulasi 30 hari maks (mm)
15	*rainfall_skewness*	Distribusi	*Skewness* distribusi curah hujan
16	*rainfall_kurtosis*	Distribusi	Kurtosis distribusi curah hujan
17	*rainfall_cv*	Distribusi	Koefisien variasi curah hujan
18	*rainfall_p90*	Distribusi	Persentil ke-90 curah hujan (mm)
19	*rainfall_p95*	Distribusi	Persentil ke-95 curah hujan (mm)

Tabel 3. 3 Daftar 8 Fitur Interaksi Drainase × Curah Hujan

No	Nama Fitur	Rumus Pembentukan	Keterangan
1	*drainage_deficit_index*	*rasio_catchment_kapasitas* × *rainfall_intensity*	Indeks defisit kapasitas drainase terhadap intensitas hujan (Chow *et al.*, 1988)
2	*flood_susceptibility*	(*heavy_rain_days* × *catchment_area*_km2) / *luas_penampang_m2*	Kerentanan banjir berbasis *hazard* dan *exposure* (Tehrany *et al_*, 2014; UNDRR, 2017)
3	*rainfall_slope_interaction*	rainfall_mean × *slope_hitung*	Adaptasi *Stream Power Index* untuk saluran drainase (Moore *et al.*, 1991; Beven & Kirkby, 1979)
4	*capacity_rainfall_ratio*	volume_tampung_est / (*rainfall_mean* × *catchment_area_m2*)	Rasio kapasitas tampung terhadap beban *design storm* (Chow *et al.*, 1988)
5	*drainage_efficiency*	*hydraulic_radius*^(2/3) × *slope_hitung*^(0,5)	Komponen efisiensi dari Persamaan Manning (Chow, 1959; Manning, 1891)
6	*heavy_rain_vulnerability*	*heavy_rain_days* / (*kondisi_fisik_encoded* + 1)	Kerentanan infrastruktur terhadap hujan lebat (UNDRR, 2017; IPCC, 2023)
7	*extreme_rain_exposure*	*rainfall_p95* / *kapasitas_per_meter*	Paparan curah hujan ekstrem P95 terhadap

No	Nama Fitur	Rumus Pembentukan	Keterangan
			kapasitas drainase (Böing *et al.*, 2020; UNDRR, 2017)
8	*drainage_stress_7d*	(*rainfall_7d_max* × *catchment_area*_m2 / 1000) / *volume_tampung_est*	Tekanan drainase berbasis *antecedent rainfall* 7 hari (Beven & Kirkby, 1979)

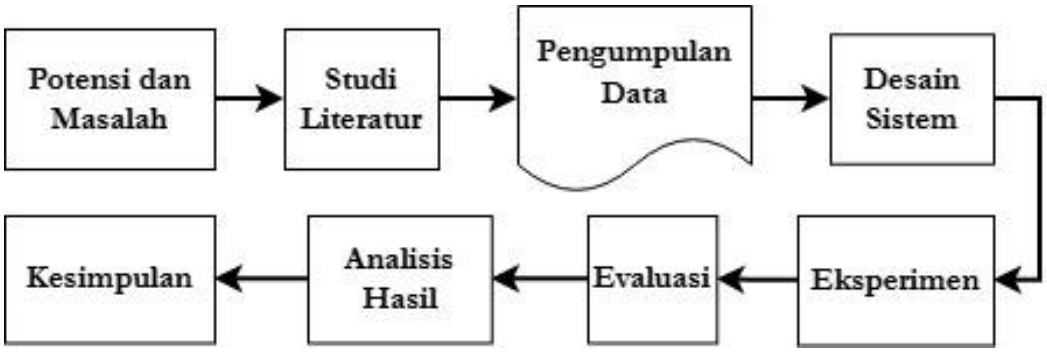
Konstruksi fitur dilakukan melalui dua tahap: *feature extraction* (mengonversi data mentah drainase dan CHIRPS menjadi 45 fitur numerik) dan *feature engineering* (merekayasa 8 fitur interaksi berbasis pengetahuan domain), menghasilkan total 53 fitur. Pada tahap *preprocessing*, dua teknik diterapkan: *feature selection* menggunakan *pipeline* MI + Pearson + RFE untuk mereduksi dimensi dari 53 fitur menjadi subset optimal, serta BorderlineSMOTE untuk menangani ketidakseimbangan distribusi kelas. Algoritma XGBoost digunakan sebagai model klasifikasi utama dengan konfigurasi *multi-class* (*softprob*) untuk tiga kelas risiko: Rendah, Sedang, dan Tinggi. Optimasi *hyperparameter* dilakukan menggunakan Optuna TPE dengan evaluasi *5-fold StratifiedKFold cross-validation* di dalam *training set*. Keluaran akhir berupa klasifikasi risiko banjir pada setiap titik drainase di wilayah Kota Malang.

Desain eksperimen menggunakan pendekatan studi ablasi dengan enam skenario (S1–S6) untuk mengisolasi kontribusi individual masing-masing komponen *preprocessing* dan optimasi. Skenario S1 merupakan *baseline* tanpa

preprocessing maupun optimasi, S2 menerapkan seleksi fitur saja, S3 menerapkan BorderlineSMOTE saja, S4 mengombinasikan seleksi fitur dan BorderlineSMOTE, S5 menerapkan Optuna saja, dan S6 merupakan *full pipeline* yang mengaktifkan seluruh komponen. Desain ablasi ini memungkinkan evaluasi kuantitatif terhadap efektivitas setiap komponen secara terpisah maupun gabungan.

3.2 Alur Penelitian

Penelitian ini dilaksanakan melalui delapan tahapan yang saling terhubung secara sekuensial, sebagaimana diilustrasikan pada Gambar 3.2.



Gambar 3. 2 Alur Penelitian

Alur penelitian pada Gambar 3.2 dimulai dari identifikasi potensi dan masalah, dilanjutkan dengan studi literatur untuk membangun landasan teoritis, pengumpulan data dari sumber primer dan sekunder, desain sistem klasifikasi, pelaksanaan eksperimen dengan enam skenario ablasi, evaluasi performa model, analisis hasil, dan diakhiri dengan penarikan kesimpulan serta rekomendasi. Setiap tahapan dijelaskan secara detail pada sub-bagian berikut.

3.2.1 Potensi dan Masalah

Kota Malang merupakan wilayah perkotaan yang mengalami peningkatan risiko banjir akibat perubahan tata guna lahan, pertumbuhan infrastruktur, dan variabilitas curah hujan. Potensi yang dapat dimanfaatkan adalah ketersediaan data drainase fisik dari survei lapangan yang mencakup seluruh jaringan saluran kota, serta data curah hujan satelit CHIRPS yang tersedia secara gratis dengan resolusi spasial 0,05 derajat dan resolusi temporal harian. Kedua sumber data ini memungkinkan pengembangan model klasifikasi risiko banjir yang mempertimbangkan aspek kapasitas infrastruktur drainase dan beban curah hujan secara simultan.

Permasalahan yang teridentifikasi meliputi tiga aspek. Pertama, bagaimana mengintegrasikan data drainase fisik yang bersifat spasial-statis dengan data curah hujan CHIRPS yang bersifat temporal menjadi satu *dataset* yang koheren untuk pelatihan model klasifikasi. Kedua, bagaimana merancang dan mengimplementasikan model klasifikasi XGBoost yang mampu mengklasifikasikan risiko banjir ke dalam tiga tingkatan (Rendah, Sedang, Tinggi) dengan akurasi tinggi pada kondisi distribusi kelas yang tidak seimbang. Ketiga, bagaimana mengevaluasi kontribusi individual dari masing-masing komponen *preprocessing* (seleksi fitur dan penanganan *imbalance*) serta optimasi *hyperparameter* terhadap performa model melalui studi ablasi yang sistematis.

3.2.2 Studi Literatur

Studi literatur dilakukan untuk mengkaji perkembangan terkini dalam bidang klasifikasi kerentanan banjir berbasis *machine learning*. Tinjauan

difokuskan pada tiga aspek: penggunaan *ensemble learning* (khususnya XGBoost) dalam pemetaan kerentanan banjir, metode *preprocessing* dan optimasi yang diterapkan, serta identifikasi *research gap*. Sepuluh jurnal ilmiah internasional yang diterbitkan antara tahun 2023-2025 dikaji, meliputi penelitian oleh Riazi *et al.* (2023), Wang *et al.* (2023), Yuan *et al.* (2024), Qin *et al.* (2025), Shrestha *et al.* (2025), Zhu *et al.* (2024), Bersabe & Jun (2025), Ren *et al.* (2024), Goumghar *et al.* (2025), dan Jang *et al.* (2025).

Hasil studi literatur menunjukkan bahwa XGBoost merupakan algoritma yang paling dominan, diadopsi oleh tujuh dari sepuluh studi yang ditinjau. Tiga *research gap* teridentifikasi: belum ada integrasi data drainase fisik dan CHIRPS pada XGBoost, belum ada studi ablasi sistematis, dan belum ada penggunaan *Bayesian optimization* (Optuna TPE) untuk XGBoost pada konteks banjir. *Gap* tersebut menjadi landasan dalam merancang metodologi penelitian ini.

3.2.3 Pengumpulan Data dan Konstruksi Fitur

Data yang digunakan dalam penelitian ini berasal dari dua sumber utama: data drainase fisik dan data curah hujan satelit CHIRPS. Proses transformasi data mentah menjadi fitur siap-pakai untuk model klasifikasi melibatkan tiga tahapan yang berbeda secara konseptual (Zheng & Casari, 2018):

1. *Feature extraction* mengonversi data mentah menjadi representasi numerik yang dapat diproses oleh model. Pada penelitian ini, *feature extraction* mencakup dua proses: derivasi 6 fitur hidraulik dari pengukuran geometri

saluran drainase (sub-bagian a), dan agregasi *time series* curah hujan harian CHIRPS selama 7 tahun menjadi 19 fitur statistik per titik drainase.

2. F*eature engineering*: mengonstruksi fitur baru dari fitur yang telah diekstraksi berdasarkan pengetahuan domain. Delapan fitur interaksi drainase × curah hujan direkayasa untuk menangkap hubungan antara kapasitas drainase dan beban curah hujan yang tidak terwakili oleh fitur individual.
3. *Feature selection*: mereduksi himpunan fitur dengan menyaring fitur yang tidak informatif dan redundan. *Pipeline* seleksi fitur MI + Pearson + RFE mereduksi 53 fitur menjadi subset.

Ketiga tahapan ini bersifat sekuensial: *feature extraction* menghasilkan 45 fitur dasar (26 drainase + 19 CHIRPS), *feature engineering* menambahkan 8 fitur interaksi menjadi total 53, dan *feature selection* memilih subset fitur yang paling informatif untuk pelatihan model.

3.2.3.1 Data Drainase Fisik Kota Malang

Data drainase diperoleh dari survei lapangan yang mencakup seluruh jaringan saluran drainase di wilayah Kota Malang. Dataset terdiri dari 3.471 titik drainase dengan atribut yang mencakup dimensi geometri saluran (panjang, tinggi, lebar atas, lebar bawah, luas penampang, keliling penampang), elevasi (hulu, hilir, selisih elevasi, *slope*), kapasitas hidraulik (*catchment area*, *hydraulic radius*, volume tampung estimasi), tipe saluran (Tertutup, Terbuka, Alami), kondisi fisik (Baik, Sedang, Buruk, Rusak), dan sisi saluran (Kiri, Kanan, Kiri dan Kanan, Tengah). Variabel kategorikal di-*encode* secara ordinal: kondisi fisik (Baik=3,

Sedang=2, Buruk=1, Rusak=0), tipe saluran (Tertutup=2, Terbuka=1, Alami=0), dan sisi saluran (Kiri=0, Kanan=1, Kiri dan Kanan=2, Tengah=3). Total 26 fitur diekstraksi dari data drainase, termasuk 6 fitur turunan yang dihitung dari variabel geometri dasar. Tiga fitur turunan utama yang menjadi komponen penting dalam fitur interaksi (sub-bagian c) didefinisikan sebagai berikut.

Hydraulic radius (Persamaan 3.1) merupakan rasio luas penampang basah terhadap keliling basah saluran, yang mencerminkan efisiensi geometris penampang dalam mengalirkan air (Chow, 1959):

$$R = \frac{A_p}{P_p} \tag{3.1}$$

dengan R adalah *hydraulic radius* (m), A_p adalah luas penampang basah (m^2), dan P_p adalah keliling penampang basah (m).

Volume tampung estimasi (Persamaan 3.2) dihitung sebagai hasil kali luas penampang dan panjang saluran, yang merepresentasikan kapasitas volumetrik total saluran:

$$V_{est} = A_p \times L \tag{3.2}$$

dengan V_{est} adalah volume tampung estimasi (m^3), A_p adalah luas penampang (m^2), dan L adalah panjang saluran (m).

Rasio *catchment*-kapasitas (Persamaan 3.3) mengukur beban area tangkapan relatif terhadap kapasitas penampang saluran (Chow *et al.*, 1988):

$$r_{ck} = \frac{A_c}{A_p} \tag{3.3}$$

dengan r_{ck} adalah rasio *catchment*-kapasitas, A_c adalah *catchment area* (m²), dan A_p adalah luas penampang saluran (m²). Nilai r_{ck} yang tinggi mengindikasikan beban tangkapan yang besar relatif terhadap kapasitas saluran.

Tiga fitur turunan lainnya (*rasio lebar*, *kapasitas per meter*, dan *rasio volume-catchment*) dihitung dari operasi aritmetika langsung antarvariabel geometri.

3.2.3.2 Data Curah Hujan Satelit CHIRPS

Data curah hujan diperoleh dari *Climate Hazards Group InfraRed Precipitation with Station data* (CHIRPS) versi 2.0 dengan resolusi spasial 0,05 derajat (~5,5 km) dan resolusi temporal harian dalam format NetCDF. Periode pengamatan mencakup tujuh tahun (2019-2025) pada area studi Kota Malang (lintang -8,05 hingga -7,90; bujur 112,56 hingga 112,70). Untuk setiap titik drainase, nilai curah hujan diekstraksi berdasarkan koordinat terdekat dan diolah menjadi 19 fitur statistik yang mencakup: statistik dasar (rata-rata, maksimum, total, standar deviasi), frekuensi kejadian (hari hujan ≥1 mm, hari hujan deras ≥20 mm, hari hujan sangat deras ≥50 mm, hari hujan berturut-turut maksimum, *dry spell* maksimum), intensitas (intensitas rata-rata, curah hujan bulanan maksimum), jendela multi-temporal (curah hujan maksimum 7, 14, dan 30 hari), serta distribusi (*skewness*, *kurtosis*, koefisien variasi, persentil ke-90 dan ke-95).

Beberapa fitur statistik distribusi yang digunakan didefinisikan secara formal sebagai berikut. Koefisien variasi (Persamaan 3.4) mengukur variabilitas curah hujan relatif terhadap rata-ratanya:

$$CV = \frac{\sigma_P}{\mu_P} \tag{3.4}$$

dengan CV adalah koefisien variasi, σ_P adalah standar deviasi curah hujan harian, dan μ_P adalah rata-rata curah hujan harian. Nilai CV yang tinggi mengindikasikan distribusi curah hujan yang tidak merata sepanjang periode pengamatan.

Skewness (Persamaan 3.5) mengukur asimetri distribusi curah hujan, di mana nilai positif yang tinggi mengindikasikan kejadian curah hujan ekstrem yang jarang namun intens:

$$\gamma = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{P_i - \mu_P}{\sigma_P}\right)^3 \tag{3.5}$$

dengan γ adalah *skewness*, N adalah jumlah hari pengamatan, P_i adalah curah hujan pada hari ke-i, μ_P adalah rata-rata, dan σ_P adalah standar deviasi.

Curah hujan maksimum jendela-w hari (Persamaan 3.6) menangkap akumulasi curah hujan pada periode berurutan yang paling ekstrem, yang relevan terhadap saturasi sistem drainase:

$$P_{w,max} = \max_{t} \sum_{i=t}^{t+w-1} P_i \tag{3.6}$$

dengan $P_{w,max}$ adalah curah hujan maksimum untuk jendela w hari (digunakan w = 7, 14, dan 30), dan P_i adalah curah hujan pada hari ke-i. Fitur ini relevan karena akumulasi curah hujan pada hari-hari berturut-turut meningkatkan saturasi tanah dan mengurangi kapasitas infiltrasi (Beven & Kirkby, 1979).

3.2.3.3 Pelabelan Kelas Risiko Banjir (*Risk Labeling*)

Setiap titik drainase diberi label kelas risiko banjir berdasarkan kedekatan spasial (*spatial proximity*) terhadap titik banjir historis. Pendekatan ini merupakan varian dari *distance-based sampling* yang lazim digunakan dalam pemodelan

kerentanan banjir (*flood susceptibility mapping*) untuk membedakan sampel positif dan negatif berdasarkan jarak spasial (Huang *et al.*, 2025; Zhang *et al.*, 2025).

Dalam literatur *flood susceptibility*, titik banjir historis diperoleh dari data kejadian banjir aktual, sedangkan titik non-banjir perlu dihasilkan melalui strategi *sampling* tertentu. Zhang *et al.* (2025) menunjukkan bahwa pemilihan sampel non-banjir merupakan sumber ketidakpastian utama dalam pemetaan kerentanan banjir, dan merekomendasikan penggunaan zona buffer di sekitar titik banjir untuk mengeliminasi titik non-banjir yang berpotensi sebagai titik banjir. Huang *et al.* (2025) mengembangkan kerangka *Distance-Based Sampling* (DBS) yang secara eksplisit menghubungkan klasifikasi sampel non-banjir dengan jarak Euclidean terhadap titik banjir terdekat, dan menemukan bahwa jarak spasial antara titik banjir dan non-banjir secara signifikan memengaruhi akurasi model kerentanan banjir perkotaan.

Penelitian ini mengadaptasi prinsip tersebut dengan menghitung jarak Haversine (d_H) dari setiap titik drainase ke titik banjir historis terdekat. Formula Haversine digunakan karena koordinat titik drainase berupa lintang-bujur pada permukaan bumi (Persamaan 3.7):

$$d_H = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \qquad (3.7)$$

dengan d_H adalah jarak Haversine (meter), R adalah jari-jari bumi (6.371.000 m), φ_1 dan φ_2 adalah lintang dua titik (radian), $\Delta\varphi = \varphi_2 - \varphi_1$, dan $\Delta\lambda = \lambda_2 - \lambda_1$ adalah selisih bujur (radian).

Berdasarkan jarak minimum ke titik banjir terdekat (d_{min}), setiap titik drainase diklasifikasikan ke dalam tiga kelas risiko:

1) **Kelas 2 (Tinggi)**: Titik banjir historis (label_banjir $= 1$), yaitu titik drainase yang terletak pada lokasi kejadian banjir aktual berdasarkan data BPBD Kota Malang.
2) **Kelas 1 (Sedang)**: Titik non-banjir dengan $d_{min} \leq$, yaitu titik drainase yang tidak tercatat mengalami banjir namun berada dalam zona pengaruh genangan. Drainase dalam radius tertentu dari titik banjir historis berbagi karakteristik topografi, elevasi, dan konektivitas jaringan drainase yang serupa, sehingga memiliki profil risiko intermediat (Teng *et al.* (2017); Merz *et al.* (2010))
3) **Kelas 0 (Rendah)**: Titik non-banjir dengan $d_{min} >$, yaitu titik drainase di luar zona pengaruh genangan yang secara spasial cukup jauh dari titik banjir historis mana pun.

Parameter δ (radius buffer zona pengaruh) ditetapkan sebesar 750 meter berdasarkan justifikasi berbasis data (*data-driven*) yang diperkuat oleh literatur kerentanan banjir.

Secara empiris, nilai δ ditentukan dari analisis distribusi jarak seluruh titik non-banjir terhadap titik banjir terdekatnya menggunakan formula Haversine (Persamaan 3.7). Tabel 3.4 menyajikan statistik deskriptif distribusi jarak tersebut.

Tabel 3. 4 Statistik Deskriptif Jarak Titik Non-Banjir ke Titik Banjir Terdekat

Statistik	**Nilai**
Jumlah titik non-banjir	3.170
Rata-rata	2.211 m
Median (P50)	1.386 m
Standar deviasi	5.495 m
Kuartil pertama (Q1/P25)	681 m
Kuartil ketiga (Q3/P75)	2.330 m
Minimum	0 m
Maksimum	49.608 m

Kuartil pertama (Q1) sebesar 681 meter menunjukkan bahwa 25% titik non-banjir memiliki jarak ≤681 meter ke titik banjir terdekat. Titik-titik pada kuartil ini merupakan titik non-banjir yang paling dekat dengan lokasi banjir historis dan paling mungkin terpengaruh oleh genangan. Nilai δ ditetapkan sebesar 750 meter, yaitu pembulatan di atas Q1, sehingga kelas Sedang mencakup 28,1% titik non-banjir terdekat (891 dari 3.170 titik). Pemilihan nilai mendekati Q1 memastikan bahwa kelas Sedang hanya merepresentasikan titik-titik non-banjir pada kuartil terdekat, sementara mayoritas (71,9%) titik non-banjir yang secara spasial lebih jauh diklasifikasikan sebagai Rendah.

Nilai δ = 750 meter ini juga konsisten dengan rentang buffer yang digunakan dalam literatur pemodelan kerentanan banjir. Zhang *et al.* (2025) menggunakan buffer 500 meter untuk mengeliminasi titik non-banjir yang berpotensi sebagai titik banjir pada pemodelan kerentanan banjir di DAS Zijiang, Tiongkok, dan menemukan bahwa kombinasi metode buffer dengan algoritma OC SVM meningkatkan akurasi model hingga 24%. Sharma & Sharma (2026) menetapkan jarak minimum 1.000 meter dari titik kejadian banjir untuk

pengambilan sampel non-banjir pada studi kerentanan banjir bandang di Himachal Pradesh, India. Huang *et al.* (2025) melalui kerangka DBS juga mengonfirmasi bahwa jarak spasial antara sampel banjir dan non-banjir berpengaruh signifikan terhadap akurasi dan estimasi kerentanan banjir. Dengan demikian, nilai δ = 750 meter berada dalam rentang 500–1.000 meter yang lazim digunakan dalam literatur kerentanan banjir, dan angka spesifiknya ditentukan secara *data-driven* berdasarkan distribusi jarak pada data penelitian ini.

Distribusi kelas yang dihasilkan menunjukkan ketidakseimbangan: Rendah mendominasi dengan 2.279 titik (65,7%), diikuti Sedang dengan 891 titik (25,7%), dan Tinggi dengan 301 titik (8,7%). Ketidakseimbangan ini ditangani pada tahap *preprocessing* menggunakan BorderlineSMOTE.

3.2.3.4 *Feature Engineering*: Fitur Interaksi Drainase × Curah Hujan

Tahap *feature engineering* mengonstruksi 8 fitur interaksi baru dari fitur-fitur yang telah diekstraksi pada tahap sebelumnya. Berbeda dengan *feature extraction* yang mengonversi data mentah menjadi representasi numerik, *feature engineering* secara eksplisit merekayasa fitur baru berdasarkan pengetahuan domain untuk menangkap hubungan antara kapasitas drainase dan beban curah hujan. Pendekatan ini mengikuti preseden konstruksi fitur turunan berbasis pengetahuan domain yang telah mapan dalam hidrologi, sebagaimana TWI (Beven & Kirkby, 1979) dan SPI (Moore *et al.*, 1991) yang menggabungkan variabel dasar menjadi indeks bermakna secara hidrologis (dibahas pada Bab II). Dalam konteks penelitian ini, variabel drainase dan curah hujan masing-masing hanya menangkap satu sisi dari sistem hidrologi; interaksi antara keduanya, yaitu seberapa besar beban

curah hujan relatif terhadap kapasitas drainase, merupakan informasi kritis yang tidak terwakili oleh fitur individual.

Berikut penjelasan dasar konseptual dan rujukan ilmiah untuk setiap fitur interaksi pada Tabel 3.3:

1) ***drainage_deficit_index*** (*rasio_catchment_kapasitas* × *rainfall_intensity*). Fitur ini mengkuantifikasi defisit kapasitas drainase relatif terhadap intensitas curah hujan. Chow *et al.* (1988) menjelaskan bahwa risiko banjir terjadi ketika debit aliran masuk melebihi kapasitas tampung dan pengaliran saluran drainase; semakin besar rasio beban tangkapan terhadap kapasitas, dan semakin tinggi intensitas hujan, maka semakin besar defisit kapasitas. Penamaan *deficit index* merujuk pada terminologi standar dalam hidrologi untuk menggambarkan ketidakcukupan kapasitas sistem (Chow *et al.*, 1988).
2) ***flood_susceptibility*** ((*heavy_rain_days* × *catchment_area_km2*) / *luas_penampang_m2*). Fitur ini mengukur kerentanan banjir berdasarkan frekuensi hujan lebat dan rasio luas tangkapan terhadap penampang saluran. Tehrany *et al.* (2014) mendefinisikan *flood susceptibility* sebagai kecenderungan inheren suatu area mengalami banjir berdasarkan *conditioning factors* yang bersifat intrinsik. Fitur ini menangkap komponen *hazard* (frekuensi hujan lebat) dan *exposure* (rasio tangkapan terhadap kapasitas penampang) dari kerangka risiko bencana United Nations Office for Disaster Risk Reduction (2017) yang menyatakan risiko sebagai fungsi dari *hazard*, *exposure*, dan *vulnerability*.

