JURNAL BUFFER INFORMATIKA
Volume 11 Nomor 1, April 2025
p-ISSN : 2527-4856, e-ISSN : 2614-5413
https://journal.fkom.uniku.ac.id/buffer

Perbandingan Kinerja Algoritma Machine Learning dalam Deteksi Potensi Risiko HIV

M. Bahril Ilmi[1)] **, Kusrini**[2)]

[1,2),] *Informatika, Amikom Yogyakarta Yogyakarta, Indonesia*
Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
Email : mbahrililmi@students.amikom.ac.id[1)],kusrini@amikom.ac.id [1)]

Abstrak

HIV/AIDS masih menjadi tantangan kesehatan masyarakat yang signifikan di seluruh dunia, yang memengaruhi jutaan orang dan menimbulkan beban substansial pada sistem perawatan kesehatan. Identifikasi dini individu yang berisiko tinggi terinfeksi HIV sangat penting untuk menerapkan *intervensi* pencegahan. Studi ini berfokus pada prediksi risiko infeksi HIV menggunakan model *machine learning* tingkat lanjut, yaitu *LightGBM, CatBoost*, dan *XGBoost*. Model-model ini dipilih karena kinerjanya yang tinggi dalam menangani kumpulan data yang kompleks dan tidak seimbang yang merupakan ciri khas studi *epidemiologi*. Penelitian ini memanfaatkan fitur klinis dan demografi untuk mengembangkan model prediktif, menggunakan teknik seperti *cross-validation, regularization, dan oversampling* (misalnya, *SMOTE*) untuk mengatasi masalah umum seperti *overfitting* dan ketidakseimbangan data. Hasilnya menunjukkan bahwa model-model ini dapat mencapai akurasi tinggi, menawarkan alat yang berharga untuk penilaian risiko dini dan membantu dalam penerapan intervensi yang ditargetkan untuk mengendalikan penyebaran HIV.

Kata Kunci : Prediksi risiko HIV; Pembelajaran mesin; LightGBM; CatBoost; XGBoost; Ketidakseimbangan data; Overfitting; Pemodelan prediktif; SMOTE

Abstract

HIV/AIDS remains a significant public health challenge worldwide, affecting millions of people and placing a substantial burden on healthcare systems. Early identification of individuals at high risk of HIV infection is critical for implementing preventive interventions. This study focuses on predicting the risk of HIV infection using advanced machine learning models, namely LightGBM, CatBoost, and XGBoost. These models were chosen because of their high performance in handling complex and imbalanced datasets that are typical of epidemiological studies. This study leverages clinical and demographic features to develop predictive models, using techniques such as cross-validation, regularization, and oversampling (e.g., SMOTE) to address common issues such as overfitting and imbalanced data. The results show that these models can achieve high accuracy, offering valuable tools for early risk assessment and aiding in the implementation of targeted interventions to control the spread of HIV.

Keywords: HIV risk prediction; Machine learning; LightGBM; CatBoost; XGBoost; Data imbalance; Overfitting; Predictive modeling; SMOTE

1. PENDAHULUAN

(Left, Bold, Times New Roman 11, UPPER CASE)

HIV/AIDS terus menjadi tantangan kesehatan global yang signifikan, yang memengaruhi jutaan orang di seluruh dunia [1]. Meskipun ada kemajuan dalam diagnosis dan pengobatan, yang telah mengubah HIV/AIDS menjadi penyakit kronis yang dapat dikelola, penyebaran virus tersebut masih menjadi masalah kritis [2]. Menurut Organisasi Kesehatan Dunia (WHO), sekitar 38 juta orang hidup dengan HIV/AIDS pada tahun 2019, dengan 1,7 juta infeksi baru dilaporkan pada tahun itu saja [3].

Mencegah penyebaran HIV sangat penting, dan mengidentifikasi individu yang berisiko tinggi terinfeksi memainkan peran penting dalam upaya intervensi yang efektif [4]. Berbagai faktor, termasuk usia, aktivitas seksual, riwayat medis, dan kesehatan umum, telah dikaitkan dengan risiko HIV.

Namun, menganalisis faktor-faktor risiko ini melibatkan hubungan yang kompleks antara berbagai variabel [5].

Teknik *machine learning* menawarkan potensi untuk mengidentifikasi pola rumit dalam data klinis dan demografis yang terkait dengan risiko infeksi HIV [6]. Melalui metode ini, model prediktif dapat dikembangkan untuk memberikan perkiraan yang lebih akurat tentang kemungkinan seseorang tertular HIV [7]. Meskipun penelitian sebelumnya telah mengeksplorasi penggunaan *machine learning* dalam prediksi risiko HIV, masih ada ruang untuk perbaikan dalam akurasi dan kecanggihan model.

Penelitian ini bertujuan untuk mengatasi kesenjangan ini dengan mengembangkan model prediktif yang lebih canggih menggunakan teknik machine learning yang canggih. Dengan mengeksplorasi hubungan antara demografi, riwayat medis, dan parameter klinis dengan risiko infeksi HIV, penelitian ini berkontribusi pada upaya global untuk mengendalikan penyebaran virus dan meningkatkan kesehatan Masyarakat [8]. Selain itu, penelitian ini akan bereksperimen dengan metode untuk meningkatkan akurasi model dan mengatasi tantangan *overfitting*.

