JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

Klasifikasi Wilayah WHO Berdasarkan Data HIV/AIDS Global Menggunakan Pendekatan Machine Learning

Muhamad Agry Sugiarto[1*], Hafiyyan Putra Pratama[2]

[1,2]Program Studi Sistem Telokumunikasi, Universitas Pendidikan Indonesia Kampus UPI di Purwakarta, Purwakarta Indonesia
Email: [1]muhamadagrysugiarto.22@upi.edu, [2]hafiyyan@upi.edu
(Email Corresponding Author: muhamadagrysugiarto.22@upi.edu)
Received: June 5, 2026 | Revision: June 9, 2026 | Accepted: June 19, 2026

Abstrak

Pemetaan epidemiologi HIV/AIDS secara global berdasarkan wilayah World Health Organization (WHO) menjadi salah satu langkah strategis dalam mendukung distribusi bantuan medis yang tepat sasaran. Namun, perbedaan karakteristik data antar-wilayah yang sangat bervariasi membuat proses klasifikasi menjadi tantangan tersendiri. Penelitian ini membandingkan tiga model machine learning, yaitu Logistic Regression, Support Vector Machine (SVM) dasar, dan SVM teroptimasi, untuk mengklasifikasikan data HIV/AIDS ke dalam enam wilayah WHO: Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, dan Western Pacific. Data yang digunakan mencakup estimasi jumlah penderita (nilai median, minimum, dan maksimum) serta persentase cakupan pengobatan antiretroviral (ART coverage %). Proses pra-pemrosesan meliputi analisis data eksploratif, imputasi median untuk menangani nilai kosong, dan normalisasi menggunakan StandardScaler. Evaluasi model dilakukan melalui validasi silang 5-fold dan matriks konfusi. Hasil menunjukkan bahwa data memiliki tingkat tumpang tindih antar-kelas yang cukup tinggi. Logistic Regression menghasilkan akurasi pengujian 39,29% dengan CV Mean 44,05%, sementara SVM teroptimasi dengan parameter C=10, gamma='scale', dan kernel RBF mencapai CV Mean tertinggi sebesar 49,05%. Analisis lebih lanjut mengungkapkan bahwa ART coverage % merupakan fitur paling dominan dalam membedakan karakteristik beban epidemiologi antar-wilayah.

Kata Kunci: *HIV/AIDS Global, Klasifikasi Wilayah WHO, Logistic Regression, Support Vector Machine, GridSearchCV*

Abstract

Global epidemiological mapping of HIV/AIDS based on World Health Organization (WHO) regional classifications is a strategic step in supporting more targeted distribution of medical assistance. However, the highly varied data characteristics across regions make the classification process particularly challenging. This study compares three machine learning models — Logistic Regression, baseline Support Vector Machine (SVM), and optimized SVM — to classify HIV/AIDS data into six WHO regions: Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, and Western Pacific. The dataset includes estimated number of cases (median, minimum, and maximum values) along with antiretroviral treatment coverage percentage (ART coverage %). Pre-processing steps involved exploratory data analysis, median imputation for handling missing values, and data normalization using StandardScaler. Model evaluation was conducted through 5-fold cross-validation and confusion matrix analysis. Results indicate that the dataset exhibits a considerably high level of inter-class overlap. Logistic Regression achieved a test accuracy of 39.29% with a CV Mean of 44.05%, while the optimized SVM with parameters C=10, gamma='scale', and RBF kernel reached the highest CV Mean of 49.05%. Further analysis revealed that ART coverage % is the most dominant feature in distinguishing the epidemiological burden characteristics across global regions

Keywords: *HIV/AIDS, WHO regional classification, machine learning, Support Vector Machine, Logistic Regression*

1. PENDAHULUAN

HIV/AIDS merupakan salah satu permasalahan kesehatan global yang paling signifikan hingga saat ini. Berdasarkan data terkini dari WHO dan UNAIDS, pada akhir tahun 2024 terdapat sekitar 40,8 juta orang di seluruh dunia yang hidup dengan HIV, di mana wilayah Afrika menjadi daerah paling terdampak dengan prevalensi sekitar 3,1% pada populasi dewasa [1]. Secara global, sekitar 77% dari total penderita atau setara 31,6 juta orang telah memperoleh akses terhadap terapi antiretroviral (ART) pada periode yang sama, meskipun tingkat cakupan tersebut masih sangat bervariasi antar kawasan [2]. Ketimpangan distribusi epidemi HIV antar wilayah WHO ini menjadikan analisis berbasis data sebagai pendekatan yang sangat relevan untuk mendukung pengambilan kebijakan kesehatan publik secara global.

