Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2973~2985
e-ISSN: 2548-6861

Anti-Data Leakage Pipeline for Differentiated Thyroid Cancer Recurrence Prediction: Integrating SMOTE, Optuna-based Optimization, and Bootstrap BCa Validation

Deri Rosadi [1]*, Sindhu Rakasiwi [2]**
* Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang
111202214191@mhs.dinus.ac.id [1], sindhu.rakasiwi@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2026-05-03
Revised 2026-06-01
Accepted 2026-06-11

Keyword:

Anti-data leakage,
Bootstrap BCa,
Calibrated pipeline,
Machine learning,
Thyroid cancer recurrence.

ABSTRACT

Thyroid cancer recurrence prediction remains a critical clinical challenge, as early identification of high-risk patients enables targeted monitoring and intervention. This study presents a comparative evaluation of six machine learning classifiers (XGBoost, LightGBM, CatBoost, Logistic Regression, Random Forest, and Decision Tree) using the UCI Differentiated Thyroid Cancer Recurrence dataset which consists of 383 patient records and 16 clinical features. To prevent performance overestimation, a rigorous anti-data leakage pipeline was implemented, encapsulating SMOTE, Optuna-based hyperparameter optimization, and Isotonic Calibration within the cross-validation process. Furthermore, model stability was assessed using Bias-Corrected and accelerated (BCa) Bootstrap validation with 2,000 iterations. Experimental results demonstrate that XGBoost achieved the best overall performance with an F1-score of 0.9545, an AUC-ROC of 0.9967, and the lowest Brier Score of 0.0183. Bootstrap BCa analysis confirmed XGBoost as the most stable model, with a 95% CI F1-score width of 0.1429 and unbiased estimation. These findings suggest that XGBoost, integrated within a zero-leakage pipeline and validated through Bootstrap BCa, is a promising candidate for post-treatment clinical decision support in differentiated thyroid cancer management.

I. PENDAHULUAN

Kanker tiroid merupakan keganasan endokrin yang paling umum di seluruh dunia, dengan lebih dari 800.000 kasus baru diperkirakan setiap tahunnya berdasarkan data Global Cancer Observatory WHO [1]. Dari seluruh tipe histologis yang ada, Differentiated Thyroid Cancer (DTC) meliputi karsinoma papilari dan folikular merupakan bentuk yang paling sering ditemukan dan mewakili lebih dari 90% seluruh kasus kanker tiroid [2]. Meskipun sebagian besar pasien DTC memiliki prognosis yang baik dengan angka kelangsungan hidup jangka panjang yang tinggi, kekambuhan penyakit ini tetap menjadi tantangan klinis yang signifikan. Sekitar 20–30% pasien DTC mengalami kekambuhan dalam 10 tahun setelah terapi awal, sehingga deteksi dini terhadap kelompok berisiko tinggi menjadi kunci keberhasilan tata laksana lanjutan [3]. Sistem stratifikasi risiko yang umum digunakan saat ini, seperti klasifikasi American Thyroid Association (ATA) dan sistem TNM, meskipun sudah terstandarisasi, memiliki keterbatasan dalam mengakomodasi kompleksitas multivariabel klinis secara individual sehingga kemampuan prediksi kekambuhan pasien masih dapat ditingkatkan [4].

Pendekatan berbasis machine learning (ML) telah menunjukkan potensi besar dalam mengatasi keterbatasan model statistik konvensional untuk prediksi kekambuhan kanker tiroid. Beberapa studi terkini telah mengeksplorasi dataset Differentiated Thyroid Cancer Recurrence dari UCI Repository yang terdiri dari 383 pasien dan 16 fitur klinis. Beberapa studi terkini pada dataset yang sama melaporkan performa tinggi: Random Forest mencapai akurasi 99% dan AUC 0,996 dengan SMOTE dan hyperparameter tuning [5], serta akurasi 98,26% dan AUC 1,00 dengan Nested CV dan kalibrasi Isotonic [6]. Studi lain menunjukkan bahwa hanya 4 fitur utama (Risk, N-stage, T-stage, Age) sudah cukup mencapai AUC 0,913 yang hampir setara model full-fitur

[7]. CatBoost dengan SHAP mengonfirmasi dominasi Treatment Response (SHAP 2,077, akurasi 97%) [8], sementara Stacking Ensemble berbasis XGBoost yang diintegrasikan dalam web app TCCheck mencapai akurasi 96,52% dan AUC 0,9921 [9].

Pendekatan serupa terbukti efektif lintas domain klinis: pada prediksi survival kanker kolorektal (72.961 pasien), CatBoost dan LightGBM dengan Optuna mencapai akurasi ~82% [10], sementara pada prediksi risiko hipertensi, CatBoost dengan SMOTE dan GridSearchCV mencapai accuracy 0,992 dan AUC 0,9987 [11].

Meski demikian, sejumlah celah metodologis yang krusial masih teridentifikasi dari studi-studi tiroid tersebut. Pertama, terkait *data leakage*: penerapan SMOTE sebelum pemisahan data atau di luar kerangka *cross-validation* berpotensi memperkenalkan leakage terselubung, karena sampel sintetis yang dibuat dari keseluruhan data akan bocor ke fold validasi sehingga mengakibatkan estimasi performa yang optimistis dan tidak dapat digeneralisasi [12][13]. Kondisi ini teridentifikasi pada dua studi sebelumnya [5][7] yang belum menerapkan protokol anti-leakage secara konsisten. Kedua, meskipun salah satu studi mencantumkan Isotonic Regression sebagai langkah kalibrasi [6], implementasinya belum terintegrasi penuh dalam pipeline CV sehingga risiko leakage kalibrasi tetap ada; sementara studi-studi lainnya [5][8] sama sekali tidak membahas kalibrasi probabilitas, padahal skor probabilitas yang tidak terkalibrasi dapat menyesatkan pengambilan keputusan klinis [13]. Ketiga, seluruh studi yang ditinjau melakukan evaluasi performa hanya pada satu partisi *test set* tunggal tanpa mengukur stabilitas estimasi melalui analisis statistik. Pelaporan metrik tunggal pada dataset berukuran kecil terbukti sangat rentan terhadap variasi partisi data (data splitting variance), sehingga tanpa analisis *confidence interval* melalui teknik bootstrap tidak dapat diketahui seberapa besar ketidakpastian dari metrik yang dilaporkan hasil akurasi tinggi pada penelitian terdahulu bisa jadi sekadar bias dari partisi data yang menguntungkan [14]. Keempat, gap signifikan antara Mean CV Accuracy (91,06%) dan *test set* accuracy (96,52%) yang dilaporkan pada salah satu studi [9] tidak dianalisis lebih lanjut, padahal selisih 5,46% ini merupakan indikasi kuat adanya *data leakage* atau instabilitas model.

Penelitian ini bertujuan membangun dan membandingkan enam model ML XGBoost, LightGBM, CatBoost, Logistic Regression, Random Forest, dan Decision Tree untuk prediksi kekambuhan DTC dengan mengatasi secara sistematis seluruh permasalahan metodologis tersebut. Keenam algoritma dipilih untuk memastikan perbandingan yang komprehensif dan adil lintas spektrum pendekatan yang luas: XGBoost, LightGBM, dan CatBoost mewakili keluarga gradient boosting modern yang dikenal unggul pada data tabular klinis namun dengan karakteristik implementasi berbeda; Random Forest mewakili pendekatan bagging ensemble yang lebih konservatif sekaligus model terbaik pada mayoritas studi sebelumnya [5][6][7] untuk menguji apakah keunggulan tersebut bertahan dalam pipeline anti-leakage yang lebih ketat; Logistic Regression sebagai pembanding model linear yang sederhana dan interpretatif; serta Decision Tree sebagai baseline untuk menetapkan batas bawah performa.

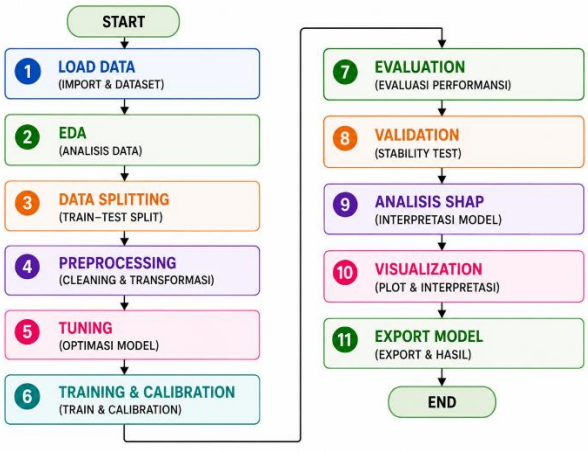
Meskipun SMOTE, Optuna, dan Bootstrap BCa masing-masing sudah dikenal dalam literatur ML medis, yang membedakan penelitian ini bukan pada penggunaan komponen-komponen tersebut secara individu, melainkan pada cara pengintegrasiannya keempatnya dijalankan sekaligus dalam satu kerangka pipeline yang koheren di mana setiap komponen saling mengunci untuk mencegah kebocoran data di setiap lapisan proses.

Kontribusi utama penelitian ini meliputi: (1) implementasi integrated pipeline berbasis ImbPipeline yang mengenkapsulasi preprocessing, SMOTE, dan classifier dalam satu kesatuan; (2) tuning hyperparameter berbasis Optuna dengan 100 trial per model di mana seluruh pipeline dioptimalkan secara bersamaan; (3) kalibrasi probabilitas menggunakan Isotonic Regression terintegrasi dalam pipeline; serta (4) validasi stabilitas model menggunakan Bootstrap BCa dengan 2.000 iterasi untuk menghasilkan *confidence interval* yang lebih akurat pada dataset berukuran kecil.