Overfitting adalah salah satu masalah utama yang sering terjadi dalam pengembangan model prediksi berbasis *machine learning*, terutama ketika model menjadi terlalu kompleks dan sangat cocok dengan data pelatihan sehingga kinerjanya menurun pada data baru. Dalam konteks prediksi risiko infeksi HIV, *overfitting* dapat menyebabkan model memberikan estimasi yang tidak akurat pada populasi yang lebih luas [9]. Untuk mengatasi masalah ini, penelitian ini menerapkan beberapa metode pengurangan *overfitting*, seperti *cross-validation*, *regularization* (L1 dan L2), dan *hyperparameter tuning* [10]. *Cross-validation* digunakan untuk mengevaluasi model dengan cara yang lebih akurat, memastikan model diuji pada berbagai subset data. *Regularization* membantu mengurangi kompleksitas model dengan menambahkan penalti terhadap parameter model yang besar [11]. Selain itu, *hyperparameter tuning* dilakukan untuk menemukan kombinasi parameter yang optimal, sehingga model dapat menghasilkan prediksi yang lebih akurat tanpa mengalami *overfitting* yang signifikan [12]. Dengan pendekatan-pendekatan ini, penelitian ini bertujuan untuk mengembangkan model prediktif yang lebih andal dan mampu mengidentifikasi individu yang berisiko tinggi terinfeksi HIV secara lebih efektif [13].

2. TINJAUAN PUSTAKA

A. Machine Learning dalam Prediksi Risiko HIV

Berbagai algoritma pembelajaran mesin (ML) telah menunjukkan potensi besar dalam mengidentifikasi risiko infeksi HIV melalui analisis data klinis dan demografi. Model seperti Regresi Logistik, Random Forest, dan SVM telah diterapkan dalam studi awal, meletakkan dasar bagi model yang lebih kompleks seperti XGBoost, *CatBoost*, dan LightGBM [14], algoritma ML memungkinkan analisis pola risiko HIV yang lebih mendalam, terutama di negara-negara dengan beban HIV yang tinggi seperti Afrika Selatan, Nigeria, dan Ethiopia. Model-model ini mengidentifikasi faktor-faktor risiko seperti usia, perilaku seksual, dan riwayat medis yang berbeda berdasarkan konteks regional dan budaya [15].

B. LIGHTGBM, CATBOOST, DAN XGBOOST dalam prediksi risiko HIV

Penelitian terkini telah menunjukkan bahwa algoritma berbasis peningkatan gradien, seperti LightGBM, *CatBoost*, dan XGBoost, memberikan akurasi tinggi dan memiliki kapasitas untuk menangani data yang kompleks [16]. Model-model ini unggul dalam mendeteksi pola risiko di antara variabel *heterogen* dan mengatasi masalah ketidakseimbangan data, yang umum terjadi dalam studi epidemiologi [16]. menunjukkan bahwa, dengan penyetelan *hiperparameter* dan teknik validasi silang, model ML mampu mengurangi risiko *overfitting* dan meningkatkan interpretabilitas model dalam konteks prediksi HIV [17].

C. Keterbatasan dan tantangan dalam implementasi model

Salah satu tantangan utama dalam mengembangkan model prediksi HIV adalah ketidakseimbangan data, di mana kelompok rentan sering kali kurang terwakili dalam kumpulan data. Penggunaan teknik *oversampling* seperti *SMOTE* telah dicoba untuk mengatasi masalah ini, tetapi masih ada ruang untuk perbaikan [18]. Selain itu, kesulitan dalam meningkatkan interpretabilitas model kompleks seperti XGBoost merupakan salah satu hambatan penggunaannya dalam keputusan klinis langsung [19]. menyoroti pentingnya pendekatan yang tidak hanya akurat tetapi juga dapat dipahami oleh petugas kesehatan sehingga hasil

model dapat diimplementasikan dalam konteks praktis.

3. METODE PENELITIAN

A. Pengumpulan dan Praproses Data

Dataset yang digunakan dalam penelitian ini adalah AIDS Virus *Infection Prediction Dataset* yang tersedia di Kaggle *https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction*. Dataset ini mencakup 23 variable, termasuk faktor demografi, riwayat medis, dan parameter klinis seperti waktu, usia, wtkg, hemo, cd40, cd420, cd80, cd820, oprior, ras, dan jenis kelamin. Variabel target dalam dataset ini adalah infection, yang menunjukkan apakah seseorang terinfeksi HIV atau tidak. Untuk mengatasi masalah ketidakseimbangan data yang sering terjadi dalam studi epidemiologi, teknik *oversampling* seperti *SMOTE* (Synthetic Minority Over-sampling Technique) digunakan sebelum melatih model. Teknik-teknik ini membantu meningkatkan representasi kelas minoritas (kasus positif HIV), sehingga meminimalkan bias dalam model prediksi.

Tabel 1. Variabel

Nama Variabel	**Keterangan**
time (hari)	waktu yang dibutuhkan hingga gagal hati
trt	indikator pengobatan (0 = ZDV hanya; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = ddI hanya)
age	usia (tahun) pada awal
wtkg	berat (kg) pada awal
hemo	hemofilia (0=tidak, 1=ya)
homo	aktivitas homoseksual (0=tidak, 1=ya)
drugs	riwayat penggunaan obat terlarang (0=tidak, 1=ya)
karnof	Skor Karnofsky (skala 0-100)
oprior	Terapi antiretroviral non-ZDV sebelum 175 (0=tidak, 1=ya)
z30	ZDV dalam 30 hari sebelum 175 (0=tidak, 1=ya)
preanti	hari pra-175 terapi antiretroviral
race	ras (0=kulit putih, 1=non-kulit putih)
gender	jenis kelamin (0=P, 1=L)
str2	riwayat antiretroviral (0=naif, 1=berpengalaman)
strat	stratifikasi riwayat antiretroviral (1='Belum pernah mendapat terapi antiretroviral',2='> 1 tetapi <= 52 minggu terapi antiretroviral sebelumnya',3='> 52 minggu)
symptom	indikator simptomatik (0=asimptot, 1=gejala)
treat	indikator pengobatan (0=hanya ZDV, 1=lainnya)
offtrt	indikator penghentian terapi sebelum 96+/-5 minggu (0=tidak,1=ya)
cd40	CD4 pada awal
cd420	CD4 pada usia 20+/-5 minggu
cd80	CD8 pada awal
cd820	CD8 pada 20+/-5 minggu
infected	terinfeksi AIDS (0=Tidak, 1=Ya)

