Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.2, April 2025, pp. 338~347
e-ISSN: 2548-6861

Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy

Abd Mizwar A. Rahim [1*], Bambang Pilu Hartato [2*], Ahmad Ridwan [3*], Firman Asharudin[4]**
[*]Informatika, Universitas Amikom Yogyakarta
[**]Teknik Informatika, Universitas AMIKOM Yogyakarta
abdulmizwar@amikom.ac.id [1], ahmadridwan@amikom.ac.id [2], bambang.pilu@amikom.ac.id [3], firman_asharudin@amikom.ac.id [4]

Article Info

Article history:

Received 2025-01-29
Revised 2025-02-19
Accepted 2025-02-20

Keyword:

Machine Learning,
Feature Selection,
Data Balancing,
HIV/AIDS Prediction,
Classification.

ABSTRACT

HIV/AIDS remains a significant global health challenge, requiring accurate predictive models for early detection and improved clinical decision-making. However, developing an effective predictive model faces challenges such as data imbalance and the presence of irrelevant features, which can compromise model accuracy. This study aims to enhance the performance of AIDS infection prediction models by integrating feature selection, data balancing, and machine learning classification techniques. Feature selection is conducted using Pearson Correlation, Mutual Information, and Chi-Square tests to retain only the most relevant features. Random Oversampling, SMOTE, and ADASYN are employed to address data imbalance and improve model robustness. Nine machine learning algorithms, including Decision Tree, Random Forest, XGBoost, LightGBM, Gradient Boosting, Support Vector Machine, AdaBoost, and Logistic Regression, are tested for classification. Performance evaluation using confusion matrix, precision, recall, F1-score, and AUC-ROC shows that tree-based models (Random Forest, Extra Trees, and XGBoost) achieve the best results, particularly in handling minority class predictions. The study concludes that combining feature selection, data balancing, and machine learning techniques significantly improves predictive performance, making it a valuable approach for early detection and clinical decision support in HIV/AIDS diagnosis. Future research may explore hyperparameter tuning and real-world clinical data integration to enhance practical applicability.

I. PENDAHULUAN

Acquired Immunodeficiency Syndrome (AIDS), yang disebabkan oleh Human Immunodeficiency Virus (HIV), tetap menjadi salah satu tantangan kesehatan global yang mendesak [1]. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa pada tahun 2022, sekitar 39 juta orang di seluruh dunia hidup dengan HIV, dengan lebih dari 1,3 juta kasus baru per tahun. Tingginya tingkat morbiditas dan mortalitas terutama terjadi di negara-negara berkembang, di mana akses terhadap layanan kesehatan dan pengobatan masih terbatas [2]. Dalam konteks ini, pengembangan model machine learning untuk prediksi HIV/AIDS menjadi penting guna mendukung upaya pencegahan dan penanganan yang lebih efektif.[3].

Salah satu tantangan utama dalam membangun model prediksi HIV/AIDS adalah ketidakseimbangan data epidemiologi, di mana jumlah kasus positif (AIDS) jauh lebih sedikit dibandingkan dengan jumlah kasus negatif [4]. Ketidakseimbangan data ini dapat menghambat kinerja model machine learning, menyebabkan bias terhadap kelas mayoritas, dan mengurangi akurasi prediksi pada kelas minoritas [5]. Oleh karena itu, diperlukan strategi penyeimbangan data, seperti Random Over Sampling dan Synthetic Minority Oversampling Technique (SMOTE), untuk meningkatkan representasi data minoritas serta memperbaiki distribusi kelas [6].

Beberapa penelitian sebelumnya telah mengeksplorasi metode prediksi HIV/AIDS menggunakan machine learning. Salah satunya menyelidiki penggunaan oversampling seperti

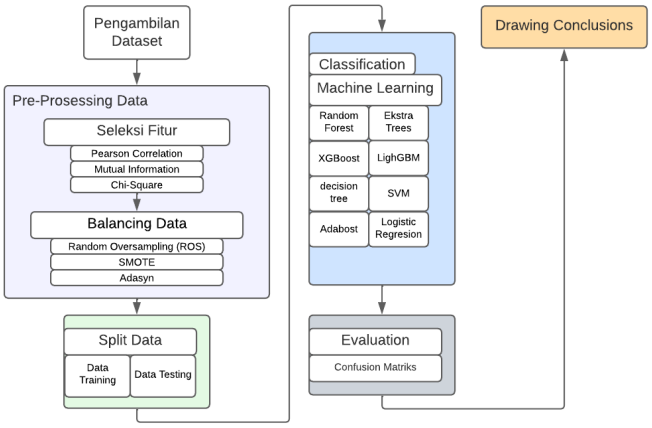
SMOTE, ADASYN, dan Random Oversampling yang dikombinasikan dengan algoritma XGBoost, menghasilkan akurasi tertinggi sebesar 94,44%, presisi 90,72%, recall 98,74%, dan skor F1 sebesar 94,65%. Penelitian lain menggunakan Random Forest untuk analisis prediksi HIV, mencapai akurasi 92,86% dengan aplikasi RapidMiner [7] [8]. Selain itu, studi yang membandingkan kinerja algoritma machine learning (SVM, Random Forest, Naive Bayes) dan deep learning (LSTM, GRU) menunjukkan bahwa model LSTM mencapai akurasi tertinggi sebesar 97,65%. Studi lainnya berfokus pada penerapan machine learning untuk memprediksi risiko infeksi HIV di kelompok berisiko tinggi, seperti pria yang berhubungan seks dengan sesama jenis, menggunakan algoritma Decision Tree, SVM, dan Random Forest. Teknik SMOTE diterapkan untuk menangani ketidakseimbangan data, dan hasilnya menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 87,1%, presisi 96,0%, recall 77,5%, dan AUC sebesar 0,942. [9][10].

Berdasarkan hasil penelitian sebelumnya, masih terdapat kebutuhan untuk mengevaluasi berbagai algoritma machine learning dalam menangani ketidakseimbangan data dan meningkatkan akurasi prediksi HIV/AIDS. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi kinerja sembilan metode machine learning, mengoptimalkan pemilihan fitur melalui analisis korelasi menggunakan Heatmap, serta menerapkan strategi balancing data untuk mengatasi ketidakseimbangan kelas. Pendekatan ini diharapkan dapat menghasilkan model prediksi yang lebih andal dan efisien, sehingga dapat mendukung pengambilan keputusan dalam bidang kesehatan secara lebih tepat.

Penelitian ini berfokus pada pengembangan model prediksi komprehensif untuk HIV/AIDS dengan pendekatan berbasis machine learning. Kami mengevaluasi kinerja berbagai algoritma, menerapkan teknik seleksi fitur, dan mengatasi ketidakseimbangan data guna menemukan metode terbaik yang dapat meningkatkan akurasi prediksi. Hasil penelitian ini diharapkan dapat berkontribusi pada peningkatan analisis data kesehatan serta pengambilan keputusan yang lebih baik dalam upaya pencegahan dan penanganan infeksi HIV/AIDS.