3) ***rainfall_slope_interaction*** (*rainfall_mean* × *slope_hitung*). Fitur ini mengadaptasi logika *Stream Power Index* (SPI = $A/s \times \tan \beta$) yang diformulasikan oleh Moore *et al.* (1991), di mana SPI mengukur daya erosif aliran berdasarkan kontribusi area tangkapan dan kemiringan. Dalam konteks saluran drainase perkotaan, luas tangkapan topografis diganti dengan curah hujan rata-rata sebagai proksi beban hidrologi, karena curah hujan merupakan variabel pengendali utama volume limpasan yang masuk ke saluran. Konsep interaksi curah hujan dan kemiringan terhadap pembangkitan limpasan juga didukung oleh Beven & Kirkby (1979) melalui *Topographic Wetness Index* (TWI = $\ln(A/s / \tan \beta)$).
4) ***capacity_rainfall_ratio*** (*volume_tampung_est* / (*rainfall_mean* × *catchment_area_m2*)). Fitur ini merupakan rasio kapasitas tampung saluran terhadap volume hujan harian rata-rata yang masuk ke sistem drainase. Chow *et al.* (1988) menjelaskan bahwa kapasitas saluran drainase harus dirancang untuk menampung debit limpasan dari curah hujan rancangan (*design storm*) pada periode ulang tertentu; rasio ini merepresentasikan *factor of safety* sistem drainase di mana nilai > 1 menunjukkan kapasitas memadai dan nilai < 1 menunjukkan kapasitas tidak memadai.
5) ***drainage_efficiency*** (*hydraulic_radius*^(2/3) × *slope_hitung*^(0,5)). Fitur ini merupakan turunan langsung dari Persamaan Manning (Manning, 1891): $V = (1/n) \times R^{(2/3)} \times S^{(1/2)}$, di mana R adalah radius hidraulik dan S adalah kemiringan dasar saluran. Chow (1959) mendefinisikan radius hidraulik sebagai rasio luas penampang basah terhadap keliling basah (R =

A/*P*), yang merupakan ukuran efisiensi geometris penampang saluran. Komponen *R*^(2/3) × *S*^(1/2) merepresentasikan efisiensi geometris dan gravitasional saluran dalam mengalirkan air tanpa koefisien kekasaran *n*, yang diasumsikan relatif konstan untuk tipe saluran yang sama.

6) ***heavy_rain_vulnerability*** (*heavy_rain_days* / (*kondisi_fisik_encoded* + 1)). Fitur ini memodelkan kerentanan titik drainase terhadap hujan lebat dengan mempertimbangkan kondisi fisik infrastruktur. UNDRR (2017) mendefinisikan *vulnerability* sebagai kondisi yang menyebabkan kerentanan terhadap bahaya, termasuk degradasi infrastruktur. Saluran dalam kondisi buruk atau rusak memiliki kapasitas fungsional yang lebih rendah akibat sedimentasi dan kerusakan struktural, sehingga frekuensi hujan lebat yang sama menghasilkan risiko yang lebih tinggi pada saluran berkondisi buruk dibandingkan saluran berkondisi baik. IPCC (2023) mengidentifikasi peningkatan frekuensi hujan ekstrem sebagai faktor utama peningkatan risiko banjir perkotaan.

7) ***extreme_rain_exposure*** (*rainfall_p95* / *kapasitas_per_meter*). Fitur ini mengukur paparan titik drainase terhadap curah hujan ekstrem (persentil ke-95) relatif terhadap kapasitas drainase per meter panjang saluran. Böing *et al.* (2020) menerapkan pendekatan berbasis persentil untuk konstruksi skenario curah hujan dalam prakiraan banjir permukaan, di mana P95 digunakan sebagai *threshold* curah hujan ekstrem. United Nations Office for Disaster Risk Reduction (2017) mendefinisikan *exposure* sebagai situasi di mana aset atau infrastruktur berada dalam zona bahaya dan berpotensi

mengalami kerugian; fitur ini mengkuantifikasi seberapa besar tekanan per unit panjang saluran saat terjadi curah hujan di atas persentil ke-95.

8) ***drainage_stress_7d*** ((*rainfall_7d_max* × *catchment_area_m2* / 1000) / *volume_tampung_est*). Fitur ini mengukur tekanan pada sistem drainase selama periode hujan puncak 7 hari berturut-turut. Beven & Kirkby (1979) menunjukkan bahwa tingkat saturasi tanah (*antecedent wetness*) secara langsung menentukan luas area yang berkontribusi terhadap limpasan permukaan; akumulasi hujan pada hari-hari sebelumnya meningkatkan saturasi dan memperbesar volume limpasan pada hujan berikutnya. Rumus fitur ini mengonversi curah hujan akumulatif 7 hari menjadi volume air masuk (m^3) dan membandingkannya dengan volume tampung saluran, di mana rasio > 1 menunjukkan volume air masuk melebihi kapasitas tampung.

Total keseluruhan fitur sebelum *feature selection* adalah 53, yang terdiri dari 45 fitur hasil *feature extraction* (26 drainase + 19 CHIRPS) dan 8 fitur hasil *feature engineering* (interaksi drainase × curah hujan). Proses *feature selection* untuk mereduksi 53 fitur ini menjadi subset optimal dibahas pada sub-bagian 3.2.4b.

3.2.4 Desain Sistem

Desain sistem klasifikasi risiko banjir terdiri dari empat komponen utama yang saling terhubung: *preprocessing*, model klasifikasi, optimasi, dan evaluasi.

3.2.4.1 Pembagian Data dan Strategi Evaluasi

Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* pada seluruh 3.471 sampel. Dataset dibagi menjadi 5 partisi (*fold*)

dengan distribusi kelas yang proporsional pada setiap fold. Pada setiap iterasi, 4 fold digunakan sebagai *training set* dan 1 fold sebagai *validation set*. Proses ini diulangi sebanyak 5 kali pengulangan (*repeat*) dengan pengacakan fold yang berbeda pada setiap pengulangan, sehingga total diperoleh 25 evaluasi (5 fold × 5 pengulangan). Seluruh tahapan *preprocessing*, yaitu seleksi fitur (MI + Pearson + RFE) dan BorderlineSMOTE, dilakukan di dalam *training fold* saja pada setiap iterasi untuk mencegah *data leakage*. Metrik performa dilaporkan sebagai rata-rata ± standar deviasi dari 25 evaluasi, yang memberikan estimasi performa yang lebih *robust* dibandingkan *single-run* 5-fold CV karena memperhitungkan variabilitas akibat pengacakan fold yang berbeda.

Konfigurasi 5-fold dipilih berdasarkan pertimbangan stabilitas estimasi kelas minoritas. Dengan 3.471 sampel dan kelas Tinggi berjumlah 301 sampel (8,7%), skema 5-fold menghasilkan *validation fold* berisi ~694 sampel dengan ~60 sampel kelas Tinggi. Ukuran ini memastikan bahwa satu kesalahan klasifikasi pada kelas Tinggi hanya mengubah *Recall* sebesar ~1,67% (1/60), sehingga estimasi metrik per kelas tetap stabil. Rasio *training-validation* 80/20 per fold juga memberikan data pelatihan yang cukup besar (2.776 sampel) agar model dapat belajar pola distribusi ketiga kelas secara memadai. Dengan 5 pengulangan, total 25 evaluasi sudah memenuhi syarat minimum untuk pengujian signifikansi statistik menggunakan uji Wilcoxon *signed-rank* antarskenario.

Selain *cross-validation*, satu kali pembagian data 80/20 (*stratified split*, *random state* 42) juga dilakukan untuk keperluan analisis diagnostik yang memerlukan satu *test set* tetap. Analisis diagnostik bertujuan mengidentifikasi pola

kesalahan model melalui *confusion matrix* detail, mengevaluasi performa pada setiap kelas risiko, dan menginterpretasi kontribusi fitur melalui SHAP. Pembagian ini menghasilkan 2.776 sampel untuk pelatihan (Rendah: 1.823, Sedang: 712, Tinggi: 241) dan 695 sampel untuk pengujian (Rendah: 456, Sedang: 179, Tinggi: 60).

3.2.4.2 *Feature Selection* (MI + Pearson + RFE)

Feature selection merupakan tahap ketiga dalam *pipeline* konstruksi fitur. Seleksi fitur dilakukan melalui *pipeline* tiga tahap untuk mereduksi dimensi dari 53 fitur menjadi subset optimal dengan menyaring fitur yang tidak informatif terhadap target, menghilangkan redundansi antar fitur, dan memilih fitur terbaik berdasarkan performa model.

Tahap pertama menghitung *Mutual Information* (MI) antara setiap fitur dan variabel target. MI mengukur besarnya ketergantungan antara dua variabel acak, yaitu seberapa banyak informasi yang diberikan suatu fitur tentang kelas target. MI didefinisikan sebagai berikut (Persamaan 3.8):

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y)\, log\, \frac{p(x,y)}{p(x)\, p(y)} \tag{3.8}$$

dengan $\mathrm{I}(\mathrm{X};\mathrm{Y})$ adalah *Mutual Information* antara fitur X dan target Y, $\mathrm{p}(\mathrm{x},\mathrm{y})$ adalah distribusi probabilitas gabungan, dan $\mathrm{p}(\mathrm{x})$ serta $\mathrm{p}(\mathrm{y})$ adalah distribusi probabilitas marginal masing-masing variabel. Nilai $\mathrm{I}(\mathrm{X};\mathrm{Y}) = 0$ berarti fitur X tidak memiliki informasi apa pun tentang target Y (independen secara statistik), sedangkan nilai yang lebih tinggi mengindikasikan ketergantungan yang lebih kuat.

Pada penelitian ini, fitur dengan $I(X;Y) \leq 0{,}05$ dieliminasi, yaitu fitur yang memiliki informasi sangat rendah terhadap variabel target. MI mengevaluasi relevansi setiap fitur terhadap target secara independen, sehingga fitur yang saling berkorelasi tinggi tetapi informatif terhadap target tetap dipertahankan.

Tahap kedua menghitung koefisien korelasi Pearson antar fitur untuk menghilangkan redundansi. Korelasi Pearson mengukur kekuatan hubungan linier antara dua variabel kontinu dan didefinisikan sebagai berikut (Persamaan 3.9):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3.9)$$

dengan r_{xy} adalah koefisien korelasi Pearson antara fitur x dan y, n adalah jumlah sampel, $\bar{x}$ dan $\bar{y}$ adalah rata-rata masing-masing fitur. Nilai r_{xy} berkisar antara -1 hingga +1, di mana nilai mendekati ±1 mengindikasikan hubungan linier yang kuat. Pada penelitian ini, salah satu fitur dari pasangan yang memiliki $|r_{xy}| > 0{,}95$ dieliminasi.

Tahap ketiga menerapkan *Recursive Feature Elimination* (RFE) menggunakan XGBoost sebagai estimator. RFE merupakan metode *wrapper-based feature selection* yang memanfaatkan performa model sebagai kriteria seleksi fitur (Guyon *et al.*, 2002). Algoritma ini bekerja secara iteratif: pada setiap iterasi, model dilatih menggunakan himpunan fitur saat ini, kemudian fitur dengan *importance* terendah dieliminasi. Proses berlanjut hingga jumlah fitur mencapai target minimum yang ditentukan. Secara formal, pada iterasi ke-t, RFE menghitung *feature importance score* $\phi_j^{(t)}$ untuk setiap fitur j dalam himpunan fitur aktif $S^{(t)}$ (Persamaan 3.10):

$$\phi_j^{(t)} = \sum_{tree} \sum_{\text{split on } j} \Delta\mathcal{L}_{split} \tag{3.10}$$

dengan $\phi_j^{(t)}$ adalah skor *importance* fitur j pada iterasi ke-t, yang dihitung sebagai total penurunan fungsi kerugian (*loss reduction*) dari seluruh pemisahan yang menggunakan fitur j di seluruh pohon keputusan dalam model XGBoost. Fitur dengan skor terendah dieliminasi pada setiap iterasi (Persamaan 3.11):

$$S^{(t+1)} = S^{(t)} \setminus \left\{ \arg\min_{j \in S^{(t)}} \phi_j^{(t)} \right\} \tag{3.11}$$

dengan $S^{(t)}$ dan $S^{(t+1)}$ adalah himpunan fitur aktif pada iterasi ke-t dan ke-$(t+1)$. Proses berlanjut hingga $|S^{(t)}|$ mencapai target minimum 15 fitur. Pendekatan RFE memastikan bahwa hanya fitur yang paling informatif terhadap prediksi model yang dipertahankan, karena seleksi dilakukan berdasarkan kontribusi aktual fitur dalam model, bukan hanya berdasarkan korelasi statistik bivariat.

3.2.4.3 Penanganan Ketidakseimbangan Kelas (BorderlineSMOTE)

Ketidakseimbangan distribusi kelas merupakan permasalahan umum pada dataset klasifikasi risiko banjir. Pada dataset penelitian ini, kelas Rendah mendominasi dengan 65,7% sedangkan kelas Tinggi hanya 8,7%, yang dapat menyebabkan model cenderung bias terhadap kelas mayoritas. *Synthetic Minority Over-sampling Technique* (SMOTE) mengatasi hal ini dengan menghasilkan sampel sintetis baru untuk kelas minoritas melalui interpolasi linier antara sampel

minoritas dan tetangga terdekatnya (Chawla *et al.*, 2002). Secara formal, proses pembentukan sampel sintetis SMOTE dinyatakan dalam Persamaan 3.12:

$$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i) \quad (3.12)$$

dengan x_{new} adalah sampel sintetis baru, x_i adalah sampel minoritas yang dipilih, x_{nn} adalah salah satu dari *k* tetangga terdekat (k= 5) dari kelas minoritas yang sama, dan λ adalah bilangan acak pada interval [0,1].

Penelitian ini menggunakan varian BorderlineSMOTE yang hanya menghasilkan sampel sintetis dari instans kelas minoritas yang berada di sekitar *decision boundary*, yaitu sampel yang setidaknya setengah dari tetangga terdekatnya berasal dari kelas yang berbeda. Pendekatan ini lebih selektif dibandingkan SMOTE standar karena memfokuskan *oversampling* pada daerah yang paling kritis bagi pemisahan kelas, sehingga mengurangi risiko *noise* dari sampel sintetis yang dihasilkan di area yang sudah jelas milik satu kelas.

3.2.4.4 Model Klasifikasi (XGBoost)

XGBoost (eXtreme Gradient Boosting) merupakan implementasi efisien dari algoritma *Gradient Boosting Machine* (GBM) yang dikembangkan oleh Chen & Guestrin (2016). Algoritma ini termasuk dalam varian yang efisien dan terukur dari GBM yang telah banyak diterapkan dalam *computer vision, data mining*, dan bidang lainnya. XGBoost membangun model prediktif secara aditif melalui sekumpulan pohon keputusan (*decision tree*) yang dapat digunakan baik untuk regresi maupun klasifikasi. XGBoost telah ditingkatkan dalam dua aspek utama dibandingkan GBM konvensional: percepatan konstruksi pohon melalui teknik

approximate greedy algorithm dan *weighted quantile sketch*, serta dukungan algoritma terdistribusi untuk pencarian pohon yang optimal (Chen & Guestrin, 2016).

Pada algoritma XGBoost, pohon keputusan dibangun secara sekuensial dengan setiap pohon berfungsi untuk memperkirakan variabel target. Pohon pertama memprediksi nilai target secara langsung, pohon kedua memprediksi selisih antara nilai aktual dan estimasi pohon pertama, pohon ketiga memprediksi selisih antara nilai aktual dan gabungan estimasi pohon pertama dan kedua, dan seterusnya hingga residual diminimalkan (Chen & Guestrin, 2016). Nilai prediksi akhir ditentukan oleh penjumlahan keluaran dari K pohon keputusan sebagaimana dinyatakan dalam Persamaan 3.13:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{3.13}$$

dengan $\widehat{y_i}$ adalah prediksi model untuk data ke-i, K adalah jumlah total pohon keputusan, $f_k(x_i)$ adalah fungsi prediksi dari pohon ke-k untuk input x_i, dan $\mathcal{F}$ adalah ruang fungsi dari semua pohon keputusan yang mungkin. Setiap pohon f_k merupakan *regression tree* yang memetakan input ke skor pada *leaf node*.

Untuk meminimalkan kesalahan prediksi, XGBoost mendefinisikan fungsi objektif (*loss function*) dengan regularisasi sebagaimana dinyatakan dalam Persamaan 3.14:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3.14)$$

Dengan $\mathcal{L}$ adalah fungsi kerugian total, $l(y_i, \hat{y}_i)$ adalah fungsi kerugian untuk sampel ke-i yang mengukur perbedaan antara nilai aktual y_i dan prediksi $\hat{y}_i$, n adalah jumlah sampel, dan $\Omega(f_k)$ adalah *penalty* regularisasi untuk pohon ke-k. *Term* regularisasi didefinisikan sebagai berikut (Persamaan 3.15):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (3.15)$$

Dengan γ adalah parameter kompleksitas minimum untuk pemisahan (*minimum loss reduction*), T adalah jumlah *leaf node* pada pohon, λ adalah koefisien regularisasi L2, dan w_j adalah bobot (*score*) pada *leaf node* ke-j. Regularisasi ini mencegah *overfitting* dengan memberikan *penalty* terhadap pohon yang terlalu kompleks (banyak *leaf*) atau memiliki bobot *leaf* yang terlalu besar.

Pada setiap iterasi, XGBoost mengoptimasi fungsi objektif menggunakan pendekatan aproksimasi orde kedua Taylor. Pendekatan ini memungkinkan perhitungan yang lebih efisien dibandingkan optimasi langsung terhadap fungsi kerugian asli. Gradien orde pertama (g_i) dan orde kedua (*Hessian*, h_i) dihitung untuk setiap sampel (Persamaan 3.16 dan 3.17):

$$g_i = \frac{\partial l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \widehat{y_i^{(t-1)}}} \tag{3.16}$$

$$h_i = \frac{\partial^2 l\left(y_i, \widehat{y_i^{(t-1)}}\right)}{\partial \left(\widehat{y_i^{(t-1)}}\right)^2} \tag{3.17}$$

dengan $\widehat{y_i^{(t-1)}}$ adalah prediksi kumulatif hingga iterasi ke-$(t-1)$. Berdasarkan *gradien* tersebut, bobot optimal untuk setiap *leaf node* j dihitung menggunakan Persamaan 3.18:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.18}$$

dengan I_j adalah himpunan indeks sampel yang jatuh pada *leaf node* ke-j, dan λ adalah koefisien regularisasi L2. Skor pemisahan (*split gain*) untuk menentukan pemisahan optimal pada setiap simpul pohon dihitung sebagai berikut (Persamaan 3.19):

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{3.19}$$

dengan I_L dan I_R adalah himpunan indeks sampel pada *child node* kiri dan kanan setelah pemisahan, $I = I_L \cup I_R$ adalah himpunan indeks pada simpul induk, dan γ adalah *minimum loss reduction* yang diperlukan agar pemisahan dilakukan. Jika $\text{Gain} \leq 0$,pemisahan tidak menguntungkan dan simpul menjadi *leaf node*.

Mekanisme ini berfungsi sebagai *built-in pruning* yang mencegah pohon tumbuh terlalu dalam.

Pada penelitian ini, XGBoost dikonfigurasi untuk klasifikasi *multi-class* dengan *objective multi:softprob* yang menghasilkan probabilitas untuk setiap kelas menggunakan fungsi *softmax* (Persamaan 3.20):

$$P(y = c \mid x) = \frac{e^{F_c(x)}}{\sum_{j=1}^{C} e^{F_j(x)}} \tag{3.20}$$

dengan $P(y = c \mid x)$ adalah probabilitas data x termasuk kelas c, $F_c(x)$ adalah skor kumulatif dari seluruh pohon untuk kelas c, dan C = 3 adalah jumlah kelas. *Default hyperparameter* yang digunakan: *n_estimators* = 200, *max_depth* = 6, *learning_rate* = 0,1, *subsample* = 0,8, *colsample_bytree* = 0,8, *min_child_weight* = 3, γ= 0,1, *reg_alpha* (α) = 0,1, dan *reg_lambda* (λ) = 1,0.

3.2.4.5 Optimasi *Hyperparameter* (Optuna TPE)

Optimasi *hyperparameter* bertujuan menemukan konfigurasi parameter model yang menghasilkan performa terbaik. Proses pencarian *hyperparameter* dapat dirumuskan sebagai masalah optimasi (Persamaan 3.21):

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta) \tag{3.21}$$

dengan θ^* adalah konfigurasi *hyperparameter* optimal, Θ adalah ruang pencarian *hyperparameter*, dan $f(\theta)$ adalah fungsi objektif (F1 Macro *cross-validation*) yang ingin dimaksimalkan.

Penelitian ini menggunakan *framework* Optuna dengan *sampler* TPE (*Tree-structured Parzen Estimator*), sebuah pendekatan *Bayesian optimization* yang dikembangkan oleh Bergstra *et al.* (2011). Berbeda dengan *Grid Search* yang mengevaluasi seluruh kombinasi secara *brute-force* atau *Random Search* yang memilih secara acak, TPE memodelkan distribusi *hyperparameter* secara probabilistik berdasarkan hasil evaluasi sebelumnya. TPE membagi hasil observasi menjadi dua kelompok berdasarkan *threshold* y^* (biasanya kuantil terbaik), kemudian membangun dua distribusi terpisah (Persamaan 3.22):

$$p(\boldsymbol{\theta} \mid y) = \begin{cases} l(\boldsymbol{\theta}) & \text{jika } y < y^* \\ g(\boldsymbol{\theta}) & \text{jika } y \geq y^* \end{cases} \tag{3.22}$$

dengan $l(\theta)$ adalah distribusi probabilitas *hyperparameter* yang menghasilkan performa baik (di atas *threshold* y^*) dan $g(\theta)$ adalah distribusi *hyperparameter* yang menghasilkan performa buruk. TPE memilih *hyperparameter* berikutnya dengan memaksimalkan rasio *Expected Improvement* (EI) yang secara proporsional sebanding dengan $l(\theta)/g(\theta)$ (Persamaan 3.23):

$$\mathrm{EI}(\theta) \propto \frac{l(\theta)}{g(\theta)} \tag{3.23}$$

Dengan memaksimalkan rasio ini, TPE secara iteratif mengarahkan pencarian ke daerah yang paling menjanjikan dalam ruang *hyperparameter*, sehingga konvergensi ke konfigurasi optimal dicapai dengan jumlah evaluasi yang lebih sedikit dibandingkan metode non-adaptif.

Evaluasi setiap *trial* menggunakan *k-fold Stratified Cross-Validation* untuk memperoleh estimasi performa yang *robust* dan mengurangi varians evaluasi. Pada pendekatan ini, *training set* dibagi menjadi k partisi (*fold*) dengan distribusi kelas yang proporsional pada setiap *fold*. Model dilatih pada $k-1$ *fold* dan dievaluasi pada 1 *fold* yang tersisa, diulangi sebanyak k kali. Skor *cross-validation* dihitung sebagai rata-rata performa dari seluruh *fold* (Persamaan 3.24):

$$\mathrm{CV}(\theta) = \frac{1}{k}\sum_{i=1}^{k} f_i(\theta) \tag{3.24}$$

Dengan $\mathrm{CV}(\theta)$ adalah skor *cross-validation* untuk konfigurasi θ, k= 5 adalah jumlah *fold*, dan $f_i(\theta)$ adalah skor F1 Macro pada *fold* ke-i. Penggunaan *StratifiedKFold* memastikan bahwa distribusi kelas pada setiap fold merepresentasikan distribusi kelas pada dataset keseluruhan, yang penting pada kondisi ketidakseimbangan kelas.

Ruang pencarian mencakup sembilan *hyperparameter*: *n_estimators* [100-500], *max_depth* [3-10], *learning_rate* [0,01-0,3] (skala logaritmik), *subsample* [0,6-1,0], *colsample_bytree* [0,6-1,0], *min_child_weight* [1-10], *gamma* [0,0-0,5], *reg_alpha* [10^{-8}-10,0] (skala logaritmik), dan *reg_lambda* [10^{-8}-10,0] (skala logaritmik). Proses optimasi dibatasi maksimum 200 *trial* atau 1.800 detik.

3.2.5 Eksperimen

Eksperimen dirancang sebagai studi ablasi dengan enam skenario (S1–S6) untuk mengevaluasi kontribusi individual dan gabungan dari setiap komponen *preprocessing* dan optimasi. Tabel 3.4 menyajikan konfigurasi setiap skenario.