Penelitian ini tidak menyertakan baseline klinis tradisional berbasis sistem skoring (seperti ATA Risk Stratification atau TNM scoring) sebagai pembanding langsung, mengingat fokus penelitian adalah perbandingan antar algoritma ML dalam kerangka metodologi yang terkontrol; perbandingan dengan baseline klinis tradisional disarankan sebagai arah penelitian lanjutan.

II. METODE

Penelitian ini dirancang dalam sepuluh tahap berurutan dari persiapan data hingga penyimpanan model, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Alur Penelitian

Gambar 1 menunjukkan 11 tahap berurutan: inisialisasi dataset → EDA → stratified train-test split (80:20) → definisi preprocessor → Optuna tuning → training dengan kalibrasi isotonic → evaluasi → analisis SHAP → Bootstrap BCa → learning curve → penyimpanan model. Sejak tahap

pemisahan data (tahap 3), X_test tidak disentuh hingga evaluasi akhir untuk mencegah data leakage.

A. *Dataset dan Lingkungan Eksperimen*

Dataset yang digunakan dalam penelitian ini adalah Differentiated Thyroid Cancer Recurrence yang bersumber dari UCI Machine Learning Repository [15], dikontribusikan oleh Borzooei dan Tarokhian (2023) dengan DOI: https://doi.org/10.24432/C5632J. Dataset ini terdiri dari 383 rekam medis pasien kanker tiroid terdiferensiasi (Differentiated Thyroid Cancer/DTC) dengan 16 fitur klinis dan patologis serta satu variabel target biner Recurred (Yes/No).

Fitur-fitur yang tersedia mencakup informasi demografis pasien (Age, Gender), riwayat medis (Smoking, Hx Smoking, Hx Radiotherapy), kondisi klinis (Thyroid Function, Physical Examination, Adenopathy), karakteristik patologis tumor (Pathology, Focality), sistem staging TNM (T, N, M, Stage), stratifikasi risiko ATA (Risk), serta respons terhadap terapi (Response). Fitur Response dilaporkan sebagai salah satu prediktor paling dominan terhadap kekambuhan DTC pada berbagai studi sebelumnya [5]-[9].

Dari keseluruhan fitur, satu fitur bersifat numerik (Age), sedangkan 15 fitur lainnya bersifat kategorikal. Dari 383 pasien, sebanyak 108 pasien (28,2%) mengalami kekambuhan (Recurred=Yes) dan 275 pasien (71,8%) tidak mengalami kekambuhan (Recurred=No), menunjukkan ketidakseimbangan kelas dengan rasio sekitar 1:2,5. Pemeriksaan menyeluruh terhadap dataset mengonfirmasi tidak adanya missing values pada seluruh 16 fitur klinis, sehingga tidak diperlukan proses imputasi data namun SimpleImputer tetap disertakan dalam pipeline sebagai *safeguard* untuk kondisi data yang tidak terduga..

Seluruh eksperimen dijalankan menggunakan Python 3.12.13 di lingkungan Google Colaboratory dengan CPU runtime standar tanpa akselerasi GPU. Pustaka utama yang digunakan meliputi scikit-learn, imbalanced-learn (ImbPipeline dan SMOTE), XGBoost, LightGBM, CatBoost, Optuna [16], dan SciPy.

B. *Eksplorasi Data (EDA)*

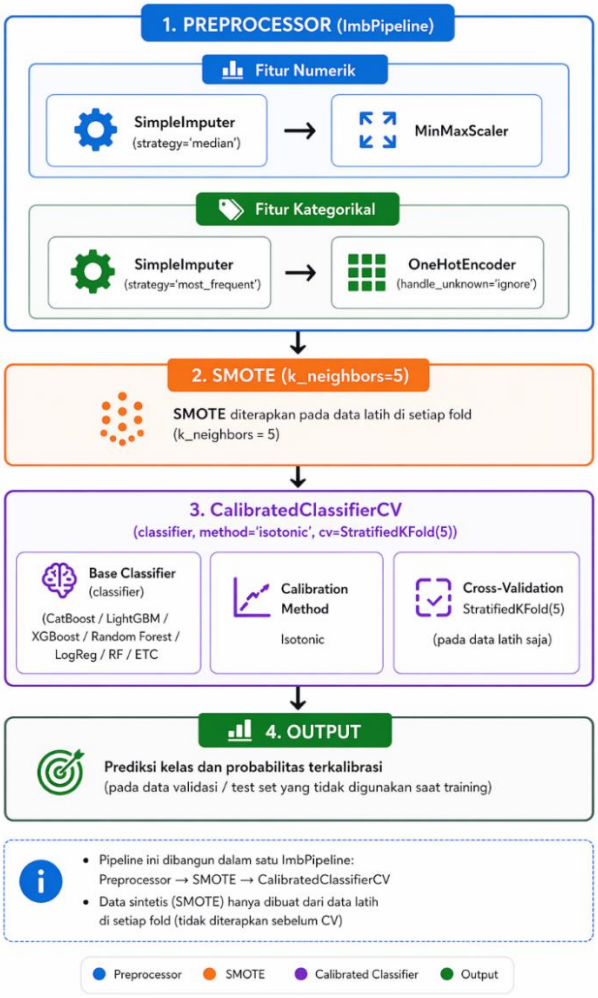
Eksplorasi data dilakukan terhadap keseluruhan dataset sebelum proses pemisahan data untuk memahami karakteristik distribusi kelas dan pola awal pada data. Hasil eksplorasi menunjukkan bahwa distribusi kelas target tidak seimbang, di mana sebanyak 275 pasien (71,8%) tidak mengalami kekambuhan dan 108 pasien (28,2%) mengalami kekambuhan. Rasio ketidakseimbangan sekitar 1:2,5 (≈0,39) mengindikasikan adanya *class imbalance* yang berpotensi menyebabkan bias model terhadap kelas mayoritas, sehingga diperlukan teknik penanganan khusus pada tahap pemodelan.

C. *Stratified Train-Test Split*

Pemisahan data dilakukan sebagai langkah pertama sebelum preprocessing apapun, menggunakan rasio 80:20 dengan parameter stratify=y untuk menjaga proporsi kelas pada kedua partisi. Dengan random_state=42, diperoleh 306 sampel data latih dan 77 sampel data uji. Pendekatan ini merupakan fondasi utama strategi *anti-data leakage.* Statistik apapun yang dipelajari (median, modus, parameter encoding, parameter scaling) hanya boleh berasal dari data latih, tidak pernah dari data uji.

D. *Arsitektur Pipeline Anti-Data leakage*

Seluruh tahapan preprocessing, *oversampling*, dan kalibrasi dibangun dalam satu ImbPipeline terintegrasi untuk memastikan tidak ada kebocoran statistik antar tahap. Arsitektur lengkap pipeline ini diilustrasikan pada Gambar 2.



Gambar 2. Arsitektur Pipeline Final

Gambar 2 memperlihatkan tiga lapis pipeline yang berjalan berurutan: (1) Preprocessor (ColumnTransformer) yang memproses fitur numerik melalui SimpleImputer(median) → MinMaxScaler, dan fitur kategorikal melalui SimpleImputer(most_frequent) → OneHotEncoder(handle_unknown='ignore'); (2) SMOTE dengan k_neighbors=5 yang hanya membangkitkan data sintetis dari fold latih di setiap iterasi CV, bukan dari keseluruhan data, mengikuti protokol anti-leakage yang direkomendasikan [12][13]. Secara eksplisit, urutan proses

dalam setiap fold cross-validation adalah: (1) data train fold di-fit oleh preprocessor; (2) SMOTE diterapkan hanya pada data train fold yang sudah dipreproses; (3) classifier dilatih pada data hasil SMOTE; (4) data validasi fold ditransform menggunakan preprocessor yang sudah di-fit pada langkah (1) tanpa SMOTE; (5) evaluasi dilakukan pada data validasi fold yang tidak mengandung sampel sintetis. Dengan urutan ini, tidak ada satupun statistik dari data validasi atau test set yang dapat bocor ke tahap pelatihan;

Perlu diakui bahwa SMOTE memiliki keterbatasan pada data klinis: data baru yang dihasilkan dibuat dengan cara mencampur nilai-nilai dari pasien yang mirip satu sama lain, sehingga hasilnya belum tentu mencerminkan kondisi klinis pasien yang sesungguhnya. Untuk mengurangi risiko ini, SMOTE hanya diterapkan pada data latih di dalam setiap fold CV tidak pernah menyentuh data validasi atau test set. Dengan cara ini, evaluasi akhir model selalu dilakukan pada data pasien nyata (n=77) tanpa campuran data buatan, sehingga angka performa yang dilaporkan mencerminkan kemampuan model yang sebenarnya pada kondisi klinis.

E. Hyperparameter Tuning dengan Optuna

Optimasi hyperparameter dilakukan menggunakan kerangka Optuna [16] dengan TPE Sampler sebanyak 100 trial per model. Pada setiap trial, full pipeline (preprocessor + SMOTE + classifier) dievaluasi menggunakan cross_val_score dengan StratifiedKFold(n_splits=5) dan metrik F1-Score bukan hanya classifier-nya saja sehingga tidak ada kebocoran statistik antar fold. Ruang pencarian hyperparameter untuk masing-masing model disajikan pada Tabel 1.