II. METODE

Penelitian ini terdiri dari beberapa tahapan penting yang dilakukan secara sistematis. Tahap pertama dimulai dengan pengumpulan data, berikutnya proses seleksi fitur, di mana fitur-fitur yang tidak memiliki hubungan signifikan dengan target variabel diidentifikasi dan dihapus untuk meningkatkan efisiensi dan akurasi model. Setelah itu, dilakukan teknik balancing data menggunakan metode Random Oversampling guna mengatasi ketidakseimbangan pada kelas target. Selanjutnya, data yang telah diproses melalui tahap ini dibagi menjadi data latih dan data uji (split data) untuk keperluan pelatihan dan pengujian model. Proses klasifikasi kemudian diterapkan dengan memanfaatkan berbagai algoritma machine learning. Akhirnya, kinerja model dievaluasi menggunakan confusion matrix, yang memungkinkan analisis mendalam terhadap metrik akurasi guna menentukan keandalan model dalam memprediksi infeksi AIDS. Keseluruhan tahapan penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Tahapan Penelitian.

1) Pengambilan Dataset.

Dataset yang digunakan berasal dari Kaggle yaitu AIDS Virus Infection Prediction milik Aadarsh velu[11], dataset ini berisi data-data pasien yang menderita penyakit AIDS yang berjumlah 2139 baris dan terdiri dari 23 fitur. Label dari dataset yaitu pada variable infected yang memiliki 2 nilai yaitu 0 (tidak terinfeksi Aids) sebanyak 1618 data dan 1 (terinfeksi Aids) sebanyak 521 data.

Berikut adalah informasi detail dari dataset, yang dapat dilihat pada table 1.

TABEL I
DATASET

No	Fitur	Deskripsi	*Kategori/Nilai*
1	time	Merepresentasikan waktu *failure* atau *censoring* pada pasien.	-
2	trt	Kategori tipe perawatan pasien	0 = ZDV, 1 = ZDV + ddl, 2 = ZDV + Zal, 3 = ddl
3	age	Umur pasien saat awal perawatan.	-
4	wtkg	Berat badan pasien saat awal perawatan (dalam kilogram).	0 = no, 1 = yes
5	hemo	Status hemofilia pasien.	0 = no, 1 = yes
6	homo	Homoseksual activity	0 = no, 1 = yes
7	drugs	History of IV drug use	0 = no, 1 = yes
8	karnof	Karnofsky score	on a scale of 0-100
9	oprior	Non-ZDV antiretrovial therapy pre-175	0 = no, 1 = yes
10	z30	ZDV in the 30 days prior to 175	0 = no, 1 = yes

11	preanti	Days pre-175 antri retroviral therapy	-
12	race	race	0=white, 1=non-white
13	gender	gender	0=F, 1=M
14	str2	antiretroviral history	0=naive, 1=experienced
15	strat	antiretroviral history stratification	1='Antiretroviral Naive',2='> 1 but <= 52 weeks of prior antiretroviral therapy',3='> 52 weeks
16	symptom	symptomatic indicator	0=asymp, 1=symp
17	treat	treatment indicator	0=ZDV only, 1=others
18	offtrt	indicator of off-trt before 96+/-5 weeks	0=no,1=yes
19	cd40	CD4 at baseline	-
20	cd420	CD4 at 20+/-5 weeks	-
21	cd80	CD8 at baseline	-
22	cd820	CD8 at 20+/-5 weeks	-
23	infected	is infected with AIDS	0=No, 1=Yes

Tabel 1 menjelaskan berbagai fitur yang terdapat dalam dataset yang digunakan untuk penelitian prediksi infeksi AIDS. Dataset ini mencakup 23 fitur utama yang dikelompokkan ke dalam beberapa kategori penting. Informasi demografi dan data dasar mencakup variabel seperti waktu (time) yang merepresentasikan durasi pengamatan, umur pasien saat awal pengobatan (age), berat badan awal pasien (wtkg), jenis kelamin (gender), ras pasien (race), serta aktivitas seksual pasien (homo). Riwayat medis mencakup status hemofilia (hemo), riwayat penggunaan obat intravena (drugs), skor Karnofsky yang mengukur tingkat kesehatan pasien secara keseluruhan (karnof), dan riwayat terapi antiretroviral sebelum masa pengamatan (oprior, str2, strat). Informasi pengobatan meliputi jenis perawatan yang diterima pasien (trt), durasi terapi antiretroviral sebelumnya (preanti), indikator penghentian pengobatan (offtrt), serta jenis terapi yang saat ini dijalani (treat). Pada kategori hasil laboratorium, fitur mencakup data jumlah CD4 dan CD8 pasien yang dicatat pada awal perawatan (cd40, cd80) serta setelah 20 minggu pengobatan (cd420, cd820). Selain itu, dataset ini juga merekam kondisi pasien melalui indikator gejala klinis (symptom) dan status infeksi AIDS pasien (infected). Seluruh fitur ini secara komprehensif menggambarkan kondisi demografi, riwayat pengobatan, dan hasil laboratorium pasien, sehingga mendukung pengembangan model prediksi yang lebih akurat untuk infeksi AIDS.

2) Seleksi fitur

Pada tahapan ini kami menerapkan beberapa metode seleksi fitur, diantaranya Pearson Correlation, Mutual Information, dan Chi-Square.

Seleksi fitur menggunakan koefisien korelasi Pearson, metode statistic ini diterapkan untuk mengidentifikasi fitur yang memiliki hubungan linier signifikan dengan target variabel[12]. Koefisien Pearson mengukur keterkaitan antara dua variabel numerik, dengan nilai berkisar dari -1 hingga 1, di mana +1 menunjukkan korelasi positif sempurna, -1 korelasi negatif sempurna, dan 0 tidak ada hubungan linier[13]. Dalam proses ini, fitur-fitur yang nilai korelasinya rendah atau mendekati nol dihapus karena dianggap kurang relevan, sementara fitur dengan korelasi tinggi dipertahankan untuk analisis lebih lanjut. Mutual Information (MI) diterapkan pada fitur kategorikal untuk menangkap hubungan non-linear dengan target. MI mengukur ketergantungan antara variabel tanpa mengasumsikan hubungan linear, sehingga efektif untuk data diskrit[14]. Berikutnya dengan metode Chi-Square digunakan untuk fitur kategorikal dengan mengevaluasi apakah distribusi nilai fitur berbeda secara signifikan berdasarkan kelas target. Fitur dengan nilai Chi-Square rendah menunjukkan kurangnya hubungan dengan target dan dapat dihapus. Dengan menggabungkan ketiga metode ini, seleksi fitur menjadi lebih optimal untuk menangani dataset dengan kombinasi variabel numerik dan kategorikal[15].

3) Balancing Data.

Ketidakseimbangan data (imbalanced data) merupakan masalah umum dalam klasifikasi, di mana jumlah sampel dalam satu kelas jauh lebih sedikit dibandingkan kelas lainnya. Dalam penelitian ini, dilakukan balancing data untuk memastikan model machine learning tidak bias terhadap kelas mayoritas dan dapat mendeteksi kelas minoritas dengan lebih baik. Untuk menangani ketidakseimbangan data, digunakan tiga teknik utama, yaitu Random Oversampling, SMOTE (Synthetic Minority Over-sampling Technique), dan ADASYN (Adaptive Synthetic Sampling).

Random oversampling, metode ini merupakan metode balancing data yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset [16]. Teknik ini bekerja dengan menambahkan salinan data dari kelas minoritas secara acak hingga jumlahnya setara dengan kelas mayoritas. Metode ini bertujuan untuk meningkatkan representasi kelas minoritas agar model machine learning tidak bias terhadap kelas mayoritas. Meskipun sederhana dan efektif, random oversampling dapat meningkatkan risiko overfitting karena duplikasi data. Oleh karena itu, metode ini sering digunakan bersama teknik lain atau validasi model untuk memastikan kinerja prediksi yang optimal [17].