Tabel 3. 5 Konfigurasi Skenario Eksperimen

Skenario	**Nama**	**Seleksi Fitur**	**Borderline SMOTE**	**Optuna TPE**
S1	*Baseline*			
S2	*Feature Selection Only*	✓		
S3	BorderlineSMOTE *Only*		✓	
S4	*Feature Selection* + BorderlineSMOTE	✓	✓	
S5	*Optuna Only*			✓
S6	*Full Pipeline*	✓	✓	✓

Skenario S1 berfungsi sebagai *baseline* yang menggunakan seluruh 53 fitur dengan *default hyperparameter* tanpa *preprocessing* apa pun. S2 mengisolasi efek seleksi fitur (MI + Pearson + RFE) terhadap performa model. S3 mengisolasi efek penanganan ketidakseimbangan kelas (BorderlineSMOTE). S4 mengevaluasi efek gabungan seleksi fitur dan BorderlineSMOTE. S5 mengisolasi efek optimasi *hyperparameter* (Optuna TPE). S6 merupakan *full pipeline* yang mengaktifkan seluruh komponen secara bersamaan. Seluruh skenario dievaluasi menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* yang identik (*random state* = 42) untuk memastikan perbandingan yang adil. Satu kali *split* 80/20 dengan *random state* yang sama juga digunakan pada setiap skenario untuk menghasilkan *confusion matrix* detail dan analisis SHAP. Seleksi fitur menggunakan *pipeline* MI + Pearson + RFE sesuai penjelasan pada §3.2.4.

3.2.6 Evaluasi

Evaluasi performa model dilakukan menggunakan tiga metrik utama (*Accuracy*, F1 Macro, dan ROC-AUC) yang dihitung melalui *Repeated 5×5 Stratified K-Fold Cross-Validation* (25 evaluasi). Setiap metrik dilaporkan sebagai rata-rata ± standar deviasi dari 25 evaluasi untuk memberikan estimasi performa yang *robust* dan mengukur variabilitas antarpartisi data. Analisis *confusion matrix* dan performa per kelas diperoleh dari satu kali evaluasi pada *test set* (695 sampel) untuk keperluan analisis diagnostik. Sebelum menjelaskan masing-masing metrik, perlu dipahami terlebih dahulu empat komponen dasar *confusion matrix* untuk setiap kelas c: *True Positive* (TP_c) yaitu sampel kelas c yang diprediksi benar sebagai kelas c; *False Positive* (FP_c) yaitu sampel bukan kelas c yang salah diprediksi sebagai kelas c; *False Negative* (FN_c) yaitu sampel kelas c yang salah diprediksi sebagai kelas lain; dan *True Negative* (TN_c) yaitu sampel bukan kelas c yang diprediksi benar sebagai bukan kelas c.

1. ***Accuracy*** mengukur proporsi prediksi yang benar terhadap total prediksi (Persamaan 3.25):

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{3.25}$$

dengan $C = 3$ adalah jumlah kelas dan N adalah total sampel pada *validation fold* (bervariasi per *fold*, rata-rata ~694 sampel). Meskipun intuitif, metrik ini dapat menyesatkan pada dataset yang tidak seimbang karena model yang selalu memprediksi kelas mayoritas tetap memperoleh

akurasi tinggi. Sebagai contoh, model yang memprediksi seluruh sampel sebagai kelas Rendah akan memperoleh akurasi ~65,7% tanpa kemampuan diskriminatif apa pun.

2. ***Precision*** mengukur ketepatan prediksi positif untuk setiap kelas, yaitu proporsi prediksi kelas c yang benar dari seluruh prediksi kelas c (Persamaan 3.26):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{3.26}$$

Precision yang tinggi menunjukkan bahwa ketika model memprediksi suatu titik drainase berisiko tinggi, prediksi tersebut memiliki kemungkinan besar benar. Metrik ini penting dalam konteks alokasi sumber daya penanggulangan banjir karena meminimalkan intervensi yang tidak diperlukan (*false alarm*).

3. ***Recall*** (*Sensitivity*) mengukur kemampuan model mendeteksi seluruh sampel kelas c yang sebenarnya (Persamaan 3.27):

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{3.27}$$

Recall yang tinggi menunjukkan bahwa model mampu menangkap sebagian besar titik drainase yang benar-benar berisiko tinggi. Dalam konteks mitigasi bencana banjir, *recall* yang tinggi pada kelas Tinggi sangat kritis karena kegagalan mendeteksi titik berisiko tinggi (*false negative*) berpotensi menyebabkan kerugian yang besar.

4. **F1 *Score*** merupakan rata-rata harmonik dari *Precision* dan *Recall* yang memberikan keseimbangan antara kedua metrik tersebut (Persamaan 3.28):

$$F1_c = 2 \cdot \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3.28}$$

Rata-rata harmonik dipilih karena memberikan penalti lebih besar terhadap nilai yang rendah dibandingkan rata-rata aritmetik, sehingga F1 *score* yang tinggi hanya tercapai jika kedua komponen (*Precision* dan *Recall*) memiliki nilai yang tinggi secara bersamaan.

5. **F1 Macro** dihitung sebagai rata-rata tidak berbobot dari F1 *score* setiap kelas (Persamaan 3.29):

$$F1_{\text{Macro}} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{3.29}$$

dengan C = 3 adalah jumlah kelas. F1 Macro memberikan bobot yang sama kepada ketiga kelas terlepas dari jumlah sampelnya, sehingga performa pada kelas minoritas (Tinggi, 8,7% dari total data) memiliki pengaruh yang sama besar dengan kelas mayoritas (Rendah, 65,7%). F1 Macro dipilih sebagai metrik utama karena mampu mengevaluasi keseimbangan performa model pada seluruh kelas, termasuk kelas minoritas.

6. **ROC-AUC** (*Receiver Operating Characteristic - Area Under Curve*) mengukur kemampuan diskriminatif model dalam membedakan antarkelas pada berbagai *threshold* klasifikasi. Kurva ROC dibentuk dengan memplot

True Positive Rate ($TPR = \text{Recall}$) terhadap *False Positive Rate* (FPR) pada berbagai nilai *threshold* (Persamaan 3.30):

$$FPR_c = \frac{FP_c}{FP_c + TN_c} \tag{3.30}$$

Nilai AUC dihitung sebagai luas area di bawah kurva ROC. AUC = 1,0 menunjukkan diskriminasi sempurna, sedangkan AUC = 0,5 menunjukkan model tidak lebih baik dari prediksi acak. Untuk mengakomodasi skenario *multi-class*, perhitungan menggunakan pendekatan *one-vs-rest* (OVR) di mana setiap kelas diperlakukan sebagai kelas positif terhadap gabungan kelas lainnya. ROC-AUC akhir dihitung sebagai rata-rata *macro* dari AUC ketiga kelas (Persamaan 3.31):

$$\text{ROC-AUC}_{\text{Macro}} = \frac{1}{C}\sum_{c=1}^{C} \text{AUC}_c \tag{3.31}$$

7. ***Confusion Matrix*** menyajikan distribusi prediksi model untuk setiap kombinasi kelas aktual dan prediksi dalam bentuk matriks berukuran $C \times C$, di mana elemen M_{ij} menunjukkan jumlah sampel kelas aktual i yang diprediksi sebagai kelas j. Elemen diagonal (M_{ii}) merepresentasikan prediksi yang benar, sedangkan elemen non-diagonal mengindikasikan kesalahan klasifikasi. *Confusion matrix* memungkinkan identifikasi pola kesalahan antarkelas secara detail, misalnya apakah model cenderung salah mengklasifikasikan kelas Tinggi sebagai Sedang atau sebaliknya.

Selain metrik keseluruhan dari *cross-validation*, evaluasi per kelas (*Precision*, *Recall*, F1 untuk Rendah, Sedang, dan Tinggi) beserta *confusion matrix* detail dilaporkan dari satu kali evaluasi pada *test set* untuk mengidentifikasi pola kesalahan klasifikasi antarkelas serta dampak setiap komponen *preprocessing* terhadap performa kelas tertentu.

3.2.7 Analisis Hasil

Analisis hasil dilakukan dalam empat dimensi. Pertama, perbandingan performa antarskenario untuk mengidentifikasi konfigurasi optimal berdasarkan metrik F1 Macro sebagai acuan utama. Kedua, analisis kontribusi individual setiap komponen melalui perbandingan selisih performa relatif terhadap *baseline* (S1): efek seleksi fitur dievaluasi dari selisih S2 vs S1, efek BorderlineSMOTE dari selisih S3 vs S1, efek Optuna dari selisih S5 vs S1, dan efek interaksi dari perbandingan S4 dan S6 terhadap komponen individualnya. Ketiga, uji signifikansi statistik dilakukan menggunakan *Wilcoxon signed-rank test* pada 25 pasangan skor CV (*paired*, *non-parametric*) untuk menentukan apakah perbedaan performa antarskenario signifikan secara statistik pada tingkat signifikansi $\alpha = 0{,}05$. *Wilcoxon signed-rank test* dipilih karena tidak mengasumsikan distribusi normal pada selisih skor, sehingga lebih robust pada ukuran sampel yang relatif kecil ($n = 25$). Keempat, analisis interpretabilitas model menggunakan SHAP (*SHapley Additive exPlanations*) TreeExplainer untuk mengkuantifikasi kontribusi setiap fitur terhadap prediksi model, sehingga dapat diidentifikasi fitur-fitur drainase dan curah hujan yang paling berpengaruh terhadap klasifikasi risiko banjir.

3.2.8 Kesimpulan

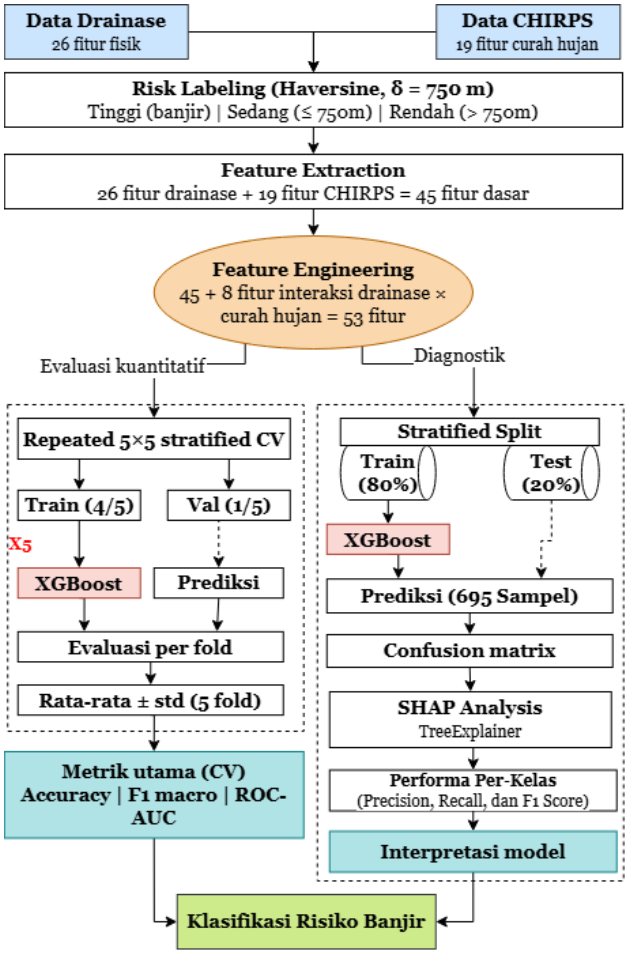
Tahap akhir penelitian menarik kesimpulan berdasarkan hasil analisis untuk menjawab rumusan masalah yang telah ditetapkan. Kesimpulan mencakup konfigurasi model optimal dari studi ablasi, kontribusi relatif masing-masing komponen *preprocessing* dan optimasi, serta rekomendasi praktis terkait penerapan model klasifikasi risiko banjir berbasis XGBoost di wilayah perkotaan. Temuan yang diperoleh diharapkan dapat menjadi acuan bagi pemerintah daerah dan pemangku kepentingan terkait dalam perencanaan pengelolaan risiko banjir di Kota Malang.

BAB IV
MODEL *BASELINE* XGBOOST

Bab ini membahas model *baseline* XGBoost yang menjadi titik acuan performa sebelum diterapkannya teknik *preprocessing* dan optimasi. Model *baseline* dilatih menggunakan seluruh 53 fitur tanpa seleksi fitur, tanpa penanganan ketidakseimbangan kelas, dan dengan *default hyperparameter*, sesuai dengan Skenario S1 pada desain eksperimen di Bab III. Performa yang dihasilkan pada bab ini menjadi pembanding bagi skenario S2–S6 pada Bab V.

4.1 Desain *Baseline* XGBoost

Desain model *baseline* mengikuti prinsip kesederhanaan: seluruh fitur yang tersedia digunakan, konfigurasi *hyperparameter* mengikuti nilai yang umum dipakai dalam literatur, dan tidak ada intervensi pada distribusi data. Performa yang dihasilkan dengan demikian merepresentasikan kemampuan "bawaan" XGBoost pada dataset klasifikasi risiko banjir Kota Malang. Gambar 4.1 menyajikan alur kerja keseluruhan model *baseline*.



Gambar 4. 1 Desain *Baseline* XGBoost

4.1.1 Arsitektur *Input-Output*

Model *baseline* menerima input berupa seluruh 53 fitur tanpa seleksi, yang terdiri dari 26 fitur drainase fisik, 19 fitur curah hujan CHIRPS, dan 8 fitur interaksi drainase-curah hujan (daftar lengkap fitur tersedia pada Tabel 3.1, 3.2, dan 3.3 di Bab III). *Output* model berupa probabilitas untuk tiga kelas risiko, yaitu Rendah (0), Sedang (1), dan Tinggi (2), yang dihasilkan melalui fungsi *softmax* (Persamaan 3.20 pada Bab III). Kelas dengan probabilitas tertinggi dipilih sebagai prediksi akhir.

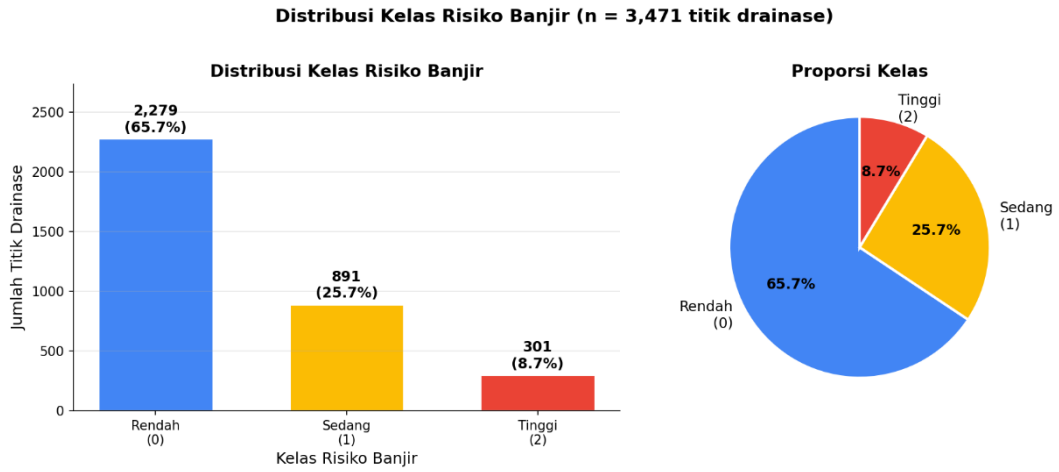
4.1.2 Pembagian Data

Dataset terdiri dari 3.471 titik drainase. Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* (25 evaluasi) pada seluruh dataset dengan setiap *fold* mempertahankan proporsi kelas yang sama. Selain itu, satu kali pembagian 80/20 (*stratified split*, *random state* 42) dilakukan untuk keperluan analisis diagnostik (*confusion matrix* detail, performa per kelas, dan analisis SHAP). Pembagian 80/20 menghasilkan 2.776 sampel pelatihan dan 695 sampel pengujian. Tabel 4.1 menyajikan distribusi kelas pada dataset keseluruhan.

Tabel 4. 1 Distribusi Kelas pada Dataset

Kelas	**Jumlah**	**Proporsi**
Rendah (0)	2.279	65,7%
Sedang (1)	891	25,7%
Tinggi (2)	301	8,7%
Total	**3.471**	**100%**

Gambar 4.2 memvisualisasikan ketidakseimbangan distribusi kelas: kelas Rendah mendominasi dengan 65,7% sampel sementara kelas Tinggi hanya mencakup 8,7%. Ketimpangan ini menjadi motivasi utama penerapan BorderlineSMOTE pada skenario S3, S4, dan S6 di Bab V.



Gambar 4. 2 Distribusi Kelas Risiko Banjir pada Dataset

4.1.3 Konfigurasi *Hyperparameter Default*

Model *baseline* menggunakan konfigurasi *hyperparameter* yang dipilih berdasarkan nilai umum yang sering direkomendasikan dalam literatur XGBoost, tanpa proses *tuning* apa pun. Tabel 4.2 menyajikan konfigurasi lengkapnya.

Tabel 4. 2 Konfigurasi *Default Hyperparameter Baseline* XGBoost

Hyperparameter	**Nilai**	**Keterangan**
n_estimators	200	Jumlah pohon keputusan
max_depth	6	Kedalaman maksimum setiap pohon
learning_rate	0,1	Laju pembelajaran (*shrinkage factor*)
subsample	0,8	Proporsi sampel per pohon
colsample_bytree	0,8	Proporsi fitur per pohon
min_child_weight	3	Minimum bobot pada *leaf node*
gamma (γ)	0,1	*Minimum loss reduction* untuk pemisahan
reg_alpha (α)	0,1	Regularisasi L1
reg_lambda (λ)	1,0	Regularisasi L2

Hyperparameter	**Nilai**	**Keterangan**
objective	*multi:softprob*	Fungsi objektif *multi-class*
eval_metric	*mlogloss*	*Multiclass log loss*

Nilai *n_estimators = 200 dipilih sebagai kompromi antara kapasitas model dan risiko* overfitting. *Max_depth* = 6 membatasi kompleksitas setiap pohon agar tidak menghafal pola *noise* pada data pelatihan. *Learning*_rate = *0,1* merupakan nilai standar yang memberikan keseimbangan antara kecepatan konvergensi dan stabilitas. *Parameter subsample* dan *colsample_bytree* yang bernilai 0,8 menerapkan *stochastic gradient boosting*: setiap pohon hanya menggunakan 80% sampel dan 80% fitur secara acak untuk mengurangi korelasi antar pohon. Regularisasi L1 (α = 0,1) dan L2 (λ= 1,0) diterapkan untuk mengontrol kompleksitas model melalui *penalty* terhadap bobot *leaf node* (Persamaan 3.15 pada Bab III).

Arsitektur *baseline* dapat diilustrasikan sebagai alur linier: 53 fitur masuk → imputasi median → pelatihan XGBoost (200 pohon, *default* HP) → prediksi probabilitas tiga kelas → evaluasi melalui *Repeated 5×5 Stratified K-Fold CV*. Tidak ada seleksi fitur, tidak ada penanganan ketidakseimbangan kelas, dan tidak ada optimasi *hyperparameter*.

4.2 Implementasi *Baseline* XGBoost

Implementasi dilakukan menggunakan bahasa pemrograman *Python* 3 dengan pustaka XGBoost versi terbaru melalui *interface XGBClassifier* dari modul *xgboost*. Proses pelatihan dan evaluasi dilaksanakan pada lingkungan komputasi dengan pustaka pendukung scikit-learn untuk pembagian data dan perhitungan metrik, serta SHAP untuk analisis interpretabilitas.

4.2.1 Pra-pemrosesan Data

Sebelum pelatihan, dua langkah pra-pemrosesan dasar diterapkan pada dataset. Pertama, nilai yang hilang (*missing values*) diisi menggunakan nilai median dari masing-masing fitur. Median dipilih karena lebih tahan terhadap pencilan dibandingkan rata-rata, terutama pada fitur drainase yang distribusinya cenderung miring (*skewed*), misalnya *catchment_area* yang nilainya tersebar dari ratusan meter persegi hingga jutaan meter persegi. Kedua, nilai tak berhingga (*infinity*) yang muncul dari operasi pembagian (misalnya pada fitur *rasio_catchment_kapasitas* ketika pembaginya nol) diganti dengan `*NaN*` terlebih dahulu, kemudian diisi kembali dengan median. Kedua langkah ini bersifat minimal dan disengaja, karena tujuan model *baseline* adalah mengukur performa XGBoost tanpa intervensi *preprocessing* yang lebih lanjut.

4.2.2 Pelatihan Model

Model diinisialisasi menggunakan *XGBClassifier* dengan parameter sesuai Tabel 4.2 dan dilatih pada *training set* (2.776 sampel, 53 fitur). Proses pelatihan membangun 200 pohon keputusan secara sekuensial; setiap pohon baru mempelajari residual dari prediksi kumulatif pohon sebelumnya (Persamaan 3.13

pada Bab III). Pada setiap iterasi pembangunan pohon, XGBoost menghitung gradien orde pertama dan *Hessian* dari fungsi *log loss* untuk menentukan pemisahan optimal pada setiap simpul (Persamaan 3.16–3.19 pada Bab III). Waktu pelatihan model *baseline* tercatat kurang dari 1 detik, yang menunjukkan efisiensi komputasi XGBoost pada dataset berukuran moderat.

Kode implementasi pelatihan model *baseline* disajikan sebagai berikut.

```
import xgboost as xgb
from sklearn.model_selection import train_test_split

# Konfigurasi hyperparameter default
params = {
 'n_estimators': 200,
 'max_depth': 6,
 'learning_rate': 0.1,
 'subsample': 0.8,
 'colsample_bytree': 0.8,
 'min_child_weight': 3,
 'gamma': 0.1,
 'reg_alpha': 0.1,
 'reg_lambda': 1.0,
 'objective': 'multi:softprob',
 'num_class': 3,
 'eval_metric': 'mlogloss',
 'random_state': 42,
 'n_jobs':-1,
}

# Pembagian data (stratified 80/20)
X_train, X_test, y_train, y_test = train_test_split(
 X, y, test_size=0.2, random_state=42, stratify=y
)

# Pelatihan model
model = xgb.XGBClassifier(**params)
model.fit(X_train, y_train)
```

4.2.3 Evaluasi *Cross-Validation*

Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* untuk mengukur performa model pada partisi data yang berbeda dengan 5 pengulangan pengacakan fold. Implementasi evaluasi *cross-validation* disajikan sebagai berikut.

```
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.metrics import accuracy_score, f1_score,
roc_auc_score
import numpy as np
# Evaluasi utama: Repeated 5×5 Stratified K-Fold Cross-Validation
rskf = RepeatedStratifiedKFold(
       n_splits=5, n_repeats=5, random_state=42
)

# Total 25 evaluasi (5 fold × 5 repeat)
for fold_idx, (train_idx, val_idx) in enumerate(rskf.split(X, y)):
          repeat_num = fold_idx // 5 + 1
          fold_num = fold_idx % 5 + 1
      # Train pada training fold, evaluasi pada validation fold
      # Preprocessing diterapkan di dalam training fold saja


# Rata-rata dan standar deviasi dari 25 evaluasi
for metric in ['accuracy', 'f1_macro', 'roc_auc']:
      scores = cv_results[f'test_{metric}']
      print(f"{metric}: {np.mean(scores):.4f} ±
      {np.std(scores):.4f}")
```

Selain *cross-validation*, prediksi pada *test set* 80/20 dilakukan untuk analisis diagnostik. Model menghasilkan prediksi dalam dua bentuk: probabilitas per kelas melalui `predict_proba()` dan label kelas melalui `predict()`. Probabilitas per kelas dibutuhkan untuk perhitungan ROC-AUC dengan pendekatan *one-vs-rest*, sedangkan label kelas digunakan untuk menghitung *accuracy*, F1 *score*, dan *confusion matrix*. Perhitungan ROC-AUC menggunakan `label_binarize` untuk mengkonversi label *multi-class* menjadi representasi biner sebelum dihitung

menggunakan `roc_auc_score` dengan parameter `multi_class='ovr'` dan `average='macro'`.

4.3 Pengujian *Baseline* XGBoost

Bagian ini menyajikan hasil pengujian model *baseline*. Evaluasi utama dilakukan melalui *Repeated 5×5 Stratified K-Fold Cross-Validation* pada seluruh 3.471 sampel (25 evaluasi), dilengkapi analisis diagnostik (*confusion matrix* detail, performa per kelas, dan SHAP) dari satu kali evaluasi pada *test set* 80/20.

4.3.1 Performa Keseluruhan

Tabel 4.3 merangkum performa keseluruhan model *baseline* berdasarkan *Repeated 5×5 Stratified K-Fold Cross-Validation*.

Tabel 4. 3 Performa Keseluruhan Model *Baseline* (S1), Repeated 5×5 CV (25 Evaluasi)

Metrik	**Rata-rata**	**Std**
Accuracy	0,8800	± 0,0123
F1 Macro	0,8152	± 0,0215
ROC-AUC (*Macro* OVR)	0,9689	± 0,0048

Model *baseline* mencapai *accuracy* rata-rata 88,00% (± 1,23%) dan F1 Macro 0,8152 (± 0,0215) dari *Repeated 5×5 cross-validation* (25 evaluasi). ROC-AUC 0,9689 (± 0,0048) menunjukkan kemampuan diskriminatif model antara ketiga kelas yang tergolong sangat baik dengan variabilitas rendah. Standar deviasi yang diperoleh dari 25 evaluasi lebih stabil dibandingkan *single-run* 5-fold CV karena memperhitungkan variabilitas antar pengacakan fold yang berbeda. Meskipun *accuracy* terlihat tinggi, F1 Macro yang lebih rendah mengisyaratkan

bahwa performa model tidak merata di seluruh kelas. Analisis per kelas pada subbab berikutnya mengkonfirmasi dugaan ini.