TABEL 1
RUANG PENCARIAN HYPERTUNING OPTUNA

Model	Hyperparameter	Rentang/Pilihan
XGBoost	n_estimators, max_depth, learning_rate, subsample, colsample_bytree, reg_alpha, reg_lambda, min_child_weight, gamma	[100–500], [3–9], [0.01–0.3], [0.6–1.0], [0.6–1.0], [1e-8–10], [1e-8–10], [1–7], [0–1]
LightGBM	n_estimators, max_depth, learning_rate, num_leaves, subsample, colsample_bytree, reg_alpha, reg_lambda, min_child_samples	[100–500], [3–9], [0.01–0.3], [20–150], [0.6–1.0], [0.6–1.0], [1e-8–10], [1e-8–10], [5–50]
CatBoost	iterations, depth, learning_rate, l2_leaf_reg, border_count, bagging_temperature, random_strength	[100–500], [3–8], [0.01–0.3], [1e-3–10], [32–128], [0–1], [1e-3–10]
Logistic Regression	C, l1_ratio	[1e-4–100], [0.0–1.0]
Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap	[100–500], [3–20], [2–20], [1–10], {sqrt, log2, 0.5}, {True, False}
Decision Tree	max_depth, min_samples_split, min_samples_leaf, max_features, criterion, splitter	[2–20], [2–30], [1–20], {sqrt, log2, None}, {gini, entropy}, {best, random}

Tabel 1 menunjukkan bahwa ruang pencarian dirancang secara berbeda untuk setiap model sesuai dengan karakteristik algoritmanya masing-masing. Model boosting (XGBoost, LightGBM, CatBoost) memiliki ruang pencarian yang lebih luas dengan parameter regularisasi seperti reg_alpha dan reg_lambda untuk mengendalikan kompleksitas. Logistic Regression difokuskan pada dua parameter kritis (C dan l1_ratio) yang mengontrol kekuatan dan jenis regularisasi. Random Forest dan Decision Tree mengeksplorasi parameter struktural pohon seperti max_depth dan min_samples_split untuk menyeimbangkan kedalaman model dengan kemampuan generalisasi.

F. Kalibrasi Probabilitas

Model ML pada umumnya menghasilkan skor probabilitas yang tidak terkalibrasi dengan baik. Kalibrasi probabilitas dilakukan menggunakan metode Isotonic Regression [17] melalui CalibratedClassifierCV(method='isotonic', cv=StratifiedKFold(5)) yang diintegrasikan langsung ke dalam ImbPipeline. Isotonic regression mencari fungsi monoton tak-turun *m* yang meminimalkan *Mean Squared Error* terhadap probabilitas aktual, seperti pada Persamaan (1).

$$\min_{m} \sum_{i=1}^{n} \left(y_i - m(f_i)\right)^2 \quad (1)$$

Dimana y_i adalah label kelas aktual dan f_i adalah probabilitas yang diprediksi oleh classifier tak-terkalibrasi. Kualitas kalibrasi dievaluasi menggunakan Reliability Diagram dan Brier Score. Brier Score mengukur keakuratan prediksi probabilitas yang diformulasikan pada Persamaan (2).

$$BS = \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2 \quad (2)$$

Dimana f_i adalah probabilitas prediksi dan o_i adalah luaran aktual (0 atau 1). Nilai Brier Score yang lebih mendekati 0 mengindikasikan kalibrasi model yang sangat baik.

G. Metrik Evaluasi

Performa model dievaluasi menggunakan delapan metrik: Accuracy, Precision, Recall, F1-Score, AUC-ROC, AUC-PR (metrik klasifikasi), serta Brier Score dan Log Loss (metrik kalibrasi probabilitas). F1-Score dipilih sebagai metrik utama perbandingan mengingat adanya ketidakseimbangan kelas.

H. Bootstrap BCa Validation

Stabilitas dan ketidakpastian estimasi performa diukur menggunakan Bootstrap BCa (Bias-Corrected and accelerated) dengan 2.000 iterasi sampling with replacement dari data test [18]. BCa dipilih karena lebih akurat dari persentil biasa pada distribusi *skewed* dan dataset kecil [19]. BCa secara otomatis menghitung faktor koreksi bias ($\hat{z}_0$) berdasarkan proporsi nilai bootstrap yang kurang dari estimasi original, yang dinyatakan dalam Persamaan (3). Pada implementasi penelitian ini, faktor akselerasi (a) ditetapkan bernilai nol karena estimasi jackknife yang umumnya digunakan untuk menghitung a tidak praktis pada ukuran *test set* yang kecil (n=77). Dengan a = 0, metode yang digunakan setara dengan Bias-Corrected (BC) bootstrap, yang tetap lebih akurat dibanding *percentile bootstrap* standar dalam mengoreksi bias distribusi, namun tanpa koreksi *skewness* tambahan dari faktor akselerasi.

$$\hat{z_0} = \Phi^{-1}\left(\frac{\sum_{b=1}^{B} I\left(\widehat{\theta_b^*} < \hat{\theta}\right)}{B}\right) \quad (3)$$

Dimana Φ^{-1} adalah fungsi kuantil dari distribusi normal standar, $\widehat{\theta_b^*}$ adalah estimasi metrik pada sampel bootstrap ke-*b*, $\hat{\theta}$ adalah estimasi pada sampel original, dan *B* adalah jumlah iterasi (2.000). Kriteria stabilitas dievaluasi berdasarkan nilai Confidence Interval (CI) Width: < 0,15 (Sangat Stabil), < 0,20 (Stabil), < 0,25 (Cukup Stabil), ≥ 0,25 (Kurang Stabil). Threshold ini ditetapkan secara kontekstual berdasarkan pertimbangan bahwa ketidakpastian estimasi meningkat signifikan pada dataset berukuran kecil [14]. CI Width ≥ 0,25 mengindikasikan rentang ketidakpastian sebesar ±12,5% yang dinilai tidak memadai untuk mendukung reliabilitas sistem pendukung keputusan klinis.

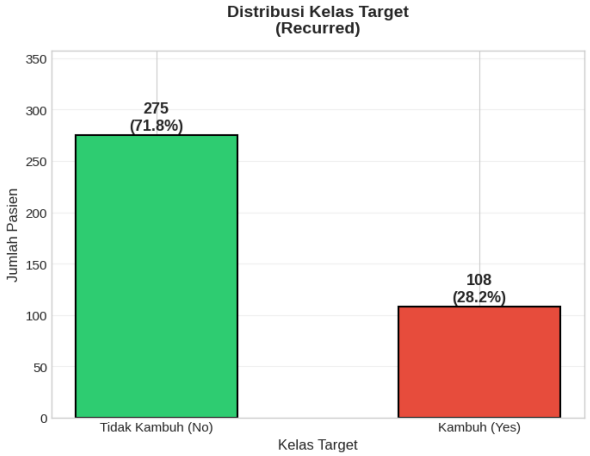
I. SHapley Additive exPlanations (SHAP)

Untuk meningkatkan relevansi klinis model terbaik, analisis SHapley Additive exPlanations (SHAP) [20] diterapkan pada model XGBoost menggunakan TreeExplainer [21]. SHAP dipilih karena memberikan penjelasan yang konsisten secara matematis berdasarkan teori permainan kooperatif, menghasilkan nilai kontribusi fitur yang bersifat aditif dan dapat diinterpretasikan baik secara global (populasi) maupun lokal (individu pasien) [20]. SHAP Beeswarm Plot dipilih sebagai visualisasi utama karena secara simultan menampilkan besaran (magnitude) dan arah (positif/negatif) kontribusi setiap fitur terhadap prediksi, serta distribusinya lintas seluruh sampel test set.

III. HASIL DAN PEMBAHASAN

A. Eksplorasi Data

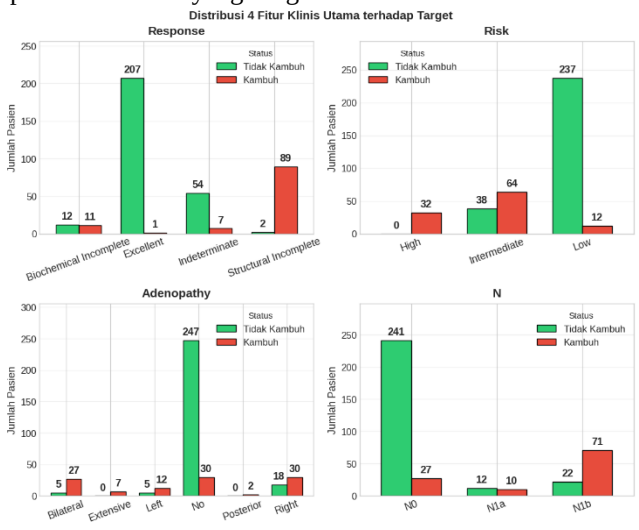
Hasil eksplorasi mengonfirmasi kelengkapan dataset tanpa adanya nilai yang hilang (*missing value*) pada keseluruhan 16 fitur dari 383 rekam medis pasien. Distribusi kelas target divisualisasikan pada Gambar 3 untuk mengidentifikasi ada tidaknya ketidakseimbangan kelas yang perlu ditangani sebelum pemodelan.