SMOTE (Synthetic Minority Over-sampling Technique) adalah teknik oversampling yang digunakan untuk menyeimbangkan dataset dengan membuat data sintetis berdasarkan interpolasi antar sampel minoritas. SMOTE menciptakan sampel baru dengan menambahkan titik-titik data di antara sampel yang berdekatan. Prosesnya dimulai dengan mengidentifikasi kelas minoritas, kemudian menentukan k-nearest neighbors (KNN) dari setiap sampel minoritas [18]. Selanjutnya, sistem memilih salah satu tetangga terdekat secara acak dan menghasilkan sampel sintetis dengan menginterpolasi nilai fitur antara dua titik

data. Data sintetis yang dihasilkan kemudian ditambahkan ke dataset hingga jumlah sampel minoritas mencapai keseimbangan yang diinginkan [19].

ADASYN adalah pengembangan dari SMOTE yang juga menghasilkan data sintetis, tetapi dengan pendekatan yang lebih adaptif. ADASYN menambahkan lebih banyak sampel sintetis pada data yang sulit diklasifikasikan, terutama di sekitar batas keputusan (decision boundary) antara kelas mayoritas dan minoritas[20]. Prosesnya dimulai dengan mengidentifikasi kelas minoritas, lalu menghitung kepadatan data menggunakan k-nearest neighbors (KNN) untuk mengetahui area yang sulit diprediksi. Setelah itu, ADASYN menentukan bobot adaptif untuk setiap sampel, di mana sampel yang lebih dekat dengan kelas mayoritas akan mendapatkan lebih banyak sampel sintetis dibandingkan sampel yang jauh dari batas keputusan[21].

4) Split Data.

Pembagian dataset penelitian ini menjadi dua bagian, yaitu 80% untuk data pelatihan (training data) dan 20% untuk data pengujian (testing data). Data pelatihan digunakan untuk melatih model machine learning agar dapat mempelajari pola dari dataset. Sementara itu, data pengujian digunakan untuk mengevaluasi performa model terhadap data yang belum pernah dilihat sebelumnya, guna mengukur akurasi dan generalisasi model[22]. Proporsi 80:20 dipilih karena memberikan jumlah data pelatihan yang cukup besar sambil menyisakan data pengujian yang memadai untuk evaluasi. Table 2 menggambarkan split data yang dilakukan.

TABEL II
SPLIT DATA

Keterangan	Data Training	Data Testing	Jumlah
Dengan ROS			
Proporsi	80%	20%	100%
Jumlah	2588	648	3236
Dengan Smote			
Proporsi	80%	20%	100%
Jumlah	2588	648	3236
Dengan Adasyn			
Proporsi	80%	20%	100%
Jumlah	2532	633	3165

Tabel II menjelaskan pembagian data (split data) setelah dilakukan balancing menggunakan tiga teknik berbeda, yaitu Random Oversampling (ROS), SMOTE, dan ADASYN. Untuk metode ROS dan SMOTE, jumlah total data setelah balancing adalah 3.236 sampel, dengan pembagian 80% data training (2.588 sampel) dan 20% data testing (648 sampel). Sementara itu, metode ADASYN menghasilkan total 3.165 sampel, dengan 2.532 sampel untuk training dan 633 sampel untuk testing. Perbedaan jumlah total data ini terjadi karena ADASYN secara adaptif menambah sampel sintetis hanya pada daerah yang sulit diklasifikasikan, sehingga jumlah sampel hasil balancing tidak selalu sama dengan metode lainnya. Meskipun jumlah data sedikit berbeda, pembagian antara training dan testing tetap konsisten dengan proporsi 80:20 untuk menjaga keseimbangan dalam pelatihan model.

5) Klasifikasi dengan metode machine learning.

Penelitian ini melakukan klasifikasi menggunakan beberapa metode machine learning diantaranya Random Fores(RF), Ekstra Trees(ET), Xgboost(XGB), Lighgbm(LG), Decision Tree(DT), Gradient Boosting(GB), Support Vektor Machine(SVM), adabost(AB), dan Logistic Regression(LR). Metode machine learning adalah proses pengelompokan data ke dalam kategori atau kelas tertentu berdasarkan pola yang dipelajari dari data latih. Dalam klasifikasi, algoritma machine learning mempelajari hubungan antara fitur-fitur independen (variabel input) dan target (variabel output) untuk membangun model prediksi[23]. Proses ini mencakup prapemrosesan data dan pelatihan model menggunakan data latih untuk mengenali pola dan hubungan antar variabel[24].

6) Evaluasi dengan Confusion Matriks.

Evaluasi model penelitian ini menggunakan confusion matrix, untuk mengukur kinerja model klasifikasi dengan membandingkan hasil prediksi model terhadap data sebenarnya. Matriks ini terdiri dari empat komponen utama yaitu True Positive (TP) Kasus positif HIV/AIDS yang diprediksi benar, True Negative (TN) Kasus negatif yang diprediksi benar, False Positive (FP) Kasus negatif yang salah diklasifikasikan sebagai positif, dan False Negative (FN) Kasus positif yang salah diklasifikasikan sebagai negatif[25]. Dari confusion matrix, berbagai metrik evaluasi seperti akurasi, presisi, recall, dan F1-score dapat dihitung untuk memberikan gambaran menyeluruh tentang kemampuan model dalam melakukan prediksi, namun pada penelitian ini melakukan perbandingan metode klasifikasi, hanya pada metrik evaluasi akurasi[26].

III. HASIL DAN PEMBAHASAN

Bagian ini kami akan menyampaikan temuan dari penelitian serta analisis terhadap hasil yang diperoleh, mulai dari tahapan pertama (Pengambilan dataset) hingga evaluasi perbandingan beberapa model machine learning. Analisis dilakukan untuk mengidentifikasi keunggulan dan kelemahan masing-masing model serta faktor yang memengaruhi performanya. Hasil ini kemudian dikaitkan dengan penelitian sebelumnya untuk melihat kesesuaian atau perbedaan temuan, sehingga dapat memberikan wawasan lebih lanjut mengenai efektivitas metode yang digunakan.

1) Pengambilan Dataset

Dataset yang digunakan dalam penelitian ini memiliki 23 fitur dataset, 1 fitur dataset yaitu infected dinilai sebagai class dataset. Fitur lainnya dinilai sebagai variable yang mempengaruhi class dataset dimulai dari time hingga cd820, berikut ini tampilan dataset yang dapat dilihat pada table 3.