4.3.2 Performa Per kelas

Untuk menganalisis pola kesalahan klasifikasi secara detail, Tabel 4.4 menyajikan *Precision*, *Recall*, dan F1 *Score* untuk masing-masing kelas risiko dari satu kali evaluasi pada *test set* (695 sampel).

Tabel 4. 4 Performa Per kelas Model *Baseline* (S1)

Kelas	***Precision***	***Recall***	**F1 *Score***	***Support***
Rendah	0,9128	0,9408	0,9266	456
Sedang	0,7513	0,7933	0,7717	179
Tinggi	0,9722	0,5833	0,7292	60

Performa model baseline (S1) pada Tabel 4.4 menunjukkan variasi yang cukup jelas antar kelas, yang mencerminkan dampak ketidakseimbangan data terhadap kemampuan klasifikasi model XGBoost. Meskipun secara umum model mampu memberikan hasil yang baik, distribusi performa pada masing-masing kelas memperlihatkan adanya perbedaan signifikan antara kelas mayoritas dan minoritas. Hal ini menjadi penting untuk dianalisis lebih lanjut guna memahami keterbatasan model sebelum dilakukan penerapan teknik penanganan *imbalance* pada tahap berikutnya.

1. **Kelas Rendah**
    a) Menunjukkan performa terbaik dengan F1-score sebesar 0,9266.
    b) Precision mencapai 0,9128 dan Recall 0,9408.

c) Performa tinggi ini dipengaruhi oleh dominasi kelas Rendah dalam dataset (65,7% atau 456 dari 695 sampel test set), sehingga model memiliki lebih banyak data untuk mempelajari pola kelas ini.

d) Hal ini membuat model lebih stabil dalam mengenali kelas Rendah dibandingkan kelas lainnya.

2. Kelas Sedang

a) Memiliki F1-score sebesar 0,7717.

b) Recall sebesar 0,7933 menunjukkan kemampuan deteksi yang cukup baik terhadap kelas Sedang.

c) Namun Precision sebesar 0,7513 menunjukkan adanya false positive, yaitu sebagian sampel dari kelas Rendah dan Tinggi salah diklasifikasikan sebagai kelas Sedang.

d) Kondisi ini menunjukkan bahwa kelas Sedang berada pada posisi transisi sehingga rentan terhadap kesalahan klasifikasi dari kedua kelas lainnya.

3. Kelas Tinggi

a) Menunjukkan karakteristik yang paling berbeda dibanding kelas lainnya.

b) Precision sangat tinggi sebesar 0,9722, yang berarti sebagian besar prediksi kelas Tinggi adalah benar.

c) Dari sekitar 36 prediksi kelas Tinggi, hanya 1 yang salah klasifikasi.

d) Namun Recall hanya sebesar 0,5833, yang berarti model hanya mampu mendeteksi 35 dari 60 sampel kelas Tinggi.

e) Sebanyak 25 sampel kelas Tinggi tidak terdeteksi, dengan 21 di antaranya salah diklasifikasikan sebagai Sedang dan 4 sebagai Rendah.
f) Dalam konteks mitigasi banjir, kondisi ini berarti sekitar 41,7% titik risiko tinggi tidak teridentifikasi oleh model, sehingga berpotensi mengurangi efektivitas peringatan dini dan alokasi sumber daya.

Secara keseluruhan, hasil pada model baseline menunjukkan adanya *trade-off* yang jelas antara Precision dan Recall, khususnya pada kelas Tinggi yang merupakan kelas minoritas. Model cenderung konservatif dalam memprediksi kelas tersebut akibat ketidakseimbangan data, sehingga menghasilkan Precision yang tinggi namun Recall yang rendah. Pola ini mengindikasikan bahwa meskipun model cukup andal ketika memberikan prediksi risiko tinggi, masih terdapat banyak kasus risiko tinggi yang tidak terdeteksi. Kondisi ini memperkuat urgensi penerapan metode penanganan data tidak seimbang pada eksperimen lanjutan untuk meningkatkan sensitivitas model terhadap kelas kritis tersebut.

4.3.3 *Confusion Matrix*

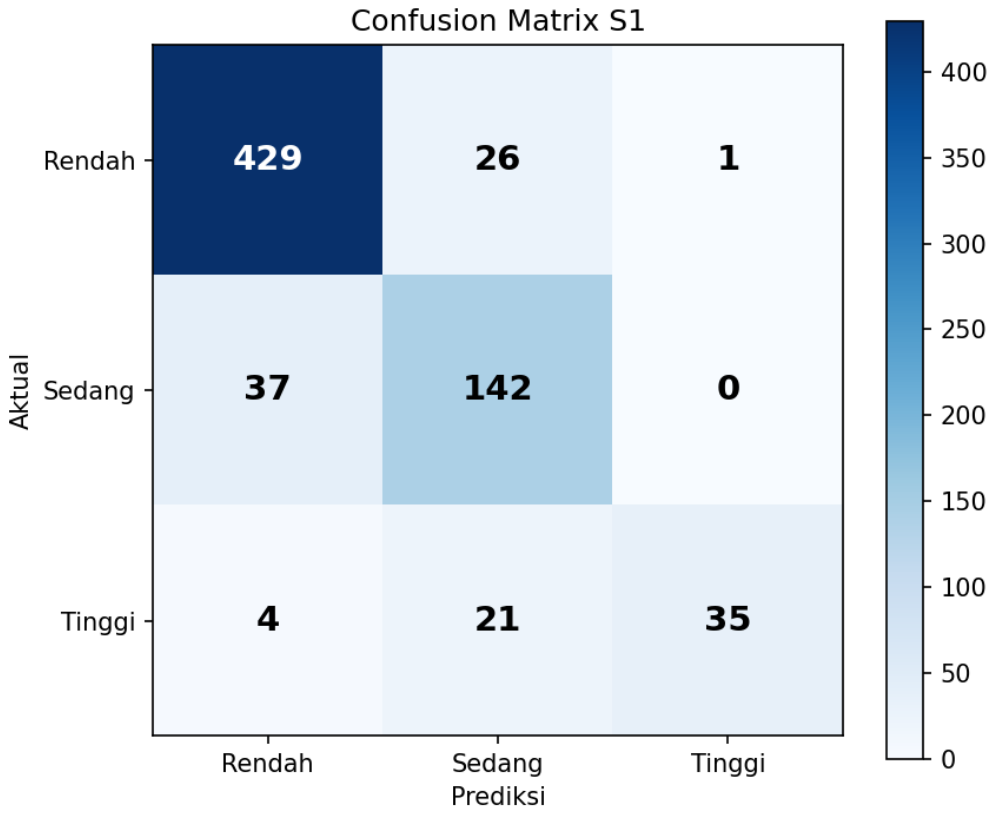
Tabel 4.5 menyajikan *confusion matrix* model *baseline* dari *test set* yang memberikan gambaran detail pola prediksi dan kesalahan klasifikasi.

Tabel 4. 5 *Confusion Matrix* Model *Baseline* XGBoost (S1)

	Prediksi Rendah	**Prediksi Sedang**	**Prediksi Tinggi**
Aktual Rendah	**429**	26	1
Aktual Sedang	37	**142**	0
Aktual Tinggi	4	21	**35**

Elemen diagonal merepresentasikan prediksi benar: 429 dari 456 titik Rendah, 142 dari 179 titik Sedang, dan 35 dari 60 titik Tinggi. Total prediksi benar adalah 606 dari 695 sampel (87,19%).

Visualisasi *confusion matrix* dalam bentuk *heatmap* disajikan pada Gambar 4.3, yang memperlihatkan pola prediksi benar pada diagonal utama dan pola kesalahan di luar diagonal.



Gambar 4. 3 *Heatmap Confusion Matrix* Model *Baseline* (S1)

Dari *confusion matrix* ini terlihat beberapa pola kesalahan. Pertama, kesalahan terbesar terjadi pada kelas Sedang ke Rendah: 37 titik Sedang yang diprediksi sebagai Rendah. Ini menunjukkan bahwa sebagian titik Sedang memiliki

karakteristik fitur yang menyerupai kelas Rendah, sehingga model kesulitan membedakannya. Kedua, 21 titik Tinggi diprediksi sebagai Sedang. Kesalahan ini merupakan *false negative* kelas Tinggi yang paling banyak. Dari sudut pandang pengelolaan risiko banjir, kesalahan ini lebih berbahaya daripada kesalahan Sedang ke Rendah, karena titik yang seharusnya mendapat perhatian prioritas justru diklasifikasikan pada tingkat risiko yang lebih rendah. Ketiga, model hampir tidak pernah melakukan kesalahan "lompat kelas": hanya 1 titik Rendah diprediksi Tinggi dan 3 titik Tinggi diprediksi Rendah. Pola ini menunjukkan bahwa meskipun batas antara kelas bersebelahan (*adjacent classes*) masih kabur, model sudah cukup mampu membedakan kelas yang berjauhan.

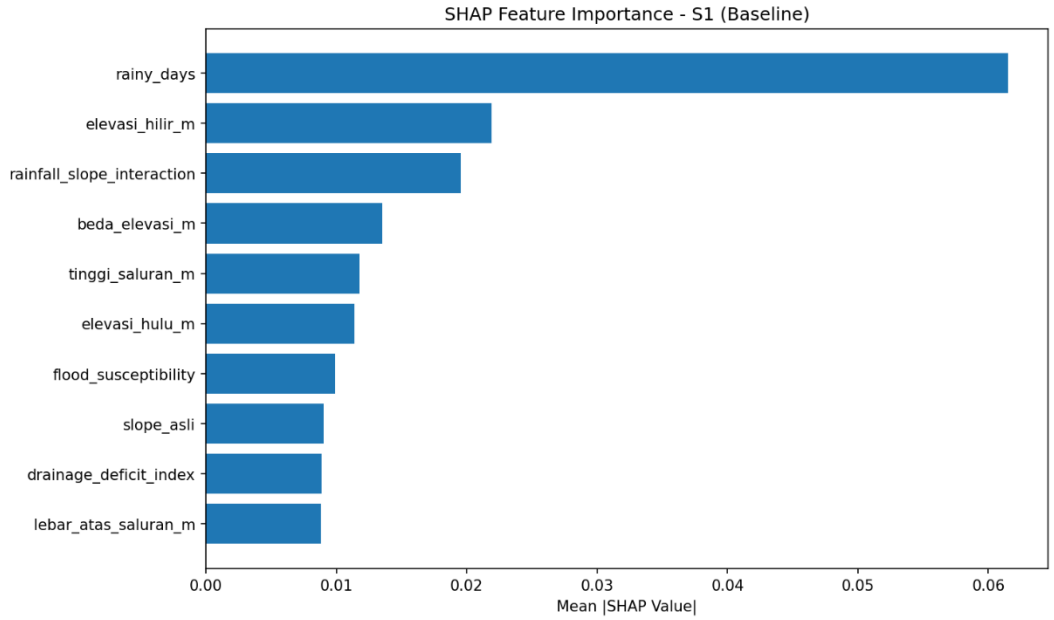
4.3.4 Analisis Feature Importance (SHAP)

Analisis *feature importance* menggunakan SHAP (*SHapley Additive exPlanations*) TreeExplainer memberikan pemahaman mengenai fitur mana yang paling berpengaruh terhadap prediksi model *baseline*. Tabel 4.6 menyajikan sepuluh fitur dengan nilai SHAP rata-rata tertinggi.

Tabel 4. 6 Top-10 Fitur Berdasarkan Nilai SHAP Rata-rata (*Baseline*)

Peringkat	**Fitur**	**Kategori**	**Mean SHAP**
1	rainy_days	Curah Hujan	0,0615
2	elevasi_hilir_m	Drainase	0,0219
3	rainfall_slope_interaction	Interaksi	0,0196
4	beda_elevasi_m	Drainase	0,0135
5	tinggi_saluran_m	Drainase	0,0118
6	elevasi_hulu_m	Drainase	0,0114
7	flood_susceptibility	Interaksi	0,0099
8	slope_asli	Drainase	0,0090
9	drainage_deficit_index	Interaksi	0,0089
10	lebar_atas_saluran_m	Drainase	0,0088

Gambar 4.4 memvisualisasikan distribusi SHAP *importance* untuk sepuluh fitur teratas dalam bentuk *bar plot*.



Gambar 4. 4 SHAP *Summary Plot* (Bar) Model *Baseline* (S1)

Fitur `rainy_days` (jumlah hari hujan selama periode pengamatan) mendominasi dengan nilai SHAP rata-rata 0,0615, sekitar tiga kali lipat fitur peringkat kedua. Dominasi ini dapat dijelaskan oleh fakta bahwa frekuensi hari hujan secara langsung menentukan seberapa sering sistem drainase menerima beban, dan secara kumulatif berkorelasi dengan risiko banjir.

Tiga fitur interaksi muncul di peringkat sepuluh besar: `rainfall_slope_interaction` (peringkat 3), `flood_susceptibility` (peringkat 7), dan `drainage_deficit_index` (peringkat 9). Fitur-fitur ini direkayasa untuk menangkap hubungan antara kapasitas drainase dan beban curah hujan, dan kemunculannya di

peringkat atas mengkonfirmasi bahwa integrasi kedua sumber data (drainase fisik dan curah hujan CHIRPS) memberikan informasi prediktif yang tidak dapat diperoleh dari masing-masing sumber data secara terpisah.

Dari sisi drainase, fitur elevasi (hulu, hilir, dan beda elevasi) serta dimensi saluran (tinggi dan lebar atas) mendominasi dibandingkan fitur kapasitas hidraulik. Pola ini mengindikasikan bahwa pada model *baseline*, faktor topografi dan geometri saluran lebih informatif daripada estimasi kapasitas yang diturunkan secara kalkulatif.

4.3.5 Ringkasan Temuan Baseline

Berdasarkan hasil pengujian model *baseline*, diperoleh beberapa temuan. Model *baseline* XGBoost dengan konfigurasi *default* sudah mampu mencapai performa yang cukup baik secara keseluruhan berdasarkan *Repeated 5×5 CV* (25 evaluasi): *Accuracy* 0,8800 ± 0,0123, F1 Macro 0,8152 ± 0,0215, dan ROC-AUC 0,9689 ± 0,0048. Kemampuan diskriminatif model antara ketiga kelas tergolong sangat baik berdasarkan nilai ROC-AUC yang mendekati 1,0 dengan variabilitas rendah antar evaluasi.

Namun demikian, terdapat kelemahan yang perlu ditangani. Performa yang tidak merata antarkelas menjadi permasalahan utama: kelas Tinggi hanya mencapai *Recall* 0,5833, yang berarti 41,7% titik berisiko tinggi tidak terdeteksi. Ketidakseimbangan distribusi kelas, dengan kelas Tinggi hanya 8,7% dari total data, menjadi penyebab utama rendahnya *Recall* tersebut. Selain itu, batas

keputusan antara kelas bersebelahan (Rendah-Sedang dan Sedang-Tinggi) masih kabur, yang terlihat dari pola kesalahan pada *confusion matrix*.

Berdasarkan temuan ini, Bab V akan menerapkan seleksi fitur untuk mereduksi dimensi dan menghilangkan fitur redundan, BorderlineSMOTE untuk menangani ketidakseimbangan kelas dengan tujuan utama meningkatkan *Recall* kelas Tinggi, serta Optuna TPE untuk mencari konfigurasi *hyperparameter* yang lebih baik daripada konfigurasi *default*.

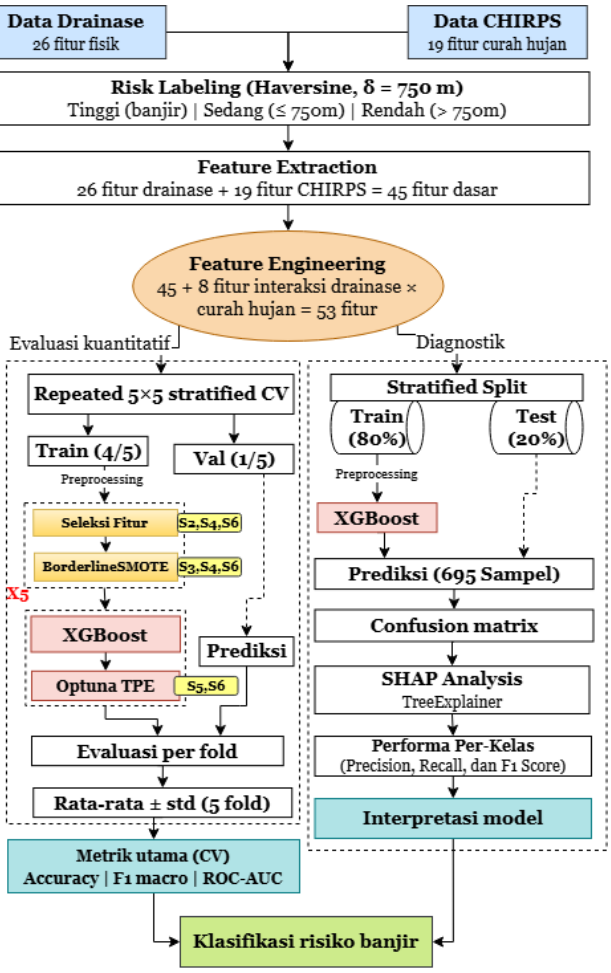
BAB V
STUDI ABLASI MODEL XGBOOST DENGAN *PREPROCESSING* DAN OPTIMASI *HYPERPARAMETER*

Bab ini membahas penerapan teknik *preprocessing* dan optimasi pada model XGBoost melalui lima skenario lanjutan (S2–S6) yang dirancang sebagai studi ablasi terhadap model *baseline* (S1) pada Bab IV. Setiap skenario mengisolasi atau mengombinasikan komponen yang berbeda, yaitu seleksi fitur (MI + Pearson + RFE), penanganan ketidakseimbangan kelas (BorderlineSMOTE), dan optimasi *hyperparameter* (Optuna TPE), untuk mengukur kontribusi masing-masing terhadap performa klasifikasi risiko banjir.

5.1 Desain Skenario Ablasi Model XGBoost dengan *Preprocessing* dan Optimasi *Hyperparameter*

Berbeda dari *baseline* (S1) yang langsung melatih XGBoost dengan 53 fitur dan *default hyperparameter*, skenario S2 hingga S6 menambahkan komponen *preprocessing* dan optimasi secara bertahap. Alur dimulai dari data drainase dan CHIRPS yang melalui *feature engineering* dan *stratified split* (identik dengan *baseline*), kemudian memasuki tahap *preprocessing* pada *training set*. Tahap ini terdiri dari dua komponen opsional yang dijalankan secara berurutan: seleksi fitur (MI + Pearson + RFE) yang mereduksi 53 fitur menjadi 19, dan BorderlineSMOTE yang menyeimbangkan distribusi kelas pada data pelatihan. Setelah *preprocessing*, proses pelatihan model dilakukan. Pada skenario yang mengaktifkan Optuna TPE (S5 dan S6), proses pelatihan dibungkus oleh optimasi *hyperparameter* melalui 5-

fold StratifiedKFold cross-validation, sehingga XGBoost dilatih berulang kali dengan konfigurasi berbeda hingga diperoleh kombinasi *hyperparameter* terbaik berdasarkan F1 Macro. Pada skenario tanpa Optuna (S2, S3, S4), XGBoost dilatih sekali dengan *default hyperparameter*. Evaluasi utama menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* dengan *preprocessing* diterapkan di dalam setiap fold pelatihan untuk mencegah kebocoran data, sedangkan satu kali *split* 80/20 digunakan untuk analisis diagnostik berupa *confusion matrix*, metrik per kelas, dan SHAP TreeExplainer. Gambar 5.1 menyajikan alur lengkap desain model dengan *preprocessing* dan optimasi.



Gambar 5. 1 Desain Model XGBoost dengan *Preprocessing* dan Optimasi *Hyperparameter*

5.1.1 Konfigurasi Skenario

Kelima skenario lanjutan mengikuti konfigurasi pada Tabel 3.5 di Bab III. Seluruh skenario menggunakan pembagian data, *random state*, dan arsitektur XGBoost yang identik dengan *baseline*, sehingga setiap perbedaan performa dapat diatribusikan secara langsung kepada komponen yang diaktifkan. Tabel 5.1 merangkum perbedaan konfigurasi antarskenario.

Tabel 5. 1 Konfigurasi Komponen Skenario S2–S6

Skenario	**Seleksi Fitur**	**BorderlineSMOTE**	**Optuna TPE**	**Jumlah Fitur**
S1 (*baseline*)				53
S2	✓			19
S3		✓		53
S4	✓	✓		19
S5			✓	53
S6 (*Full Pipeline*)	✓	✓	✓	19

5.1.2 Komponen Seleksi Fitur

Seleksi fitur diterapkan pada skenario S2, S4, dan S6 melalui *pipeline* tiga tahap (MI + Pearson + RFE) sebagaimana dijelaskan pada Bab III §3.2.4. Tahap pertama menghitung *Mutual Information* antara setiap fitur dan variabel target, kemudian mengeliminasi fitur dengan MI $\leq 0{,}05$ (Persamaan 3.8). Tahap kedua menghilangkan salah satu fitur dari pasangan yang memiliki korelasi Pearson $|r| > 0{,}95$ (Persamaan 3.9). Tahap ketiga menerapkan RFE dengan XGBoost sebagai estimator untuk memilih fitur terbaik (Persamaan 3.10–3.11). Proses seleksi dilakukan hanya pada *training set* untuk mencegah kebocoran informasi dari *test set*.

Dari 53 fitur awal, *pipeline* seleksi fitur memilih 19 fitur berikut: 10 fitur drainase (*panjang_saluran_m*, *tinggi_saluran_m*, *luas_penampung_m2*, *elevasi_hulu_m*, *slope_asli*, *catchment_area_m2*, *catchment_density_m2perm*, *kapasitas_per_meter*, *rasio_lebar*, *selisih_elevasi_m*), 2 fitur curah hujan CHIRPS (*rainy_days*, *rainfall_30d_max*), dan 7 fitur interaksi (*capacity_rainfall_ratio*, *drainage_stress_7d*, *rainfall_slope_interaction*, *flood_susceptibility*, *drainage_deficit_index*, *heavy_rain_vulnerability*, *extreme_rain_exposure*). *pipeline* MI + Pearson + RFE mempertahankan 2 fitur curah hujan karena MI mengevaluasi relevansi setiap fitur terhadap target secara independen, bukan redundansi antar sesama fitur. Tahap MI mengeliminasi 5 fitur (*sisi_saluran_encoded*, *tipe_saluran_encoded*, *kondisi_fisik_encoded*, *drainage_efficiency*, *lebar_bawah_saluran_m*) yang memiliki MI $\leq$ 0,05, tahap Pearson mengeliminasi 9 fitur dengan korelasi > 0,95, dan RFE memilih 19 fitur terbaik dari 39 fitur yang tersisa.

5.1.3 Komponen Penanganan Ketidakseimbangan Kelas

BorderlineSMOTE diterapkan pada skenario S3, S4, dan S6 untuk menangani distribusi kelas yang tidak seimbang pada *training set* (Rendah: 1.823, Sedang: 712, Tinggi: 241). Teknik ini menghasilkan sampel sintetis hanya dari instans kelas minoritas yang berada di sekitar *decision boundary* (Persamaan 3.12 pada Bab III), dengan parameter k = 5 tetangga terdekat. BorderlineSMOTE diterapkan setelah seleksi fitur (pada S4 dan S6) dan hanya pada *training set*; *test set* tidak dimodifikasi agar evaluasi tetap merepresentasikan distribusi data asli.

5.1.4 Komponen Optimasi *Hyperparameter*

Optuna TPE diterapkan pada skenario S5 dan S6 untuk mencari konfigurasi *hyperparameter* yang lebih baik daripada konfigurasi *default* pada *baseline*. Proses optimasi menggunakan *5-fold StratifiedKFold cross-validation* dengan F1 Macro sebagai fungsi objektif (Persamaan 3.21-3.24 pada Bab III). Ruang pencarian mencakup sembilan *hyperparameter* sesuai konfigurasi pada Bab III §3.2.4.

5.2 Implementasi Skenario Ablasi Model XGBoost dengan *Preprocessing* dan Optimasi *Hyperparameter*

Bagian ini menjabarkan implementasi teknis dari masing-masing skenario S2 hingga S6. Setiap skenario dijalankan menggunakan bahasa pemrograman Python 3 dengan pustaka yang sama seperti *baseline*: XGBoost melalui `XGBClassifier`, scikit-learn untuk pembagian data dan metrik evaluasi, imblearn untuk BorderlineSMOTE, serta Optuna untuk optimasi *hyperparameter*. Pra-pemrosesan dasar (pengisian *missing values* dengan median dan penggantian nilai tak berhingga) diterapkan pada seluruh skenario sebagaimana pada *baseline* di Bab IV. Perbedaan antarskenario terletak pada komponen tambahan yang diaktifkan sebelum dan selama pelatihan model.