B. Pengembangan Model dan Pengurangan Overfitting

Dalam studi ini, kami menerapkan proses komprehensif untuk mengembangkan dan mengoptimalkan model pembelajaran mesin untuk memprediksi risiko infeksi HIV. Kami menggunakan tiga algoritma canggih: LightGBM, *CatBoost*, dan XGBoost. Awalnya, kami mengatasi masalah data yang tidak seimbang dalam kumpulan data dengan menerapkan teknik *oversampling* seperti *SMOTE* (Synthetic Minority Over-sampling Technique). Metode ini menghasilkan sampel sintetis dari kelas minoritas, sehingga meningkatkan kemampuan model untuk belajar dari titik data yang kurang terwakili. Setelah praproses, kami membagi kumpulan data menjadi set pelatihan dan pengujian dan menerapkan validasi silang untuk memastikan evaluasi kinerja model yang kuat. Kami kemudian melakukan penyetelan hiperparameter menggunakan *GridSearchCV* untuk menemukan set parameter yang optimal untuk setiap model, yang bertujuan untuk meningkatkan kinerjanya dan meminimalkan *overfitting*. Selain itu, kami menggabungkan teknik regularisasi (L1 dan L2) untuk menyesuaikan model yang terlalu rumit dan selanjutnya mengurangi risiko *overfitting*. Akhirnya, kami mengevaluasi model menggunakan metrik seperti akurasi, *presisi*, *recall*, skor F1, dan AUC untuk menilai kinerja dan kemampuan prediktifnya. Proses ini menghasilkan model yang dioptimalkan dengan baik, dengan akurasi tertinggi dicapai oleh model XGBoost.

Penelitian ini menggunakan tiga algoritma peningkatan gradien yang populer, yaitu LightGBM, *CatBoost*, dan XGBoost. Alasan pemilihan algoritma ini adalah karena kemampuannya dalam menangani kumpulan data yang kompleks, kinerja yang unggul dalam klasifikasi, dan fleksibilitas dalam penyesuaian hiperparameter.

1) LightGBM (*Light Gradient Boosting Machine*) dikenal karena efisiensi dan kecepatan komputasinya. Algoritma ini menggunakan pertumbuhan pohon per cabang yang memungkinkan pertumbuhan pohon lebih optimal dibandingkan dengan metode per tingkat [20].
2) *CatBoost* (*Categorical Boosting*) dirancang khusus untuk menangani fitur kategoris dan mengurangi *overfitting* dengan menerapkan penguatan terurut. Algoritma ini sangat efektif dalam kumpulan data yang memiliki banyak fitur kategoris, seperti kumpulan data yang digunakan dalam penelitian ini [21].

3) XGBoost (*Extreme Gradient Boosting*) adalah salah satu algoritma boosting yang paling banyak digunakan karena stabilitas dan kinerjanya yang tinggi. XGBoost menawarkan kemampuan untuk menangani data yang jarang dan mengurangi kompleksitas model melalui regularisasi L1 dan L2 [22].

C. Evaluasi Model

Evaluasi model dilakukan dengan menggunakan beberapa metrik, yaitu:

1) Akurasi adalah persentase prediksi yang benar dari total pengamatan.
2) Presisi adalah proporsi prediksi positif yang benar-benar positif.
3) Recall adalah proporsi kasus positif yang berhasil diidentifikasi oleh model.
4) F1-Score adalah rata-rata harmonik presisi dan recall, yang memberikan keseimbangan antara keduanya.
5) AUC-ROC adalah area di bawah kurva ROC, yang mengukur kemampuan model untuk membedakan antara kelas positif dan negatif.
6) Average Cross-Validation Accuracy adalah rata-rata akurasi dari validasi silang (cross-validation), atau performa model pada data uji sebenarnya

D. Penyetelan Hiperparameter

Untuk meningkatkan kinerja model, penyesuaian hiperparameter dilakukan menggunakan teknik Grid Search dan Random Search. Teknik ini membantu menemukan kombinasi parameter yang optimal untuk setiap model, sehingga mengurangi risiko *overfitting* dan meningkatkan akurasi prediksi.

Khususnya, untuk LightGBM, *CatBoost*, dan XGBoost, berbagai hiperparameter disetel, termasuk:

1) Laju Pembelajaran
   Ini mengendalikan ukuran langkah selama pelatihan model. Laju pembelajaran yang lebih rendah dapat mengurangi risiko melampaui kerugian minimum tetapi mungkin memerlukan lebih banyak iterasi untuk konvergen. Ini membantu mengurangi *overfitting* dengan memungkinkan model belajar secara perlahan dan bertahap.
2) Jumlah Estimator (*n_estimators*)
   Ini menunjukkan jumlah iterasi atau pohon boosting. Meningkatkan nilai ini dapat meningkatkan kinerja model, tetapi juga meningkatkan risiko *overfitting*. Menemukan angka optimal membantu menyeimbangkan akurasi dan generalisasi.
3) Kedalaman Maksimum (*max_depth*)
   Parameter ini mengendalikan kedalaman maksimum setiap pohon. Pohon yang lebih dalam dapat menangkap pola yang lebih kompleks dalam data tetapi juga dapat menyebabkan *overfitting*. Menetapkan kedalaman maksimum yang tepat membantu membatasi kompleksitas model.
4) Parameter *Subsample* dan *Colsample* (*subsample, colsample_bytree*)
   Parameter ini menentukan fraksi sampel dan fitur yang digunakan dalam setiap iterasi boosting. Dengan menggunakan sebagian kecil data, varians model berkurang, yang membantu mencegah *overfitting*.
5) Parameter Regularisasi (*reg_lambda*, *reg_alpha* untuk XGBoost dan *l2_leaf_reg* untuk *CatBoost*):
   a) Regularisasi L2 (*reg_lambda* atau *l2_leaf_reg*) menambahkan penalti pada koefisien besar, membantu mengurangi kompleksitas model dan mencegah *overfitting*.
   b) Regularisasi L1 (reg_alpha) dapat digunakan untuk mengurangi dampak fitur yang kurang penting dengan mengecilkan bobotnya menjadi nol, sehingga menghasilkan model yang lebih sederhana dan lebih tangguh.
6) Jumlah Daun (*num_leaves* untuk LightGBM)