TABEL III
NILAI SETIAP FITUR DATASET

No	Fitur	*Nilai*	*Nilai*	*Nilai*
Data Baris Ke-1			Data Baris Ke-2	Data Baris Ke-3
1	time	948	1002	961
2	trt	2	3	3
3	age	48	61	45
4	wtkg	90	49	88
5	hemo	0	0	0
6	homo	0	0	1
7	drugs	0	0	1
8	karnof	100	90	90
9	oprior	0	0	0
10	z30	0	1	1
11	preanti	0	895	707
12	race	0	0	0
13	gender	0	0	1
14	str2	0	1	1
15	strat	1	3	3
16	symptom	0	0	0
17	treat	1	1	1
18	offtrt	0	0	1
19	cd40	422	162	326
20	cd420	477	218	274
21	cd80	566	392	2063
22	cd820	324	564	1893
23	infected	0	1	0

Tabel 3 menyajikan nilai dari setiap fitur dalam dataset berdasarkan tiga baris pertama. Data ini mencakup berbagai informasi pasien, seperti waktu pengamatan (time), jenis perawatan (trt), usia (age), berat badan (wtkg), serta status hemofilia (hemo), homoseksualitas (homo), dan riwayat penggunaan obat (drugs). Selain itu, tabel ini juga menampilkan skor kesehatan pasien (karnof), riwayat terapi sebelumnya (oprior, preanti), serta indikator ras (race) dan gender (gender). Fitur lain yang dicantumkan mencakup riwayat pengobatan antiretroviral (str2, strat), kondisi simptomatik (symptom), dan jenis pengobatan saat ini (treat). Hasil laboratorium seperti jumlah CD4 dan CD8 pada berbagai periode juga disertakan (cd40, cd420, cd80, cd820). Variabel terakhir, infected, menunjukkan status infeksi AIDS, dengan nilai 0 untuk tidak terinfeksi dan 1 untuk terinfeksi. Informasi dalam tabel ini memberikan gambaran awal tentang variasi nilai dalam dataset yang digunakan untuk analisis prediksi infeksi AIDS.

2) Seleksi Fitur

Proses seleksi fitur dalam penelitian ini menggunakan tiga metode utama, yaitu Pearson Correlation untuk fitur numerik, Mutual Information untuk fitur kategorikal, dan Chi-Square untuk fitur kategorikal.

Hasil seleksi menggunakan Pearson Correlation menunjukkan bahwa fitur time, cd40, cd420, preanti, dan karnof memiliki korelasi yang cukup kuat dengan variabel target (infected). Fitur-fitur ini dipertahankan karena memiliki hubungan linear yang signifikan dengan probabilitas infeksi.

Untuk menangkap hubungan non-linear, dilakukan seleksi fitur menggunakan Mutual Information (MI), yang mengukur ketergantungan antara variabel tanpa mengasumsikan hubungan linear. Hasilnya menunjukkan bahwa fitur trt, symptom, z30, treat, dan strat memiliki nilai Mutual Information yang cukup tinggi, sehingga dianggap berkontribusi dalam prediksi infeksi AIDS.

Selain itu, seleksi fitur menggunakan Chi-Square diterapkan pada fitur kategorikal untuk mengukur hubungan statistik antara variabel dengan target. Hasil analisis menunjukkan bahwa fitur offtrt, symptom, trt, strat, z30, dan str2 memiliki nilai Chi-Square yang tinggi, menunjukkan bahwa distribusi kategori dalam fitur ini berpengaruh terhadap status infeksi.

Berdasarkan hasil keseluruhan seleksi fitur dari ketiga metode ini, diperoleh 12 fitur utama yang paling berpengaruh terhadap prediksi infeksi AIDS, yaitu time, cd40, cd420, preanti, karnof, trt, symptom, offtrt, strat, z30, str2, dan treat. Hasil keseluruhan dari proses seleksi fitur ini dapat dilihat pada table 4.

TABEL IV
HASIL SELEKSI FITUR

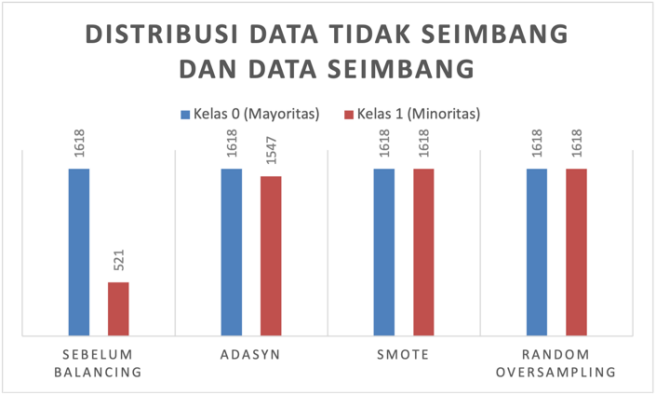
No	Feature	Pearson_ Correlation	Mutual_ Information	Chi_ Square
1	offtrt	0.000000	0.000000	11.7506
2	symptom	0.000000	0.009255	29.5446
3	trt	0.000000	0.024447	12.7273
4	time	0.574989	0.000000	0.00000
5	cd40	0.185647	0.000000	0.00000
6	strat	0.000000	0.005131	30.4692
7	preanti	0.128453	0.000000	0.00000
8	karnof	0.102944	0.000000	0.00000
9	z30	0.000000	0.009952	15.1489
10	str2	0.000000	0.000000	13.4894
11	treat	0.000000	0.008940	8.92998
12	cd420	0.345908	0.000000	0.00000

Tabel 4 menunjukkan hasil seleksi fitur berdasarkan tiga metode utama, yaitu Pearson Correlation, Mutual Information, dan Chi-Square, untuk menentukan fitur yang berpengaruh terhadap prediksi infeksi AIDS. Fitur numerik seperti time (0.5749), cd420 (0.3459), cd40 (0.1856), dan preanti (0.1284) memiliki nilai Pearson Correlation yang cukup tinggi, menunjukkan adanya hubungan linear yang signifikan dengan target variabel, di mana fitur dengan korelasi di atas 0.1 dianggap memiliki kontribusi penting terhadap prediksi. Sementara itu, fitur kategorikal seperti trt (0.0244), symptom (0.0092), z30 (0.0099), treat (0.0089), dan strat (0.0051) memiliki nilai Mutual Information yang menunjukkan adanya hubungan non-linear dengan target, yang berarti fitur ini tetap memiliki informasi penting meskipun tidak memiliki hubungan linear langsung. Selain itu, fitur seperti strat (30.4692), symptom (29.5446), z30 (15.1489), trt (12.7273), offtrt (11.7506), dan str2 (13.4894) memiliki nilai Chi-Square yang tinggi, menunjukkan bahwa

distribusi kategori dalam fitur tersebut memiliki perbedaan yang signifikan terhadap status infeksi, dengan nilai di atas 5 dianggap sebagai indikator yang kuat.

3) Balancing Data.

Dataset ini awalnya mengalami ketidakseimbangan dataset, dimana terdapat class mayoritas (class 0) dan minoritas (class 1) pada dataset penelitian. Barikut ini hasil data setelah diseimbangkan dengan tiga Teknik balancing yaitu Random Oversampling, SMOTE, dan Adasyn yang dapat dilihat pada gambar 2.



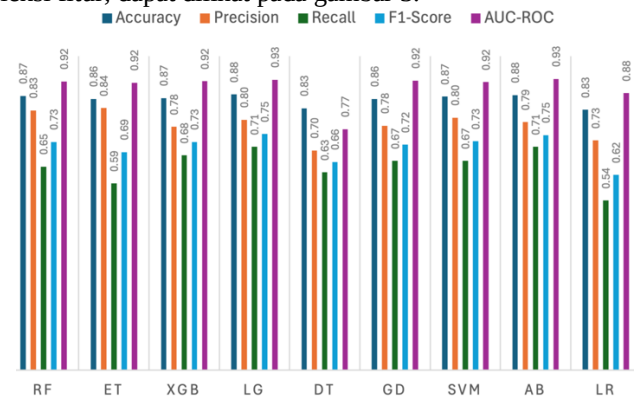
Gambar 2. Data seimbang dan tidak seimbang.

