Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

Perbandingan Kinerja Algoritma CatBoost, XGBoost, LightGBM dan Random Forest Dalam Memprediksi Risiko Infeksi Aids Dalam Dataset Kesehatan

Pramudya Ridwan Yulianto, Yani Parti Astuti*

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarng, Indonesia
Email: [1]111202113680@mhs.dinus.ac.id, [2,*]yanipartiastuti@dns.dinus.ac.id
Email Penulis Korespondensi: yanipartiastuti@dns.dinus.ac.id
Submitted: **18/12/2025**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstrak−Penelitian ini mengkaji prediksi risiko infeksi AIDS menggunakan algoritma berbasis pohon keputusan, yaitu CatBoost, XGBoost, LightGBM, dan Random Forest, pada dataset medis dan demografis sebanyak 2.139 observasi dengan 23 variabel. Tahapan penelitian meliputi eksplorasi data, pembersihan, penanganan nilai ekstrem dengan metode interkuartil, normalisasi menggunakan RobustScaler, serta penyeimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE). Mengingat karakteristik dataset yang tidak seimbang, evaluasi model tidak hanya menitikberatkan pada akurasi, tetapi juga pada Recall, F1-Score, dan AUC-ROC untuk menilai kemampuan deteksi kelas terinfeksi. Sebelum penerapan SMOTE, seluruh model menunjukkan akurasi tinggi namun recall kelas positif relatif rendah. Setelah penyeimbangan data, CatBoost menunjukkan peningkatan paling signifikan dengan recall meningkat dari 63% menjadi 77% dan F1-Score dari 72% menjadi 79%, serta akurasi mencapai 90%. Sebagai perbandingan, XGBoost mencapai akurasi 88,63% dengan peningkatan recall yang lebih moderat, sementara LightGBM dan Random Forest menunjukkan perbaikan yang konsisten namun tidak setinggi CatBoost. Hasil ini menunjukkan bahwa kombinasi SMOTE dan CatBoost lebih efektif dalam meminimalkan False Negatives pada kasus infeksi AIDS. Kontribusi penelitian ini terletak pada integrasi teknik penanganan outlier, normalisasi robust, dan penyeimbangan kelas dalam satu kerangka eksperimen yang terstruktur, serta penekanan pada peningkatan sensitivitas model sebagai aspek krusial dalam mendukung deteksi dini risiko AIDS secara lebih akurat dan aplikatif dalam konteks klinis.

Kata Kunci: AIDS; Machine Learning; SMOTE; CatBoost; Klasifikasi Medis

Abstract−This study investigates the prediction of AIDS infection risk using tree-based algorithms CatBoost, XGBoost, LightGBM, and Random Forest applied to a medical and demographic dataset consisting of 2,139 observations and 23 variables. The research process includes data exploration, cleaning, handling extreme values using the interquartile range (IQR) method, normalization with RobustScaler, and class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). Due to the imbalanced nature of the dataset, model evaluation emphasizes not only accuracy but also Recall, F1-Score, and AUC-ROC to better assess infected class detection. Prior to SMOTE implementation, all models achieved high accuracy but relatively low recall for the positive class; after resampling, CatBoost demonstrated the most significant improvement, with recall increasing from 63% to 77% and F1-Score from 72% to 79%, achieving an overall accuracy of 90%. In comparison, XGBoost reached an accuracy of 88.63% with a more moderate recall improvement, while LightGBM and Random Forest showed consistent yet smaller gains, indicating that the combination of SMOTE and CatBoost is more effective in minimizing False Negatives in AIDS infection cases. The main contribution of this study lies in the integration of robust outlier handling, feature normalization, and class balancing within a structured experimental framework, with a specific emphasis on sensitivity optimization to enhance early detection reliability in clinical screening contexts.

Keywords: AIDS; Machine Learning; SMOTE; CatBoost; Medical Classification

1. PENDAHULUAN

HIV (Human Immunodeficiency Virus) adalah virus yang menyerang sistem kekebalan tubuh manusia dengan menargetkan sel-sel yang mengandung antigen CD4, terutama limfosit T[1]. Sel-sel ini memiliki peran penting dalam mengatur dan mempertahankan respons imun tubuh. Infeksi HIV menyebabkan kerusakan bertahap pada sel-sel tersebut, yang pada gilirannya melemahkan kemampuan sistem kekebalan tubuh untuk melawan infeksi lainnya[2]. Pada akhir tahun 2024, diperkirakan sekitar 40,8 juta orang di seluruh dunia hidup dengan HIV. Di kawasan Asia Tenggara, jumlah orang yang hidup dengan HIV diperkirakan mencapai 3,5 juta orang[3]. Sementara itu, Kementerian Kesehatan Republik Indonesia (Kemenkes RI) memperkirakan jumlah Orang Dengan HIV (ODHIV) di Indonesia pada triwulan pertama tahun 2025 sebanyak 564.000 orang, dengan kelompok umur terbanyak pada 25-49 tahun (61%), diikuti oleh 20-24 tahun (20%), dan ≥50 tahun (11%)[4].

Permasalahan utama dari kondisi tersebut adalah masih tingginya tingkat keterlambatan diagnosis, sehingga banyak kasus baru ditemukan ketika telah memasuki fase lanjut. Rendahnya tingkat deteksi dini menyebabkan pasien tidak segera memperoleh terapi antiretroviral yang efektif[5]. Oleh karena itu, diperlukan pendekatan berbasis teknologi yang mampu memprediksi risiko infeksi lebih awal, sehingga intervensi dapat dilakukan lebih cepat dan tepat[6]. Prediksi dini juga membantu pemerintah dan tenaga kesehatan dalam memetakan kelompok risiko tinggi yang membutuhkan pemantauan khusus.

Pencegahan yang optimal memerlukan deteksi sejak dini, pengobatan menggunakan ARV, serta pemberian edukasi yang tepat kepada masyarakat[7]. Meskipun HIV dapat dikelola dengan pengobatan antiretroviral (ARV), apabila tidak segera ditangani, HIV dapat berkembang menjadi AIDS (Acquired Immunodeficiency Syndrome), yang merupakan tahap paling parah dari infeksi HIV[8]. AIDS terjadi ketika sistem kekebalan tubuh sangat rusak akibat

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

infeksi HIV dan tubuh menjadi sangat rentan terhadap infeksi oportunistik dan kanker tertentu. Pencegahan **AIDS** sangat penting, dan ini membutuhkan deteksi dini serta pengobatan yang tepat untuk menghambat perkembangan infeksi menjadi tahap ini[9]. Pencegahan AIDS sangat penting, berbagai faktor, seperti usia, perilaku seksual, riwayat medis, dan kondisi kesehatan secara keseluruhan, terkait dengan peningkatan risiko AIDS, namun menganalisisnya melibatkan hubungan yang rumit antara berbagai variabel[10]. Kompleksitas interaksi antar faktor risiko tersebut membuat metode analisis tradisional sulit menghasilkan prediksi yang akurat. Oleh karena itu, penggunaan *machine learning* menjadi alternatif yang sangat potensial karena mampu memodelkan hubungan non-linear dan interaksi multivariat dalam dataset medis yang kompleks.