Gambar 3. Visualisasi Distribusi Kelas Target

Gambar 3 menampilkan bar chart distribusi kelas target pada keseluruhan dataset. Kelas "Tidak Kambuh" (No) berjumlah 275 pasien (71,8%), sedangkan kelas "Kambuh" (Yes) berjumlah 108 pasien (28,2%). Rasio ketidakseimbangan sebesar 0,39 mengonfirmasi perlunya penanganan *class imbalance* dalam penelitian ini diatasi melalui SMOTE yang diintegrasikan ke dalam pipeline CV, bukan diterapkan secara global sebelum split.

Gambar 4 memfokuskan visualisasi pada empat prediktor klinis paling dominan berdasarkan literatur: Treatment Response, Risk Stratification, N-Stage, dan Adenopathy [5][6][8][9], yang kesemuanya menunjukkan perbedaan proporsi antar kelas yang sangat mencolok.



Gambar 4. Visualisasi 4 Fitur Utama

Gambar 4 menyajikan distribusi silang (*crosstab*) empat fitur klinis utama terhadap kelas target. Pada fitur Response, kategori "Excellent" didominasi kelas Tidak Kambuh, sementara "Structural Incomplete" didominasi kelas Kambuh menunjukkan daya diskriminatif tertinggi di antara keempat fitur. Pada fitur Risk, kategori "Low" hampir sepenuhnya berasosiasi dengan Tidak Kambuh, sedangkan "High" dan "Intermediate" lebih banyak berkaitan dengan kekambuhan. Fitur Adenopathy menunjukkan bahwa pasien tanpa adenopathy (kategori "No") mayoritas tidak kambuh, sementara kategori "Bilateral" dan "Extensive" lebih sering ditemukan pada kelas Kambuh. Pada fitur N (nodal stage), N1b menunjukkan proporsi kambuh yang jauh lebih tinggi dibanding N0 dan N1a.

B. Hasil Hyperparameter Tuning OPTUNA

Proses Optuna berhasil diselesaikan untuk keenam model masing-masing sebanyak 100 trial. Tabel 2 menyajikan ringkasan hasil CV tuning beserta perbandingannya dengan Test F1 untuk mengukur konsistensi generalisasi. XGBoost dan CatBoost menunjukkan nilai CV F1 tertinggi di antara model boosting, sementara Logistic Regression menunjukkan performa CV yang kompetitif. Decision Tree memberikan nilai CV F1 terendah, konsisten dengan kapasitas generalisasinya yang terbatas. Kolom Gap menunjukkan selisih CV F1 dan Test F1 sangat kecil (−0,017 hingga +0,023), dengan empat dari enam model mencatat Test F1 lebih tinggi dari CV F1, mengonfirmasi estimasi yang jujur dan tidak optimistis. Konfigurasi optimal XGBoost: n_estimators=421, max_depth=9, learning_rate=0,088, subsample=0,952, colsample_bytree=0,600, min_child_weight=1, gamma=0,223, reg_alpha=0,00224, reg_lambda=$8{,}59\times10^{-7}$ (seluruh parameter tersedia di best_params.json).

TABEL 2.
TABEL HASIL OPTUNA TUNING

Model	Best F1-Score (CV 5-Fold)	N Trials	Test F1	Gap
XGBoost	0.958246	100	0.9545	-0.0037
LightGBM	0.947553	100	0.9302	-0.0174
CatBoost	0.951839	100	0.9524	+0.0006
Logistic Regression	0.929167	100	0.9524	+0.0232
Random Forest	0.951839	100	0.9524	+0.0006
Decision Tree	0.921410	100	0.9268	+0.0054

Berdasarkan Tabel 2, terlihat bahwa XGBoost mencapai CV F1 tertinggi (0,9582) diikuti CatBoost dan Random Forest (0,9518). Yang lebih penting, kolom Gap menunjukkan tidak ada satu pun model yang mengalami penurunan signifikan antara performa CV dan test bahkan empat model mencatat gap positif yang berarti test F1 lebih tinggi dari CV F1. Hal ini secara empiris membuktikan bahwa pipeline anti-leakage yang diterapkan berhasil menghasilkan estimasi CV yang konservatif dan tidak optimistis, berbeda dengan fenomena yang dikritisi pada studi terdahulu [9].

C. Perbandingan Performa Model

Memasuki tahap pemodelan, hasil evaluasi performa dari keenam model diukur menggunakan berbagai metrik utama yaitu Accuracy, Precision, Recall, F1-Score, AUC, PR-AUC, Brier Score, dan Log Loss. Rangkuman hasil performa dari seluruh model tersebut disajikan secara lengkap pada Tabel 3.

TABEL 3.
TABEL PERBANDINGAN PERFORMA MODEL

Model	Accuracy	Precision	Recall	F1-Score	AUC	PR-AUC	Brier Score	Log Loss
XGBoost	0.9740	0.9545	0.9545	0.9545	0.9967	0.9930	0.0183	0.0662
LightGBM	0.9610	0.9524	0.9091	0.9302	0.9917	0.9847	0.0259	0.1001
CatBoost	0.9740	1.0000	0.9091	0.9524	0.9946	0.9876	0.0251	0.0898
Logistic Regression	0.9740	1.0000	0.9091	0.9524	0.9942	0.9876	0.0223	0.0815
Random Forest	0.9740	1.0000	0.9091	0.9524	0.9926	0.9831	0.0217	0.0806
Decision Tree	0.9610	1.0000	0.8636	0.9268	0.9744	0.9655	0.0288	0.1221

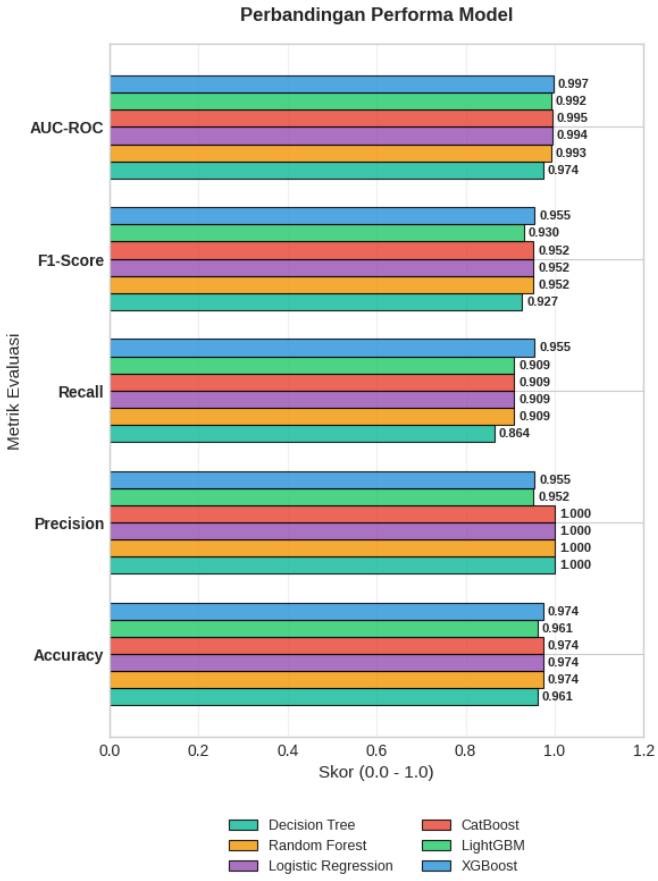
Tabel 3 menyajikan perbandingan performa keenam model pada data uji setelah melalui pipeline terkalibrasi. Dari kedelapan metrik tersebut, enam di antaranya (Accuracy, Precision, Recall, F1-Score, AUC-ROC, PR-AUC) mengukur ketepatan prediksi kelas, sedangkan dua sisanya yaitu Brier Score dan Log Loss mengukur seberapa dapat dipercaya angka probabilitas yang dikeluarkan model. Brier Score menghitung rata-rata kuadrat selisih antara probabilitas prediksi model dengan kondisi aktual pasien nilainya berkisar 0 hingga 1, semakin mendekati 0 semakin baik, dan nilai 0,25 setara tebakan acak. Penjelasan lebih lengkap disajikan pada Sub-bagian 3F.

Nilai yang dicetak tebal menunjukkan hasil terbaik pada setiap metrik. XGBoost tampil sebagai model terbaik dengan F1-Score 0,9545 dan AUC-ROC 0,9967, diikuti CatBoost, Logistic Regression, dan Random Forest yang sama-sama mencapai F1-Score 0,9524. LightGBM memperoleh F1-Score 0,9302, sedangkan Decision Tree memberikan performa terendah dengan F1-Score 0,9268.

Nilai AUC-ROC 0,9967 yang mendekati sempurna perlu dikontekstualisasikan secara kritis. Nilai ini konsisten dengan seluruh studi terdahulu pada dataset yang sama, di

mana AUC tertinggi yang dilaporkan mencapai 0,996 [5], 1,00 [6], dan 0,99 [8]. Performa tinggi ini dapat dijelaskan oleh kekuatan diskriminatif fitur Treatment Response yang sangat dominan dalam dataset fitur ini secara klinis merupakan prediktor kekambuhan yang sangat kuat karena langsung mencerminkan respons biologis pasien terhadap terapi awal, sebagaimana dikonfirmasi oleh analisis SHAP pada sub-bagian berikutnya. Validasi melalui Bootstrap BCa mengonfirmasi bahwa nilai AUC ini bukan artefak partisi data yang menguntungkan, melainkan estimasi yang stabil dan tidak bias (Bias mendekati nol pada seluruh model).