Parameter ini mengontrol kompleksitas setiap pohon. Jumlah cabang yang lebih banyak dapat menangkap lebih banyak pola tetapi dapat meningkatkan risiko *overfitting*. Penyetelan parameter ini membantu mencapai keseimbangan antara akurasi dan generalisasi.

E. Diagram Alir Penelitian(flowchart)

Untuk memperjelas tahapan penelitian, berikut adalah diagram alir yang menggambarkan alur proses dari tahap awal hingga evaluasi model.

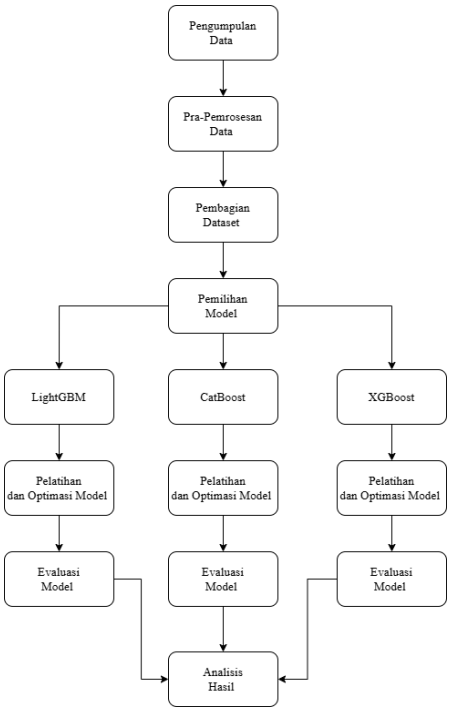


Diagram alir ini menunjukkan langkah-langkah utama dalam penelitian ini, yang meliputi:

1) Pengumpulan Data – Mengambil dataset dari sumber yang relevan, yaitu dataset AIDS Virus *Infection Prediction* dari Kaggle.
2) Pra-Pemrosesan Data – Melakukan pembersihan data, menangani nilai yang hilang, dan menangani ketidakseimbangan kelas menggunakan teknik *oversampling* (*SMOTE, ADASYN*).
3) Pembagian Dataset – Memisahkan data menjadi training set dan testing set untuk keperluan pelatihan dan evaluasi model.
4) Pemilihan Model – Menggunakan tiga algoritma gradient boosting: LightGBM, *CatBoost*, dan XGBoost.
5) Pelatihan dan Optimasi Model – Melatih model dengan teknik validasi silang, regulasi, serta tuning *hyperparameter*.
6) Evaluasi Model – Menggunakan metrik seperti akurasi, *presisi, recall, F1-score*, dan AUC-ROC untuk menilai kinerja masing-masing model.
7) Analisis Hasil – Membandingkan hasil dari ketiga model dan mengidentifikasi faktor yang mempengaruhi kinerja mereka.

4. HASIL DAN PEMBAHASAN

A. Perbandingan Kinerja Model

Hasil evaluasi menunjukkan bahwa ketiga model gradient boosting memberikan kinerja tinggi dalam memprediksi risiko infeksi HIV. *CatBoost* mencatat akurasi tertinggi pada data uji dengan 89,01%, diikuti oleh XGBoost dengan 88,08%, dan LightGBM dengan 87,61%. Dari hasil cross-validation, *CatBoost* juga menunjukkan performa terbaik dengan rata-rata akurasi 92,81%, disusul oleh LightGBM (92,54%) dan XGBoost (92,16%).

Selain itu, *CatBoost* memiliki nilai precision tertinggi (0.80), menunjukkan keandalan model dalam memprediksi individu yang berisiko terinfeksi HIV. Dari sisi recall, *CatBoost* juga unggul dengan 0.73, diikuti oleh XGBoost (0.72) dan LightGBM (0.71), menandakan kemampuannya dalam mendeteksi lebih banyak kasus positif. F1-score, yang merupakan keseimbangan antara precision dan recall, juga tertinggi pada *CatBoost* (0.76), menunjukkan bahwa model ini memiliki performa yang lebih stabil dibandingkan lainnya.

Dari segi AUC-ROC, yang mengukur kemampuan model dalam membedakan antara individu yang terinfeksi dan tidak terinfeksi, *CatBoost* kembali menunjukkan keunggulan dengan nilai 0.93, sementara LightGBM dan XGBoost masing-masing mencapai 0.92. Selain itu, nilai Matthews Correlation Coefficient (MCC), yang mengukur korelasi antara prediksi dan label sebenarnya, menunjukkan hasil tertinggi pada *CatBoost* (0.69), diikuti oleh XGBoost (0.67) dan LightGBM (0.66), mengindikasikan keseimbangan yang lebih baik dalam prediksi kelas positif dan negatif.