Gambar 2 menampilkan perbandingan distribusi data sebelum dan sesudah balancing menggunakan tiga teknik: ADASYN, SMOTE, dan Random Oversampling. Sebelum dilakukan balancing, dataset mengalami ketidakseimbangan kelas, di mana kelas mayoritas (0) memiliki 1.618 sampel, sedangkan kelas minoritas (1) hanya 521 sampel. Setelah dilakukan balancing, teknik ADASYN meningkatkan jumlah kelas minoritas menjadi 1.547 sampel, sementara metode SMOTE dan Random Oversampling membuat jumlah kelas minoritas dan mayoritas menjadi seimbang, masing-masing sebanyak 1.618 sampel. Hal ini menunjukkan bahwa SMOTE dan Random Oversampling menyeimbangkan kelas secara penuh, sementara ADASYN lebih adaptif dengan menambahkan jumlah sampel sintetis berdasarkan tingkat kesulitan klasifikasi.

4) Klasifikasi dengan metode machine learning.

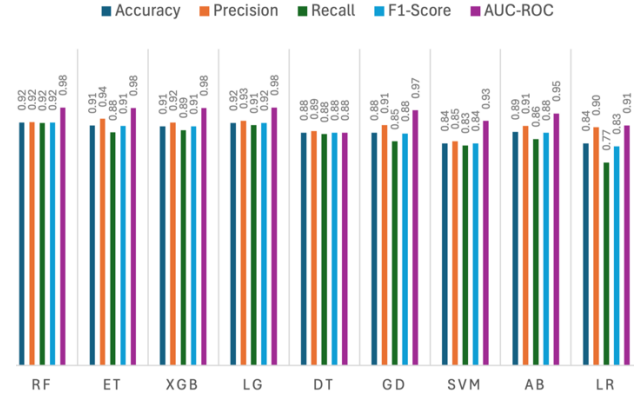
Tahap ini melakukan klasifikasi dengan sembilan metode machine learning yaitu Random Fores, Ekstra Trees, Xgboost, Lighgbm, Decision Tree, Gradient Boosting, Support Vektor Machine, adabost, dan Logistic Regression. Proses klasifikasi ini pada empat kondisi dataset yang berbeda, pertama pada kondisi dataset tidak seimbang, kedua pada dataset seimbang dengan hasil Teknik smote, ketiga dengan hasil teknik adasyn, dan ke empat dengan hasil teknik random oversampling. Klasifikasi ini dilakukan tanpa menerapkan Hyperparameter Tuning, melainkan hanya menguji berbagai metode machine learning menggunakan pengaturan parameter default pada masing-masing model.

Berikut ini hasil keseluruhan model dalam melakukan klasifikasi pada dataset tidak seimbang yang telah dilakukan seleksi fitur, dapat dilihat pada gambar 3.



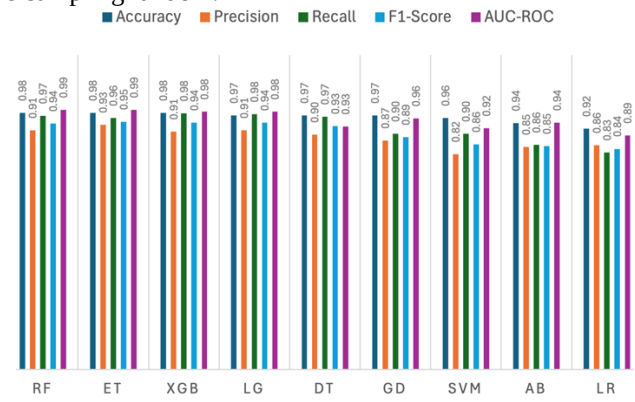
Gambar 3. Hasil Klasifikasi model machine learning pada data tidak seimbang.

Selanjutnya hasil keseluruhan model dalam melakukan klasifikasi pada dataset seimbang dengan smote, dapat dilihat pada gambar 4.



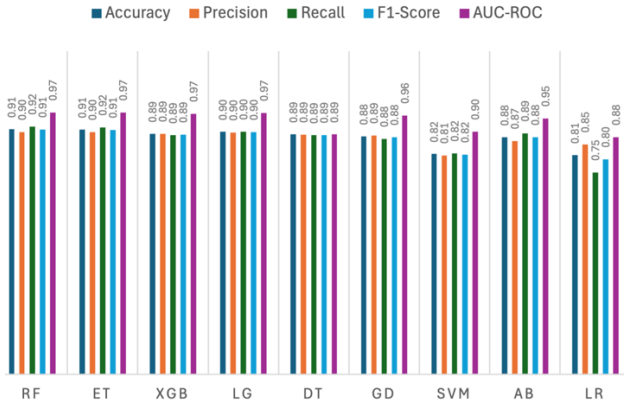
Gambar 4. Hasil Klasifikasi model machine learning pada data seimbang dengan smote.

Gambar 5 menunjukkan hasil keseluruhan model saat melakukan klasifikasi pada dataset yang seimbang dari oversampling random.



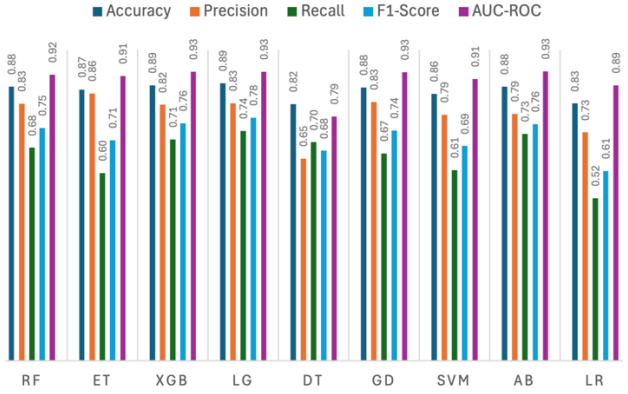
Gambar 5. Hasil Klasifikasi model machine learning pada data seimbang dengan random oversampling.

Gambar 6 menunjukkan hasil keseluruhan model dalam melakukan klasifikasi pada dataset seimbang dari adasyn.



Gambar 6. Hasil Klasifikasi model machine learning pada data seimbang dari Adasyn.

Terakhir, hasil keseluruhan dari model klasifikasi pada dataset tidak seimbang dan tanpa diterapkannya seleksi fitur, dapat dilihat pada gambar 7.



Gambar 7. Hasil Klasifikasi model machine learning pada data tidak seimbang dan tanpa seleksi fitur.

Gambar 7 menunjukkan hasil klasifikasi tanpa balancing data, di mana model machine learning mengalami bias terhadap kelas mayoritas. Model dengan performa terbaik berdasarkan AUC-ROC dan F1-Score adalah Random Forest (AUC-ROC: 0.91, F1-Score: 0.72), Extra Trees (AUC-ROC: 0.91, F1-Score: 0.69), dan XGBoost (AUC-ROC: 0.92, F1-Score: 0.72). Namun, model seperti Support Vector Machine (AUC-ROC: 0.91, F1-Score: 0.72) dan Logistic Regression (AUC-ROC: 0.88, F1-Score: 0.62) memiliki performa lebih rendah, terutama dalam Recall, yang menunjukkan kesulitan dalam mengenali kelas minoritas. Secara keseluruhan, tanpa balancing data, model cenderung memiliki nilai Recall yang lebih rendah, yang berarti model lebih sulit dalam mendeteksi kasus dari kelas minoritas. Akurasi model cukup tinggi, tetapi kurang mencerminkan kemampuan model dalam mengklasifikasikan seluruh data dengan baik karena ketidakseimbangan kelas.