Kemajuan pesat dalam bidang ilmu data dan pembelajaran mesin (*machine learning*) telah membuka peluang baru dalam analisis prediktif untuk berbagai permasalahan kesehatan. Studi-studi terdahulu menunjukkan bahwa algoritma machine learning mampu menghasilkan performa yang kompetitif dalam tugas klasifikasi data medis, termasuk prediksi penyakit jantung, diabetes, hingga infeksi menular [3][4]. Dalam konteks HIV/AIDS secara khusus, pendekatan machine learning telah dimanfaatkan untuk tujuan yang beragam, mulai dari prediksi mortalitas, identifikasi komunitas dengan prevalensi tinggi, hingga estimasi kebutuhan layanan kesehatan di wilayah terbatas sumber daya [5][6]. Dua algoritma yang paling umum digunakan dalam klasifikasi data kesehatan adalah *Logistic Regression* dan *Support Vector Machine* (SVM), keduanya dikenal karena kemampuannya menangani data multikelas serta sifat interpretabilitasnya yang relatif baik dibandingkan model yang lebih kompleks [7][8].

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

Salah satu tantangan utama yang kerap muncul dalam dataset kesehatan berskala global adalah permasalahan ketidakseimbangan kelas (*class imbalance*), di mana sejumlah kelompok atau wilayah tertentu memiliki representasi sampel yang jauh lebih sedikit dibandingkan kelompok lainnya [9]. Kondisi ini berpotensi menyebabkan model machine learning menjadi bias terhadap kelas mayoritas sehingga performa prediksi pada kelas minoritas menjadi sangat buruk [10]. Berbagai solusi telah diusulkan untuk mengatasi masalah ini, salah satunya adalah teknik *Synthetic Minority Oversampling Technique* (SMOTE), yang bekerja dengan membangkitkan sampel sintetis pada kelas yang memiliki jumlah data lebih sedikit [11][12]. Di samping itu, tahap normalisasi fitur menggunakan *StandardScaler* juga berperan krusial dalam menjamin bahwa tidak ada fitur tunggal yang mendominasi proses pelatihan model secara tidak proporsional [13].

Penelitian terdahulu oleh Rasheed et al. [3] menggunakan GridSearchCV dengan validasi silang 5-fold untuk mengoptimalkan berbagai algoritma seperti SVM, Logistic Regression, dan Random Forest pada dataset penyakit jantung, dan menemukan bahwa hyperparameter tuning secara signifikan meningkatkan akurasi model. Sementara itu, Khalid et al. [4] membandingkan enam algoritma machine learning dengan dan tanpa tuning hyperparameter pada dataset yang sama, di mana SVM dengan tuning berhasil mencapai akurasi pengujian tertinggi sebesar 87,91%. Pada domain HIV/AIDS, Domínguez-Rodríguez et al. [5] membangun framework machine learning yang skalabel untuk klasifikasi infeksi HIV menggunakan data klinis dan laboratorium, sedangkan penelitian lain memanfaatkan data sosial-ekonomi dan perilaku untuk mengidentifikasi komunitas dengan prevalensi HIV tinggi di wilayah dengan keterbatasan sumber daya [6]. Lebih lanjut, permasalahan class imbalance pada dataset multikelas juga dibahas oleh Zhang et al. [8] yang mengusulkan metode i-SVM-DE untuk menangani ketidakseimbangan kelas pada SVM, serta oleh Wongvorachan et al. [7] yang membandingkan berbagai teknik resampling termasuk SMOTE pada konteks klasifikasi data tidak seimbang.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi berbasis machine learning guna memprediksi wilayah WHO suatu negara menggunakan dataset HIV/AIDS Global Statistics. Proses penelitian mencakup eksplorasi dan visualisasi data, praproses meliputi imputasi, normalisasi, dan encoding, pelatihan model Logistic Regression dan SVM, evaluasi menggunakan validasi silang 5-fold, serta optimasi model melalui hyperparameter tuning dengan GridSearchCV [14][15]. Hasil penelitian ini diharapkan dapat memberikan wawasan tentang karakteristik distribusi HIV/AIDS antar wilayah WHO dan potensi penerapan machine learning dalam analisis data kesehatan global.