Penelitian ini mengaplikasikan beberapa algoritma berbasis pohon keputusan seperti CatBoost, XGBoost, LightGBM, dan Random Forest untuk memprediksi risiko infeksi AIDS menggunakan dataset tabular karena algoritma-algoritma tersebut terbukti efektif dalam menangani hubungan non-linear dan kompleksitas tinggi pada data medis[11],[12]. Karakteristik data klinis umumnya memiliki variasi nilai ekstrem atau pencilan (outlier) yang muncul akibat perbedaan kondisi biologis, tingkat keparahan penyakit, serta respons fisiologis antar pasien. Kondisi ini dapat memengaruhi stabilitas model apabila tidak ditangani dengan tepat[13]. Oleh karena itu, penelitian ini menggunakan RobustScaler untuk proses normalisasi data karena metode ini berbasis rentang interkuartil (Interquartile Range/IQR) sehingga lebih tahan terhadap pengaruh pencilan dibandingkan metode scaling tradisional seperti MinMaxScaler atau StandardScaler[14]. Selain itu, permasalahan umum berupa ketidakseimbangan kelas pada dataset HIV ditangani dengan menerapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) guna meningkatkan representasi kelas minoritas dan memperbaiki sensitivitas model terhadap kasus positif HIV[12]. Evaluasi performa dilakukan secara komprehensif menggunakan metrik akurasi, precision, recall, dan F1-score untuk mengidentifikasi algoritma yang paling efektif dalam klasifikasi risiko infeksi[11]. Setiap algoritma yang digunakan menawarkan keunggulan spesifik, mulai dari kemampuan CatBoost dalam menangani fitur kategorikal secara efisien, kecepatan komputasi tinggi pada LightGBM, hingga stabilitas dan robustitas prediksi pada Random Forest dalam berbagai kondisi distribusi data[11].

Penelitian serupa telah dilakukan oleh beberapa peneliti yang fokus pada penerapan algoritma machine learning untuk klasifikasi AIDS. Salah satu penelitian yang berjudul "Perbandingan Kinerja Algoritma Machine Learning dalam Deteksi Potensi Risiko HIV" membandingkan beberapa algoritma machine learning dalam prediksi AIDS, termasuk XGBoost, Random Forest, dan teknik SMOTE untuk menyeimbangkan data. Hasil penelitian mereka menunjukkan bahwa CatBoost menghasilkan akurasi tertinggi sebesar 89,01%, diikuti oleh XGBoost dengan 88,08%, dan LightGBM dengan 87,61%[12]. Penelitian lainnya, berjudul "Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy" juga melakukan perbandingan antara Decision Tree, Random Forest, XGBoost, dan LightGBM dengan menggunakan SMOTE untuk mengatasi ketidakseimbangan kelas. Hasil mereka menunjukkan bahwa Random Forest mencatatkan AUC-ROC sebesar 0.99 dan F1-Score 0.93, sementara Decision Tree memperoleh AUC-ROC sebesar 0.99 dan F1-Score 0.94, serta XGBoost dan LightGBM masing-masing dengan AUC-ROC 0.98 dan F1-Score 0.94[15].

Penelitian lain dalam kurun waktu lima tahun terakhir juga mulai menyoroti tingginya variabilitas performa model akibat teknik preprocessing yang berbeda. Beberapa penelitian menekankan pentingnya penanganan outlier, normalisasi fitur, serta pemilihan parameter model untuk meningkatkan akurasi prediksi. Meskipun demikian, masih sedikit penelitian yang secara khusus mengombinasikan penanganan outlier, normalisasi robust, dan oversampling SMOTE sekaligus pada dataset HIV/AIDS. Hal inilah yang menjadi *research gap* utama yang ingin dijawab dalam penelitian ini. Walaupun sejumlah penelitian telah membahas penerapan machine learning untuk memprediksi risiko AIDS, namun akurasi dan tingkat kecanggihan model yang dihasilkan masih dapat ditingkatkan[16]. Hasil penelitian ini diharapkan dapat memberikan wawasan tentang algoritma yang lebih efektif dan efisien dalam klasifikasi AIDS, serta berkontribusi pada pengembangan sistem pendukung keputusan di bidang kesehatan masyarakat[16]. Dengan memanfaatkan data terbaru dan teknologi machine learning, penelitian ini bertujuan untuk meningkatkan akurasi diagnosis AIDS, mengurangi beban ekonomi dan sosial yang ditimbulkan oleh penyakit ini, serta memperluas akses layanan kesehatan di Indonesia[17].

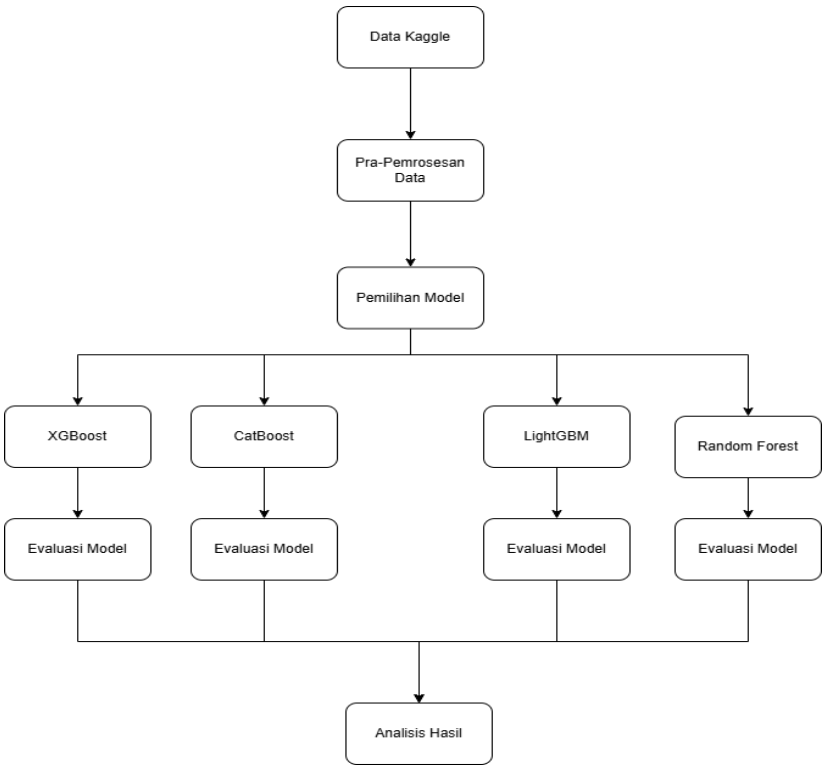
Berdasarkan uraian tersebut, penelitian ini berkontribusi dengan menawarkan pendekatan prediktif berbasis machine learning untuk mendukung deteksi dini risiko infeksi AIDS melalui integrasi penanganan outlier berbasis IQR, normalisasi menggunakan RobustScaler, dan penyeimbangan kelas dengan SMOTE dalam satu alur pra-pemrosesan yang terstruktur. Penelitian ini juga melakukan perbandingan empat algoritma berbasis pohon keputusan guna mengidentifikasi model paling optimal, dengan penekanan pada peningkatan akurasi, sensitivitas (recall) kelas minoritas untuk meminimalkan kesalahan False Negative dalam konteks klinis.

2. METODOLOGI PENELITIAN

Penelitian ini membandingkan kinerja algoritma CatBoost, XGBoost, LightGBM, dan Random Forest dalam klasifikasi risiko infeksi AIDS berdasarkan data klinis dan demografis. Tahapan penelitian meliputi pengumpulan data, prapemrosesan (EDA, pembersihan data, penanganan outlier, normalisasi, dan penyeimbangan kelas), pelatihan model menggunakan parameter default masing-masing algoritma, serta evaluasi performa menggunakan metrik

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

klasifikasi standar dengan cross-validation. Alur tahapan penelitian yang digunakan dalam studi ini disajikan pada Gambar 1.