Meskipun CatBoost, Logistic Regression, Random Forest, dan Decision Tree mencapai Precision sempurna (1,0000) yang menunjukkan tidak adanya false positive, terdapat trade-off berupa nilai Recall yang lebih rendah. Dilihat dari sisi klinis, Recall terhadap kelas Kambuh merupakan metrik yang paling kritis karena kesalahan prediksi berupa false negative (pasien kambuh yang tidak terdeteksi) memiliki risiko yang jauh lebih berbahaya bagi keselamatan pasien. Perbandingan performa kelima metrik utama antar model secara visual disajikan pada Gambar 5.

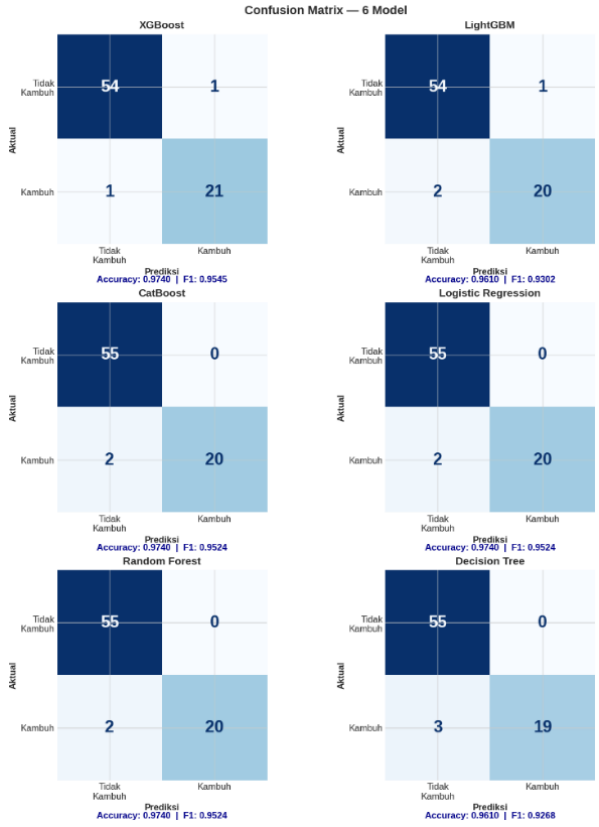


Gambar 5. Visualisasi Perbandingan Performa Model

Gambar 5 menampilkan perbandingan lima metrik utama (Accuracy, Precision, Recall, F1-Score, AUC-ROC) keenam model dalam format horizontal bar chart. Secara visual terlihat bahwa XGBoost unggul pada metrik Recall dan F1-Score sekaligus, yang menjadi indikator penting dalam konteks klinis. CatBoost, Logistic Regression, dan Random Forest menunjukkan Precision sempurna (1,000) namun dengan Recall lebih rendah (0,909), sehingga F1-Score ketiganya sedikit di bawah XGBoost. Decision Tree konsisten berada di posisi terbawah di seluruh metrik.

D. Analisis Confusion Matrix

Analisis confusion matrix dilakukan untuk menelaah pola kesalahan klasifikasi tiap model secara lebih rinci pada data test (n=77: 55 tidak kambuh, 22 kambuh), sebagaimana ditampilkan pada Gambar 6.



Gambar 6. Confusion Matrix Model

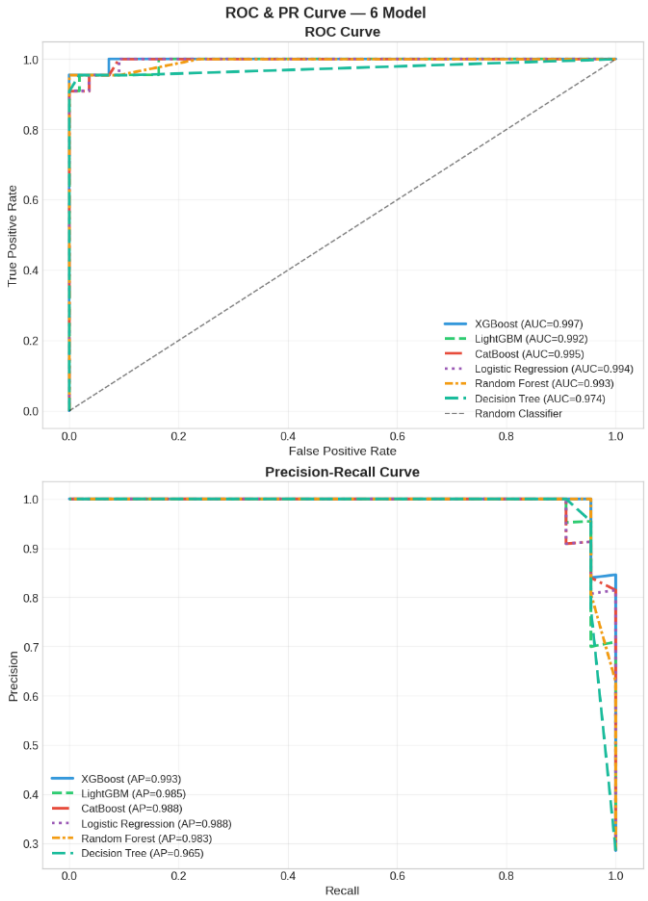
Gambar 6 digunakan untuk menelaah pola kesalahan klasifikasi tiap model secara lebih rinci terhadap 22 pasien kambuh pada data uji. Gambar tersebut memperlihatkan bahwa XGBoost hanya menghasilkan 2 kesalahan klasifikasi: 1 false positive dan 1 false negative. Hal ini menjadikan XGBoost sebagai model dengan tingkat deteksi tertinggi, di mana model berhasil mendeteksi 21 dari 22 pasien kambuh.

Sebagai perbandingan, model LightGBM, CatBoost, Logistic Regression, dan Random Forest menunjukkan pola serupa dengan meluputkan deteksi pada 2 pasien kambuh (Recall 0,9091). Sementara itu, Decision Tree memiliki performa terburuk dengan meluputkan 3 pasien kambuh

(Recall 0,8636). Dari perspektif klinis, false negative jauh lebih berbahaya karena pasien kambuh yang tidak terdeteksi berisiko tidak mendapatkan penanganan medis lanjutan yang diperlukan. Oleh karena itu, XGBoost dengan tingkat false negative terendah menjadi pilihan paling aman untuk sistem pendukung keputusan klinis.

E. ROC Curve dan Precision-Recall Curve

Semua model mencapai AUC-ROC di atas 0,97, dengan XGBoost tertinggi (0,9967) dan Decision Tree terendah (0,9744). Nilai AUC-ROC yang tinggi secara konsisten mengindikasikan bahwa fitur-fitur klinis dalam dataset memiliki daya diskriminatif yang sangat baik. Pada Precision-Recall Curve, XGBoost kembali mencatat Average Precision tertinggi sebesar 0,9930. ROC Curve dan Precision-Recall Curve keenam model disajikan secara berdampingan pada Gambar 7 untuk memvisualisasikan kemampuan diskriminasi dan performa pada kondisi kelas tidak seimbang.

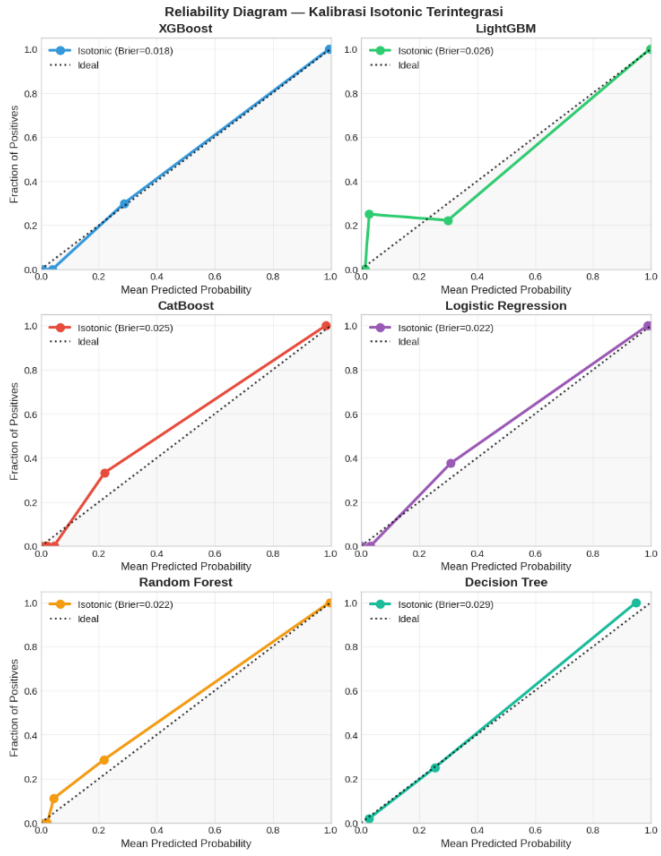


Gambar 7. Visualisasi ROC Curve dan Precision-Recall Curve Model

Gambar 7 menampilkan ROC Curve (atas) dan Precision-Recall Curve (bawah) keenam model. Pada ROC Curve, seluruh model membentuk kurva yang sangat mendekati sudut kiri atas, mencerminkan kemampuan diskriminasi yang sangat baik. XGBoost mencatat AUC-ROC tertinggi (0,9967), diikuti CatBoost (0,9946), Logistic Regression (0,9942), Random Forest (0,9926), LightGBM (0,9917), dan Decision Tree (0,9744). Pada Precision-Recall Curve, XGBoost kembali unggul dengan Average Precision (AP) tertinggi sebesar 0,993, sedangkan Decision Tree memiliki AP terendah (0,965). Seluruh model mempertahankan Precision=1,000 pada nilai Recall rendah hingga menengah, namun mulai menurun saat Recall mendekati 1,0 konsisten dengan *trade-off* Precision-Recall yang diobservasi pada tabel performa.