5.2.1 Implementasi Seleksi Fitur (S2)

Pada skenario S2, *pipeline* seleksi fitur dijalankan pada *training set* (2.776 sampel, 53 fitur). Tahap MI menyaring fitur yang tidak informatif terhadap target, tahap korelasi Pearson menghilangkan fitur yang redundan satu sama lain, dan tahap RFE memilih 19 fitur final menggunakan XGBoost sebagai estimator. Setelah seleksi, model dilatih dengan 19 fitur dan *default hyperparameter* yang sama dengan *baseline*. Implementasi *pipeline* seleksi fitur disajikan sebagai berikut.

```
import numpy as np
from sklearn.feature_selection import RFE, mutual_info_classif

# Tahap 1: Eliminasi fitur dengan MI <= 0,05 terhadap target
mi_scores = mutual_info_classif(X_train, y_train, random_state=42)
mi_selected = [f for f, mi in zip(features, mi_scores) if mi > 0]
X_train_fs = X_train[mi_selected]

# Tahap 2: Eliminasi fitur dengan korelasi Pearson > 0,95
corr_matrix = X_train_fs.corr().abs()
upper = corr_matrix.where(
    np.triu(np.ones(corr_matrix.shape), k=1).astype(bool)
)
high_corr = [c for c in upper.columns if any(upper[c] > 0.95)]
X_train_fs = X_train_fs.drop(columns=high_corr)

# Tahap 3: RFE dengan XGBoost (target 15 fitur minimum)
n_select = max(15, len(X_train_fs.columns) // 2)
rfe = RFE(estimator=xgb.XGBClassifier(**params),
          n_features_to_select=n_select, step=1)
rfe.fit(X_train_fs, y_train)
selected_features = X_train_fs.columns[rfe.support_].tolist()
```

5.2.2 Implementasi BorderlineSMOTE (S3)

Pada skenario S3, BorderlineSMOTE diterapkan pada *training set* sebelum pelatihan model. Proses ini menghasilkan sampel sintetis untuk kelas Sedang dan Tinggi sehingga distribusi kelas pada *training set* menjadi lebih seimbang. Model kemudian dilatih menggunakan seluruh 53 fitur dengan *default hyperparameter* pada *training set* yang sudah diseimbangkan. Implementasi penerapan BorderlineSMOTE disajikan sebagai berikut.

```
from imblearn.over_sampling import BorderlineSMOTE

# Penerapan BorderlineSMOTE pada training set
smote = BorderlineSMOTE(k_neighbors=5, random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(
      X_train, y_train
)

# Pelatihan model pada data yang sudah diseimbangkan
model_s3 = xgb.XGBClassifier(**params)
model_s3.fit(X_train_resampled, y_train_resampled)
```

5.2.3 Implementasi Seleksi Fitur + BorderlineSMOTE (S4)

Skenario S4 mengombinasikan kedua komponen secara berurutan: seleksi fitur terlebih dahulu, lalu BorderlineSMOTE. Urutan ini dipilih agar seleksi fitur bekerja pada distribusi data asli, bukan distribusi yang sudah dimodifikasi oleh *oversampling*. Seleksi fitur menghasilkan 19 fitur yang identik dengan S2, kemudian BorderlineSMOTE diterapkan pada *training set* yang sudah direduksi menjadi 19 fitur.

5.2.4 Implementasi Optuna TPE (S5)

Pada skenario S5, Optuna TPE dijalankan pada seluruh 53 fitur tanpa seleksi fitur dan tanpa BorderlineSMOTE. Proses optimasi mengevaluasi berbagai konfigurasi *hyperparameter* melalui *5-fold StratifiedKFold cross-validation*. Implementasi fungsi objektif dan proses optimasi Optuna TPE disajikan sebagai berikut.

```
import optuna
from sklearn.model_selection import StratifiedKFold,
cross_val_score
from sklearn.metrics import make_scorer, f1_score
from xgboost import XGBClassifier

def objective(trial):
    params = {
        'n_estimators': trial.suggest_int('n_estimators',
100,500),
        'max_depth': trial.suggest_int('max_depth', 3, 10),
        'learning_rate': trial.suggest_float(
            'learning_rate', 0.01, 0.3, log=True
        ),
        'subsample': trial.suggest_float('subsample', 0.6, 1.0),
        'colsample_bytree': trial.suggest_float(
            'colsample_bytree', 0.6, 1.0
        ),
        'min_child_weight': trial.suggest_int('min_child_weight',
1, 10),
        'gamma': trial.suggest_float('gamma', 0.0,0.5),
        'reg_alpha': trial.suggest_float(
            'reg_alpha', 1e-8, 10.0, log=True
        ),
        'reg_lambda': trial.suggest_float(
            'reg_lambda', 1e-8, 10.0, log=True
        ),
        'objective': 'multi:softprob',
        'eval_metric': 'mlogloss',
        'random_state': 42,
    }
    model = XGBClassifier(**params)
    cv = StratifiedKFold(n_splits=5, shuffle=True,
random_state=42)
    f1_macro = make_scorer(f1_score, average='macro')
    scores = cross_val_score(model, X_train, y_train, cv=cv,
                             scoring=f1_macro)
    return scores.mean()

study = optuna.create_study(
    direction='maximize',
sampler=optuna.samplers.TPESampler(seed=42)
)
study.optimize(objective, n_trials=200,timeout=1800)

best_params = study.best_trial.params
model_s5 = XGBClassifier(**best_params,
objective='multi:softprob',
                         eval_metric='mlogloss', random_state=42)
model_s5.fit(X_train, y_train)
```

Tabel 5.2 menyajikan perbandingan *hyperparameter default* (S1) dan hasil optimasi Optuna (S5).

Tabel 5. 2 Perbandingan *Hyperparameter Default* (S1) vs Optuna (S5)

Hyperparameter	**S1 (*Default*)**	**S5 (Optuna)**
n_estimators	200	163
max_depth	6	4
learning_rate	0,1	0,1312
subsample	0,8	0,7087
colsample_bytree	0,8	0,7687
min_child_weight	3	1
gamma	0,1	0,0613
reg_alpha	0,1	$2{,}67 \times 10^{-5}$
reg_lambda	1,0	$4{,}46 \times 10^{-8}$

Optuna mengurangi jumlah pohon menjadi 163 sambil menurunkan kedalaman menjadi 4, sehingga kombinasi ini menghasilkan *ensemble* yang lebih kecil dari pohon yang lebih dangkal. *learning rate* sedikit naik (0,1312 vs 0,1). Perubahan yang paling terlihat ada pada regularisasi: *reg_alpha* dan *reg_lambda* diturunkan beberapa orde magnitudo, yang berarti Optuna menemukan bahwa regularisasi yang sangat ringan lebih optimal pada dataset ini. *gamma* juga diturunkan dari 0,1 menjadi 0,0613, yang berarti pemisahan pada pohon memerlukan *gain* minimum yang lebih kecil.

5.2.5 Implementasi *Full Pipeline* (S6)

Skenario S6 mengaktifkan seluruh komponen secara berurutan: seleksi fitur (19 fitur) → BorderlineSMOTE → Optuna TPE. Tabel 5.3 menyajikan *hyperparameter* yang dihasilkan Optuna pada S6, yang berbeda dari S5 karena ruang fitur dan distribusi *training set* yang berbeda.

Tabel 5. 3 *Hyperparameter* Hasil Optuna pada S6 (*Full Pipeline*)

Hyperparameter	**S6 (Optuna)**
n_estimators	351
max_depth	10
learning_rate	0,0849
subsample	0,8289
colsample_bytree	0,9561
min_child_weight	1
gamma	0,2325
reg_alpha	$1{,}05 \times 10^{-7}$
reg_lambda	$1{,}93 \times 10^{-3}$

Konfigurasi Optuna pada S6 berbeda dari S5 pada beberapa aspek. *max_depth* naik ke 10 (dari 4 pada S5), kemungkinan karena dengan hanya 19 fitur, pohon yang lebih dalam diperlukan untuk menangkap interaksi antar fitur yang tersisa. *Learning_rate* turun ke 0,0849 (vs 0,1312 pada S5), yang dikompensasi oleh jumlah pohon yang lebih banyak (351 vs 163). *Gamma* naik ke 0,2325 (vs 0,0613 pada S5), mengindikasikan kebutuhan regularisasi yang lebih ketat pada pohon yang lebih dalam untuk mencegah *overfitting*.

5.3 Pengujian Skenario Ablasi Model XGBoost dengan *Preprocessing* dan Optimasi *Hyperparameter*

Bagian ini menyajikan hasil pengujian kelima skenario lanjutan (S2–S6) secara komparatif terhadap *baseline* (S1). Evaluasi utama dilakukan melalui *Repeated 5×5 Stratified K-Fold Cross-Validation* pada seluruh 3.471 sampel (25 evaluasi), dengan *preprocessing* diterapkan di dalam setiap fold pelatihan untuk mencegah kebocoran data. Analisis diagnostik berupa *confusion matrix* detail, performa per kelas, dan SHAP TreeExplainer dilakukan dari satu kali evaluasi pada

test set 80/20 (695 sampel). Penyajian dimulai dari metrik keseluruhan, dilanjutkan analisis per kelas, *confusion matrix*, dan *feature importance* SHAP.

5.3.1 Performa Keseluruhan

Tabel 5.4 merangkum performa seluruh skenario berdasarkan *Repeated 5×5 Stratified K-Fold Cross-Validation* (rata-rata ± standar deviasi dari 25 evaluasi).

Tabel 5. 4 Perbandingan Performa Seluruh Skenario (*Repeated* 5×5 CV, 25 Evaluasi)

Skenario	***Accuracy* (Rata-rata ± Std)**	**F1 Macro (Rata-rata ± Std)**	**ROC-AUC (Rata-rata ± Std)**
S1 (*Baseline*)	0,8800 ± 0,0123	0,8152 ± 0,0215	0,9689 ± 0,0048
S2 (Seleksi Fitur)	0,8716 ± 0,0129	0,7978 ± 0,0249	0,9653 ± 0,0057
S3 (BorderlineSMOTE)	0,8713 ± 0,0120	0,8077 ± 0,0222	0,9658 ± 0,0053
S4 (FS + BorderlineSMOTE)	0,8655 ± 0,0131	0,7990 ± 0,0203	0,9607 ± 0,0062
S5 (Optuna)	0,8819 ± 0,0140	0,8193 ± 0,0234	0,9695 ± 0,0050
S6 (*Full Pipeline*)	0,8658 ± 0,0130	0,8011 ± 0,0193	0,9629 ± 0,0062

Dari tabel di atas, S5 (Optuna) sedikit mengungguli *baseline* pada ketiga metrik (*Accuracy* +0,0019; F1 Macro +0,0041; ROC-AUC +0,0006), sementara skenario lainnya menurunkan metrik keseluruhan. S4 (FS + BorderlineSMOTE) mencatat penurunan terbesar pada *Accuracy* (0,8655 ± 0,0131). Standar deviasi antarskenario relatif rendah (0,0120–0,0140 untuk *Accuracy*), yang menunjukkan stabilitas model di seluruh evaluasi.

Untuk menentukan apakah perbedaan performa antarskenario signifikan secara statistik, *Wilcoxon signed-rank test* dilakukan pada 25 pasangan skor CV

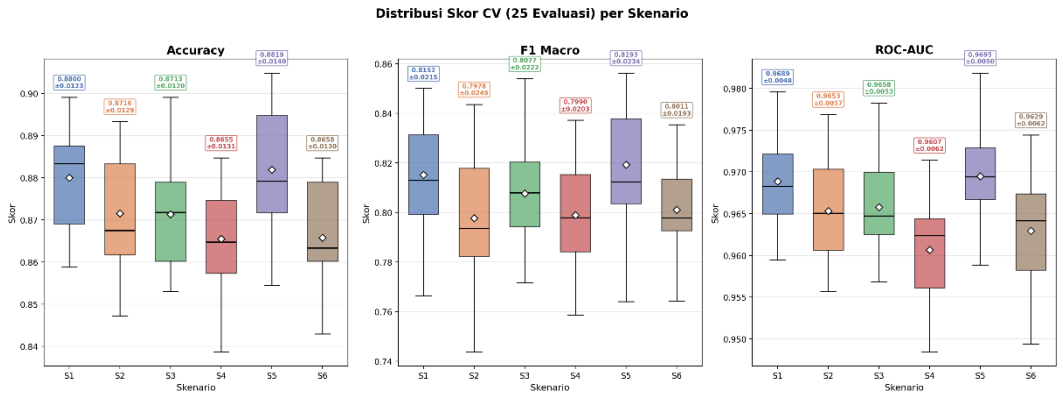
dengan tingkat signifikansi α = 0,05. Tabel 5.5 menyajikan hasil uji untuk ketiga metrik.

Tabel 5. 5 Hasil Uji *Wilcoxon Signed-Rank Test* terhadap *Baseline* (S1)

Perbandingan	***Accuracy* (p)**	**F1 Macro (p)**	**ROC-AUC (p)**
S2 vs S1	< 0,001*	< 0,001*	< 0,001*
S3 vs S1	0,001*	0,127	< 0,001*
S4 vs S1	< 0,001*	< 0,001*	< 0,001*
S5 vs S1	0,084	0,008*	0,034*
S6 vs S1	< 0,001*	< 0,001*	< 0,001*

***Keterangan**: *signifikan pada α = 0,05; uji dua sisi pada 25 pasangan skor CV.*

Gambar 5.2 memvisualisasikan distribusi skor dari 25 evaluasi CV untuk setiap skenario dalam bentuk *box plot*, yang menampilkan median, kuartil, dan *outlier* dari ketiga metrik.



Gambar 5. 2 Distribusi Skor CV (25 Evaluasi) per Skenario

Dari Gambar 5.2, terlihat bahwa distribusi skor S1 dan S5 saling tumpang tindih pada ketiga metrik, yang mengonfirmasi bahwa perbedaan performa keduanya sangat kecil. Sebaliknya, S4 secara konsisten memiliki distribusi terendah pada *Accuracy*, sementara distribusi F1 Macro S3 hampir seluruhnya tumpang

tindih dengan S1, yang sejalan dengan hasil Wilcoxon bahwa penurunan F1 Macro S3 tidak signifikan (p = 0,127).

Hasil uji statistik memberikan nuansa yang lebih presisi dibandingkan perbandingan rentang standar deviasi. Seleksi fitur (S2) dan kombinasi FS + BorderlineSMOTE (S4, S6) menghasilkan penurunan yang signifikan secara statistik pada ketiga metrik (p < 0,001). BorderlineSMOTE (S3) menurunkan *Accuracy* secara signifikan (p = 0,001), tetapi penurunan F1 Macro tidak signifikan (p = 0,127), yang mengindikasikan bahwa peningkatan *Recall* kelas Tinggi oleh BorderlineSMOTE cukup mengompensasi penurunan pada kelas lain sehingga F1 Macro tetap terjaga. Optuna (S5) tidak meningkatkan *Accuracy* secara signifikan (p = 0,084), tetapi peningkatan F1 Macro (p = 0,008) dan ROC-AUC (p = 0,034) signifikan, yang menunjukkan bahwa Optuna memperbaiki keseimbangan performa antarkelas meskipun perbaikannya kecil (+0,0041 F1 Macro).

Metrik keseluruhan didominasi oleh kelas Rendah yang proporsinya paling besar (65,7%). Analisis per kelas dari satu kali evaluasi pada *test set* (695 sampel) pada subbab berikutnya menunjukkan bahwa beberapa skenario justru memperbaiki deteksi kelas Tinggi secara nyata, meskipun dengan *trade-off* pada kelas lain.

5.3.2 Kontribusi *Feature Engineering*

Sebagai analisis pendukung, dilakukan pengujian tambahan untuk mengukur kontribusi 8 fitur interaksi hasil *feature engineering* terhadap performa model. Model *baseline* dilatih ulang menggunakan hanya 45 fitur dasar (26 drainase

+ 19 CHIRPS) tanpa 8 fitur interaksi, dengan konfigurasi *hyperparameter* dan skema evaluasi yang identik. Tabel 5.6 menyajikan perbandingan hasilnya.

Tabel 5. 6 Kontribusi *Feature Engineering*: 45 Fitur Dasar vs 53 Fitur (Repeated 5×5 CV)

Konfigurasi	***Accuracy* (Rata-rata ± Std)**	**F1 Macro (Rata-rata ± Std)**	**ROC-AUC (Rata-rata ± Std)**
45 fitur (tanpa interaksi)	0,8785 ± 0,0119	0,8154 ± 0,0199	0,9700 ± 0,0045
53 fitur / S1 (*Baseline*)	0,8800 ± 0,0123	0,8152 ± 0,0215	0,9689 ± 0,0048
Delta (53 − 45)	+0,0014	+0,0007	+0,0003

Wilcoxon signed-rank test menunjukkan bahwa perbedaan *Accuracy* signifikan secara marginal (p = 0,045), sedangkan perbedaan F1 Macro (p = 0,578) dan ROC-AUC (p = 0,063) tidak signifikan. Hasil ini mengindikasikan bahwa 8 fitur interaksi memberikan kontribusi yang sangat kecil terhadap performa model. XGBoost, sebagai algoritma berbasis pohon keputusan, mampu menangkap pola interaksi antar fitur secara implisit melalui kombinasi *split* pada berbagai kedalaman pohon, sehingga fitur interaksi eksplisit tidak banyak menambah informasi baru. Meskipun demikian, fitur interaksi tetap dipertahankan pada seluruh skenario utama (S1–S6) karena memberikan interpretabilitas yang lebih baik terhadap hubungan drainase-curah hujan dan konsisten dengan praktik *feature engineering* dalam literatur hidrologi.

5.3.3 Performa Per kelas

Tabel 5.7 menyajikan *Precision*, *Recall*, dan F1 *Score* per kelas untuk seluruh skenario dari satu kali evaluasi pada *test set* (695 sampel).

Tabel 5. 7 Performa Per kelas Seluruh Skenario

Skenario	**Kelas**	***Precision***	***Recall***	**F1 *Score***
S1	Rendah	0,9128	0,9408	0,9266
S1	Sedang	0,7513	0,7933	0,7717
S1	Tinggi	0,9722	0,5833	0,7292
S2	Rendah	0,9068	0,9386	0,9224
S2	Sedang	0,7556	0,7598	0,7577
S2	Tinggi	0,8372	0,6000	0,6990
S3	Rendah	0,9350	0,8838	0,9087
S3	Sedang	0,6919	0,7654	0,7268
S3	Tinggi	0,7424	0,8167	0,7778
S4	Rendah	0,9330	0,8860	0,9089
S4	Sedang	0,7143	0,7263	0,7202
S4	Tinggi	0,6625	0,8833	0,7571
S5	Rendah	0,9066	0,9364	0,9213
S5	Sedang	0,7316	0,7765	0,7534
S5	Tinggi	1,0000	0,5667	0,7234
S6	Rendah	0,9419	0,8882	0,9142
S6	Sedang	0,7120	0,7318	0,7218
S6	Tinggi	0,6543	0,8833	0,7518

Analisis per kelas memperlihatkan gambaran yang berbeda dibandingkan metrik keseluruhan.

1. **Kelas Tinggi**

   a) Skenario tanpa BorderlineSMOTE (S1, S2, S5) menghasilkan Precision tinggi, yaitu 0,8372–1,0000, tetapi Recall relatif rendah, yaitu 0,5667–0,6000.

b) Hasil tersebut menunjukkan bahwa model cenderung berhati-hati dalam memprediksi kelas Tinggi. Ketika model memberikan prediksi kelas Tinggi, prediksi tersebut umumnya benar, namun masih banyak sampel kelas Tinggi yang tidak berhasil terdeteksi.
c) S5 (Optuna) mencapai Precision sempurna sebesar 1,0000. Seluruh 34 sampel yang diprediksi sebagai kelas Tinggi merupakan prediksi yang benar, tetapi 26 dari 60 sampel kelas Tinggi tidak terdeteksi.
d) Skenario dengan BorderlineSMOTE (S3, S4, S6) menunjukkan pola yang berlawanan, dengan Recall meningkat menjadi 0,8167–0,8833.
e) Peningkatan Recall menunjukkan bahwa model mampu mendeteksi sekitar 49–53 dari 60 sampel kelas Tinggi.
f) Peningkatan tersebut diikuti penurunan Precision menjadi 0,6543–0,7424 karena model menjadi lebih agresif dalam memprediksi kelas Tinggi sehingga sebagian sampel kelas Sedang ikut terklasifikasi sebagai kelas Tinggi.
g) S4 (FS + BorderlineSMOTE) dan S6 (Full Pipeline) mencapai Recall tertinggi sebesar 0,8833 dengan F1-score kelas Tinggi masing-masing sebesar 0,7571 dan 0,7518.
h) Hasil ini menunjukkan bahwa BorderlineSMOTE efektif meningkatkan kemampuan model dalam mendeteksi sampel kelas Tinggi yang sebelumnya sulit dikenali.

2. Kelas Sedang

a) Precision kelas Sedang menurun pada skenario yang menggunakan BorderlineSMOTE, yaitu S3 (0,6919), S4 (0,7143), dan S6 (0,7120).
b) Penurunan Precision terjadi karena sebagian sampel yang sebelumnya diprediksi sebagai kelas Sedang beralih diprediksi sebagai kelas Tinggi.
c) Kondisi ini merupakan konsekuensi dari meningkatnya sensitivitas model terhadap kelas Tinggi setelah penerapan BorderlineSMOTE.
d) Dengan kata lain, peningkatan kemampuan deteksi kelas Tinggi diperoleh melalui pengorbanan sebagian performa pada kelas Sedang.

3. Kelas Rendah

a) Performa kelas Rendah relatif stabil pada seluruh skenario.
b) Recall kelas Rendah pada skenario BorderlineSMOTE berada pada rentang 0,8838–0,8882, lebih rendah dibandingkan baseline yang mencapai 0,9408.
c) Penurunan Recall tersebut diimbangi oleh peningkatan Precision sehingga kualitas prediksi kelas Rendah secara umum tetap terjaga.
d) Hasil ini menunjukkan bahwa pengaruh BorderlineSMOTE terhadap kelas Rendah relatif kecil dibandingkan pengaruhnya terhadap kelas Tinggi dan kelas Sedang.

Secara keseluruhan, penerapan BorderlineSMOTE efektif meningkatkan kemampuan model dalam mendeteksi kelas Tinggi, yang ditunjukkan oleh kenaikan Recall dari 0,5667–0,6000 menjadi 0,8167–0,8833. Peningkatan tersebut diikuti oleh penurunan Precision pada kelas Tinggi dan kelas Sedang karena model menjadi lebih agresif dalam memberikan prediksi kelas Tinggi. Sementara itu, kelas

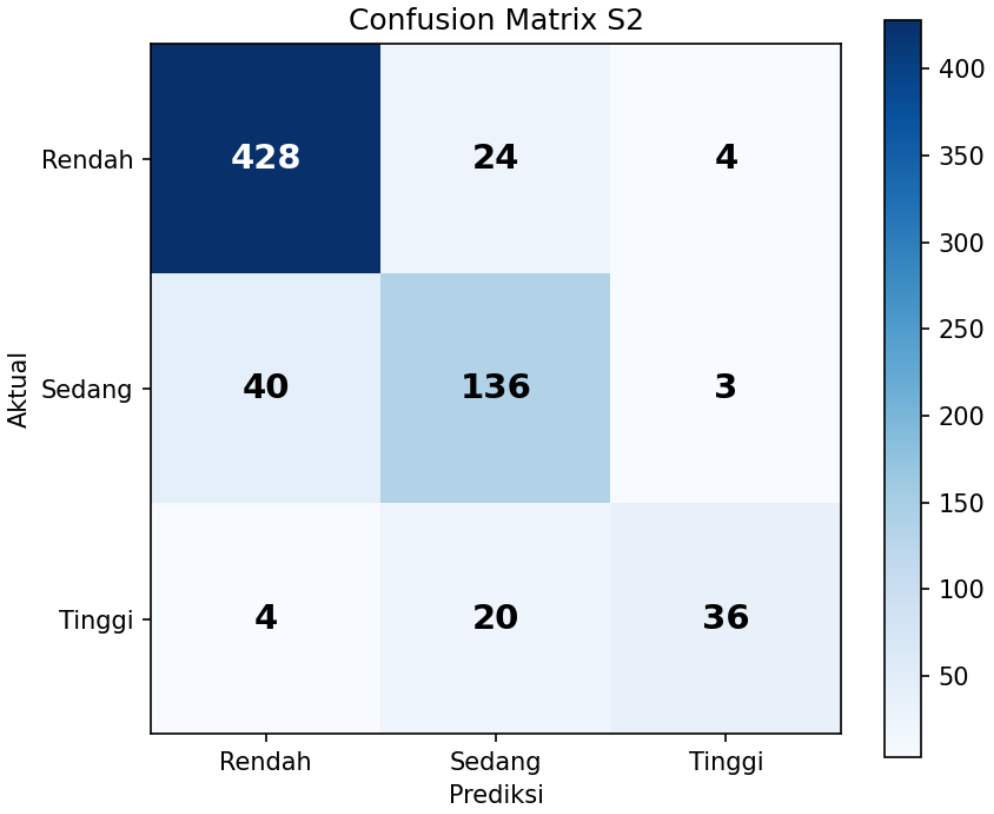
Rendah tetap menunjukkan performa yang relatif stabil dengan sedikit penurunan Recall yang diimbangi oleh peningkatan Precision. Temuan ini menunjukkan bahwa BorderlineSMOTE meningkatkan kemampuan deteksi kelas Tinggi dengan konsekuensi penurunan Precision pada kelas Tinggi dan kelas Sedang serta sedikit penurunan Recall pada kelas Rendah.