Gambar 1. Alur Program

Berdasarkan Gambar 1, penelitian diawali dengan pengambilan dataset dari Kaggle yang kemudian melalui tahap pra-pemrosesan data, meliputi eksplorasi data (EDA), pembersihan data, penanganan outlier, normalisasi, dan penyeimbangan kelas menggunakan SMOTE. Selanjutnya dilakukan pemilihan dan pelatihan model menggunakan empat algoritma, yaitu XGBoost, CatBoost, LightGBM, dan Random Forest. Masing-masing model kemudian dievaluasi menggunakan metrik klasifikasi melalui skema cross-validation. Tahap akhir penelitian adalah analisis hasil untuk membandingkan performa setiap algoritma dalam mendeteksi risiko infeksi AIDS.

2.1 Pengumpulan Data

Penelitian ini menggunakan dataset AIDS Virus Infection Prediction yang diperoleh dari platform Kaggle.Dataset ini mencakup 23 fitur yang melibatkan data medis dan demografis pasien dengan HIV, yang digunakan untuk memprediksi kemungkinan infeksi AIDS. Fitur-fitur yang terkandung dalam dataset meliputi parameter medis seperti Age (Usia), Wtkg (Berat badan), Hemo (Riwayat hemofilia), Homo (Riwayat homoseksual), Drugs (Riwayat penggunaan narkoba suntik), dan Karnof (Skor Karnofsky), serta informasi klinis seperti Cd40, Cd420, Cd80, dan Cd820. Fitur-fitur ini digunakan untuk membangun model klasifikasi yang dapat memprediksi kemungkinan seseorang terinfeksi AIDS berdasarkan data medis dan faktor risiko yang ada. Dataset dapat diakses di https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction/data

2.2 Preprocessing Data

Preprocessing data adalah tahap yang sangat penting dalam proses machine learning untuk memastikan bahwa data yang digunakan dalam model adalah bersih, terstruktur dengan baik, dan siap digunakan[18]. Tahapan preprocessing ini meliputi beberapa langkah penting yang dirancang untuk mengatasi masalah-masalah seperti distribusi data yang tidak merata, ketidakseimbangan kelas, dan deteksi outlier[19]. Dalam penelitian ini, proses preprocessing dilakukan melalui serangkaian tahapan yang dijelaskan sebagai berikut:

a. Eksplorasi Data (Exploratory Data Analysis - EDA)
   Pada tahap ini, dilakukan analisis statistik deskriptif untuk memeriksa distribusi data, memeriksa adanya nilai yang hilang, dan mendeteksi outlier. Fungsi df.describe() digunakan untuk menghitung statistik dasar seperti rata-rata (mean), standar deviasi, nilai minimum, nilai maksimum, dan kuartil. Selain itu, visualisasi distribusi data menggunakan histogram atau Kernel Density Estimation (KDE) untuk memeriksa apakah data terdistribusi normal atau miring (skewed). Rumus untuk menghitung rata-rata (mean):
   Mean (Rata-rata)

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

Di mana $x_i$adalah nilai data dan $n$adalah jumlah data.
Standar Deviasi:

$$\text{Standar Deviasi} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad (2)$$

Di mana μ adalah nilai rata-rata dari data.

b. Penanganan Missing Value dan Data Duplikat
Missing value diatasi dengan menghapus baris yang mengandung nilai yang hilang (NaN) menggunakan df.dropna(inplace=True). Selain itu, data duplikat dihapus dengan menggunakan df.drop_duplicates(inplace=True), yang memastikan bahwa dataset tidak memiliki data yang terulang, sehingga model dapat mempelajari pola yang lebih beragam. Imputasi Missing Value dapat dilakukan dengan menggantikan nilai yang hilang menggunakan rata-rata (mean) atau modus (mode), khususnya jika missing value ditemukan dalam jumlah yang signifikan. Rumus untuk imputasi rata-rata adalah:

$$\text{Imputasi} = \frac{\sum x}{n} \quad (3)$$

Di mana x adalah nilai-nilai yang ada dan n adalah jumlah nilai yang ada dalam kolom.

c. Deteksi Outlier
Deteksi outlier dilakukan menggunakan boxplot dan Interquartile Range (IQR). Outlier yang terdeteksi digantikan dengan nilai median dari kolom terkait. Rumus untuk menghitung IQR adalah:

$$\text{IQR} = Q3 - Q1 \quad (4)$$

Di mana $Q1$ adalah kuartil pertama (25th percentile), $Q3$adalah kuartil ketiga (75th percentile). Batas bawah dan batas atas untuk mendeteksi outlier dihitung dengan rumus berikut:

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} \quad (5)$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} \quad (6)$$

Data yang berada di luar batas ini dianggap sebagai outlier dan digantikan dengan nilai median dari kolom tersebut. Meskipun algoritma berbasis pohon seperti CatBoost, XGBoost, LightGBM, dan Random Forest secara umum cukup robust terhadap nilai ekstrem, penanganan outlier tetap dilakukan untuk mengurangi potensi distorsi distribusi fitur numerik, khususnya pada tahap normalisasi. Dalam penelitian ini, penggantian outlier dengan nilai median dilakukan secara selektif pada fitur dengan distribusi sangat skewed dan nilai ekstrem yang signifikan. Pendekatan ini dipilih dibandingkan teknik clipping karena median lebih stabil terhadap pencilan dan tetap mempertahankan karakteristik pusat distribusi data. Namun demikian, karena model berbasis pohon tidak sensitif terhadap skala fitur, dampak reduksi variansi terhadap performa model relatif minimal.

d. Normalisasi Data
Normalisasi digunakan untuk mengubah data agar berada dalam rentang yang konsisten, terutama ketika ada nilai ekstrem (outliers). Anda menggunakan RobustScaler untuk menangani data dengan outliers. RobustScaler mengurangi pengaruh outliers dengan menggunakan median dan interquartile range (IQR). Rumus untuk normalisasi dengan RobustScaler:

$$X_{\text{norm}} = \frac{X - \text{median}(X)}{\text{IQR}} \quad (7)$$

Di mana Xadalah nilai data, median(X)adalah nilai median dari fitur, IQR adalah Interquartile Range.

e. SMOTE (Synthetic Minority Over-sampling Technique)
Pada dataset ini, SMOTE digunakan untuk menangani ketidakseimbangan kelas dalam dataset, khususnya pada variabel target Infected. SMOTE meningkatkan representasi kelas minoritas (pasien yang terinfeksi AIDS) dengan menghasilkan data sintetis berdasarkan kedekatan data yang ada. Teknik ini membantu model untuk mempelajari pola dari kedua kelas (terinfeksi dan tidak terinfeksi) secara seimbang. Untuk menghindari data leakage, proses SMOTE dilakukan setelah pembagian data menjadi data pelatihan dan data pengujian. Teknik oversampling hanya diterapkan pada data pelatihan, sementara data pengujian tetap menggunakan distribusi asli. Pendekatan ini memastikan bahwa model tidak memperoleh informasi dari data pengujian selama proses pelatihan, sehingga hasil evaluasi tetap realistis dan mencerminkan performa model pada data yang belum pernah dilihat sebelumnya.