F. Verifikasi Kalibrasi Probabilitas

Kualitas kalibrasi probabilitas dievaluasi melalui Reliability Diagram dan nilai Brier Score. Brier Score mengukur rata-rata kuadrat selisih antara probabilitas prediksi model dan label aktual (0 atau 1) semakin rendah nilainya semakin baik, dengan nilai 0 berarti prediksi sempurna dan nilai 0,25 setara dengan prediksi acak pada dataset seimbang.



Gambar 8. Reliability Diagram Model

Dalam konteks klinis, Brier Score yang rendah berarti probabilitas yang diberikan model kepada klinisi dapat dipercaya: jika model menyatakan risiko kambuh 80%, maka memang sekitar 80% pasien dengan profil klinis serupa benar-benar mengalami kekambuhan. Kalibrasi yang baik sangat penting untuk sistem pendukung keputusan klinis karena klinisi tidak hanya membutuhkan prediksi biner

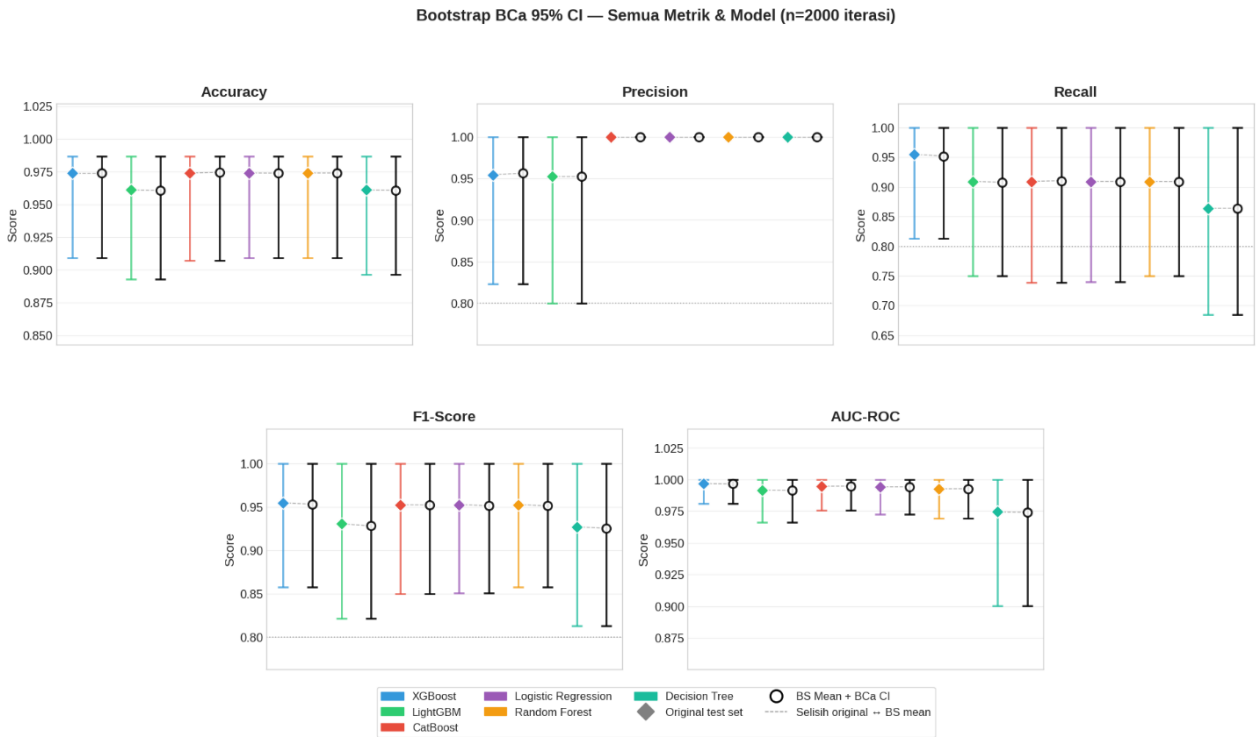
(kambuh/tidak), tetapi juga estimasi probabilitas yang dapat diandalkan untuk menentukan intensitas pemantauan pasien. Reliability Diagram keenam model disajikan Gambar 8.

Gambar 8 menampilkan Reliability Diagram (diagram keandalan kalibrasi) untuk keenam model. Sumbu-x merepresentasikan rata-rata probabilitas prediksi, sedangkan sumbu-y merepresentasikan proporsi kejadian aktual (fraction of positives). Model yang terkalibrasi sempurna akan menghasilkan kurva yang tepat berimpit dengan garis diagonal putus-putus (ideal). Dari gambar, XGBoost menunjukkan kurva yang paling dekat dengan diagonal sepanjang rentang probabilitas 0,0–1,0, konsisten dengan Brier Score terendahnya (0,018). LightGBM memperlihatkan sedikit penyimpangan di rentang probabilitas rendah (0,0–0,3), yang ditunjukkan oleh Brier Score-nya yang lebih tinggi (0,026). CatBoost dan Random Forest menunjukkan pola serupa kurva hampir linear namun sedikit di bawah diagonal pada probabilitas menengah. Logistic Regression dan Decision Tree juga terkalibrasi cukup baik, meskipun Decision Tree mencatat Brier Score tertinggi (0,029). Secara keseluruhan, kalibrasi Isotonic yang terintegrasi dalam pipeline CV berhasil menghasilkan probabilitas yang cukup andal pada keenam model.

G. Bootstrap BCa Validation

Validasi Bootstrap BCa dilakukan dengan 2.000 iterasi sampling with replacement dari data test (n=77) untuk menghasilkan confidence interval yang lebih akurat dibandingkan percentile bootstrap standar pada distribusi skewed dan dataset kecil [19]. Karena nilai akselerasi (a) yang dihitung mendekati nol pada seluruh model, koreksi BCa yang diterapkan setara dengan Bias-Corrected (BC) bootstrap artinya distribusi bootstrap sudah cukup simetris dan tidak memerlukan koreksi skewness tambahan. Pendekatan ini sangat direkomendasikan untuk evaluasi model ML pada dataset berukuran kecil karena mampu mengukur ketidakpastian estimasi performa secara lebih andal dibandingkan pelaporan metrik tunggal pada satu partisi test set [14].

Setiap kolom pada Tabel 4 merepresentasikan informasi yang berbeda dan saling melengkapi: kolom Original adalah nilai F1-Score aktual pada data test asli sebagai titik referensi utama; kolom BS Mean adalah rata-rata F1-Score dari 2.000 sampel bootstrap yang diharapkan mendekati Original; kolom Bias (BS Mean − Original) yang mendekati nol mengonfirmasi tidak adanya estimasi yang terlalu optimistis; kolom Std mengukur konsistensi estimasi antar iterasi; kolom CI Low dan CI High adalah batas 95% confidence interval BCa [18]; kolom CI Width merepresentasikan lebar ketidakpastian estimasi; serta kolom Stabilitas yang dikategorikan sebagai Sangat Stabil (CI Width < 0,15), Stabil (< 0,20), Cukup Stabil (< 0,25) dan Kurang Stabil (≥ 0,25). Ringkasan hasil Bootstrap BCa untuk F1-Score keenam model serta visualisasi confidence interval seluruh metrik disajikan pada Tabel 4 dan Gambar 9.



Gambar 9. CI All Metrics

TABEL 4.

TABEL STABILITAS BOOTSTRAP BCA— F1-SCORE (N=2.000 ITERASI)

Model	Original	BS Mean	Bias	Std	CI Low	CI High	CI Width	Stabilitas
XGBoost	0.9545	0.9531	-0.0014	0.0345	0.8571	1.0	0.1429	Sangat Stabil
LightGBM	0.9302	0.9280	-0.0023	0.0415	0.8213	1.0	0.1787	Stabil
CatBoost	0.9524	0.9519	-0.0005	0.0355	0.8500	1.0	0.1500	Stabil
Logistic Regression	0.9524	0.9512	-0.0012	0.0354	0.8505	1.0	0.1495	Sangat Stabil
Random Forest	0.9524	0.9511	-0.0013	0.0356	0.8571	1.0	0.1429	Sangat Stabil
Decision Tree	0.9268	0.9252	-0.0017	0.0431	0.8125	1.0	0.1875	Stabil

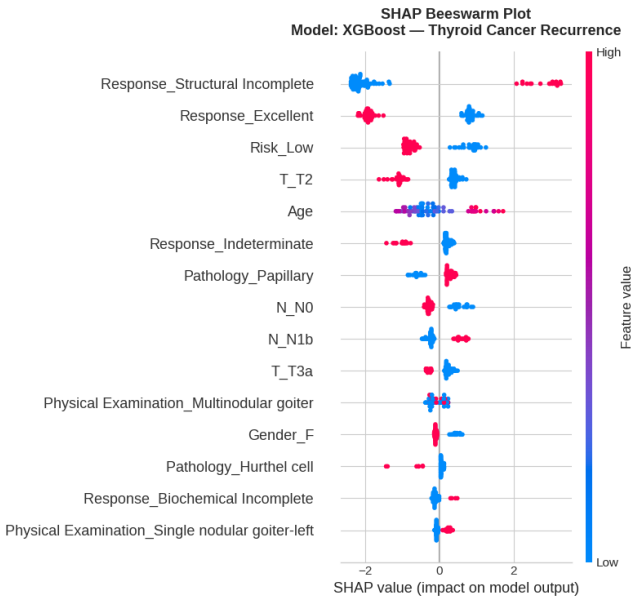
Berdasarkan Tabel 4 dan Gambar 9, seluruh model mencatat nilai Bias dengan nilai absolut berkisar antara 0,0005 hingga 0,0023 mendekati nol membuktikan bahwa pipeline anti-data leakage yang diterapkan menghasilkan estimasi generalisasi yang jujur dan tidak mengalami inflasi akibat kebocoran data. XGBoost dan Random Forest mencatat CI Width terendah (0,1429) dan masuk kategori Sangat Stabil, diikuti Logistic Regression (CI Width 0,1495, Sangat Stabil). LightGBM, CatBoost, dan Decision Tree masuk kategori Stabil dengan CI Width berkisar antara 0,1500 hingga 0,1875. Nilai Std yang kecil pada seluruh model (berkisar 0,034 hingga 0,043) mengonfirmasi bahwa estimasi F1-Score sangat konsisten lintas 2.000 iterasi bootstrap tidak ada model yang menunjukkan fluktuasi besar antar sampel.