5.3.4 *Confusion Matrix*

Tabel 5.8-5.12 dan Gambar 5.3-5.7 menyajikan *confusion matrix* beserta visualisasi *heatmap* dari *test set* untuk setiap skenario, yang memungkinkan analisis pola kesalahan secara lebih detail.

Tabel 5. 8 *Confusion Matrix* S2 (Seleksi Fitur)

	Prediksi Rendah	**Prediksi Sedang**	**Prediksi Tinggi**
Aktual Rendah	**428**	24	4
Aktual Sedang	40	**136**	3
Aktual Tinggi	4	20	**36**

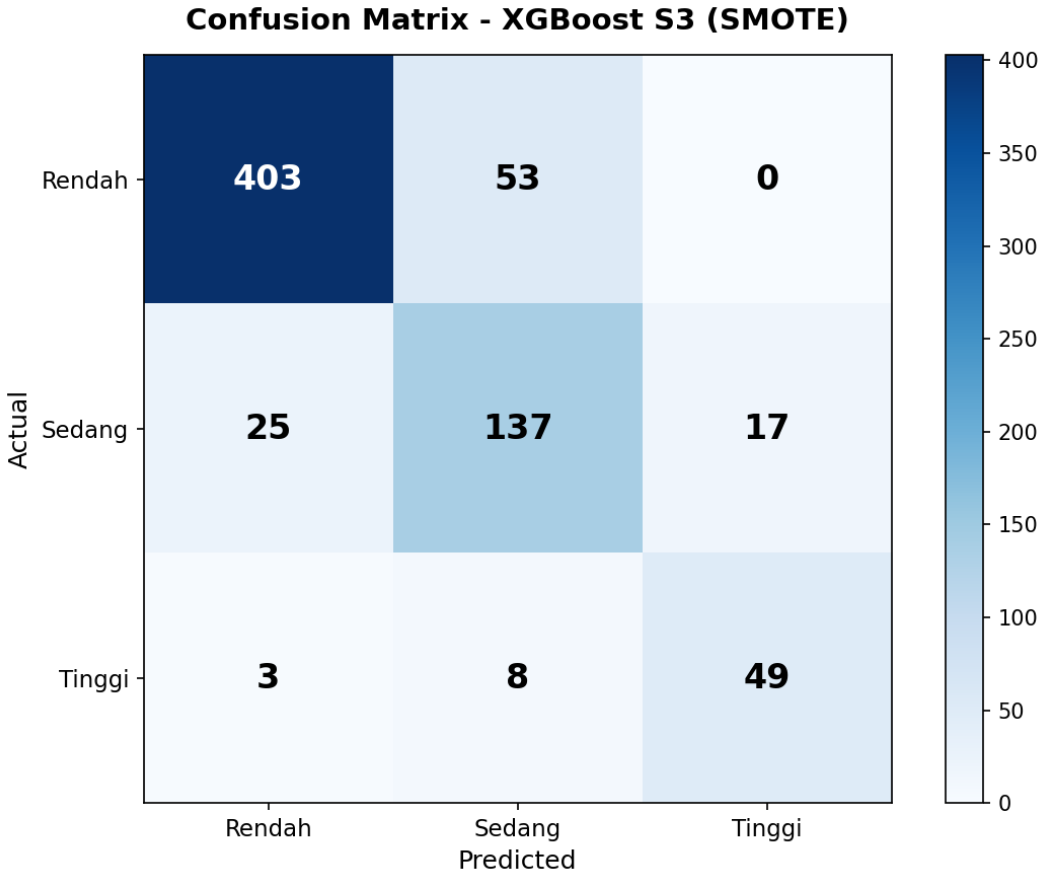


Gambar 5. 3 *Heatmap Confusion Matrix* S2 (Seleksi Fitur)

Pada S2, pola klasifikasi masih mirip dengan *baseline*: kelas Rendah terklasifikasi dengan baik (428/456), namun 20 dari 60 titik Tinggi masih salah diklasifikasikan sebagai Sedang. Seleksi fitur MI + Pearson + RFE (19 fitur) mempertahankan fitur curah hujan *rainy_days* dan *rainfall_30d_max*, sehingga *Recall* kelas Tinggi (0,6000) sedikit lebih baik dibandingkan *baseline* (0,5833).

Tabel 5. 9 *Confusion Matrix* S3 (BorderlineSMOTE)

	Prediksi Rendah	**Prediksi Sedang**	**Prediksi Tinggi**
Aktual Rendah	**403**	53	0
Aktual Sedang	25	**137**	17
Aktual Tinggi	3	8	**49**



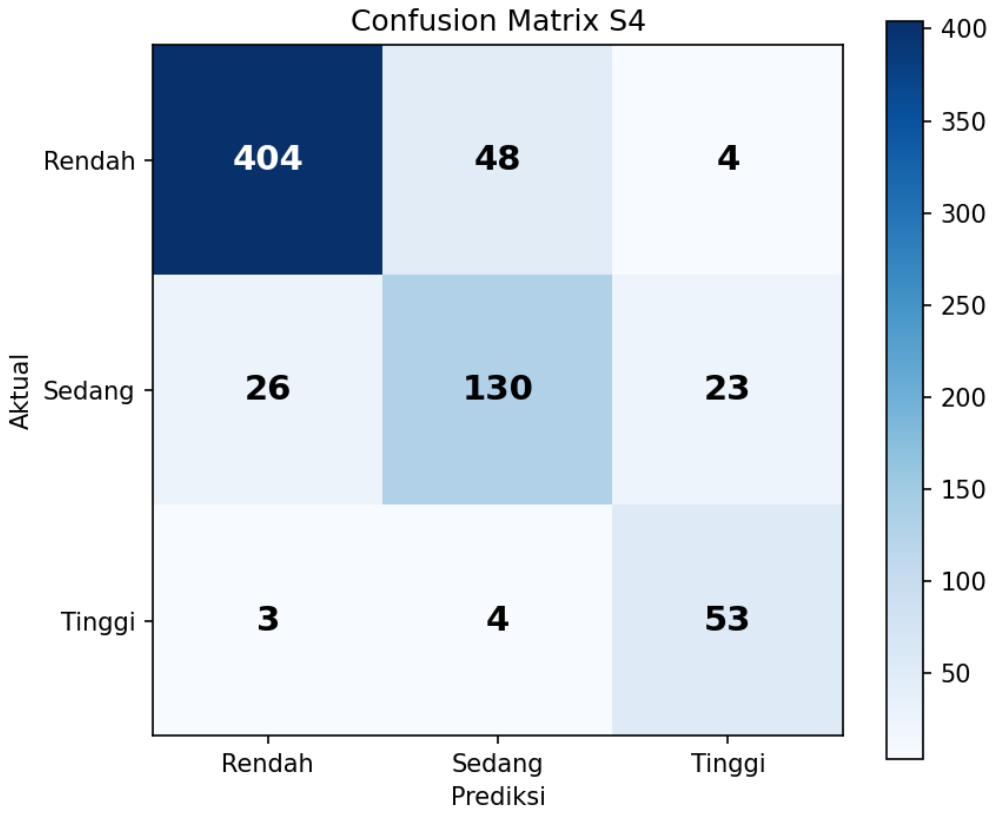
Gambar 5. 4 *Heatmap Confusion Matrix* S3 (BorderlineSMOTE)

Penerapan BorderlineSMOTE pada S3 memberikan perubahan paling mencolok pada kelas Tinggi: misklasifikasi Tinggi→Sedang turun drastis dari 21 (S1) menjadi 8 titik. Sebagai kompensasi, 53 titik Rendah salah diklasifikasikan sebagai Sedang (vs 26 pada S1) karena batas keputusan bergeser akibat sampel sintetis pada kelas Sedang dan Tinggi.

Tabel 5. 10 *Confusion Matrix* S4 (FS + BorderlineSMOTE)

	Prediksi Rendah	**Prediksi Sedang**	**Prediksi Tinggi**
Aktual Rendah	**404**	48	4
Aktual Sedang	26	**130**	23

Aktual Tinggi	3	4	**53**

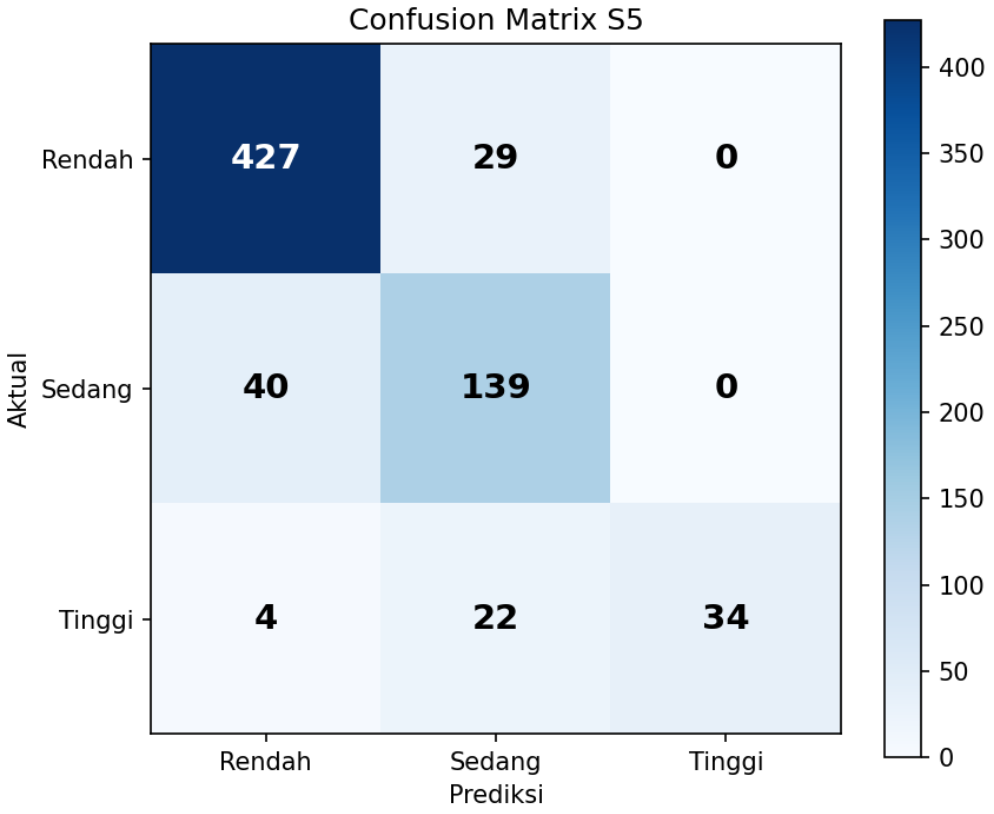


Gambar 5. 5 *Heatmap Confusion Matrix* S4 (FS + BorderlineSMOTE)

Kombinasi seleksi fitur dan BorderlineSMOTE pada S4 menghasilkan *Recall* kelas Tinggi sebesar 53/60 = 0,8833, hanya 4 salah ke Sedang dan 3 ke Rendah. Dibandingkan S3 yang juga menggunakan BorderlineSMOTE, pengurangan fitur dari 53 ke 19 justru meningkatkan kemampuan model mendeteksi kelas Tinggi karena fitur curah hujan yang dipertahankan (*rainy_days*) berinteraksi lebih efektif dengan fitur drainase setelah fitur redundan dihilangkan.

Tabel 5. 11 *Confusion Matrix* S5 (Optuna)

	Prediksi Rendah	Prediksi Sedang	Prediksi Tinggi
Aktual Rendah	**427**	29	0
Aktual Sedang	40	**139**	0
Aktual Tinggi	4	22	**34**

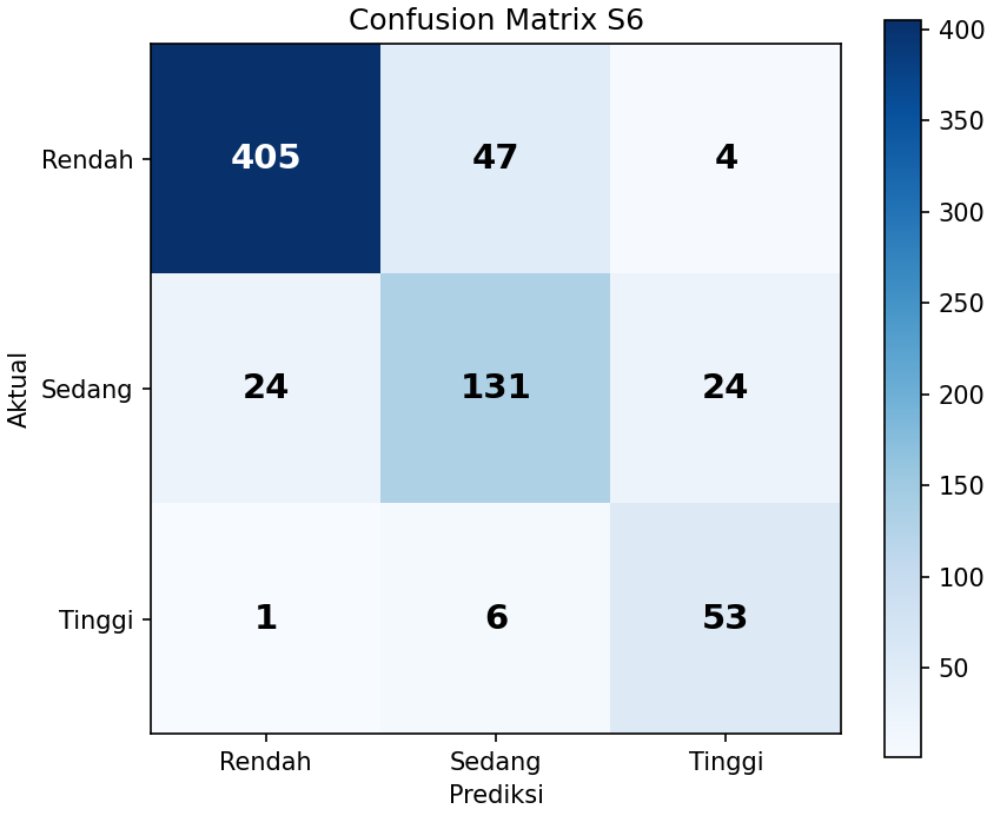


Gambar 5. 6 *Heatmap Confusion Matrix* S5 (Optuna)

S5 (Optuna tanpa BorderlineSMOTE) menunjukkan pola yang mirip dengan *baseline*: 22 titik Tinggi tetap salah diklasifikasikan sebagai Sedang, dan 0 sampel Sedang yang diprediksi Tinggi. Hal ini menegaskan bahwa optimasi *hyperparameter* saja tidak cukup untuk mengatasi bias terhadap kelas minoritas; penanganan distribusi data melalui BorderlineSMOTE diperlukan secara eksplisit.

Tabel 5. 12 *Confusion Matrix* S6 (*Full Pipeline*)

	Prediksi Rendah	Prediksi Sedang	Prediksi Tinggi
Aktual Rendah	**405**	47	4
Aktual Sedang	24	**131**	24
Aktual Tinggi	1	6	**53**



Gambar 5. 7 *Heatmap Confusion Matrix* S6 (Full Pipeline)

S6 menghasilkan *Recall* kelas Tinggi 0,8833 (53/60), dengan hanya 6 titik salah ke Sedang dan 1 ke Rendah. Hasil ini identik dengan S4 pada *Recall* kelas Tinggi (0,8833), namun S6 memiliki lebih sedikit misklasifikasi Tinggi→Rendah (1 vs 3 pada S4), meskipun misklasifikasi Tinggi→Sedang sedikit lebih banyak (6 vs 4). Dari perspektif ablasi, *confusion matrix* lintas skenario memperlihatkan BorderlineSMOTE sebagai komponen yang paling menentukan kemampuan model

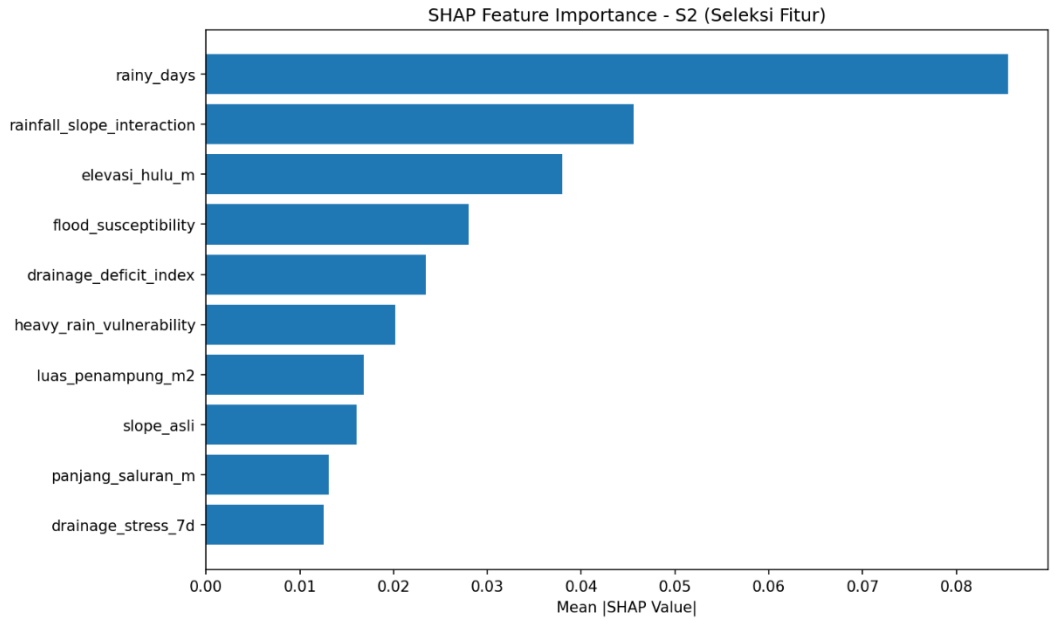
mengenali kelas Tinggi, sementara optimasi *hyperparameter* (S5) tidak memberikan perubahan berarti pada pola kesalahan klasifikasi.

5.3.5 Analisis *Feature Importance* (SHAP)

Tabel 5.13-5.17 dan Gambar 5.8-5.12 menyajikan lima fitur teratas berdasarkan nilai SHAP rata-rata beserta visualisasi *bar plot* untuk masing-masing skenario S2–S6. Penyajian dibatasi pada lima fitur teratas (berbeda dari sepuluh fitur pada *baseline* di Bab IV) karena skenario dengan seleksi fitur (S2, S4, S6) hanya memiliki 19 fitur, sehingga konsentrasi *importance* pada fitur-fitur teratas lebih tinggi dan lima fitur sudah cukup merepresentasikan pola utama. Analisis SHAP untuk *baseline* (S1) telah dibahas pada Bab IV (Gambar 4.4), sehingga di sini fokus diberikan pada perubahan *feature importance* akibat perbedaan konfigurasi *preprocessing*.

Tabel 5. 13 Top-5 Fitur SHAP S2 (Seleksi Fitur)

Peringkat	**Fitur**	**Nilai SHAP**
1	*rainy_days*	0,0855
2	*rainfall_slope_interaction*	0,0456
3	*elevasi_hulu_m*	0,0380
4	*flood_susceptibility*	0,0280
5	*drainage_deficit_index*	0,0234

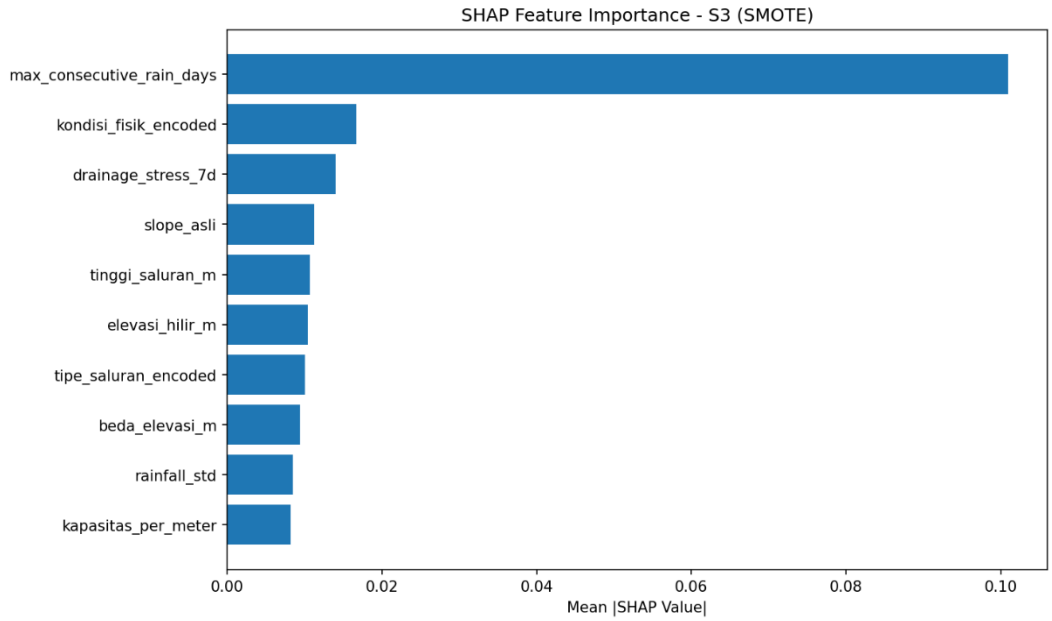


Gambar 5. 8 SHAP *Summary Plot* (Bar) S2 (Seleksi Fitur)

Pada S2, fitur curah hujan *rainy_days* tetap mendominasi dengan nilai SHAP 0,0855, diikuti fitur interaksi *rainfall_slope_interaction* (0,0456) dan *flood_susceptibility* (0,0280). *pipeline* MI + Pearson + RFE mempertahankan *rainy_days* dan *rainfall_30d_max*, sehingga model tetap memiliki akses terhadap informasi temporal curah hujan. Distribusi *importance* lebih merata dibandingkan *baseline* karena fitur redundan sudah dihilangkan.

Tabel 5. 14 Top-5 Fitur SHAP S3 (BorderlineSMOTE)

Peringkat	**Fitur**	**Nilai SHAP**
1	*max_consecutive_rain_days*	0,1010
2	*kondisi_fisik_encoded*	0,0167
3	*drainage_stress_7d*	0,0140
4	*slope_asli*	0,0113
5	*tinggi_saluran_m*	0,0107

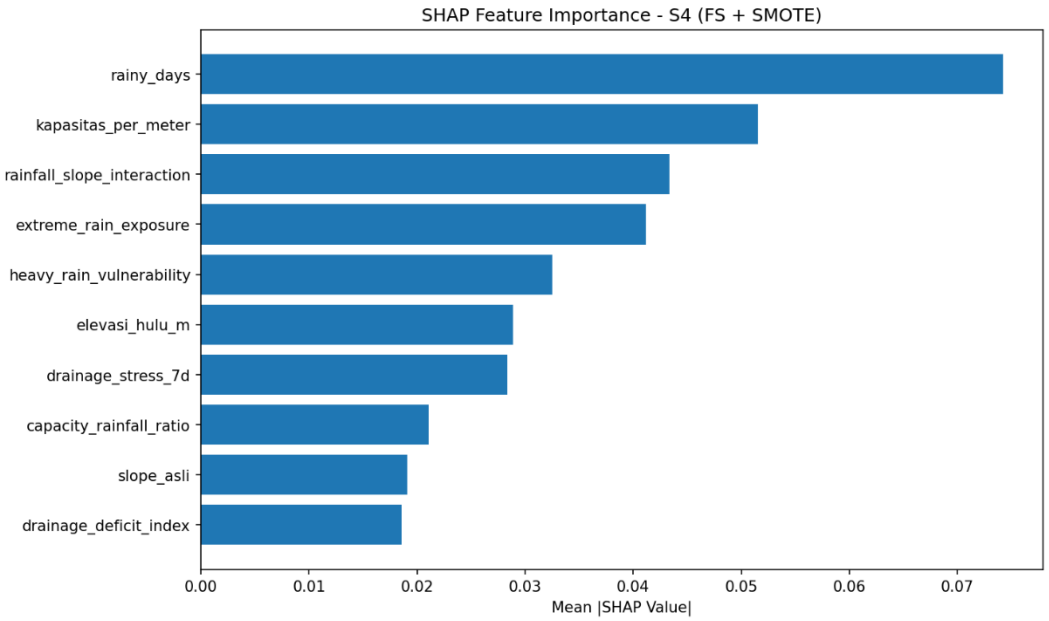


Gambar 5. 9 SHAP *Summary Plot* (Bar) S3 (BorderlineSMOTE)

Pada S3 (BorderlineSMOTE tanpa seleksi fitur), fitur teratas bergeser ke *max_consecutive_rain_days* (0,1010), yaitu fitur curah hujan temporal yang mengukur durasi hujan berturut-turut terpanjang. Ini berbeda dari *baseline* yang didominasi *rainy_days* (total hari hujan). BorderlineSMOTE mengubah distribusi sampel di sekitar *decision boundary*, yang tampaknya membuat fitur durasi hujan berturut-turut lebih informatif untuk membedakan kelas Tinggi dari kelas lain. Fitur *kondisi_fisik_encoded* naik ke peringkat 2 (0,0167), diikuti *drainage_stress_7d* (0,0140).