2.3 Evaluasi Model

Evaluasi model dilakukan dengan menggunakan cross-validation, Akurasi, Precision, Recall, F1-Score, AUC-ROC, dan Confusion Matrix[20]. Berikut adalah rumus yang digunakan untuk menghitung metrik evaluasi

a. Akurasi (Accuracy)
Akurasi mengukur seberapa banyak prediksi yang benar dari total prediksi yang dibuat oleh model. Rumus untuk menghitung akurasi adalah sebagai berikut:

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Dimana TP adalah True Positives (positif yang benar), TN adalah True Negatives (negatif yang benar), FP adalah False Positives (negatif yang salah diprediksi sebagai positif), FN adalahFalse Negatives (positif yang salah diprediksi sebagai negatif). Akurasi memberikan gambaran umum tentang kinerja model dalam mengklasifikasikan kedua kelas (terinfeksi dan tidak terinfeksi).

b. Presisi (Precision)
Precision mengukur proporsi prediksi positif yang benar-benar positif. Rumus untuk menghitung precision adalah sebagai berikut:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

Precision sangat penting ketika kita ingin menghindari False Positives, misalnya dalam kasus diagnosis medis di mana kita tidak ingin memberikan diagnosis positif yang salah.

c. Recall (Sensitivitas atau True Positive Rate)
Recall mengukur seberapa baik model mendeteksi kelas positif yang sebenarnya. Rumus untuk menghitung recall adalah sebagai berikut:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

Recall sangat penting ketika kita ingin meminimalkan False Negatives, misalnya untuk memastikan bahwa sebanyak mungkin kasus positif (terinfeksi AIDS) terdeteksi oleh model.

d. F1-Score
F1-Score adalah rata-rata harmonis antara precision dan recall, yang memberikan gambaran seimbang antara keduanya, terutama dalam dataset yang tidak seimbang. Rumus untuk menghitung F1-Score adalah:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (11)$$

F1-Score memberikan keseimbangan yang lebih baik ketika kita menangani dataset dengan distribusi kelas yang tidak seimbang, seperti dalam prediksi infeksi AIDS.

e. AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)
AUC-ROC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC yang lebih tinggi menunjukkan bahwa model lebih baik dalam memisahkan kedua kelas. Rumus untuk menghitung AUC-ROC adalah:

$$TPR = \frac{TP}{TP+FN} \quad (12)$$

$$FPR = \frac{FP}{FP+TN} \quad (13)$$

AUC-ROC memberikan gambaran bagaimana model mampu memisahkan kelas positif dan negatif di berbagai threshold.

f. Confusion Matrix
Confusion matrix adalah alat evaluasi yang memberikan gambaran lebih rinci mengenai prediksi model. Matrix ini menampilkan jumlah prediksi yang benar dan salah untuk masing-masing kelas. Sebuah confusion matrix akan terlihat seperti berikut:

Tabel 1. Confusion Matrix

	Prediksi Positif (1)	Prediksi Negatif (0)
Aktual Positif (1)	True Positive (TP)	False Negative (FN)
Aktual Negatif (0)	False Positive (FP)	True Negative (TN)

Confusion matrix sangat berguna untuk memahami bagaimana model memprediksi setiap kelas dan mengidentifikasi jenis kesalahan yang terjadi, seperti kesalahan dalam mendeteksi False Positives atau False Negatives.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil yang diperoleh dari setiap tahapan yang telah dijelaskan sebelumnya dalam metodologi penelitian. Setiap langkah yang diuraikan, mulai dari pengumpulan data, preprocessing, pemilihan model, hingga evaluasi model, memiliki kontribusi yang sangat penting terhadap pencapaian hasil akhir. Penjelasan mendalam mengenai hasil-hasil yang diperoleh pada setiap tahap ini akan memberikan gambaran yang lebih jelas tentang bagaimana proses eksperimen dijalankan dan bagaimana data yang terkumpul dapat memberikan pemahaman yang lebih baik terhadap topik yang diteliti. Berdasarkan hasil evaluasi model, akan dibahas pula perbandingan kinerja

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

masing-masing algoritma serta penyesuaian yang dilakukan untuk meningkatkan performa model.Berikut adalah uraian lebih rinci mengenai hasil yang diperoleh dari masing-masing tahapan tersebut.

3.1 Pengumpulan Data

Pada penelitian ini digunakan dataset AIDS Virus Infection Prediction dari platform Kaggle yang berisi informasi medis dan demografis pasien dengan HIV untuk memprediksi risiko infeksi AIDS. Dataset tersebut terdiri dari 2.139 baris dan 23 kolom, dengan 22 kolom sebagai fitur dan 1 kolom sebagai target (Infected). Fitur-fitur di dalamnya mencakup variabel klinis seperti jumlah sel CD4 dan CD8, serta faktor demografis dan riwayat medis pasien yang menjadi dasar pembangunan model prediksi.

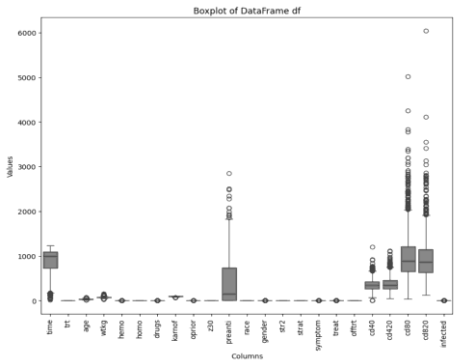
Cuplikan struktur dataset ditampilkan pada Tabel 2. Tabel tersebut menunjukkan beberapa variabel numerik seperti time, Age, dan Wtkg, serta variabel klinis seperti Cd40, Cd420, Cd80, dan Cd820. Kolom Infected berperan sebagai variabel target yang menunjukkan status infeksi (0 = tidak terinfeksi, 1 = terinfeksi). Cuplikan ini memberikan gambaran awal mengenai format data, tipe variabel, serta distribusi nilai yang akan diproses lebih lanjut pada tahap pra-pemrosesan.

Tabel 2. Cuplikan Dataset

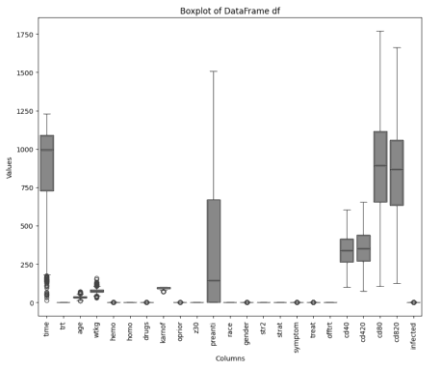
Nama Variabel	1	2	3	4	5
time	948	1002	961	1166	1090
Trt	2	3	3	3	0
Age	48	61	45	47	43
Wtkg	89	49	88	85	66
Hemo	0	0	0	0	0
…	…	…	…	…	…
Cd40	422	162	326	287	504
Cd420	477	218	274	394	353
Cd80	566	392	893	1590	870
Cd820	324	564	1893	966	782
Infected	0	1	0	0	0