Gambar 4 menunjukkan hasil klasifikasi setelah menerapkan SMOTE (Synthetic Minority Oversampling Technique) untuk menangani ketidakseimbangan data. Dengan metode ini, nilai Recall dan F1-Score meningkat secara signifikan, menunjukkan bahwa model lebih baik dalam mengenali kelas minoritas dibandingkan dengan data tanpa balancing. Model dengan performa terbaik berdasarkan AUC-ROC dan F1-Score adalah Random Forest (AUC-ROC: 0.97, F1-Score: 0.92), Extra Trees (AUC-ROC: 0.97, F1-Score: 0.90), XGBoost (AUC-ROC: 0.97, F1-Score: 0.90), dan LightGBM (AUC-ROC: 0.97, F1-Score: 0.92).

Sementara itu, model dengan performa lebih rendah adalah Support Vector Machine (AUC-ROC: 0.92, F1-Score: 0.84) dan Logistic Regression (AUC-ROC: 0.91, F1-Score: 0.83), yang masih menunjukkan kesulitan dalam menangani kelas minoritas meskipun mengalami peningkatan dibandingkan sebelumnya. SMOTE terbukti meningkatkan keseimbangan klasifikasi dengan meningkatkan Recall dan F1-Score tanpa menurunkan akurasi secara signifikan. Oleh karena itu, metode ini sangat efektif dalam mengatasi ketidakseimbangan data dan meningkatkan kinerja model.

Gambar 5 menunjukkan hasil klasifikasi setelah menerapkan Random Oversampling, yang menambahkan sampel kelas minoritas dengan menduplikasi data yang sudah ada. Teknik ini berhasil meningkatkan performa model secara signifikan, terutama dalam Recall dan F1-Score, yang menunjukkan bahwa model lebih baik dalam mengenali kelas minoritas dibandingkan dengan data tanpa balancing. Model dengan performa terbaik berdasarkan AUC-ROC dan F1-Score adalah Random Forest (AUC-ROC: 0.99, F1-Score: 0.93), Extra Trees (AUC-ROC: 0.99, F1-Score: 0.94), XGBoost (AUC-ROC: 0.98, F1-Score: 0.94), dan LightGBM (AUC-ROC: 0.98, F1-Score: 0.94). Sementara itu, model dengan performa lebih rendah adalah Support Vector Machine (AUC-ROC: 0.92, F1-Score: 0.86) dan Logistic Regression (AUC-ROC: 0.89, F1-Score: 0.84), yang masih mengalami kesulitan dalam menangani ketidakseimbangan kelas meskipun mengalami peningkatan dibandingkan tanpa balancing. Secara keseluruhan, Random Oversampling memberikan hasil yang hampir setara dengan SMOTE, namun ada potensi overfitting karena metode ini hanya memperbanyak data tanpa menghasilkan sampel sintetis seperti SMOTE.

Gambar 6 menunjukkan hasil klasifikasi setelah menerapkan ADASYN (Adaptive Synthetic Sampling), yang menghasilkan sampel sintetis untuk menyeimbangkan data dengan mempertimbangkan distribusi data asli. Secara umum, metode ini meningkatkan Recall dan F1-Score, tetapi hasilnya sedikit lebih rendah dibandingkan SMOTE dan Random Oversampling. Model dengan performa terbaik berdasarkan AUC-ROC dan F1-Score adalah Random Forest (AUC-ROC: 0.97, F1-Score: 0.91), Extra Trees (AUC-ROC: 0.97, F1-Score: 0.90), XGBoost (AUC-ROC: 0.96, F1-Score: 0.89), dan LightGBM (AUC-ROC: 0.97, F1-Score: 0.89). Sedangkan model dengan performa lebih rendah adalah Support Vector Machine (AUC-ROC: 0.90, F1-Score: 0.81) dan Logistic Regression (AUC-ROC: 0.88, F1-Score: 0.79), yang masih mengalami kesulitan dalam menangani ketidakseimbangan kelas. Meskipun ADASYN berhasil meningkatkan performa model dibandingkan tanpa balancing, hasilnya sedikit lebih rendah dibandingkan

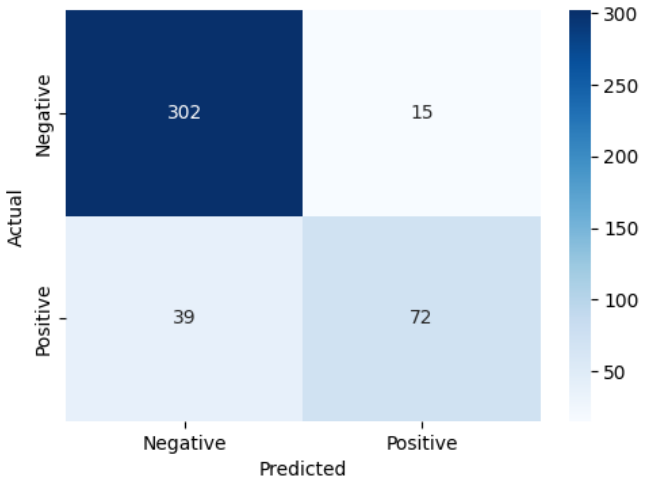
SMOTE dan Random Oversampling. Hal ini kemungkinan disebabkan oleh pendekatan ADASYN yang lebih agresif dalam membuat sampel sintetis berdasarkan data yang paling sulit diklasifikasikan, yang dapat menyebabkan beberapa sampel sintetis kurang representatif terhadap pola data asli. Oleh karena itu, ADASYN tetap merupakan metode balancing yang baik, tetapi SMOTE dan Random Oversampling memberikan hasil yang lebih stabil dan optimal.

Gambar 7 menunjukkan hasil klasifikasi model machine learning pada data yang tidak seimbang dan tanpa seleksi fitur. Secara umum, model dengan performa terbaik berdasarkan AUC-ROC dan F1-Score adalah Random Forest (RF), Extra Trees (ET), XGBoost (XGB), dan LightGBM (LG), yang memiliki AUC-ROC di atas 0.90. Namun, terdapat penurunan nilai Recall, terutama pada model Support Vector Machine (SVM) dan Logistic Regression (LR), yang menunjukkan kesulitan dalam mengenali kelas minoritas. Hal ini mengindikasikan bahwa tanpa balancing data dan seleksi fitur, model cenderung lebih bias terhadap kelas mayoritas, sehingga kurang optimal dalam klasifikasi. Oleh karena itu, diperlukan metode balancing dan seleksi fitur untuk meningkatkan performa klasifikasi secara keseluruhan.