Tabel 5. 15 Top-5 Fitur SHAP S4 (FS + BorderlineSMOTE)

Peringkat	Fitur	Nilai SHAP
1	*rainy_days*	0,0742
2	*kapasitas_per_meter*	0,0516
3	*rainfall_slope_interaction*	0,0433
4	*extreme_rain_exposure*	0,0412
5	*heavy_rain_vulnerability*	0,0325

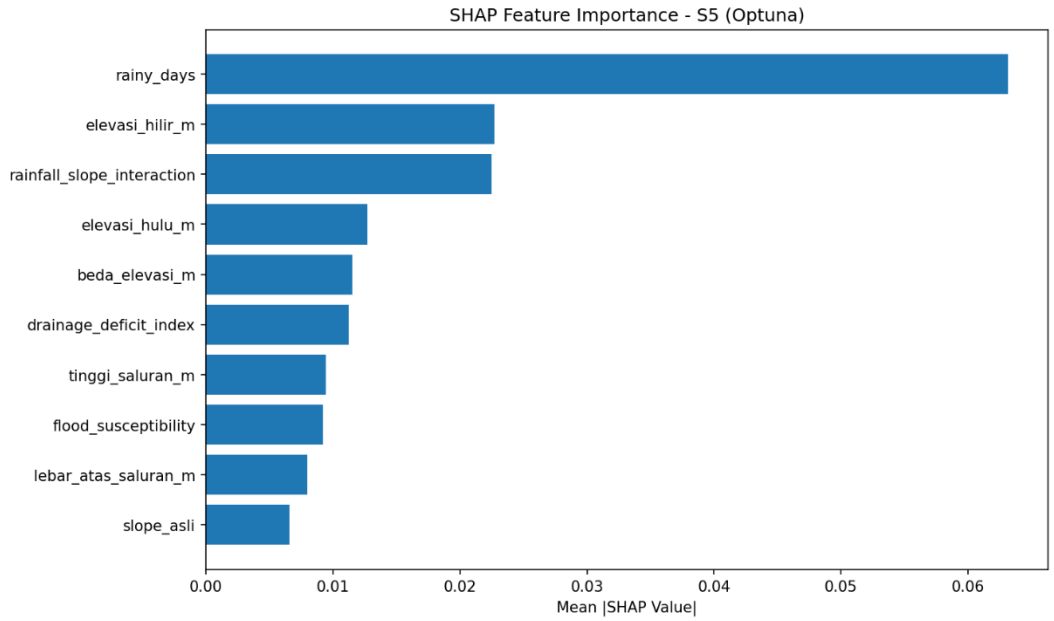


Gambar 5. 10 SHAP *Summary Plot* (Bar) S4 (FS + BorderlineSMOTE)

S4 menggabungkan seleksi fitur dan BorderlineSMOTE. *Rainy_days* tetap di posisi teratas (0,0742), namun fitur drainase *kapasitas_per_meter* naik ke peringkat 2 (0,0516), diikuti fitur interaksi *rainfall_slope_interaction* (0,0433) dan *extreme_rain_exposure* (0,0412). Munculnya *kapasitas_per_meter* di peringkat tinggi pada S4 mengindikasikan bahwa BorderlineSMOTE memperkuat peran fitur kapasitas drainase dalam membedakan kelas Tinggi dari kelas lain.

Tabel 5. 16 Top-5 Fitur SHAP S5 (Optuna)

Peringkat	Fitur	Nilai SHAP
1	*rainy_days*	0,0631
2	*elevasi_hilir_m*	0,0227
3	*rainfall_slope_interaction*	0,0225
4	*elevasi_hulu_m*	0,0127
5	*beda_elevasi_m*	0,0115

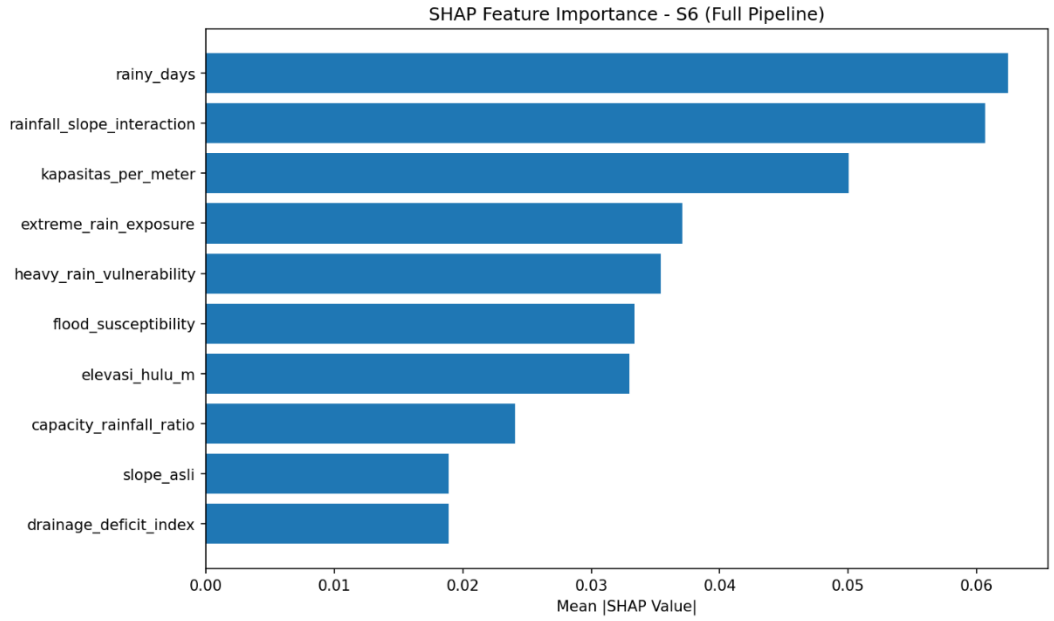


Gambar 5. 11 SHAP *Summary Plot* (Bar) S5 (Optuna)

S5 (Optuna tanpa seleksi fitur dan BorderlineSMOTE) menunjukkan pola yang mirip dengan *baseline*: *rainy_days* tetap di posisi teratas (0,0631), diikuti *elevasi_hilir_m* (0,0227) dan *rainfall_slope_interaction* (0,0225). Meskipun konfigurasi *hyperparameter* berubah (163 pohon vs 200,*max_depth* 4 vs 6), ranking fitur hampir tidak berubah. Ini mengonfirmasi bahwa *feature importance* lebih ditentukan oleh komposisi fitur dan distribusi data daripada konfigurasi *hyperparameter*.

Tabel 5. 17 Top-5 Fitur SHAP S6 (*Full Pipeline*)

Peringkat	Fitur	Nilai SHAP
1	*rainy_days*	0,0624
2	*rainfall_slope_interaction*	0,0607
3	*kapasitas_per_meter*	0,0500
4	*extreme_rain_exposure*	0,0371
5	*heavy_rain_vulnerability*	0,0354



Gambar 5. 12 SHAP *Summary Plot* (Bar) S6 (*Full Pipeline*)

S6 (*Full Pipeline*) menghasilkan pola *importance* yang konsisten dengan S4: *rainy_days* tetap di posisi teratas (0,0624), diikuti *rainfall_slope_interaction* (0,0607) dan *kapasitas_per_meter* (0,0500). Distribusi *importance* pada S6 lebih merata dibandingkan skenario lain, dengan selisih antara fitur teratas dan kelima hanya 0,0270, yang mengindikasikan bahwa kombinasi seleksi fitur, BorderlineSMOTE, dan Optuna menghasilkan model yang memanfaatkan lebih banyak fitur secara seimbang.

Dari tabel dan gambar di atas, dapat diidentifikasi pola lintas skenario yang konsisten. Fitur *rainy_days* mendominasi lima dari enam skenario (S1, S2, S4, S5, S6) dengan nilai SHAP 0,0615–0,0855. Pada S3 (BorderlineSMOTE tanpa seleksi fitur), *max_consecutive_rain_days* mengambil alih posisi teratas, tetapi ini tetap merupakan fitur curah hujan. Pola ini membuktikan bahwa *pipeline* MI + Pearson + RFE berhasil mempertahankan fitur curah hujan yang informatif.

5.3.6 Analisis Kontribusi Komponen

Berdasarkan seluruh hasil pengujian, kontribusi masing-masing komponen dapat dirangkum sebagai berikut. Angka performa keseluruhan mengacu pada hasil *Repeated 5×5 CV* (Tabel 5.4), sedangkan angka per kelas mengacu pada analisis *test set* (Tabel 5.6).

Seleksi fitur MI + Pearson + RFE (S2 vs S1) mereduksi jumlah fitur dari 53 menjadi 19 (pengurangan 64,2%), dengan penurunan F1 Macro CV sebesar 0,0174 (dari 0,8152 menjadi 0,7978). Penurunan ini signifikan secara statistik berdasarkan *Wilcoxon signed-rank test* ($p < 0,001$). *Pipeline* MI + Pearson + RFE mempertahankan 2 fitur curah hujan (*rainy_days* dan *rainfall_30d_max*) yang terbukti informatif, sehingga penurunan performa tetap terkendali meskipun 64,2% fitur dieliminasi.

BorderlineSMOTE (S3 vs S1) meningkatkan *Recall* kelas Tinggi dari 0,5833 menjadi 0,8167 (+0,2334) pada *test set*, yang berarti model berhasil mendeteksi 14 titik berisiko tinggi tambahan yang sebelumnya terlewat. *Accuracy* CV turun secara signifikan dari 0,8800 menjadi 0,8713 ($p = 0,001$), namun

penurunan F1 Macro (0,8152 menjadi 0,8077) tidak signifikan secara statistik (p = 0,127). Temuan ini mengindikasikan bahwa peningkatan *Recall* kelas Tinggi oleh BorderlineSMOTE cukup mengompensasi penurunan pada kelas lain, sehingga keseimbangan performa lintas kelas (F1 Macro) tetap terjaga. *Trade-off* pada *test set* berupa penurunan *Precision* kelas Tinggi dari 0,9722 menjadi 0,7424 (−0,2298), namun dari perspektif mitigasi bencana, kemampuan mendeteksi lebih banyak titik berisiko tinggi umumnya lebih diutamakan daripada menghindari *false positive*.

Optimasi *hyperparameter* Optuna TPE (S5 vs S1) memberikan perbaikan yang kecil berdasarkan CV: *Accuracy* naik dari 0,8800 menjadi 0,8819 (+0,0019), F1 Macro naik dari 0,8152 menjadi 0,8193 (+0,0041), dan ROC-AUC naik dari 0,9689 menjadi 0,9695 (+0,0006). Uji *Wilcoxon signed-rank test* menunjukkan bahwa peningkatan *Accuracy* tidak signifikan (p = 0,084), tetapi peningkatan F1 Macro (p = 0,008) dan ROC-AUC (p = 0,034) signifikan pada α = 0,05. Meskipun peningkatan absolut kecil, Optuna berhasil memperbaiki keseimbangan performa antarkelas secara terukur. Pada analisis *test set*, perubahan yang paling mencolok adalah pada kelas Tinggi: *Precision* meningkat menjadi 1,0000, tetapi *Recall* turun menjadi 0,5667 (−0,0166), yang berarti model hanya mendeteksi 34 dari 60 titik berisiko tinggi.

Kombinasi seleksi fitur dan BorderlineSMOTE (S4 vs S1) menghasilkan *Recall* kelas Tinggi sebesar 0,8833 (+0,3000) pada *test set*, lebih tinggi dari S3 yang hanya menggunakan BorderlineSMOTE (0,8167). *Accuracy* CV turun menjadi 0,8655 (−0,0145 dari *baseline*), karena fitur curah hujan yang dipertahankan oleh *pipeline* MI + Pearson + RFE membantu menstabilkan performa.

Pipeline lengkap S6 (seleksi fitur + BorderlineSMOTE + Optuna TPE) mencapai *Recall* kelas Tinggi sebesar 0,8833 pada *test set*, yang berarti 53 dari 60 titik berisiko tinggi berhasil terdeteksi. *Accuracy* CV S6 (0,8658) sedikit di atas S4 (0,8655), artinya penambahan Optuna pada *pipeline* lengkap memberikan kontribusi yang marginal.

Dari enam skenario ini, BorderlineSMOTE merupakan komponen dengan dampak terbesar terhadap deteksi titik berisiko tinggi, seleksi fitur MI + Pearson + RFE berhasil mereduksi dimensi sambil mempertahankan fitur curah hujan yang informatif, dan Optuna TPE memberikan perbaikan yang kecil secara absolut namun signifikan secara statistik pada F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$), meskipun tidak signifikan pada *Accuracy* ($p = 0,084$).

5.3.7 Ringkasan Temuan

Berdasarkan *Repeated 5×5 Stratified K-Fold Cross-Validation*, S5 (Optuna) sedikit mengungguli *baseline* pada ketiga metrik (*Accuracy* 0,8819 vs 0,8800; F1 Macro 0,8193 vs 0,8152; ROC-AUC 0,9695 vs 0,9689). *Wilcoxon signed-rank test* menunjukkan peningkatan F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$) signifikan pada $\alpha = 0,05$, meskipun peningkatan *Accuracy* tidak signifikan ($p = 0,084$). Skenario lain (S2, S3, S4, S6) menurunkan *Accuracy* CV secara signifikan ($p \leq 0,001$), dengan penurunan terbesar pada S4 (0,8655).

Pada *test set*, BorderlineSMOTE meningkatkan *Recall* kelas Tinggi dari 0,5833 menjadi 0,8167–0,8833, sehingga 49–53 dari 60 titik berisiko tinggi

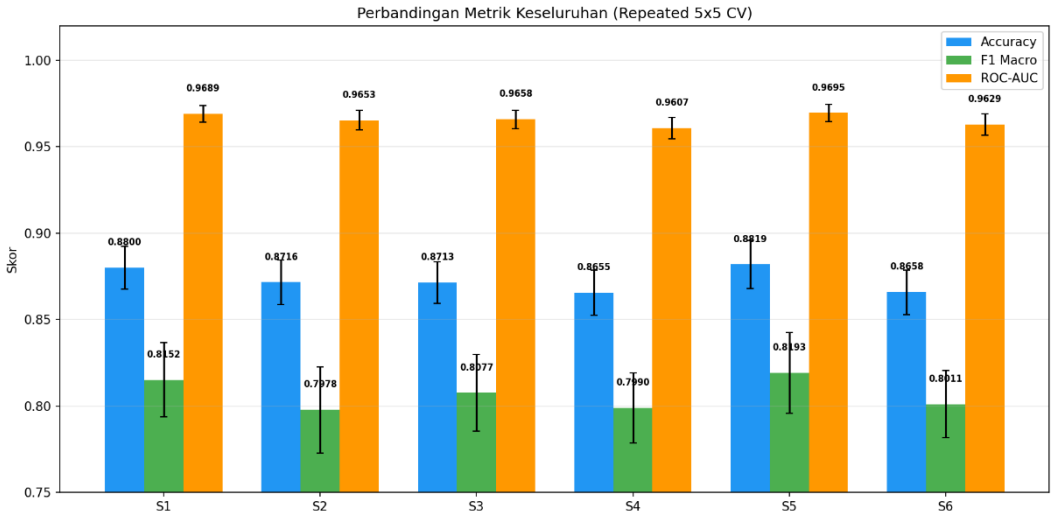
terdeteksi. Seleksi fitur MI + Pearson + RFE mereduksi dimensi dari 53 menjadi 19 fitur (pengurangan 64,2%) dengan penurunan F1 Macro CV sebesar 0,0174.

BAB VI
PEMBAHASAN

Bab IV dan V telah menyajikan hasil pengujian enam skenario klasifikasi risiko banjir menggunakan XGBoost pada dataset drainase Kota Malang. Bab ini membahas temuan-temuan tersebut secara lebih luas: kontribusi studi ablasi ini, posisinya terhadap penelitian terdahulu, serta implikasinya bagi pengelolaan risiko banjir di lapangan.

6.1 Interpretasi Hasil Studi Ablasi

Gambar 6.1 memvisualisasikan perbandingan metrik keseluruhan (*Accuracy*, F1 Macro, dan ROC-AUC) dari keenam skenario untuk memberikan gambaran umum mengenai hasil studi ablasi.



Gambar 6. 1 Perbandingan Metrik Keseluruhan (*Accuracy*, F1 Macro, ROC-AUC) antarskenario

Hasil studi ablasi berdasarkan *Repeated 5×5 Stratified K-Fold Cross-Validation* menunjukkan bahwa model *baseline* (S1) hanya sedikit terlampaui oleh

S5 (Optuna) pada metrik keseluruhan (*Accuracy* 0,8819 vs 0,8800; F1 Macro 0,8193 vs 0,8152). *Wilcoxon signed-rank test* pada 25 pasangan skor CV menunjukkan bahwa peningkatan F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$) signifikan pada $\alpha = 0,05$, meskipun peningkatan *Accuracy* tidak signifikan ($p = 0,084$). Skenario lainnya menurunkan kedua metrik tersebut secara signifikan ($p \leq 0,001$ untuk *Accuracy*), kecuali penurunan F1 Macro pada S3 yang tidak signifikan ($p = 0,127$). Hasil ini berbeda dari ekspektasi umum dalam literatur, mengingat *preprocessing* dan optimasi yang umumnya dianggap meningkatkan performa justru memberikan efek yang minimal atau negatif pada metrik agregat.

Terdapat beberapa penjelasan yang dapat dikemukakan. Pertama, *default hyperparameter* yang digunakan pada *baseline* (200 pohon, *max_depth* 6, *learning_rate* 0,1) ternyata sudah cukup baik untuk dataset ini. S5 (Optuna), yang mengevaluasi ratusan konfigurasi alternatif, hanya menghasilkan selisih +0,0019 pada *Accuracy* CV dan +0,0006 pada ROC-AUC CV. Meskipun peningkatan absolut kecil, *Wilcoxon signed-rank test* mengonfirmasi bahwa peningkatan F1 Macro dan ROC-AUC signifikan secara statistik, yang menunjukkan bahwa Optuna berhasil memperbaiki keseimbangan performa antarkelas secara konsisten meskipun ruang perbaikan sangat sempit.

Kedua, seleksi fitur melalui *pipeline* MI + Pearson + RFE mereduksi dimensi dari 53 menjadi 19 fitur (64,2%). *Mutual Information* mengevaluasi dependensi setiap fitur terhadap target secara independen, sehingga fitur curah hujan tidak otomatis tereliminasi hanya karena saling berkorelasi tinggi. Hasilnya, *pipeline* ini berhasil mempertahankan 2 fitur curah hujan CHIRPS (*rainy_days* dan

rainfall_30d_max) beserta 7 fitur interaksi, dengan penurunan F1 Macro CV sebesar 0,0174. Fitur *rainy_days* yang dipertahankan ini terbukti tetap mendominasi analisis SHAP bahkan pada skenario dengan seleksi fitur.

Ketiga, BorderlineSMOTE mengubah perilaku model secara mendasar terhadap kelas Tinggi. Dari analisis *test set*, *Recall* kelas Tinggi naik dari 0,5833 (S1) menjadi 0,8167 (S3) dan 0,8833 (S4), yang berarti 49–53 dari 60 titik berisiko tinggi terdeteksi, dibandingkan hanya 35 pada *baseline*. Peningkatan ini datang dengan menurunnya *Accuracy* CV (0,8713 pada S3 dan 0,8655 pada S4, dibandingkan 0,8800 pada S1), karena batas keputusan bergeser dan menyebabkan lebih banyak titik Rendah salah terprediksi sebagai Sedang. Metrik keseluruhan yang turun, pada kasus ini, tidak selalu berarti model menjadi lebih buruk; interpretasinya bergantung pada prioritas dalam konteks aplikasi.

6.2 Perbandingan dengan Penelitian Terdahulu

Temuan dari studi ablasi pada bagian sebelumnya perlu ditempatkan dalam konteks penelitian yang lebih luas. Bagian ini membandingkan hasil penelitian dengan beberapa studi terdahulu pada empat aspek: performa model XGBoost, efektivitas BorderlineSMOTE dalam penanganan ketidakseimbangan kelas, peran seleksi fitur dan data drainase, serta interpretabilitas model melalui SHAP. Perbandingan langsung antar penelitian tentu memiliki keterbatasan karena perbedaan dataset, definisi kelas target, dan konfigurasi eksperimen, tetapi pola umum yang ditemukan dapat memberikan perspektif mengenai posisi penelitian ini relatif terhadap literatur.

6.2.1 Performa Model XGBoost

Perbandingan langsung performa antar penelitian perlu dilakukan dengan hati-hati karena perbedaan dataset, jumlah kelas, dan definisi target. Meskipun demikian, beberapa pola dapat diidentifikasi.

Accuracy baseline penelitian ini (88,00% berdasarkan *Repeated 5×5 CV*) berada di bawah beberapa studi yang menggunakan klasifikasi biner: Qin *et al.* (2025) mencatat *overall accuracy* 0,96 untuk XGBoost dengan BorderlineSMOTE, dan Shrestha *et al.* (2025) mencapai 95,83% untuk XGBoost (meskipun Logistic Regression lebih unggul dengan 97,92%). Perbedaan ini wajar mengingat penelitian ini menggunakan tiga kelas (Rendah, Sedang, Tinggi) yang secara inheren lebih sulit daripada klasifikasi biner (banjir vs non-banjir), karena model harus membedakan batas antar tiga kelas yang distribusinya tidak seimbang.

ROC-AUC *baseline* (0,9689 ± 0,0048 berdasarkan *Repeated 5×5 CV*) lebih tinggi dibandingkan Goumghar *et al.* (2025) yang mencatat AUC = 0,727 untuk XGBoost di Upper Drâa Basin, dan sebanding dengan Bersabe & Jun (2025) yang melaporkan AUC = 0,902 untuk Random Forest di Seoul. Perbedaan ini memperkuat observasi Goumghar *et al.* bahwa performa relatif antar algoritma sangat bergantung pada karakteristik dataset dan konteks wilayah.

6.2.2 Efektivitas BorderlineSMOTE

Penerapan BorderlineSMOTE pada penelitian ini memberikan pola yang konsisten dengan Qin *et al.* (2025), yang melaporkan bahwa SMOTE dengan *StratifiedKFold* meningkatkan performa XGBoost pada dataset banjir yang tidak

seimbang. Perbedaannya adalah Qin *et al.* tidak melaporkan *trade-off* terhadap kelas mayoritas, sementara pada penelitian ini efek samping terhadap kelas Rendah dan Sedang terukur dengan jelas. Sebagai contoh, titik Rendah yang salah terprediksi sebagai Sedang naik dari 26 (S1) menjadi 53 (S3). Studi ablasi pada penelitian ini memungkinkan pengukuran *trade-off* ini secara eksplisit, sesuatu yang tidak tersedia pada penelitian terdahulu yang menggunakan BorderlineSMOTE dalam konfigurasi tunggal.

6.2.3 Seleksi Fitur dan Peran Data Drainase

Bersabe & Jun (2025) menunjukkan bahwa penambahan variabel drainase (*Sewer Pipe Density* dan *Drain Station Density*) meningkatkan akurasi sebesar 7,58% dan AUC sebesar 3,80%. Penelitian ini memperkuat temuan tersebut dari sudut pandang yang berbeda: dari 19 fitur yang dipertahankan oleh *pipeline* MI + Pearson + RFE, 10 merupakan fitur drainase fisik, 2 fitur curah hujan CHIRPS (*rainy_days* dan *rainfall_30d_max*), dan 7 fitur interaksi yang menggabungkan informasi drainase dan curah hujan. Dominasi fitur drainase dalam hasil seleksi menunjukkan bahwa fitur-fitur ini memiliki informasi yang lebih unik dan kurang redundan dibandingkan fitur curah hujan lainnya.

Jang *et al.* (2025) melaporkan bahwa fitur statistik curah hujan (*mean*, standar deviasi, *skewness*, *kurtosis*) menghasilkan prediksi yang lebih baik dibandingkan total curah hujan (R^2 = 0,9761 vs 0,9682). Pada penelitian ini, fitur curah hujan CHIRPS juga mencakup fitur-fitur statistik serupa (*rainfall_mean*, *rainfall_std*, *rainfall_skewness*, *rainfall_kurtosis*). Sebagian besar fitur statistik ini tereliminasi oleh tahap korelasi Pearson karena saling berkorelasi tinggi, tetapi

pipeline MI + Pearson + RFE tetap mempertahankan *rainy_days* (jumlah hari hujan) dan *rainfall_30d_max* (curah hujan maksimum 30 hari) yang merepresentasikan aspek frekuensi dan intensitas curah hujan.

6.2.4 Interpretabilitas Model

Wang *et al.* (2023) dan Qin *et al.* (2025) menggunakan SHAP untuk menginterpretasi model XGBoost dalam konteks kerentanan banjir. Pendekatan serupa diterapkan pada penelitian ini, tetapi dengan penambahan analisis SHAP lintas skenario yang belum dilakukan oleh studi-studi sebelumnya. Perbandingan SHAP antarskenario mengungkap bahwa fitur curah hujan *rainy_days* mendominasi lima dari enam skenario (S1, S2, S4, S5, S6) dengan nilai SHAP 0,0615–0,0855, termasuk skenario dengan seleksi fitur. Konsistensi ini dimungkinkan karena *pipeline* MI + Pearson + RFE berhasil mempertahankan *rainy_days* dalam set fitur terpilih. Skenario S3 (BorderlineSMOTE tanpa seleksi fitur) merupakan satu-satunya pengecualian, di mana *max_consecutive_rain_days* menggantikan posisi teratas. Temuan ini menunjukkan bahwa *ranking* fitur tetap relatif stabil ketika fitur informatif tidak tereliminasi oleh tahap seleksi, berbeda dari situasi di mana seleksi fitur berbasis multikolinearitas menghapus fitur penting dan menyebabkan pergeseran *ranking* yang drastis.