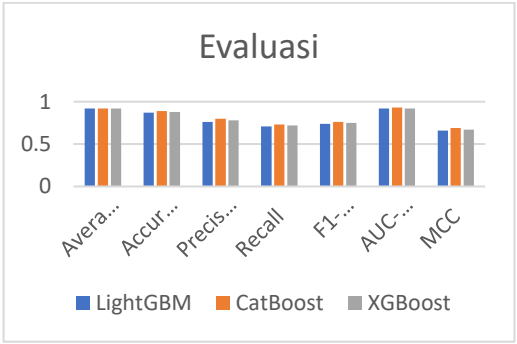
Secara keseluruhan, hasil ini menunjukkan bahwa *CatBoost* memiliki performa terbaik di antara ketiga model, baik dalam akurasi, precision, recall, dan keseimbangan klasifikasi. Penggunaan teknik oversampling seperti SMOTE juga membantu meningkatkan deteksi individu yang berisiko terinfeksi HIV, memastikan model lebih efektif dalam menangani ketidakseimbangan data.

Tabel 2. Evaluasi

Evaluation	LightGBM	*CatBoost*	XGBoost
Average CV Accuracy	92.54%	92.81%	92.16%
Accuracy	87.61%	89.01%	88.08%
Precision	0.76	0.80	0.78
Recall	0.71	0.73	0.72
F1-Score	0.74	0.76	0.75
AUC-ROC	0.92	0.93	0.92
MCC	0.66	0.69	0.67

Tabel ini menunjukkan hasil evaluasi dari tiga model machine learning—LightGBM, *CatBoost*, dan XGBoost—dalam memprediksi risiko infeksi HIV. Evaluasi dilakukan menggunakan enam metrik utama:

1) *Average Cross-Validation Accuracy* mengukur rata-rata akurasi model pada beberapa subset data yang berbeda dalam proses *cross-validation*. Ini memberikan gambaran tentang seberapa baik model dapat menggeneralisasi ke data yang tidak terlihat. Semakin tinggi nilai ini, semakin stabil dan dapat diandalkan performa model di berbagai pembagian data.
2) *Accuracy* menunjukkan proporsi prediksi yang benar dari keseluruhan data uji. Metrik ini memberikan gambaran umum tentang kinerja model dalam mengklasifikasikan individu sebagai terinfeksi atau tidak terinfeksi HIV.
3) *Precision* mengukur keandalan model dalam memprediksi kasus positif dengan membandingkan jumlah prediksi positif yang benar terhadap total prediksi positif yang dibuat oleh model. *Precision* yang tinggi menunjukkan bahwa model jarang memberikan prediksi positif yang salah (false positives).
4) *Recall* mengukur kemampuan model dalam menemukan semua kasus positif yang sebenarnya. Nilai *recall* yang lebih tinggi menunjukkan bahwa model berhasil mengidentifikasi lebih banyak individu yang benar-benar berisiko terinfeksi HIV, mengurangi jumlah kasus positif yang terlewat (*false negatives*).
5) *F1-Score* merupakan rata-rata harmonik dari *precision* dan *recall*, memberikan keseimbangan antara keduanya. Metrik ini sangat berguna ketika terdapat ketidakseimbangan kelas, seperti dalam prediksi risiko HIV, di mana jumlah individu yang terinfeksi jauh lebih sedikit dibandingkan yang tidak terinfeksi.
6) *AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)* mengukur kemampuan model dalam membedakan kelas positif dan negatif pada berbagai *threshold* keputusan. Nilai yang lebih tinggi menunjukkan bahwa model lebih efektif dalam membedakan individu yang berisiko dan tidak berisiko terinfeksi HIV.
7) *Matthews Correlation Coefficient (MCC)* memberikan ukuran korelasi antara prediksi model dan label sebenarnya. Berbeda dengan accuracy, MCC lebih informatif dalam situasi dengan ketidakseimbangan kelas karena mempertimbangkan semua elemen dari matriks konfusi (*true positives, true negatives, false positives, false negatives*). *MCC* berkisar antara -1 hingga 1, dengan nilai lebih tinggi menunjukkan kinerja model yang lebih seimbang.



Hasil pada Tabel 2 menunjukkan bahwa model *CatBoost* memiliki performa terbaik dengan akurasi sebesar 89,01%, diikuti oleh XGBoost (88,08%) dan LightGBM (87,61%).

Tabel 3. *Hyperparameter* LIGHTGBM

Learning rate	Max depth	N estimators	Num leaves	Reg Alpha	Reg Lambda
0.1	10.0	300.0	50.0	0.01	0.01

Tabel ini merangkum nilai *hyperparameter* yang digunakan untuk model LightGBM, yang berpengaruh terhadap performa model:

1) *Learning rate (0.1)* Mengontrol kecepatan pembelajaran model, menentukan seberapa besar perubahan bobot setelah setiap iterasi. Nilai yang lebih kecil membuat model belajar lebih lambat tetapi lebih stabil.
2) *Max depth (10.0)* Kedalaman maksimum setiap pohon dalam model. Semakin besar nilai ini, semakin kompleks model, tetapi juga berisiko menyebabkan *overfitting*.
3) *N estimators (300.0)* Jumlah total pohon dalam *ensemble*. Lebih banyak pohon dapat meningkatkan akurasi, tetapi juga meningkatkan waktu pelatihan.

4) *Num leaves (50.0)* Jumlah maksimum daun dalam setiap pohon. Semakin banyak daun, semakin kompleks pohon dan semakin banyak pola yang dapat ditangkap oleh model.
5) *Reg Alpha (0.01)* Parameter regulasi L1 yang membantu mengurangi kompleksitas model dengan membuat beberapa bobot menjadi nol, sehingga meningkatkan generalisasi.
6) *Reg Lambda (0.01)* Parameter regulasi L2 yang membantu mengurangi *overfitting* dengan mencegah bobot menjadi terlalu besar.