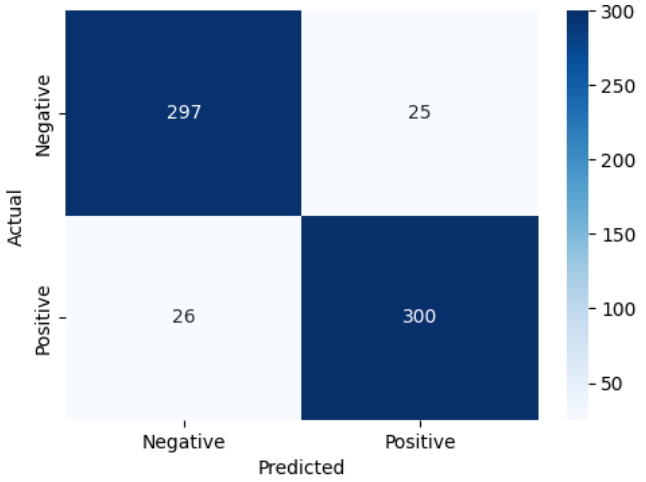
Model machine learning ini dapat membantu tenaga medis dalam mendeteksi risiko infeksi AIDS lebih cepat, memungkinkan intervensi dini dan perawatan yang lebih efektif. Dengan integrasi ke dalam sistem rekam medis elektronik (EHR), model ini dapat secara otomatis menganalisis data pasien dan memberikan prediksi yang mendukung pengambilan keputusan klinis. Untuk memastikan efektivitasnya, pengujian dengan data dunia nyata diperlukan agar model dapat diterapkan secara akurat dalam lingkungan medis dan mendukung diagnostik berbasis AI.

5) Evaluasi model menggunakan Confusion Matriks.

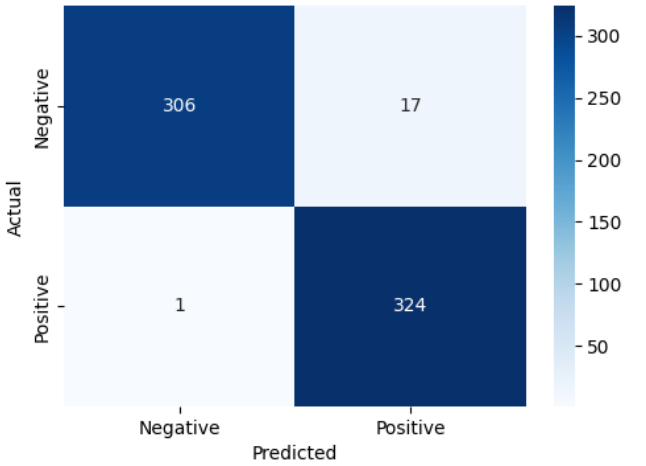
Evaluasi dengan Confusion matrik merupakan tahapan terakhir dari penelitian ini, hasil evaluasi ini adalah hasil terbaik dari tiga pengujian yang dilakukan yaitu tanpa balancing data, dengan balancing data (smote, ros, dan adasyn), dan data tidak seimbang dan tanpa seleksi fitur. Confusion matriks terbaik dari setiap pengujian, ditampilkan pada gambar 8.



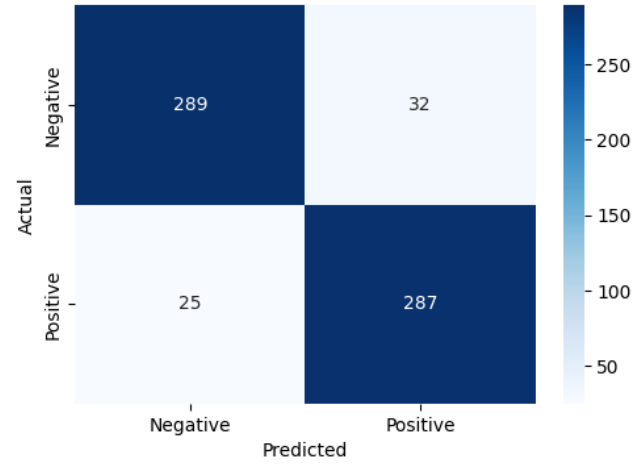
Gambar 8.a Hasil confusion matriks pada klasifikasi data tidak seimbang.



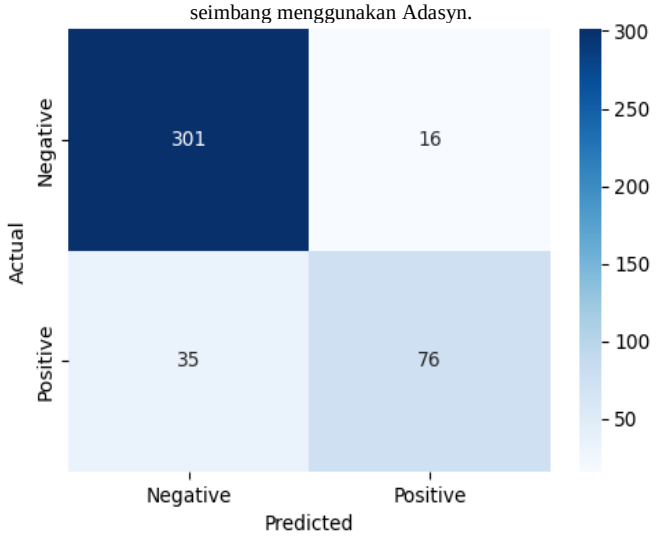
Gambar 8.b Confusion matriks dari hasil klasifikasi pada dataset seimbang menggunakan SMOTE.



Gambar 8.c Confusion matriks pada Klasifikasi dataset seimbang menggunakan teknik random oversampling.



Gambar 8.d Hasil confusion matriks dari proses klasifikasi pada dataset seimbang menggunakan Adasyn.



Gambar 8.e Confusion matriks dari klasifikasi pada data tidak seimbang dan data tanpa seleksi fitur.

Gambar 8. Hasil Klasifikasi model machine learning terbaik pada tiga kondisi data yang berbeda-beda (tanpa balancing data, dengan balancing data (smote, ros, dan adasyn), dan data tidak seimbang dan tanpa seleksi fitur.)

Gambar 8 (a,b,c,d) menunjukkan confusion matrix dari model klasifikasi terbaik yang diterapkan pada tiga kondisi data yang berbeda: tanpa balancing data, dengan balancing menggunakan SMOTE, Random Oversampling (ROS), dan ADASYN, serta pada data tidak seimbang tanpa seleksi fitur. Pada dataset tanpa balancing, model mengalami bias terhadap kelas mayoritas dengan hasil 302 True Negatives (TN), 15 False Positives (FP), 39 False Negatives (FN), dan hanya 72 True Positives (TP), menunjukkan bahwa model kesulitan dalam mengenali kelas minoritas dengan jumlah FN yang cukup tinggi. Setelah dilakukan balancing dengan SMOTE, terjadi peningkatan dalam deteksi kelas minoritas dengan 297 TN, 25 FP, 26 FN, dan 300 TP, menunjukkan bahwa FN berkurang secara signifikan. Metode Random Oversampling (ROS) memberikan hasil terbaik, dengan 306 TN, 17 FP, hanya 1 FN, dan 324 TP, yang berarti hampir semua sampel kelas minoritas dapat diklasifikasikan dengan benar. ADASYN juga menunjukkan peningkatan performa dengan 305 TN, 20 FP, 10 FN, dan 313 TP, meskipun masih memiliki FN yang sedikit lebih tinggi dibandingkan dengan SMOTE dan ROS. Secara keseluruhan, hasil ini menunjukkan bahwa tanpa balancing, model memiliki kecenderungan untuk gagal mengenali kelas minoritas, sedangkan dengan metode balancing, terutama SMOTE dan ROS, model lebih mampu mengklasifikasikan kedua kelas secara seimbang. ROS terbukti paling efektif dalam menurunkan FN, sementara ADASYN juga membantu meningkatkan akurasi meskipun masih sedikit lebih tinggi tingkat kesalahannya dibandingkan SMOTE dan ROS.