Pada Gambar 9, titik berlian (◆) yang merepresentasikan nilai original test set terlihat sangat berdekatan dengan titik lingkaran (●) BS Mean pada seluruh model di semua metrik mengonfirmasi minimnya bias secara visual dan konsisten dengan nilai Bias pada Tabel 4 yang mendekati nol. Garis hitam vertikal yang tampak pada setiap model adalah error bar yang merepresentasikan rentang 95% BCa confidence interval dari CI Low hingga CI High. Semakin pendek garis hitam tersebut, semakin sempit rentang ketidakpastian dan semakin stabil estimasi performa model. Error bar yang lebih lebar pada metrik Recall mencerminkan variabilitas yang lebih tinggi akibat sedikitnya jumlah kasus kambuh pada data test (n=22) kelas minoritas yang kecil secara inheren menghasilkan estimasi yang lebih tidak pasti dibandingkan metrik lain.

Sebaliknya, metrik AUC-ROC menunjukkan error bar paling sempit dan konsisten di seluruh model, mencerminkan bahwa kemampuan diskriminasi model sangat stabil lintas partisi bootstrap. Pada subplot Precision, beberapa model menunjukkan error bar yang cukup panjang ke bawah hingga ~0,90 hal ini wajar karena Precision sangat sensitif terhadap jumlah false positive yang dapat berfluktuasi besar pada dataset kecil. CI High yang mencapai 1,000 pada beberapa model dan metrik adalah hal wajar pada dataset kecil (n=77) dan mencerminkan ketidakpastian estimasi pada batas atas distribusi bootstrap, bukan indikasi kualitas model yang buruk. Secara keseluruhan, hasil Bootstrap BCa mengonfirmasi bahwa XGBoost tidak hanya unggul dalam performa point estimate, tetapi juga paling stabil dan konsisten dalam estimasi ketidakpastiannya dibandingkan kelima model lainnya.

H. Analisis Interpretabilitas Model (SHAP)

Untuk menganalisis kontribusi fitur terhadap prediksi model XGBoost secara interpretatif, SHAP Beeswarm Plot disajikan pada Gambar 10. Sumbu horizontal merepresentasikan nilai SHAP (dampak terhadap output model), di mana nilai positif mendorong prediksi ke arah kambuh dan nilai negatif bersifat protektif. Warna titik mencerminkan nilai fitur asli: merah (nilai tinggi) dan biru (nilai rendah).



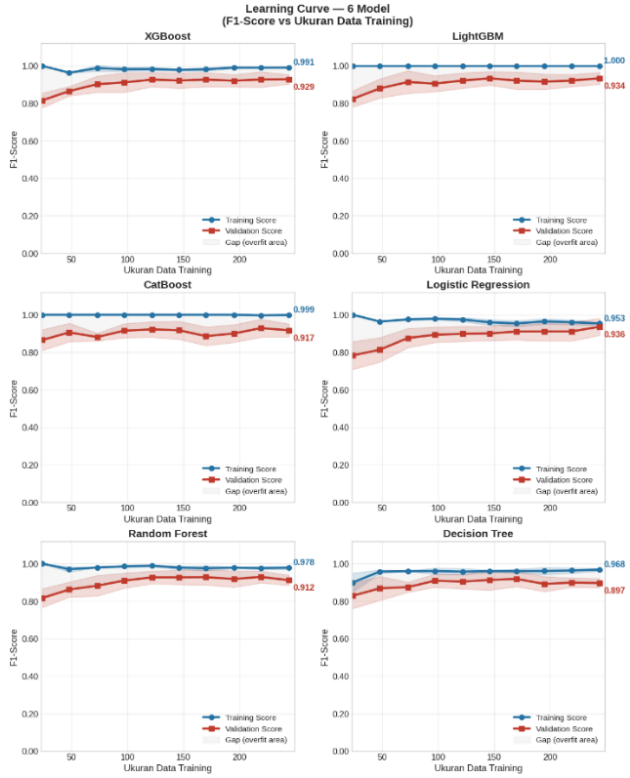
Gambar 10. SHAP Beeswarm Plot XGBoost

Berdasarkan Gambar 10, Response_Structural Incomplete merupakan fitur paling dominan (SHAP +2 hingga +4), mengindikasikan kontribusi prediksi kambuh yang sangat kuat, sementara Response_Excellent bersifat sebaliknya (SHAP −2 hingga −3) sebagai faktor protektif. Kedua fitur ini bersama-sama mengonfirmasi Treatment Response sebagai prediktor klinis paling informatif, konsisten dengan studi terdahulu yang melaporkan Treatment Response sebagai prediktor dominan dengan nilai SHAP 2,077 [8] dan fitur yang sama mendominasi pada pendekatan ensemble [9].

Fitur staging T_T2 dan N_N1b turut berkontribusi positif terhadap prediksi kambuh, mencerminkan relevansi sistem TNM dalam stratifikasi risiko DTC. Fitur Age menunjukkan distribusi gradasi warna yang tersebar di sekitar nol mengindikasikan hubungan non-linear antara usia dan risiko kekambuhan yang tidak dapat ditangkap oleh model statistik linear konvensional. Dari 55 fitur hasil OneHotEncoding, hanya 26 fitur yang memiliki kontribusi non-zero, sementara 29 fitur sisanya termasuk riwayat merokok, riwayat radioterapi, dan sebagian besar kategori fungsi tiroid tidak berkontribusi sama sekali terhadap prediksi, mengindikasikan bahwa model XGBoost secara efektif hanya memanfaatkan fitur-fitur yang relevan secara klinis. Konsistensi temuan SHAP ini dengan studi terdahulu [5]-[9] serta dengan pengetahuan klinis yang mapan memperkuat kredibilitas model XGBoost sebagai kandidat sistem pendukung keputusan klinis [20][21].

I. Analisis Learning Curve

Untuk mengonfirmasi bahwa performa tinggi yang dicapai bukan merupakan artefak *overfitting*, learning curve keenam model dianalisis. Learning curve dihasilkan dengan memvariasikan ukuran data latih dan mengukur F1-Score training maupun validasi pada setiap ukuran, sehingga dapat diamati apakah model mengalami *overfitting* atau masih dapat memperoleh manfaat dari data tambahan. Hasil learning curve keenam model disajikan pada Gambar 11.



Gambar 11. Learning Curve Model

Gambar 11 menunjukkan learning curve keenam model yang diplot dengan memvariasikan ukuran data latih dari ~10% hingga ~80% dari X_train, dengan F1-Score dihitung pada setiap ukuran menggunakan 5-fold CV.

Hasil analisis menunjukkan tiga pola konsisten yang mengindikasikan generalisasi yang sehat: (1) Seluruh validation curve bergerak naik seiring bertambahnya training size, mengindikasikan bahwa model belum mencapai saturasi dan masih dapat memperoleh manfaat dari data tambahan pola ini berlawanan dengan *overfitting* klasik di mana validation score justru menurun; (2) Gap antara training score dan validation score mengecil secara bertahap seiring training size bertambah pada semua model, mengindikasikan konvergensi yang normal; (3) Variance validation score (area yang diarsir) mengecil seiring bertambahnya data, menunjukkan stabilitas estimasi yang meningkat. Training score yang tinggi (mendekati 1,0) pada model ensemble seperti XGBoost, LightGBM, dan Random Forest adalah perilaku yang diharapkan ensemble kuat memang cenderung fit sempurna pada data latih, dan ini tidak bermasalah selama validation score terus naik.

Performa tinggi pada dataset ini konsisten dengan seluruh studi terdahulu pada dataset yang sama [5]-[9], dan dijelaskan secara substantif oleh kekuatan diskriminatif fitur-fitur klinis yang tersedia terutama Treatment Response yang secara konsisten tampil sebagai prediktor dominan dengan nilai SHAP tertinggi di berbagai studi [8][9]. Learning curve ini memberikan bukti empiris bahwa performa model yang dicapai merupakan refleksi nyata dari pola dalam data, bukan artefak teknis.

J. Perbandingan dengan Peneliti Terkait

Perbandingan komprehensif antara performa model usulan dengan lima studi terdahulu pada dataset Differentiated Thyroid Cancer Recurrence yang sama dirangkum dalam Tabel 5.