6.3 Peran Fitur Drainase dan Curah Hujan dalam Klasifikasi

Analisis SHAP pada enam skenario memberikan gambaran mengenai kontribusi relatif fitur drainase fisik dan curah hujan CHIRPS. Fitur curah hujan *rainy_days* mendominasi peringkat teratas pada lima dari enam skenario (S1, S2, S4, S5, S6) dengan nilai SHAP berkisar 0,0615–0,0855. Dominasi ini mencakup

skenario dengan seleksi fitur (S2, S4, S6), yang dimungkinkan karena *pipeline* MI + Pearson + RFE mempertahankan *rainy_days* dalam set fitur terpilih. Skenario S3 (BorderlineSMOTE tanpa seleksi fitur) merupakan pengecualian, di mana *max_consecutive_rain_days* menempati posisi teratas, kemungkinan karena pergeseran batas keputusan akibat *oversampling* menjadikan durasi hujan berturut-turut lebih diskriminatif dibandingkan frekuensi hari hujan.

Fitur *kondisi_fisik_encoded* (kondisi fisik saluran: Baik, Sedang, Buruk, Rusak) muncul sebagai prediktor konsisten di skenario S3. Dari perspektif pengelolaan, variabel ini memiliki keunggulan praktis karena dapat diubah melalui intervensi; perbaikan saluran berpotensi menurunkan risiko banjir, berbeda dari fitur topografi seperti elevasi yang bersifat tetap.

Kemampuan *pipeline* MI + Pearson + RFE mempertahankan fitur curah hujan merupakan temuan penting. Dengan *rainy_days* tetap tersedia pada skenario seleksi fitur, model tidak perlu bergantung sepenuhnya pada fitur interaksi sebagai pengganti informasi curah hujan. Hal ini menghasilkan distribusi kontribusi SHAP yang lebih merata antar fitur dan mengurangi risiko model menjadi terlalu bergantung pada satu fitur tunggal.

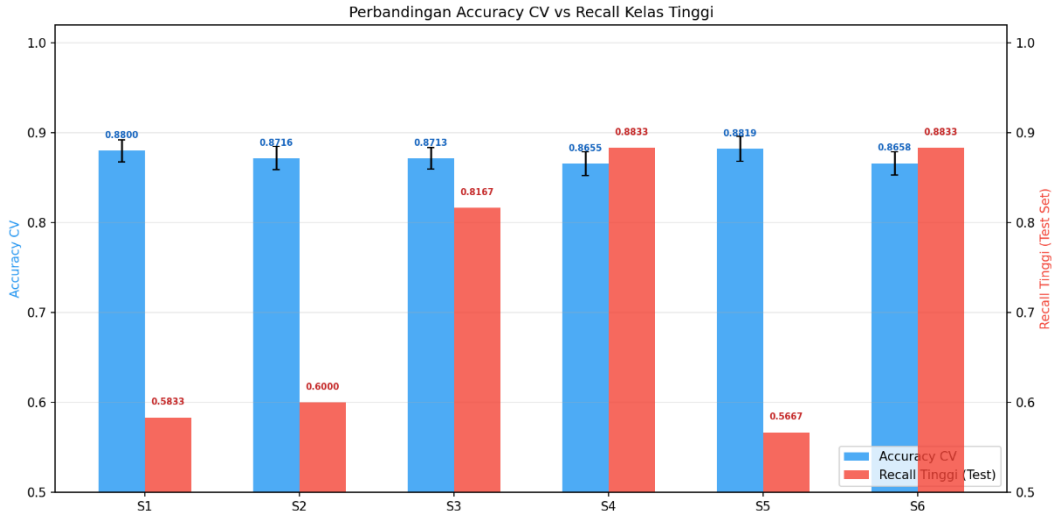
Pengujian tambahan yang membandingkan model dengan 45 fitur dasar (tanpa 8 fitur interaksi) terhadap *baseline* S1 (53 fitur) menunjukkan bahwa kontribusi *feature engineering* terhadap performa model sangat kecil (delta *Accuracy* +0,0014; F1 Macro +0,0007; ROC-AUC +0,0003), dengan hanya *Accuracy* yang signifikan secara marginal ($p = 0,045$). Temuan ini menunjukkan

bahwa XGBoost mampu menangkap pola interaksi drainase-curah hujan secara implisit melalui kombinasi *split* pada pohon keputusan. Meskipun kontribusi kuantitatifnya kecil, fitur interaksi tetap bernilai dari sisi interpretabilitas karena merepresentasikan hubungan hidrologis yang bermakna secara domain, seperti rasio kapasitas drainase terhadap beban curah hujan (*capacity_rainfall_ratio*) dan indeks kerentanan terhadap hujan lebat (*heavy_rain_vulnerability*).

Temuan ini mendukung argumen bahwa integrasi data drainase fisik dan curah hujan menghasilkan informasi prediktif yang lebih baik dibandingkan penggunaan salah satu sumber data secara terpisah, sejalan dengan Bersabe & Jun (2025) yang melaporkan peningkatan akurasi setelah penambahan variabel drainase.

6.4 *Trade-off* Performa Keseluruhan dan Deteksi Kelas Minoritas

Gambar 6.2 memvisualisasikan perbandingan *Recall* kelas Tinggi antarskenario untuk mengilustrasikan *trade-off* antara deteksi kelas minoritas dan performa keseluruhan.



Gambar 6. 2 Perbandingan *Recall* Kelas Tinggi dan *Accuracy* antarskenario

Dari seluruh hasil eksperimen, *trade-off* antara performa keseluruhan dan kemampuan mendeteksi kelas Tinggi merupakan temuan yang paling relevan untuk aplikasi di lapangan.

Dari analisis *test set*, pada *baseline* 25 dari 60 titik berisiko tinggi (41,7%) tidak terdeteksi. Dalam konteks mitigasi banjir, ini berarti 25 lokasi yang seharusnya mendapat prioritas penanganan justru diklasifikasikan pada tingkat risiko yang lebih rendah. Setelah BorderlineSMOTE diterapkan dengan seleksi fitur (S4), jumlah ini berkurang menjadi 7 titik (11,7%), tetapi sebagai gantinya lebih banyak titik Rendah salah terprediksi sebagai Sedang atau Tinggi, dan *Accuracy* CV turun dari 0,8800 menjadi 0,8655 (−0,0145).

Pertanyaannya kemudian: apakah penurunan ~1,5 poin *Accuracy* CV sepadan dengan pendeteksian 18 titik berisiko tinggi tambahan? Jawaban atas pertanyaan ini bergantung pada konteks. Jika biaya gagal mendeteksi titik berisiko tinggi jauh lebih besar daripada biaya *false alarm* (misalnya, kerusakan infrastruktur atau korban jiwa), maka S4 atau S6 lebih tepat meskipun *Accuracy*-nya lebih rendah. Sebaliknya, jika sumber daya penanganan terbatas dan *false alarm* menyebabkan pemborosan alokasi, maka S1 atau S5 yang lebih konservatif dapat menjadi pilihan yang lebih tepat.

Trade-off serupa dilaporkan pada konteks berbeda oleh Ren *et al.* (2024) yang menunjukkan bahwa strategi *sampling* memengaruhi distribusi prediksi antarkelas, dan Goumghar *et al.* (2025) yang mencatat bahwa performa algoritma

boosting sangat bergantung pada karakteristik dataset. Penelitian ini menambahkan bukti bahwa pada dataset dengan ketidakseimbangan kelas yang cukup besar (kelas Tinggi hanya 8,7%), teknik *oversampling* terbukti efektif meningkatkan *Recall* kelas minoritas tetapi selalu disertai penurunan pada metrik keseluruhan.

6.5 Implikasi Praktis

Hasil penelitian ini memiliki beberapa implikasi bagi pengelolaan risiko banjir di Kota Malang.

Pertama, model XGBoost *baseline* dengan 53 fitur sudah mampu mengklasifikasikan risiko banjir dengan *Accuracy* 0,8800 ± 0,0123 dan ROC-AUC 0,9689 ± 0,0048 berdasarkan *Repeated 5×5 Stratified K-Fold CV*, tanpa memerlukan *preprocessing* yang kompleks. Ini berarti implementasi awal sistem klasifikasi risiko banjir berbasis *machine learning* dapat dilakukan dengan konfigurasi yang sederhana dan waktu komputasi kurang dari 1 detik.

Kedua, jika prioritas utama adalah memaksimalkan deteksi titik berisiko tinggi, konfigurasi S4 (FS + BorderlineSMOTE) dapat dipertimbangkan karena mampu mendeteksi 88,3% titik berisiko tinggi (53 dari 60), meskipun dengan penurunan *Accuracy* keseluruhan. Dalam konteks perencanaan mitigasi banjir, lebih baik memeriksa beberapa titik yang ternyata tidak berisiko tinggi (*false positive*) daripada melewatkan titik yang benar-benar berisiko tinggi (*false negative*).

Ketiga, fitur *kondisi_fisik_encoded* yang muncul sebagai prediktor konsisten di skenario S3 memberi sinyal bahwa pemeliharaan dan perbaikan

kondisi fisik saluran drainase dapat menjadi intervensi yang efektif. Berbeda dari faktor topografi (elevasi, *slope*) yang tidak dapat diubah, kondisi fisik saluran merupakan variabel yang dapat diintervensi oleh pemerintah daerah melalui program pemeliharaan rutin.

Keempat, *pipeline* seleksi fitur MI + Pearson + RFE berhasil mempertahankan 2 fitur curah hujan CHIRPS (*rainy_days* dan *rainfall_30d_max*) dalam set 19 fitur terpilih. Fitur *rainy_days* tetap mendominasi analisis SHAP bahkan pada skenario dengan seleksi fitur, yang menegaskan pentingnya informasi curah hujan dalam klasifikasi risiko banjir. Keberhasilan mempertahankan fitur curah hujan ini merupakan keunggulan pendekatan MI dibandingkan metode berbasis multikolinearitas yang cenderung mengeliminasi fitur-fitur berkorelasi tinggi meskipun informatif.

6.6 Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang memengaruhi interpretasi hasil.

Pertama, dataset hanya mencakup wilayah Kota Malang, sehingga generalisasi temuan ke wilayah lain perlu dilakukan dengan hati-hati. Karakteristik topografi, pola curah hujan, dan infrastruktur drainase yang berbeda antarwilayah dapat menghasilkan pola *feature importance* dan efektivitas *preprocessing* yang berbeda pula.

Kedua, label kelas risiko (Rendah, Sedang, Tinggi) pada dataset berasal dari penilaian yang sudah ada sebelumnya, bukan dari observasi langsung kejadian

banjir. Akurasi label ini akan memengaruhi performa model, dan ketidakakuratan pada label (jika ada) tidak dapat dideteksi oleh model.

Ketiga, data curah hujan CHIRPS merupakan data satelit dengan resolusi spasial sekitar 5,5 km × 5,5 km. Pada skala kota, variasi curah hujan lokal yang lebih halus mungkin tidak tertangkap oleh resolusi ini. Data curah hujan dari stasiun pengamatan lokal dengan resolusi lebih tinggi berpotensi memberikan informasi yang lebih detail, meskipun ketersediaan dan cakupan spasialnya sering kali terbatas.

Keempat, studi ablasi menggunakan BorderlineSMOTE sebagai satu-satunya teknik penanganan ketidakseimbangan kelas. Teknik lain seperti SMOTE-ENN, ADASYN, atau pendekatan *cost-sensitive learning* melalui *class_weight* pada XGBoost belum dievaluasi. Teknik-teknik ini mungkin menghasilkan *trade-off* yang berbeda antara *Recall* kelas Tinggi dan *Accuracy* keseluruhan.

Kelima, seleksi fitur menggunakan satu *pipeline* (MI + Pearson + RFE). Meskipun *pipeline* ini berhasil mempertahankan 2 fitur curah hujan, sebagian besar fitur curah hujan CHIRPS tetap tereliminasi karena korelasi Pearson yang tinggi antar fitur curah hujan. Metode seleksi fitur alternatif seperti seleksi berbasis SHAP *importance* atau pendekatan *wrapper* yang berbeda mungkin mempertahankan lebih banyak fitur curah hujan dan menghasilkan performa yang berbeda.

Keenam, meskipun evaluasi utama telah menggunakan *Repeated 5×5 Stratified K-Fold Cross-Validation* (25 evaluasi) untuk mengukur variabilitas performa, analisis diagnostik berupa *confusion matrix*, metrik per kelas, dan SHAP

tetap didasarkan pada satu kali *split* 80/20 dengan *random state* tetap. Hasil analisis diagnostik ini dapat bervariasi pada *split* yang berbeda, sehingga interpretasinya perlu dikontekstualisasikan sebagai ilustrasi dari satu partisi data, bukan estimasi definitif.

BAB VII
PENUTUP

7.1 Kesimpulan

Berdasarkan hasil eksperimen dan pembahasan pada bab-bab sebelumnya, penelitian ini menghasilkan beberapa kesimpulan sebagai berikut.

1. Integrasi data curah hujan satelit CHIRPS yang bersifat temporal dengan data karakteristik drainase yang bersifat spasial dilakukan melalui *feature engineering* berbasis koordinat titik drainase. Data CHIRPS diekstraksi menjadi 19 fitur statistik untuk setiap titik drainase berdasarkan koordinat terdekat, kemudian digabungkan dengan 26 fitur drainase fisik dan 8 fitur interaksi yang merepresentasikan hubungan antara kapasitas drainase dan beban curah hujan. Dataset akhir terdiri dari 53 fitur dan 3.471 sampel dengan tiga kategori risiko (Rendah 65,7%, Sedang 25,7%, Tinggi 8,7%). Fitur curah hujan *rainy_days* terbukti mendominasi analisis SHAP pada lima dari enam skenario (SHAP = 0,0615–0,0855), yang membuktikan penggabungan kedua sumber data menghasilkan informasi yang tidak dapat diperoleh dari masing-masing sumber secara terpisah.
2. Model XGBoost *baseline* (S1) dengan 53 fitur dan *default hyperparameter* mencapai *Accuracy* 0,8800 ± 0,0123, F1 Macro 0,8152 ± 0,0215, dan ROC-AUC 0,9689 ± 0,0048 berdasarkan *Repeated 5×5 Stratified K-Fold Cross-Validation* (5 fold × 5 pengulangan = 25 evaluasi). Studi ablasi pada lima skenario lanjutan (S2–S6) menunjukkan bahwa S5 (Optuna) sedikit

mengungguli *baseline* (*Accuracy* 0,8819 ± 0,0140; F1 Macro 0,8193 ± 0,0234), dan *Wilcoxon signed-rank test* mengonfirmasi bahwa peningkatan F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$) signifikan pada $\alpha = 0,05$, meskipun peningkatan *Accuracy* tidak signifikan ($p = 0,084$). Kontribusi masing-masing komponen terukur secara terpisah: seleksi fitur (MI + Pearson + RFE) mereduksi dimensi dari 53 menjadi 19 fitur (pengurangan 64,2%) sambil mempertahankan 2 fitur curah hujan CHIRPS (*rainy_days* dan *rainfall_30d_max*) beserta 7 fitur interaksi, dengan penurunan F1 Macro CV sebesar 0,0174. BorderlineSMOTE meningkatkan *Recall* kelas Tinggi dari 0,5833 (S1) menjadi 0,8167 (S3) dan 0,8833 (S4) pada *test set*, sehingga 49–53 dari 60 titik berisiko tinggi terdeteksi, dengan konsekuensi penurunan *Accuracy* CV dari 0,8800 menjadi 0,8713 (S3) dan 0,8655 (S4). Optuna TPE memberikan perbaikan yang kecil secara absolut (+0,0019 *Accuracy* CV, +0,0006 ROC-AUC CV) namun signifikan secara statistik pada F1 Macro ($p = 0,008$) dan ROC-AUC ($p = 0,034$). Pemilihan konfigurasi optimal bergantung pada prioritas aplikasi: S1 atau S5 untuk performa keseluruhan yang tinggi, atau S4/S6 untuk memaksimalkan deteksi titik berisiko tinggi.

3. Analisis SHAP menunjukkan dominasi fitur curah hujan *rainy_days* pada lima dari enam skenario (S1, S2, S4, S5, S6) dengan nilai SHAP berkisar 0,0615–0,0855. Dominasi ini mencakup skenario dengan seleksi fitur, yang dimungkinkan karena *pipeline* MI + Pearson + RFE berhasil mempertahankan *rainy_days* dalam set fitur terpilih. Skenario S3

(BorderlineSMOTE tanpa seleksi fitur) merupakan pengecualian, di mana *max_consecutive_rain_days* menempati posisi teratas. Fitur *kondisi_fisik_encoded* muncul sebagai prediktor konsisten di skenario S3; kondisi fisik infrastruktur drainase dengan demikian merupakan faktor prediktif yang relevan dan dapat diintervensi melalui program pemeliharaan.

7.2 Saran

Berdasarkan keterbatasan yang teridentifikasi pada penelitian ini, berikut beberapa saran untuk penelitian selanjutnya.

1. Menguji generalisasi model pada wilayah lain dengan karakteristik topografi, pola curah hujan, dan infrastruktur drainase yang berbeda dari Kota Malang, untuk mengetahui sejauh mana temuan ini berlaku di luar konteks wilayah studi.
2. Mengevaluasi teknik penanganan ketidakseimbangan kelas alternatif seperti SMOTE-ENN, ADASYN, atau pendekatan *cost-sensitive learning* melalui *class_weight* pada XGBoost, yang mungkin menghasilkan *trade-off* yang berbeda antara *Recall* kelas Tinggi dan *Accuracy* keseluruhan.
3. Mengevaluasi metode seleksi fitur alternatif seperti seleksi berbasis SHAP *importance* atau pendekatan *embedded* langsung dari XGBoost, untuk membandingkan hasilnya dengan *pipeline* MI + Pearson + RFE yang digunakan pada penelitian ini dan mengidentifikasi apakah konfigurasi fitur yang berbeda dapat meningkatkan performa.

4. Mengintegrasikan data curah hujan dari stasiun pengamatan lokal dengan resolusi spasial yang lebih tinggi sebagai pelengkap data CHIRPS, terutama untuk menangkap variasi curah hujan lokal pada skala kota yang mungkin tidak tertangkap oleh resolusi CHIRPS (5,5 km × 5,5 km).
5. Mengembangkan prototipe sistem informasi geografis (SIG) berbasis *web* yang menyajikan peta risiko banjir per titik drainase menggunakan model XGBoost S1, dilengkapi fitur pembaruan otomatis data curah hujan CHIRPS dan visualisasi nilai SHAP per titik, sehingga Dinas PUPR Kota Malang dapat mengidentifikasi titik-titik prioritas pemeliharaan saluran drainase secara periodik.

DAFTAR PUSTAKA

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'11*, 2546–2554.

Bersabe, J. T., & Jun, B.-W. (2025). The Machine Learning-Based Mapping of Urban Pluvial Flood Susceptibility in Seoul Integrating Flood Conditioning Factors and Drainage-Related Data. *ISPRS International Journal of Geo-Information*, *14*(2). https://doi.org/10.3390/ijgi14020057

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Böing, S. J., Birch, C. E., Rabb, B. L., & Shelton, K. L. (2020). A percentile-based approach to rainfall scenario construction for surface-water flood forecasts. *Meteorological Applications*, *27*(6), e1963. https://doi.org/10.1002/met.1963

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 785–794. https://doi.org/10.1145/2939672.2939785

Chow, V. Te. (1959). *Open-Channel Hydraulics*. McGraw-Hill.

Chow, V. Te, Maidment, D. R., & Mays, L. W. (1988). *Applied Hydrology*. McGraw-Hill.

Goumghar, L., Hajaj, S., Haida, S., Kili, M., Mridekh, A., Khandouch, Y., Jari, A., El Harti, A., & El Mansouri, B. (2025). Analysis of Baseline and Novel Boosting Models for Flood-Prone Prediction and Explainability: Case from the Upper Drâa Basin (Morocco). *Earth*, *6*(3). https://doi.org/10.3390/earth6030069

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Huang, H., Tao, Z., Zhan, J., & Wang, C. (2025). Contrast or Diversity: Non-Flood sampling in urban flood susceptibility modelling. *Journal of Hydrology*, *656*, 133053. https://doi.org/10.1016/j.jhydrol.2025.133053

(IPCC), I. P. on C. C. (2023). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/DOI: 10.1017/9781009157896

Jang, S.-D., Yoo, J.-H., Lee, Y.-S., & Kim, B. (2025). Flood prediction in urban areas based on machine learning considering the statistical characteristics of rainfall. *Progress in Disaster Science*, *26*, 100415. https://doi.org/10.1016/j.pdisas.2025.100415

Manning, R. (1891). On the Flow of Water in Open Channels and Pipes. *Transactions of the Institution of Civil Engineers of Ireland*, *20*, 161–207.

Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "Assessment of economic flood damage." *Natural Hazards and Earth System Sciences*, *10*(8), 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, *5*(1), 3–30. https://doi.org/10.1002/hyp.3360050103

Pourzangbar, A., Oberle, P., Kron, A., & Franca, M. J. (2025). Analysis of the Utilization of Machine Learning to Map Flood Susceptibility. *Journal of Flood Risk Management*, *18*(2), e70042. https://doi.org/10.1111/jfr3.70042

Qin, Y., Yu, Y., Liu, J., & Liu, R. (2025). Machine learning-based identification of key factors and spatial heterogeneity analysis of urban flooding: a case study of the central urban area of Ordos. *Scientific Reports*, *15*(1), 24749. https://doi.org/10.1038/s41598-025-08162-4

Ren, H., Pang, B., Bai, P., Zhao, G., Liu, S., Liu, Y., & Li, M. (2024). Flood Susceptibility Assessment with Random Sampling Strategy in Ensemble Learning (RF and XGBoost). *Remote Sensing*, *16*(2). https://doi.org/10.3390/rs16020320

Riazi, M., Khosravi, K., Shahedi, K., Ahmad, S., Jun, C., Bateni, S. M., & Kazakis, N. (2023). Enhancing flood susceptibility modeling using multi-temporal SAR images, CHIRPS data, and hybrid machine learning algorithms. *Science of The Total Environment*, *871*, 162066. https://doi.org/10.1016/j.scitotenv.2023.162066

Sharma, P., & Sharma, S. (2026). *Flood Risk Follows Valleys, Not Grids: Graph Neural Networks for Flash Flood Susceptibility Mapping in Himachal Pradesh with Conformal Uncertainty Quantification*. https://arxiv.org/abs/2603.15681

Shrestha, S., Dahal, D., Bhattarai, N., Regmi, S., Sewa, R., & Kalra, A. (2025). Machine Learning-Based Flood Risk Assessment in Urban Watershed: Mapping Flood Susceptibility in Charlotte, North Carolina. *Geographies*, *5*(3). https://doi.org/10.3390/geographies5030043

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood Susceptibility Mapping Using a Novel Ensemble Weights-of-Evidence and Support Vector Machine Models in GIS. *Journal of Hydrology*, *512*, 332–343. https://doi.org/10.1016/j.jhydrol.2014.03.008

Teng, J., Jakeman, A. J., Vaze, J., Croke, B. F. W., Dutta, D., & Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental Modelling & Software*, *90*, 201–216. https://doi.org/10.1016/j.envsoft.2017.01.006

United Nations Office for Disaster Risk Reduction. (2017). *The Sendai Framework Terminology on Disaster Risk Reduction: Disaster Risk Reduction*. https://www.undrr.org/terminology/disaster-risk-reduction

Wang, M., Li, Y., Yuan, H., Zhou, S., Wang, Y., Adnan Ikram, R. M., & Li, J. (2023). An XGBoost-SHAP approach to quantifying morphological impact on urban flooding susceptibility. *Ecological Indicators*, *156*, 111137. https://doi.org/10.1016/j.ecolind.2023.111137

Yuan, H., Wang, M., Zhang, D., Muhammad Adnan Ikram, R., Su, J., Zhou, S., Wang, Y., Li, J., & Zhang, Q. (2024). Data-driven urban configuration optimization: An XGBoost-based approach for mitigating flood susceptibility and enhancing economic contribution. *Ecological Indicators*, *166*, 112247. https://doi.org/10.1016/j.ecolind.2024.112247

Zhang, Y., Wei, Y., Yao, R., Sun, P., Zhen, N., & Xia, X. (2025). Data Uncertainty of Flood Susceptibility Using Non-Flood Samples. *Remote Sensing*, *17*(3). https://doi.org/10.3390/rs17030375

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

Zhu, X., Guo, H., & Huang, J. J. (2024). Urban flood susceptibility mapping using remote sensing, social sensing and an ensemble machine learning model. *Sustainable Cities and Society*, *108*, 105508. https://doi.org/10.1016/j.scs.2024.105508