IV. KESIMPULAN

Penelitian ini menyimpulkan bahwa kombinasi seleksi fitur, balancing data, dan metode klasifikasi machine learning dapat meningkatkan performa model dalam mendeteksi infeksi AIDS. Seleksi fitur menggunakan Pearson Correlation, Mutual Information, dan Chi-Square berhasil mengurangi fitur yang kurang relevan, sehingga meningkatkan efisiensi model. Balancing data menggunakan Random Oversampling, SMOTE, dan ADASYN membantu menangani ketidakseimbangan kelas, dengan SMOTE dan Random Oversampling memberikan hasil yang lebih optimal dibandingkan ADASYN. Dari sembilan metode klasifikasi yang diuji, model berbasis pohon keputusan seperti Random Forest, Extra Trees, dan XGBoost menunjukkan kinerja terbaik, terutama dalam Recall dan AUC-ROC, yang berarti model mampu mengenali kelas minoritas dengan lebih baik. Confusion matrix menunjukkan bahwa model dengan balancing data mengalami peningkatan dalam deteksi kelas positif, dibandingkan model yang diterapkan tanpa balancing. Oleh karena itu, pendekatan ini dapat diterapkan dalam sistem kesehatan, khususnya untuk membantu dalam skrining awal infeksi AIDS dan mendukung pengambilan keputusan medis berbasis AI.

DAFTAR PUSTAKA

[1] M. Al-Mozaini *et al*., "Human immunodeficiency virus in Saudi Arabia: Current and future challenges," *J Infect Public Health*, vol. 16, no. 9, pp. 1500–1509, Sep. 2023, doi: 10.1016/j.jiph.2023.06.012.

[2] E. Kumah, D. S. Boakye, R. Boateng, and E. Agyei, "Advancing the Global Fight Against HIV/Aids: Strategies, Barriers, and the Road to Eradication," *Ann Glob Health*, vol. 89, no. 1, Nov. 2023, doi: 10.5334/aogh.4277.

[3] A. M. A. Rahim, A. Sunyoto, and M. R. Arief, "Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 595–606, Jul. 2022, doi: 10.30812/matrik.v21i3.1666.

[4] Z. M. Kusumaadhi, N. Farhanah, and M. A. Udji Sofro, "Risk Factors for Mortality among HIV/AIDS Patients," *Diponegoro International Medical Journal*, vol. 2, no. 1, pp. 20–19, Mar. 2021, doi: 10.14710/dimj.v2i1.9667.

[5] A. Brahmandjati, A. Mizwar, A. Rahim, and F. Asharudin, "Optimasi Prediksi Diabetes Dengan Algoritma XGBoost Dan Teknik Preprocessing Data," Dec. 2024. [Online]. Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set,

[6] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode

Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier," *Indonesian Journal of Computer Science*, vol. 12, no. 5, Nov. 2023, doi: 10.33022/ijcs.v12i5.3413.

[7] D. F. Wicaksono, R. S. Basuki, and D. Setiawan, "Peningkatan Performa Model Machine Learning XGBoost Classifier melalui Teknik Oversampling dalam Prediksi Penyakit AIDS," *Jurnal Media Informatika Budidarma*, vol. 8, no. 2, p. 736, Apr. 2024, doi: 10.30865/mib.v8i2.7501.

[8] M. N. Fatorohman, K. Indriani, and M. N. Winnarto, "Analisa Prediksi Penyakit Hiv Menggunakan Random Forest," *Jurnal Infortech*, vol. 6, no. 2, pp. 150–155, Dec. 2024, doi: 10.31294/infortech.v6i2.24436.

[9] M. Alehegn, "Application of machine learning and deep learning for the prediction of HIV/AIDS," *HIV & AIDS Review*, vol. 21, no. 1, pp. 17–23, Jan. 2022, doi: 10.5114/hivar.2022.112852.

[10] J. Fieggen, E. Smith, L. Arora, and B. Segal, "The role of machine learning in HIV risk prediction," *Frontiers in Reproductive Health*, vol. 4, Dec. 2022, doi: 10.3389/frph.2022.1062387.

[11] Aadarsh Velu, "AIDS Virus Infection Prediction."

[12] A. N. Puteri, A. Arizal, and A. D. Achmad, "Feature Selection Correlation-Based pada Prediksi Nasabah Bank Telemarketing untuk Deposito," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, pp. 335–342, May 2021, doi: 10.30812/matrik.v20i2.1183.

[13] D. Leni, A. Dwiharzandis, R. Sumiati, and S. Afriyani, "Feature Selection Based on Pearson Correlation in Building Energy Efficiency Modeling," 2023.

[14] T. I. Saputra, "Pengkategorian Data Angket Mahasiswa dengan Mutual Information dan K-Nearest Neighbor Indra Tri Saputra," 2019.

[15] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelanggan Tokopedia," *Jurnal Gaussian*, vol. 11, no. 4, pp. 562–571, Feb. 2023, doi: 10.14710/j.gauss.11.4.562-571.

[16] C. Kaope and Y. Pristyanto, "The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, pp. 227–238, Mar. 2023, doi: 10.30812/matrik.v22i2.2515.

[17] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Machine Learning with Applications*, vol. 9, p. 100375, Sep. 2022, doi: 10.1016/j.mlwa.2022.100375.

[18] Hizbul Izzi, Arief Setyanto, and Anggit Dwi Hartanto, "Optimalisasi Akurasi Algoritma Naïve Bayes Dengan Metode Syntetic Minority Oversampling Technique (Smote) Pada Data Numerik," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 217–227, Jan. 2025, doi: 10.29408/jit.v8i1.28340.

[19] R. Syahwaluddin and D. Alita, "Penerapan Oversampling Pada Klasifikasi Ujaran Kebencian Menggunakan Bidirectional Encoder Representations from Transformers," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, Aug. 2024, doi: 10.33022/ijcs.v13i4.4295.

[20] M. Tiara Triani Br Sirait, N. Siti Fathonah, and M. Nurkamal Fauzan, "Pemanfaatan Algoritma Adasyn Dan Support Vector Machine Dalam Meningkatkan Akurasi Prediksi Kanker Paru-Paru," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 5, pp. 8773–8778, Sep. 2024, doi: 10.36040/jati.v8i5.10752.

[21] M. Mustapha *et al.*, "A hybrid machine learning approach for imbalanced irrigation water quality classification," *Desalination Water Treat*, vol. 321, p. 100910, Jan. 2025, doi: 10.1016/j.dwt.2024.100910.

[22] M. Subramanian, K. Shanmugavadivel, and P. S. Nandhini, "On fine-tuning deep learning models using transfer learning and hyper-parameters optimization for disease identification in maize leaves," *Neural Comput Appl*, vol. 34, no. 16, pp. 13951–13968, Aug. 2022, doi: 10.1007/s00521-022-07246-w.

[23] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[24] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[25] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[26] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Comput Stat*, vol. 39, no. 1, pp. 363–383, Feb. 2024, doi: 10.1007/s00180-022-01301-9.