3.2 Preprocessing Data

Pada tahap ini, penggunaan fungsi df.describe().T untuk menampilkan ringkasan statistik deskriptif dari kolom-kolom numerik mencakup metrik kunci seperti jumlah data (count), rata-rata (mean), standar deviasi (std), nilai minimum (min), kuartil (25%, 50%, 75%), dan nilai maksimum (max), yang memberikan gambaran cepat mengenai distribusi, pusat data, dan sebaran nilai untuk setiap fitur. Setelah itu, untuk memeriksa kelengkapan data, digunakan fungsi df.isnull().sum() untuk mengecek apakah ada nilai yang hilang (missing values) pada setiap kolom. Selain itu, dilakukan juga pemeriksaan terhadap data duplikat dengan menggunakan df.duplicated().sum(). Dalam analisis ini, tidak ditemukan missing values atau data duplikat, sehingga langkah ini tidak memerlukan penanganan lebih lanjut. Pada tahap selanjutnya, dilakukan deteksi outlier untuk mengidentifikasi nilai ekstrem yang dapat memengaruhi kinerja model. Berdasarkan visualisasi boxplot, beberapa kolom, seperti "time", "preanti", "cd40", "cd420", "cd80", dan "infected", mengandung outlier signifikan. Untuk mendeteksi outlier, digunakan fungsi detect_outlier_iqr_all_columns(df), yang mengaplikasikan metode Interquartile Range (IQR) untuk menghitung batas bawah dan atas pada setiap kolom numerik. Outlier yang terdeteksi kemudian diganti dengan nilai median kolom terkait. Penanganan outlier ini penting untuk menghindari pengaruhnya terhadap akurasi model dalam memprediksi infeksi AIDS.



Gambar 2. Hasil Outlier Sebelum Normalisasi

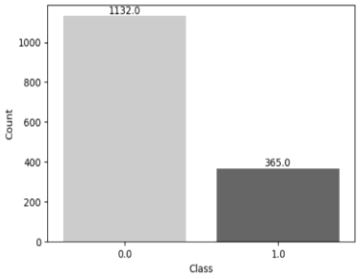


Gambar 3. Hasil Outlier Setelah Normalisasi

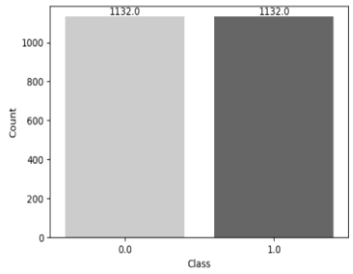
Boxplot pada Gambar 2 memperlihatkan distribusi setiap kolom setelah penanganan outlier pada variabel "preanti", "cd40", "cd420", dan "cd80" dengan mengganti nilai ekstrem menggunakan median, sehingga sebaran data menjadi lebih terpusat dan wajar. Tampilan awal dataset setelah penanganan outlier menunjukkan bahwa fitur-fitur

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

medis dan demografis seperti "time", "age", "wtkg", dan "hemo" masih memiliki perbedaan skala yang cukup besar, sehingga normalisasi atau standardisasi tetap diperlukan agar model machine learning dapat mempelajari pola data secara lebih optimal dan tidak bias terhadap fitur dengan rentang nilai lebih besar.Selanjutnya, hasil normalisasi menggunakan RobustScaler ditunjukan pada Gambar 3. Setelah penanganan outlier, data dinormalisasi dengan RobustScaler sehingga seluruh fitur berada pada skala yang lebih seragam dan terkontrol. RobustScaler mengurangi pengaruh nilai ekstrem dengan memanfaatkan median dan IQR, sehingga fitur seperti time, wtkg, preanti, age, dan cd40 menjadi lebih konsisten. Normalisasi ini membantu model machine learning mengolah data secara lebih efisien dan menghindari bias akibat perbedaan skala antar fitur.



Gambar 4. Data Sebelum Smote



Gambar 5. Data Setelah Smote

Tahap berikutnya adalah penerapan SMOTE untuk mengatasi ketidakseimbangan kelas. Pada Gambar 4, sebelum SMOTE, kelas 0 (tidak terinfeksi AIDS) memiliki 1.132 sampel dan kelas 1 (terinfeksi AIDS) hanya 365 sampel, sehingga model cenderung bias pada kelas mayoritas. Pada Gambar 5, setelah SMOTE, kedua kelas menjadi seimbang, masing-masing 1.132 sampel, sehingga model dapat mempelajari pola kedua kelas secara lebih adil dan meningkatkan akurasi deteksi infeksi AIDS.

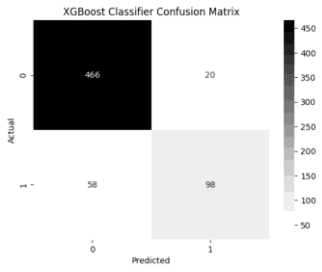
3.3 Evaluasi Model

Setelah data dinormalisasi menggunakan RobustScaler, langkah selanjutnya adalah melakukan pemodelan dengan berbagai algoritma machine learning untuk memprediksi risiko infeksi AIDS. Dengan data yang telah dinormalisasi, model-model seperti XGBoost, CatBoost, LightGBM, dan Random Forest diterapkan untuk membandingkan kinerja masing-masing dalam mendeteksi infeksi AIDS. Berikut adalah hasil evaluasi model yang mencakup metrik Akurasi, Precision, Recall, dan F1-Score untuk setiap model sebagaimana ditampilkan pada Tabel 3.

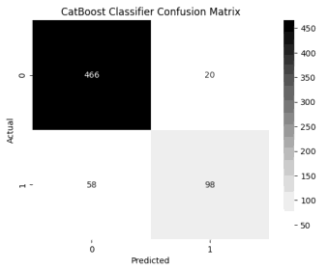
Tabel 3. Hasil Evaluasi Model Sebelum Smote

Model	Accuracy	Precision (Class 0.0)	Recall (Class 0.0)	F1-Score (Class 0.0)	Precision (Class 0.1)	Recall (Class 0.1)	F1-Score (Class 0.1)
XGBoost	87%	89%	94%	92%	79%	63%	70%
CatBoost	88%	89%	96%	92%	83%	63%	72%
LightGBM	88%	90%	96%	93%	84%	65%	73%
Random Forest	88%	89%	96%	92%	83%	63%	72%

Keempat algoritma menunjukkan akurasi yang relatif serupa, yaitu 87% untuk XGBoost dan 88% untuk CatBoost, LightGBM, serta Random Forest. Kinerja pada kelas 0 (tidak terinfeksi) sangat baik dengan precision 89–90% dan recall 94–96% pada seluruh model. Namun, performa pada kelas 1 (terinfeksi) masih terbatas, dengan precision 79–84% dan recall hanya 63–65%; LightGBM mencatat F1-Score tertinggi untuk kelas 1 (73%), disusul CatBoost dan Random Forest (72%) serta XGBoost (70%). Hal ini menunjukkan bahwa meskipun akurasi keseluruhan tinggi, kemampuan model dalam mendeteksi kelas minoritas masih perlu ditingkatkan, khususnya dari sisi recall.