2. METODOLOGI PENELITIAN

2.1 Dataset dan Sumber Data

Dataset yang digunakan dalam penelitian ini adalah *HIV/AIDS Global Statistics* yang tersedia secara publik melalui platform Kaggle (https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset). Dataset ini terdiri dari 170 baris data yang merepresentasikan negara-negara di seluruh dunia, dengan 11 kolom fitur mencakup jumlah penerima ART yang dilaporkan, estimasi jumlah orang yang hidup dengan HIV, cakupan ART dalam persentase, serta nilai median, minimum, dan maksimum dari estimasi tersebut. Variabel target yang digunakan adalah kolom *WHO Region* yang terdiri dari 6 kelas: Africa, Europe, Americas, Eastern Mediterranean, Western Pacific, dan South-East Asia.:

2.2 Tahapan Penelitan

Penelitian ini dilakukan melalui lima tahapan utama yang saling berkesinambungan, sebagaimana digambarkan dalam alur berikut:.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

Gambar 1 *Tahapan Alur*

3. HASIL DAN PEMBAHASAN

3.1 Performa Model awal

Kedua model menghasilkan akurasi pada data uji yang identik, yaitu sebesar **0,3929** (39,29%). Meskipun angka ini tergolong rendah, kondisi tersebut dapat dipahami mengingat karakteristik dataset yang kompleks: distribusi kelas yang tidak seimbang, tumpang tindih nilai fitur antar wilayah, serta tingginya korelasi antar sebagian besar fitur.

Tabel 1 menyajikan perbandingan performa lengkap kedua model.

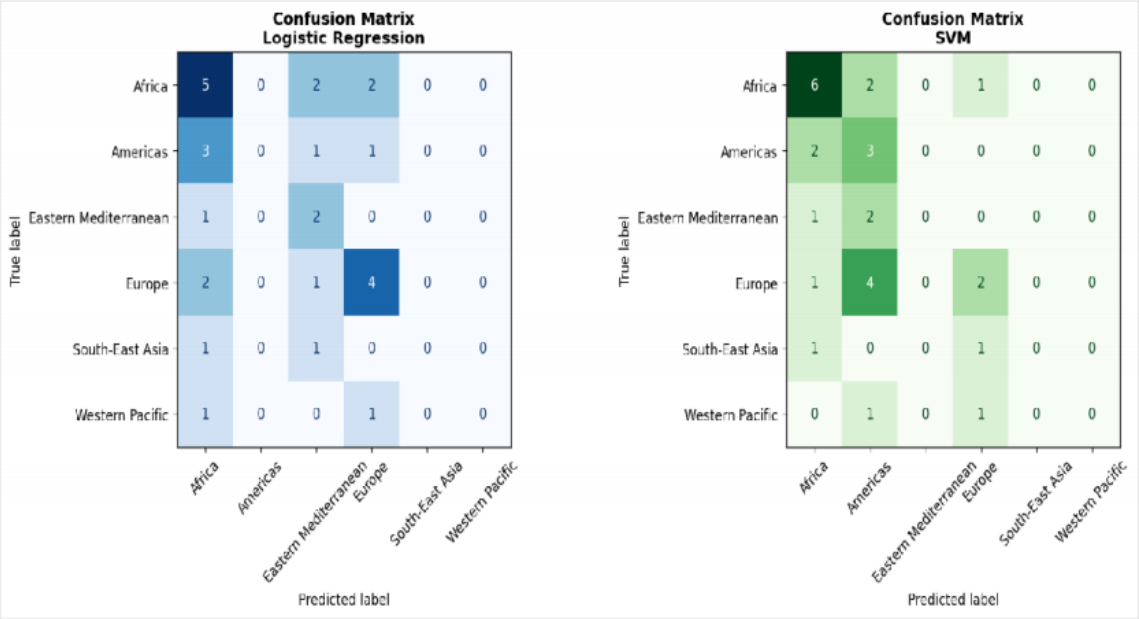
Model	Akurasi	Precision (W)	Recall (W)	F1-Score (W)	CV Mean	CV Std
Logistic Regression	0,3929	0,2792	0,3929	0,3223	0,4405	0,0722
SVM Dasar	0,3929	0,3200	0,3929	0,3392	0,3968	0,0522