Berdasarkan Tabel 5, performa XGBoost (AUC-ROC 0,9967) tampil sangat kompetitif dibandingkan studi terdahulu. Berbeda dengan studi sebelumnya yang rentan mengalami data leakage akibat penerapan SMOTE di luar cross-validation [5], ketiadaan pelaporan metrik kalibrasi kuantitatif (seperti Brier Score) [6], berfokus pada subset fitur [7], atau belum melakukan validasi stabilitas [8][9], penelitian ini menawarkan empat keunggulan metodologis komprehensif: (1) pipeline zero-leakage berbasis ImbPipeline; (2) optimasi Optuna 100 trial; (3) kalibrasi Isotonic terintegrasi (Brier Score 0,0183); dan (4) validasi stabilitas Bootstrap BCa 2.000 iterasi.

TABEL 5.
TABEL PERBANDINGAN DENGAN PENELITI TERKAIT

Penelitian	Model	Model Terbaik	Accuracy	Precision	Recall	F1-Score	AUC
Clark et al. [5] (2024)	KNN, SVM, DT, RF, AdaBoost, XGBoost	Random Forest	0,99	0,99	0,97	0,98	0,992
Thakur et al. [6] (2025)	LR, DT, RF, GB, SVM, KNN	Random Forest	0,9826	0,9784	0,9784	0,9784	1,00
Hu et al. [7] (2025)	RF, XGBoost, LightGBM	Random Forest	0,869	-	0,722	0,757	0,931
Hanani et al. [8] (2025)	CatBoost, ET, LGBM, XGB, RF, DT, LR, NB, KNN, SVM	CatBoost	0,97	0,97	0,97	0,97	0,99
Wen et al. [9] (2026)	Stacking (SGD+ET+DT+XGB)	Stacking Ensemble	0,9652	0,9667	0,9062	0,9355	0,9921
Penelitian ini	XGB, LGBM, CB, LR, RF, DT	XGBoost	0,9740	0,9545	0,9545	0,9545	0,9967

IV. KESIMPULAN

Studi ini mengevaluasi enam algoritma machine learning (ML) untuk prediksi kekambuhan Differentiated Thyroid Cancer (DTC) menggunakan pipeline anti-leakage. XGBoost terbukti sebagai model terbaik dan sangat stabil (F1-Score 0,9545; AUC-ROC 0,9967; Brier Score 0,0183; CI Width 0,1429). Performa stabil yang turut dicapai oleh Logistic Regression (LR) dan Random Forest (RF) menegaskan bahwa pipeline terintegrasi (SMOTE, tuning, kalibrasi dalam CV) jauh lebih esensial dibandingkan kompleksitas model.

Analisis SHAP mengonfirmasi bahwa Treatment Response (khususnya Structural Incomplete) merupakan prediktor paling dominan dengan mean absolute SHAP value 2,264, diikuti fitur staging TNM dan usia memvalidasi bahwa prediksi model didasarkan pada faktor-faktor yang relevan secara patofisiologis dan konsisten dengan literatur klinis. Secara klinis, model XGBoost yang dikembangkan menunjukkan potensi awal sebagai kandidat komponen sistem pendukung keputusan pasca-terapi (post-treatment clinical decision support) untuk membantu klinisi mengidentifikasi pasien DTC berisiko tinggi berdasarkan respons awal terapi. Namun demikian, validasi prospektif pada lingkungan klinis nyata masih diperlukan sebelum implementasi aktual, mengingat performa yang dilaporkan berasal dari satu dataset retrospektif dari satu institusi. Keterbatasan lain meliputi ukuran dataset yang relatif kecil (383 pasien) yang meningkatkan risiko ketidakpastian estimasi pada model boosting seperti XGBoost, meskipun learning curve dan Bootstrap BCa mengonfirmasi tidak adanya indikasi overfitting yang signifikan. Generalisasi hasil penelitian ini terhadap populasi pasien di luar dataset UCI Machine Learning Repository perlu dikaji lebih lanjut, mengingat variasi demografis, praktik klinis, dan definisi respons terapi yang mungkin berbeda antar institusi dan populasi.

Penelitian lanjutan disarankan berfokus pada: (1) validasi eksternal multisenter menggunakan data dari institusi dan populasi yang berbeda untuk mengonfirmasi generalisabilitas model; (2) perbandingan dengan baseline klinis tradisional seperti sistem skoring ATA dan TNM untuk mengukur nilai tambah pendekatan ML secara klinis; (3) eksplorasi stacking ensemble keenam model sebagai base learner; dan (4) integrasi sistem ke dalam alur kerja klinis nyata dengan pengujian penerimaan klinisi. Selain itu, tidak adanya pembanding baseline klinis tradisional (sistem skoring ATA/TNM) dalam penelitian ini merupakan keterbatasan yang perlu diatasi pada penelitian lanjutan untuk mengkuantifikasi nilai tambah model ML dibandingkan praktik klinis standar yang sudah ada.

DAFTAR PUSTAKA

[1] J. Ferlay *et al.*, "Cancer site ranking," Lyon, France, 2024. Accessed: Apr. 25, 2026. [Online]. Available: https://gco.iarc.who.int/today

[2] L. Davies and H. G. Welch, "Current thyroid cancer trends in the United States," *JAMA Otolaryngol. Head Neck Surg.*, vol. 140, no. 4, pp. 317–322, 2014, doi: 10.1001/jamaoto.2014.1.

[3] R. M. Tuttle *et al.*, "Estimating risk of recurrence in differentiated thyroid cancer after total thyroidectomy and radioactive iodine remnant ablation: Using response to therapy variables to modify the initial risk estimates predicted by the new American thyroid association staging system," *Thyroid*, vol. 20, no. 12, pp. 1341–1349, Dec. 2010, doi: 10.1089/thy.2010.0178.

[4] Y. Li *et al.*, "A machine learning-based model for predicting recurrence in intermediate- and high-risk differentiated thyroid cancer: insights from a retrospective single-center study of 2388 patients," *Front. Endocrinol. (Lausanne).*, vol. 16, 2025, doi: 10.3389/fendo.2025.1552479.

[5] E. Clark, S. Price, T. Lucena, B. Haberlein, A. Wahbeh, and R. Seetan, "Predictive Analytics for Thyroid Cancer Recurrence: A Machine Learning Approach," *Knowledge*, vol. 4, no. 4, pp. 557–570, Nov. 2024, doi: 10.3390/knowledge4040029.

[6] D. Thakur, T. Gera, V. Bhardwaj, R. Mazen, A. Lasisi, and T. Engida, "A comparative study on advanced predictive modeling of thyroid cancer recurrence using multi algorithmic machine learning frameworks," *Sci. Rep.*, Dec. 2025, doi: 10.1038/s41598-025-33396-7.

[7] G. Hu *et al.*, "Machine learning prediction of thyroid cancer recurrence for early screening and clinical decision pathways: a retrospective cohort study," *Discover Oncology*, Jan. 2026, doi: 10.1007/s12672-025-04293-2.

[8] A. A. Hanani, T. B. Donmez, M. Kutlu, and M. Mansour, "Predicting thyroid cancer recurrence using supervised CatBoost A SHAP-based explainable AI approach," *Medicine (United*

States), vol. 104, no. 22, May 2025, doi: 10.1097/MD.0000000000042667.

[9] H. Wen, X. Li, and X. Zhao, "TC check: a web app for thyroid cancer recurrence prediction using explainable machine learning," *J. Cancer Res. Clin. Oncol.*, vol. 152, no. 1, Jan. 2026, doi: 10.1007/s00432-025-06377-6.

[10] A. Woźniacki, W. Książek, and P. Mrowczyk, "A Novel Approach for Predicting the Survival of Colorectal Cancer Patients Using Machine Learning Techniques and Advanced Parameter Optimization Methods," *Cancers (Basel).*, vol. 16, no. 18, Sep. 2024, doi: 10.3390/cancers16183205.

[11] D. A. Murtiningsih, W. Sari, and I. N. Fajri, "Comparison of Light Gradient Boosting Machine, eXtreme Gradient Boosting, and CatBoost with Balancing and Hyperparameter Tuning for Hypertension Risk Prediction on Clinical Dataset," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[12] A. Ichwani, R. Indra Kesuma, A. Setiawan, E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," *Jurnal Teknik Informatika (JUTIF)*, vol. 7, no. 1, 2026, doi: 10.52436/1.jutif.2026.7.2.5490.

[13] P. Netayawijit, W. Chansanam, and K. Sorn-In, "Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support," *Healthcare (Switzerland)*, vol. 13, no. 20, Oct. 2025, doi: 10.3390/healthcare13202588.

[14] G. Varoquaux and O. Colliot, "Evaluating Machine Learning Models and Their Diagnostic Value," in *Neuromethods*, vol. 197, Humana Press Inc., 2023, pp. 601–630. doi: 10.1007/978-1-0716-3195-9_20.

[15] S. Borzooei and A. Tarokhian, "Differentiated Thyroid Cancer Recurrence," UCI Machine Learning Repository. Accessed: May 02, 2026. [Online]. Available: https://doi.org/10.24432/C5632J

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[17] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, in ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.

[18] Bradley. Efron and Robert. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall/CRC, 1998.

[19] B. Efron and B. Narasimhan, "The Automatic Construction of Bootstrap Confidence Intervals," *Journal of Computational and Graphical Statistics*, vol. 29, no. 3, pp. 608–619, Jul. 2020, doi: 10.1080/10618600.2020.1714633.

[20] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[21] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.