Tabel 4. *Hyperparameter CATBOOST*

depth	iterations	l2_leaf_reg	learning_rate
10.0	500.0	1.0	0.05

Tabel ini menunjukkan hyperparameter yang digunakan dalam model *CatBoost*, yang menentukan kompleksitas dan generalisasi model:

1) *Depth (10)* Kedalaman maksimum pohon dalam model. Semakin dalam pohon, semakin kompleks modelnya, tetapi juga berisiko mengalami *overfitting.*
2) *Iterations (500)* Jumlah *iterasi boosting* yang dilakukan. Semakin banyak iterasi, semakin lama waktu pelatihan, tetapi dapat meningkatkan akurasi model.
3) *L2 leaf reg (1) Regularisasi* L2 yang membantu mengurangi *overfitting* dengan menekan nilai bobot yang terlalu besar, sehingga model menjadi lebih *generalizable*.
4) *Learning rate (0.05)* Mengontrol kecepatan pembelajaran model. Nilai yang lebih kecil membuat model belajar lebih lambat tetapi lebih stabil dan mengurangi risiko *overfitting.*

Tabel 5. Hyperparameter XGBOOST

Learning rate	Max depth	n_estimators	reg_alpha	reg_lambda
0.1	10.0	100.0	0	5.0

Tabel ini merangkum *hyperparameter* yang digunakan dalam model XGBoost:

1) *Learning rate* (0.1) Kecepatan pembelajaran model. Nilai ini menentukan seberapa besar perubahan bobot model dalam setiap iterasi.
2) *Max depth* (10) Kedalaman maksimum setiap pohon dalam model. Semakin dalam pohon, semakin kompleks modelnya, tetapi juga meningkatkan risiko *overfitting*.
3) *N estimators* (100) Jumlah total pohon dalam ensemble. Semakin banyak pohon, semakin baik model dapat menangkap pola dalam data, tetapi dengan peningkatan waktu komputasi.
4) *Reg alpha* (0) Regularisasi L1 yang membantu mengurangi kompleksitas model dengan membuat beberapa bobot menjadi nol, meskipun dalam hal ini tidak diterapkan (nilai 0).
5) *Reg lambda* (5) Regularisasi L2 yang membantu mencegah *overfitting* dengan menekan bobot besar, sehingga model lebih generalizable.

Hasil evaluasi menunjukkan bahwa ketiga model *gradient boosting LightGBM*, *CatBoost*, dan *XGBoost* memiliki kinerja yang baik dalam memprediksi risiko infeksi HIV, dengan nilai akurasi di atas 87%. Model *CatBoost* mencatat akurasi tertinggi sebesar 89,01%, diikuti oleh *XGBoost* dengan 88,08%, dan *LightGBM* dengan 87,61%. Sementara itu, *CatBoost* juga memiliki nilai AUC-ROC tertinggi (0.93), menunjukkan kemampuan diskriminatif yang sangat baik dalam membedakan antara individu yang berisiko terinfeksi HIV dan yang tidak.

1) Efek Penyetelan *Hyperparameter*

a) *CatBoost*
Penggunaan parameter *depth* = 10 dan *iterations* = 500 memberikan keseimbangan antara kompleksitas dan *generalisasi* model. *Regularisasi* L2 (*l2_leaf_reg* = 1.0) membantu mengurangi risiko *overfitting* dengan menjaga *fleksibilitas* model tetap optimal. *Learning rate* sebesar 0.05 memastikan pembelajaran dilakukan secara bertahap untuk meningkatkan stabilitas model.

b) XGBoost
Pengaturan *max_depth* = 10, *learning_rate* = 0.1, dan *n_estimators* = 100 menghasilkan model yang cukup fleksibel dalam menangkap pola data tanpa mengalami *overfitting* yang signifikan. *Regularisasi reg_alpha* = 0 dan *reg_lambda* = 5 menunjukkan bahwa model memanfaatkan *regularisasi* L2 untuk menghindari bobot yang terlalu besar, membantu meningkatkan generalisasi model terhadap data baru.

c) LightGBM
Kombinasi parameter *learning_rate* = 0.1, *max_depth* = 10, dan *n_estimators* = 300 memberikan keseimbangan antara kecepatan

pembelajaran dan kompleksitas model. *Num_leaves* = 50 memberikan fleksibilitas dalam membentuk keputusan, sementara *regularisasi reg_alpha* = 0.01 dan *reg_lambda* = 0.01 membantu mencegah *overfitting*.

2) Efektivitas Metode dalam Mengurangi *Overfitting*

a) Validasi silang

Average Cross-Validation Accuracy menunjukkan bahwa semua model mampu mempertahankan performa yang stabil di berbagai subset data, mengurangi bias dalam proses penyetelan *hiperparameter*. *CatBoost* memiliki rata-rata CV *Accuracy* tertinggi (92.81%), menunjukkan kemampuannya dalam menangani variasi data yang lebih baik dibandingkan LightGBM dan XGBoost.

b) Regularisasi

Regularisasi L2 (*reg_lambda*) digunakan dalam XGBoost dan LightGBM, dengan nilai yang lebih tinggi pada XGBoost (*reg_lambda* = 5) dibandingkan LightGBM (*reg_lambda* = 0.01). Pada *CatBoost*, *regularisasi l2_leaf_reg* = 1 digunakan untuk menyeimbangkan kompleksitas model. Penggunaan regularisasi ini membantu mengurangi *overfitting* dengan membatasi kompleksitas pohon keputusan yang dibentuk.

c) Penyetelan hiperparameter

Penyetelan hiperparameter terbukti menjadi faktor paling signifikan dalam mengurangi *overfitting*. Parameter seperti *depth*, *learning_rate*, dan *regularisasi* memainkan peran penting dalam menjaga keseimbangan antara bias dan varians model. *CatBoost*, dengan *depth* = 10 dan *iterations* = 500, menunjukkan kinerja terbaik secara keseluruhan, sementara XGBoost dengan regularisasi lebih tinggi memiliki daya generalisasi yang lebih baik dibandingkan LightGBM.