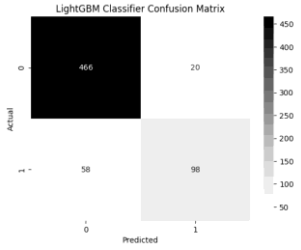


Gambar 6. Confusion Matriks XGBoost Sebelum Smote

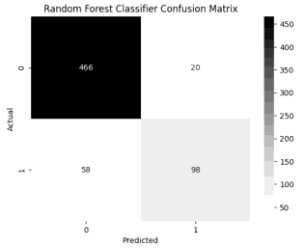


Gambar 7. Confusion Matriks CatBoost Sebelum Smote

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975



Gambar 8. Confusion Matriks LightGBM Sebelum Smote



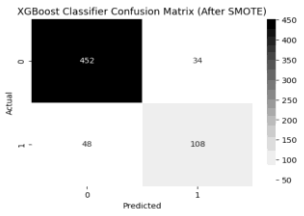
Gambar 9. Confusion Matriks Random Forest Sebelum Smote

Pada Gambar 6 menampilkan confusion matrix XGBoost sebelum penerapan SMOTE, Gambar 7 menampilkan confusion matrix CatBoost, Gambar 8 menampilkan confusion matrix LightGBM, dan Gambar 9 menampilkan confusion matrix Random Forest. Keempat confusion matrix untuk XGBoost, CatBoost, LightGBM, dan Random Forest menunjukkan pola yang sama, dengan 466 True Negatives (TN), 20 False Positives (FP), 98 True Positives (TP), dan 58 False Negatives (FN). Hasil ini mengindikasikan bahwa seluruh model sangat baik dalam mengklasifikasikan data tidak terinfeksi, tetapi masih menghadapi keterbatasan dalam mendeteksi kasus terinfeksi, yang tercermin dari jumlah FN yang relatif tinggi. Dengan demikian, meskipun performa terhadap kelas mayoritas kuat, kemampuan deteksi kelas minoritas (terinfeksi) masih perlu ditingkatkan.Setelah penerapan SMOTE (Synthetic Minority Over-sampling Technique) pada data pelatihan, distribusi kelas dalam dataset menjadi lebih seimbang, yang memungkinkan model untuk belajar lebih baik dari kedua kelas (terinfeksi dan tidak terinfeksi). Pada tahap ini, dilakukan evaluasi terhadap performa model XGBoost, CatBoost, LightGBM, dan Random Forest setelah SMOTE diterapkan, menggunakan metrik Akurasi, Precision, Recall, F1-Score, dan AUC-ROC. Berikut Tabel 4 hasil evaluasi masing-masing model setelah SMOTE diterapkan :

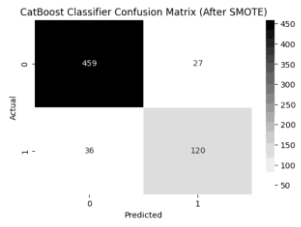
Tabel 4. Hasil Evaluasi Model Setelah Smote

Model	Accuracy	Precision (Class 0.0)	Recall (Class 0.0)	F1-Score (Class 0.0)	Precision (Class 0.1)	Recall (Class 0.1)	F1-Score (Class 0.1)
XGBoost	87%	90%	93%	92%	76%	69%	72%
CatBoost	90%	93%	94%	94%	82%	77%	79%
LightGBM	88%	91%	93%	92%	78%	72%	75%
Random Forest	89%	92%	93%	93%	77%	75%	76%

Setelah penerapan SMOTE, seluruh model menunjukkan peningkatan kemampuan dalam mendeteksi kelas positif (terinfeksi). XGBoost mencapai akurasi 87% dengan F1-Score kelas 1 sebesar 72%. CatBoost memberikan kinerja terbaik dengan akurasi 90% serta F1-Score kelas 1 sebesar 79%, disertai precision dan recall yang lebih seimbang antar kelas. LightGBM mencatat akurasi 88% dengan F1-Score kelas 1 sebesar 75%, sementara Random Forest mencapai akurasi 89% dengan F1-Score kelas 1 sebesar 76%. Secara keseluruhan, SMOTE berhasil memperbaiki performa deteksi kelas minoritas, khususnya pada CatBoost dan Random Forest. Keunggulan CatBoost dalam penelitian ini juga dipengaruhi oleh mekanisme internalnya yang menggunakan pendekatan Ordered Boosting, yang dirancang untuk mengurangi bias prediksi dan overfitting selama proses boosting. Selain itu, CatBoost memiliki kemampuan dalam menangani fitur kategorikal secara otomatis tanpa memerlukan proses encoding manual yang kompleks. Hal ini sangat relevan dengan karakteristik dataset medis yang mengandung banyak fitur biner dan kategorikal seperti Hemo, Homo, dan Drugs. Kemampuan tersebut memungkinkan CatBoost menangkap interaksi antar variabel risiko secara lebih efektif dibandingkan algoritma boosting lainnya, sehingga menghasilkan performa recall dan F1-Score yang lebih stabil pada kelas minoritas.



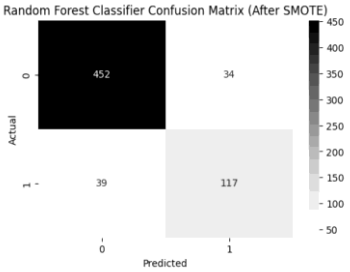
Gambar 10. Confusion Matriks XGBoost Setelah Smote



Gambar 11. Confusion Matriks CatBoost Setelah Smote

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

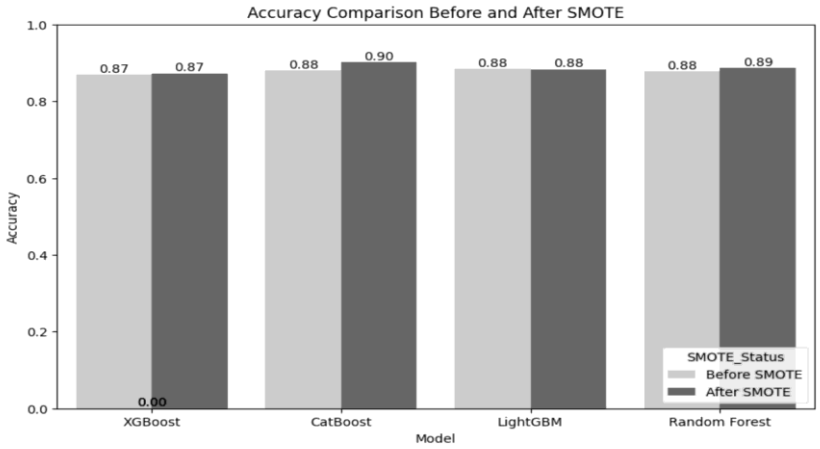
Gambar 12. Confusion Matriks LightGBM Setelah Smote



Gambar 13. Confusion Matriks Random Forest Setelah Smote

Untuk mengevaluasi kinerja model setelah penerapan SMOTE, ditampilkan confusion matrix pada Gambar 10, Gambar 11, Gambar 12, dan Gambar 13. Setelah penerapan SMOTE, keempat confusion matrix pada Gambar 10 (XGBoost), Gambar 11 (CatBoost), Gambar 12 (LightGBM), dan Gambar 13 (Random Forest) menunjukkan peningkatan deteksi kelas positif dibandingkan sebelum oversampling, dengan True Positives berkisar 108–120 dan False Negatives 36–48. XGBoost, CatBoost, LightGBM, dan Random Forest sama-sama mampu mempertahankan kemampuan yang baik dalam mengklasifikasikan kelas negatif (TN 452–459), namun masih terdapat kesalahan pada kelas positif (FP 27–34 dan FN yang relatif tinggi). Dalam konteks klasifikasi AIDS, Recall (sensitivitas) merupakan metrik yang sangat krusial karena False Negative berarti pasien terinfeksi diklasifikasikan sebagai tidak terinfeksi, yang berpotensi menunda penanganan medis. Oleh karena itu, peningkatan recall pada CatBoost dari 63% menjadi 77% setelah penerapan SMOTE merupakan kontribusi klinis yang lebih signifikan dibandingkan sekadar peningkatan akurasi dari 88% menjadi 90%, karena secara langsung mengurangi risiko kesalahan diagnosis pada pasien yang benar-benar terinfeksi.