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

SVM Tuned	0,3571	-	-	-	0,4905	-

3.2 Validasi Silang 5-Fold

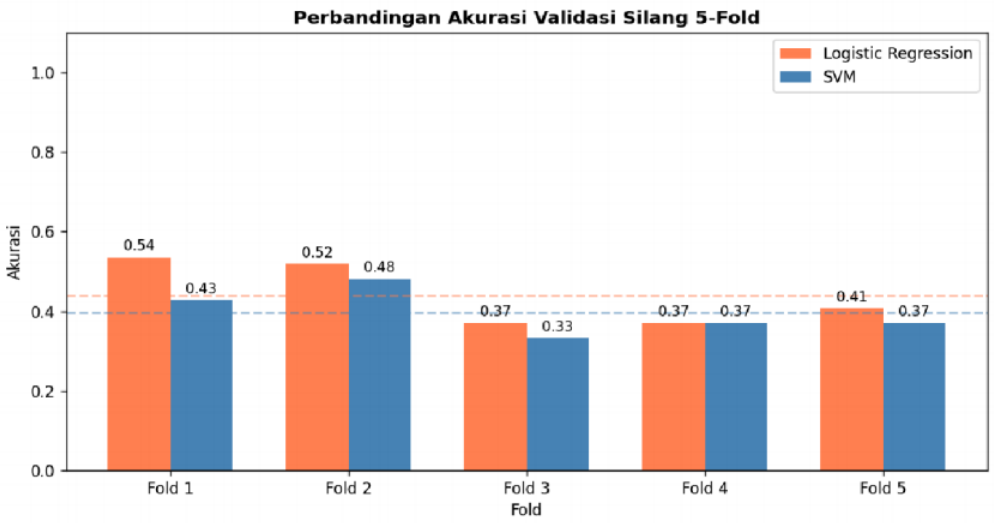
Hasil validasi silang menunjukkan bahwa Logistic Regression memiliki rata-rata akurasi yang lebih unggul (0,4405) dibandingkan SVM (0,3968). Secara per-fold, Logistic Regression menghasilkan nilai [0,5357; 0,5185; 0,3704; 0,3704; 0,4074], sementara SVM menghasilkan [0,4286; 0,4815; 0,3333; 0,3704; 0,3704]. Hampir seluruh kesalahan prediksi terjadi antara kelas-kelas yang memiliki karakteristik fitur serupa, sesuai temuan EDA yang mengungkapkan tumpang tindih nilai fitur antar wilayah.