B. Pembahasan Hasil

Berdasarkan hasil evaluasi, *CatBoost* menunjukkan kinerja terbaik dalam semua metrik evaluasi, dengan akurasi tertinggi (89,01%) dan *AUC-ROC* tertinggi (0.93), diikuti oleh XGBoost dan LightGBM. Keunggulan *CatBoost* dalam penelitian ini dapat dikaitkan dengan kemampuannya menangani data yang tidak seimbang secara lebih efektif serta mekanisme regularisasi *l2_leaf_reg* = 1, yang membantu mengurangi kompleksitas model dan mencegah *overfitting*.

XGBoost menunjukkan performa yang kompetitif dengan akurasi 88,08% dan *AUC-ROC* 0.92, hanya sedikit lebih rendah dari *CatBoost*. Model ini memiliki keunggulan dalam menangani fitur *numerik* dan menerapkan *regularisasi* yang lebih kuat dengan *reg_lambda* = 5, yang membantu mengurangi *overfitting* dan meningkatkan daya generalisasi model.

LightGBM, meskipun memiliki akurasi sedikit lebih rendah (87,61%) dibandingkan *CatBoost* dan XGBoost, tetap menunjukkan kinerja yang baik dengan waktu komputasi yang lebih cepat. Penggunaan *max_depth* = 10 dan *num_leaves* = 50 memungkinkan model menangkap kompleksitas data tanpa *overfitting* yang signifikan. Kecepatan dan efisiensi LightGBM menjadikannya pilihan yang lebih sesuai untuk aplikasi waktu nyata atau sistem dengan keterbatasan sumber daya.

Selain faktor-faktor di atas, keunggulan *CatBoost* dalam penelitian ini dapat dikaitkan dengan beberapa aspek utama:

1) Penanganan Fitur Kategoris

CatBoost menggunakan metode target *encoding* yang menghindari kebocoran informasi dan memungkinkan model untuk lebih efektif memanfaatkan informasi dari variabel kategoris. Sebaliknya, XGBoost dan LightGBM memerlukan pra-pemrosesan tambahan untuk mengonversi fitur kategoris menjadi numerik sebelum digunakan dalam model.

2) Regularisasi yang Lebih Efektif

Regularisasi L2 (*l2_leaf_reg* = 1) dalam *CatBoost* membantu mengontrol kompleksitas pohon keputusan dan mencegah *overfitting*, terutama dalam dataset dengan dimensi yang relatif besar. Penggunaan nilai regularisasi yang optimal memungkinkan *CatBoost* mempertahankan keseimbangan antara bias dan varians, sehingga menghasilkan prediksi yang lebih stabil dan *generalizable*.

3) Struktur Pohon yang Lebih Stabil

CatBoost menerapkan *symmetric tree growth*, yang memastikan pertumbuhan pohon terjadi secara seimbang di setiap tingkat. Pendekatan ini berbeda dari metode *greedy-based tree growth* yang digunakan oleh XGBoost dan LightGBM, yang dapat menyebabkan ketidakseimbangan dalam struktur pohon dan mengurangi efisiensi model.

5. KESIMPULAN

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa:

a) Model *Gradient Boosting seperti XGBoost, CatBoost,* dan *LightGBM* mampu memberikan hasil prediksi yang akurat dalam mengidentifikasi risiko infeksi HIV.
b) *CatBoost* mencatat kinerja terbaik dengan akurasi 89,01%, menjadikannya model yang paling potensial untuk aplikasi klinis dalam memprediksi risiko HIV.
c) *XGBoost* dan *LightGBM* juga menunjukkan performa yang kompetitif, dengan keunggulan masing-masing dalam regularisasi (*XGBoost*) dan efisiensi komputasi (*LightGBM*).
d) Teknik *oversampling* seperti *SMOTE* terbukti efektif dalam menangani ketidakseimbangan data, yang berkontribusi pada peningkatan recall pada kelas minoritas (HIV-positif).
e) *Interpretabilitas* model masih menjadi tantangan, terutama untuk model yang lebih kompleks seperti *XGBoost* dan *CatBoost*, sehingga diperlukan pendekatan yang lebih baik untuk menjelaskan keputusan model dalam konteks medis.
f) Penelitian lebih lanjut diperlukan untuk meningkatkan *interpretabilitas* model dan mengintegrasikan hasil prediksi dengan keputusan klinis yang lebih informatif.

6. SARAN

Untuk pengembangan lebih lanjut, penelitian ini dapat ditingkatkan dengan:

a) Mengeksplorasi metode lain dalam menangani ketidakseimbangan data, seperti teknik *cost-sensitive learning* atau *ensemble* model yang lebih kompleks.
b) Melakukan eksperimen dengan dataset yang lebih besar dan beragam untuk meningkatkan daya generalisasi model, sehingga hasilnya dapat diterapkan dalam skenario yang lebih luas.
c) Menggunakan teknik *interpretabilitas* model seperti SHAP (*SHapley Additive exPlanations*) untuk memahami kontribusi setiap fitur terhadap prediksi risiko HIV, sehingga dapat memberikan wawasan yang lebih mendalam bagi tenaga medis.
d) Mengintegrasikan model prediksi dengan sistem klinis untuk mendukung pengambilan keputusan yang lebih akurat dan berbasis data dalam pencegahan serta penanganan infeksi HIV.
e) Meneliti pendekatan hibrida dengan mengombinasikan metode berbasis *deep learning* dan *machine learning* untuk meningkatkan akurasi serta memahami pola yang lebih kompleks dalam data medis.