3.4 Pembandingan Model



Gambar 14. Diagrma Sebelum dan Sesudah Smote

Pada Gambar 14 menunjukkan perbandingan akurasi model XGBoost, CatBoost, LightGBM, dan Random Forest sebelum dan setelah penerapan SMOTE (Synthetic Minority Over-sampling Technique). Perbandingan ini memberikan gambaran tentang bagaimana penerapan SMOTE mempengaruhi kinerja masing-masing model dalam mendeteksi infeksi AIDS, terutama dalam hal akurasi, precision, recall, dan F1-Score.

a. Pada model XGBoost, akurasi tetap stabil pada 0.87 sebelum dan setelah penerapan SMOTE. Precision untuk kelas 0 (tidak terinfeksi) adalah 90% sebelum dan 93% setelah SMOTE, sedangkan recall untuk kelas yang sama adalah 94% sebelum dan 93% setelah SMOTE. Untuk kelas 1 (terinfeksi), precision adalah 76% sebelum dan 69% setelah SMOTE, dengan recall 63% sebelum dan 72% setelah SMOTE. Penerapan SMOTE memberikan sedikit perubahan pada kinerja model, dengan peningkatan recall untuk kelas positif meskipun perubahan pada precision relatif kecil.
b. Pada model CatBoost, akurasi meningkat dari 0.88 menjadi 0.90 setelah SMOTE diterapkan. Precision untuk kelas 0 meningkat dari 93% menjadi 94%, sedangkan recall kelas 0 sedikit menurun dari 96% menjadi 94% setelah SMOTE. Untuk kelas 1, precision sedikit menurun dari 83% menjadi 82%, namun recall meningkat signifikan dari 63% menjadi 77%, dengan F1-Score yang meningkat dari 72% menjadi 79% setelah penerapan SMOTE. Hasil ini menunjukkan bahwa SMOTE secara efektif meningkatkan sensitivitas model dalam mendeteksi kelas positif (terinfeksi), meskipun terdapat penurunan kecil pada precision, yang secara keseluruhan menghasilkan peningkatan keseimbangan performa antar kelas.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

c. Pada model LightGBM, akurasi tetap stabil di 0.88 sebelum dan setelah SMOTE diterapkan. Precision untuk kelas 0 adalah 90% sebelum dan 91% setelah SMOTE, sementara recall untuk kelas yang sama adalah 96% sebelum dan 93% setelah SMOTE. Untuk kelas 1, precision adalah 84% sebelum dan 78% setelah SMOTE, dengan recall 65% sebelum dan 72% setelah SMOTE. Meskipun ada sedikit penurunan dalam precision untuk kelas positif setelah SMOTE, recall menunjukkan perbaikan yang cukup signifikan, mengindikasikan peningkatan deteksi untuk kelas 1.
d. Pada model Random Forest, akurasi meningkat sedikit dari 0.88 menjadi 0.89 setelah penerapan SMOTE. Precision untuk kelas 0 adalah 92% sebelum dan 93% setelah SMOTE, dengan recall yang meningkat dari 96% menjadi 93% setelah SMOTE. Untuk kelas 1, precision meningkat dari 83% menjadi 77% dan recall dari 63% menjadi 75% setelah SMOTE, dengan F1-Score yang sedikit meningkat dari 72% menjadi 76%. Penerapan SMOTE memberikan perbaikan pada recall kelas 1, meskipun masih ada penurunan kecil pada precision.

Tabel 5. Split Data Sebelum dan Sesudah Smote

Model	70:30 (Sebelum SMOTE)	80:20 (Sebelum SMOTE)	90:10 (Sebelum SMOTE)	70:30 (Setelah SMOTE)	80:20 (Setelah SMOTE)	90:10 (Setelah SMOTE)
XGBoost	86.92%	86.92%	87.23%	87.23%	88.63%	87.54%
CatBoost	88.01%	88.63%	88.32%	90.19%	89.10%	89.25%
LightGBM	88.47%	88.32%	87.85%	88.32%	88.16%	87.54%
Random Forest	87.85%	87.85%	88.32%	88.63%	87.85%	88.47%

Tabel 5 menyajikan perbandingan akurasi dari empat algoritma berbasis pohon keputusan, yaitu XGBoost, CatBoost, LightGBM, dan Random Forest, sebelum dan setelah penerapan SMOTE (Synthetic Minority Over-sampling Technique) pada berbagai rasio pembagian data (70:30, 80:20, dan 90:10). Analisis ini bertujuan untuk mengevaluasi pengaruh oversampling terhadap performa model dalam menghadapi ketidakseimbangan kelas pada dataset klasifikasi risiko AIDS. XGBoost menunjukkan peningkatan akurasi yang moderat setelah penerapan SMOTE, dengan akurasi tertinggi tercatat pada rasio 80:20 (88,63%). CatBoost mengalami peningkatan yang lebih signifikan, dari 88,01% menjadi 90,19% pada rasio 70:30, dan tetap tinggi pada rasio 90:10 (89,25%), menunjukkan respons yang kuat terhadap penambahan sampel sintetis. LightGBM relatif stabil, dengan perubahan akurasi yang kecil, meskipun metrik precision dan recall untuk kelas minoritas meningkat, menandakan kemampuan deteksi kasus minoritas yang lebih baik. Random Forest juga menunjukkan peningkatan akurasi, tertinggi pada rasio 70:30 (88,63%), dengan performa yang konsisten dan stabil. Secara keseluruhan, penerapan SMOTE terbukti meningkatkan kinerja model, terutama pada CatBoost dan XGBoost, sementara LightGBM dan Random Forest tetap mempertahankan performa yang solid. Hasil ini menekankan pentingnya penerapan strategi oversampling dalam menangani ketidakseimbangan kelas pada dataset medis, sehingga dapat meningkatkan akurasi prediksi risiko AIDS dan mendukung pengambilan keputusan yang lebih efektif di bidang kesehatan.