UCAPAN TERIMA KASIH

Saya Selaku Penulis mengucapkan terima kasih kepada Kusrini yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

[1] 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. Heal.*, vol. 89, pp. 1–6, 2023, doi: 10.5334/aogh.4277.

[2] S. Shao and Y. Shao, "From a deadly disease to a manageable chronic disease, HIV/AIDS remains a challenge for mankind," *Infect. Microbes Dis.*, vol. 5, pp. 41–43, 2023, doi: 10.1097/IM9.0000000000000121.

[3] D. E. Andarge, H. E. Hailu, and T. Menna, "Incidence, survival time and associated factors of virological failure among adult HIV/AIDS patients on first line antiretroviral therapy in St. Paul's Hospital Millennium Medical College—A retrospective cohort study," *PLoS One*, vol. 17, pp. 1–12, 2022, doi: 10.1371/journal.pone.0275204.

[4] G. P. Garnett, "Reductions in HIV incidence are likely to increase the importance of key population programmes for HIV control in sub-Saharan Africa," *J. Int. AIDS Soc.*, vol. 24, pp. 5–9, 2021, doi: 10.1002/jia2.25727.

[5] A. O. Soogun, A. B. M. Kharsany, T. Zewotir, D. North, and R. E. Ogunsakin, "Identifying potential factors associated with high HIV viral load in KwaZulu-Natal, South Africa using multiple correspondence analysis and random forest analysis," *BMC Med. Res. Methodol.*, vol. 22, pp. 1–16, 2022, doi: 10.1186/s12874-022-01625-6.

[6] J. Fieggen, E. Smith, L. Arora, and B. Segal, "The role of machine learning in HIV risk prediction," *Front. Reprod. Heal.*, vol. 4, pp. 1–8, 2022, doi: 10.3389/frph.2022.1062387.

[7] K. M. Jia et al., "Risk scores for predicting HIV incidence among adult heterosexual populations in sub-Saharan Africa: A systematic review and meta-analysis," *J. Int. AIDS Soc.*, vol. 25, pp. 1–17, 2022, doi: 10.1002/jia2.25861.

[8] N. Peiffer-Smadja et al., "Machine learning for clinical decision support in infectious diseases: A narrative review of current applications," *Clin. Microbiol. Infect.*, vol. 26, pp. 584–95, 2020, doi: 10.1016/j.cmi.2019.09.009.

[9] C. O. S. Sorzano et al., "On bias, variance, *overfitting*, gold standard and consensus in single-particle analysis by cryo-electron microscopy," *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 78, pp. 410–23, 2022, doi: 10.1107/S2059798322001978.

[10] G. A. Saputra, I. M. A. Agastya, and B. Splendens, "Betta fish identification system based on convolutional neural network," *J. AI Res.*, vol. 8, pp. 443–52, 2024.

[11] S. Wang, Y. Chen, Z. Cui, L. Lin, and Y. Zong, "Diabetes risk analysis based on machine learning LASSO regression model," *J. Theory Pract. Eng. Sci.*, vol. 4, pp. 58–64, 2024, doi: 10.53469/jtpes.2024.04(01).08.

[12] M. Doddy, A. Pranatha, G. H. Setiawan, and M. A. Maricar, "Utilization of ResNet architecture and transfer learning method in the classification of faces of individuals with Down syndrome," *J. AI Res.*, vol. 8, pp. 434–42, 2024.

[13] Y. Dong et al., "Prediction model for the risk of HIV infection among MSM in China: Validation and stability," *Int. J. Environ. Res. Public Health*, vol. 19, 2022, doi: 10.3390/ijerph19021010.

[14] H. Du, Q. Yang, A. Ge, C. Zhao, Y. Ma, and S. Wang, "Explainable machine learning models for early gastric cancer diagnosis," *Sci. Rep.*, vol. 14, pp. 1–15, 2024, doi: 10.1038/s41598-024-67892-z.

[15] M. Jaiteh, E. Phalane, Y. A. Shiferaw, K. A. Voet, and R. N. Phaswana-Mafuya, "Utilization of machine learning algorithms for the strengthening of HIV testing: A systematic review," *Algorithms*, vol. 17, pp. 1–24, 2024, doi: 10.3390/a17080362.

[16] I. B. Mustapha et al., "Comparative analysis of gradient-boosting ensembles for estimation of compressive strength of quaternary blend concrete," *Int. J. Concr. Struct. Mater.*, vol. 18, 2024, doi: 10.1186/s40069-023-00653-w.

[17] S. D. Rodriguez et al., "Machine learning outperformed logistic regression classification even with limit sample size: A model to predict pediatric HIV mortality and clinical progression to AIDS," *PLoS One*, vol. 17, pp. 1–13, 2022, doi: 10.1371/journal.pone.0276116.

[18] M. Mujahid et al., "Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, 2024, doi: 10.1186/s40537-024-00943-4.

[19] J. Gao et al., "Prediction of sepsis mortality in ICU patients using machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 24, pp. 1–11, 2024, doi: 10.1186/s12911-024-02630-z.

[20] M. Imani and H. R. Arabnia, "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis," *Technologies*, vol. 11, 2023, doi: 10.3390/technologies11060167.

[21] J. T. Hancock and T. M. Khoshgoftaar, "*CatBoost* for big data: An interdisciplinary review," *J. Big Data*, vol. 7, 2020, doi: 10.1186/s40537-020-00369-8.

[22] A. I. O. Osman et al., "Extreme gradient boosting (XGBoost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, pp. 1545–56, 2021, doi: 10.1016/j.asej.2020.11.011.