4. KESIMPULAN

Berdasarkan keseluruhan rangkaian eksperimen, penelitian ini menegaskan bahwa tantangan utama dalam klasifikasi risiko infeksi AIDS bukan terletak pada pencapaian akurasi tinggi, melainkan pada kemampuan model dalam mendeteksi kelas minoritas secara konsisten dan sensitif. Ketidakseimbangan distribusi kelas pada dataset awal menyebabkan bias terhadap kelas mayoritas, yang tercermin dari tingginya akurasi namun rendahnya recall pada kelas terinfeksi. Penerapan SMOTE terbukti secara signifikan memperbaiki kondisi tersebut dengan meningkatkan representasi kelas minoritas pada data pelatihan, sehingga model mampu mempelajari pola yang lebih seimbang. Perbaikan paling substansial terlihat pada model CatBoost, di mana recall kelas positif meningkat dari 63% menjadi 77% dan F1-Score dari 72% menjadi 79%. Peningkatan ini menunjukkan reduksi False Negatives yang bermakna, yang dalam konteks klinis memiliki implikasi langsung terhadap penurunan risiko keterlambatan diagnosis dan intervensi medis. Dengan demikian, keberhasilan SMOTE dalam penelitian ini tidak hanya tercermin pada peningkatan akurasi, tetapi terutama pada peningkatan sensitivitas model terhadap kasus infeksi yang sebelumnya sulit terdeteksi. Secara metodologis, keunggulan CatBoost dapat dikaitkan dengan mekanisme Ordered Boosting yang membantu mengurangi bias prediktif serta meningkatkan stabilitas model pada data tidak seimbang, serta kemampuannya dalam menangani fitur kategorikal secara intrinsik. Karakteristik ini membuat CatBoost lebih adaptif terhadap kompleksitas variabel klinis dan faktor risiko yang berinteraksi secara non-linear dibandingkan XGBoost, LightGBM, maupun Random Forest. Temuan ini menunjukkan bahwa kombinasi SMOTE dan CatBoost berpotensi dikembangkan sebagai sistem pendukung keputusan klinis untuk skrining awal dan deteksi dini kelompok berisiko tinggi dalam upaya pengendalian HIV/AIDS berbasis data. Secara keseluruhan, kontribusi penelitian ini terletak pada integrasi strategi pra-pemrosesan yang komprehensif dengan analisis komparatif algoritma berbasis pohon keputusan pada dataset HIV/AIDS yang tidak seimbang, serta penekanan pada peningkatan sensitivitas model sebagai indikator yang lebih relevan secara klinis dalam mendukung deteksi dini dan pengambilan keputusan kesehatan masyarakat.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2383−2393
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.8975

REFERENCES

[1] R. S. Gumarianto, S. Lardo, and A. Chairani, "Hubungan Antara Hitung Jumlah CD4 dengan Kejadian Wasting Syndrome pada Pasien HIV/AIDS di RSPAD Gatot Soebroto Periode Januari–Desember 2020," *Jurnal Kedokteran dan Kesehatan : Publikasi Ilmiah Fakultas Kedokteran Universitas Sriwijaya*, vol. 9, no. 2, pp. 133–142, May 2022, doi: 10.32539/jkk.v9i2.16975.

[2] Jocelyn *et al.*, "HIV/AIDS in Indonesia: current treatment landscape, future therapeutic horizons, and herbal approaches," *Front. Public Health*, vol. 12, Feb. 2024, doi: 10.3389/fpubh.2024.1298297.

[3] World Health Organization, "HIV Statistics, Globally and by WHO Region, 2025," Geneva, 2025. Accessed: Feb. 19, 2026. [Online]. Available: https://share.google/Xb6HLrQ2I2s8TkzZL

[4] Kemenkes RI, "Perkembangan HIV AIDS dan Penyakit Infeksi Menular Seksual (PIMS) Semester I Tahun 2025," Jun. 2025. Accessed: Feb. 19, 2026. [Online]. Available: https://share.google/sKhjSlrkuYg2l2ake

[5] A. Phuphuakrat, K. Khamnurak, S. Srichatrapimuk, and W. Wangsomboonsiri, "Missed opportunities for earlier diagnosis of HIV infection in people living with HIV in Thailand," *PLOS Global Public Health*, vol. 2, no. 7, Jul. 2022, doi: 10.1371/journal.pgph.0000842.

[6] X. Hu *et al.*, "Development and application of an early prediction model for risk of bloodstream infection based on real-world study," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, May 2025, doi: 10.1186/s12911-025-03020-9.

[7] A. Hidayani and C. Florency, "Faktor Risiko HIV pada Anak dengan Ibu Penderita HIV Positif Ditinjau dari Berbagai Literatur di Palangka Raya," *Jurnal Analis Kesehatan Klinikal Sains*, vol. 13, no. 1, Jun. 2025, doi: https://doi.org/10.36341/klinikalsains.v13i1.6345.

[8] World Health Organization, "HIV and AIDS," Jul. 2025. Accessed: Feb. 19, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[9] I. Y. Mauleti *et al.*, "Rapid Antiretroviral Therapy Initiation Reduces Mortality Among People Living With HIV in Indonesia: A Retrospective Observational Study," *Journal of Preventive Medicine and Public Health*, vol. 58, no. 4, pp. 360–369, Jan. 2025, doi: 10.3961/jpmph.24.622.

[10] D. Ayu Novita Prameswari, "Faktor Risiko yang Berhubungan dengan HIV/AIDS di Indonesia: Literature Review," *Jurnal Kesehatan Tambusai*, vol. 5, no. 3, Sep. 2024, doi: https://doi.org/10.31004/jkt.v5i3.31350.

[11] P. Pramita Izati, N. Aniniyah, and D. P. Isnawaty, "Comparison Between XGBoost, CatBoost, Random Forest, and LightGBM in Indonesian Women's Breast Cancer Dataset," *Parameter: Journal of Statistics*, vol. 5, no. 2, pp. 76–88, Dec. 2025, doi: 10.22487/27765660.2025.v5.i2.17658.

[12] M. Bahril Ilmi and Kusrini, "Perbandingan Kinerja Algoritma Machine Learning dalam Deteksi Potensi Risiko HIV," *Buffer Informatika*, vol. 11, no. 1, Apr. 2025, doi: https://doi.org/10.25134/buffer.v11i1.355.

[13] B. L. Ortiz *et al.*, "Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care," *JMIR Mhealth Uhealth*, vol. 12, Sep. 2024, doi: 10.2196/59587.

[14] J. M. H. Pinheiro *et al.*, "The Impact of Feature Scaling in Machine Learning: Effects on Regression and Classification Tasks," *IEEE Access*, vol. 13, pp. 199903–199931, Nov. 2025, doi: 10.1109/ACCESS.2025.3635541.

[15] A. Mizwar, A. Rahim, P. Hartato, A. Ridwan, and F. Asharudin, "Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 2, pp. 338–347, Apr. 2025, doi: https://doi.org/10.30871/jaic.v9i2.9125.

[16] Y. Li *et al.*, "The Predictive Accuracy of Machine Learning for the Risk of Death in HIV Patients: A Systematic Review and Meta-Analysis," *BMC Infect. Dis.*, vol. 24, no. 1, May 2024, doi: 10.1186/s12879-024-09368-z.

[17] M. F. H. Lamem, M. I. Sahid, and A. Ahmed, "Artificial intelligence for access to primary healthcare in rural settings," *Journal of Medicine, Surgery, and Public Health*, vol. 5, no. 2, Apr. 2025, doi: 10.1016/j.glmedi.2024.100173.

[18] S. Adhikari, "Importance of Data Preprocessing and Parameters Tuning for Supervised Machine Learning Models on Tweets Sentiment Analysis," *THE BATUK : A Peer Reviewed Journal of Interdisciplinary Studies*, vol. 10, no. 1, pp. 133–151, Jan. 2024, doi: 10.3126/batuk.v10i1.62303.

[19] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced Data Preprocessing Techniques for Machine Learning: A Systematic Mapping Study," *Knowl. Inf. Syst.*, vol. 65, no. 1, pp. 31–57, Jan. 2023, doi: 10.1007/s10115-022-01772-8.

[20] M. D. Teja and G. M. Rayalu, "Optimizing Heart Disease Diagnosis with Advanced Machine Learning Models: A Comparison of Predictive Performance," *BMC Cardiovasc. Disord.*, vol. 25, Mar. 2025, doi: 10.1186/s12872-025-04627-6.