Gambar 2 Confusion Matrix Logistic Regression dan SVM

Menampilkan dua confusion matrix berdampingan (6×6). Matriks kiri (Logistic Regression) menggunakan skema Warna biru, matriks kanan (SVM) menggunakan skema warna hijau. Diagonal utama menunjukkan prediksi benar. Terlihat baris South-East Asia dan Western Pacific memiliki nilai nol pada diagonal, artinya kedua kelas minoritas Tersebut tidak berhasil diprediksi dengan benar oleh model manapun.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

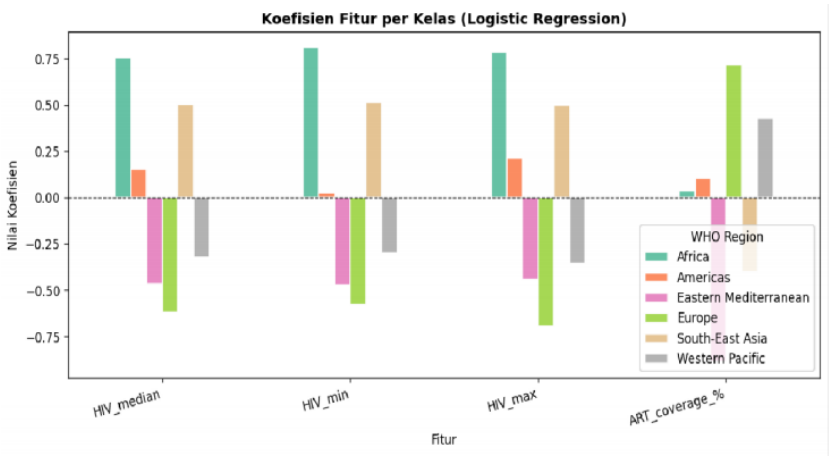


Gambar 3 *Perbandingan Akurasi Validasi Silang 5-Fold*

Menampilkan diagram batang kelompok (grouped bar chart) dengan 5 kelompok fold pada sumbu X dan nilai akurasi (0–1,0) pada sumbu Y. Bar merah mewakili Logistic Regression, bar biru mewakili SVM. Logistic Regression secara Konsisten lebih tinggi di Fold 1 (0,54 vs 0,43) dan Fold 2 (0,52 vs 0,48).

3.2 Interpretasi Koefisien dan Hyperparameter Tuning

Salah satu keunggulan Logistic Regression adalah kemampuannya untuk memberikan interpretasi model melalui nilai koefisien setiap fitur. Hasil analisis koefisien menunjukkan bahwa wilayah Afrika memiliki koefisien positif yang tinggi pada fitur-fitur HIV (sekitar 0,75–0,81), mengindikasikan bahwa jumlah penderita HIV yang besar merupakan karakteristik khas kawasan ini. Sebaliknya, wilayah Europe menunjukkan koefisien negatif pada fitur HIV namun positif pada ART_coverage_% (0,7171), yang mencerminkan cakupan pengobatan tinggi sebagai ciri khas Eropa. Eastern Mediterranean memiliki koefisien negatif pada semua fitur, menunjukkan bahwa wilayah ini sulit dibedakan hanya berdasarkan fitur yang tersedia.



Gambar 4 Koefisien Fitur per Kelas (Logistic Regression)

Menampilkan grouped bar chart horizontal atau vertikal dengan 4 kelompok fitur (HIV_median, HIV_min, HIV_max, ART_coverage_%) pada sumbu X dan nilai koefisien pada sumbu Y (rentang -1,0 hingga +1,0). Setiap bar

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

mewakili satu wilayah WHO dengan warna berbeda. Bar Africa Tertinggi positif pada fitur HIV, bar Europe tertinggi positif pada ART_coverage_%.

Untuk meningkatkan performa SVM, dilakukan optimasi hyperparameter menggunakan GridSearchCV dengan ruang pencarian: C ∈ {0,1; 1; 10; 100}, gamma ∈ {scale, auto, 0,01; 0,001}, dan kernel ∈ {rbf, linear}. Validasi silang 5-fold diterapkan dalam proses pencarian. Kombinasi parameter terbaik yang ditemukan adalah C=10, gamma='scale', kernel=rbf, Menghasilkan akurasi validasi silang sebesar 0,4905. Namun demikian, akurasi pada data uji turun Menjadi 0,3571, mengindikasikan adanya *overfitting* ringan. Secara keseluruhan, Logistic Regression tetap terbukti sebagai model yang lebih stabil dengan akurasi validasi silang tertinggi (0,4405).

4. KESIMPULAN

Penelitian ini berhasil membangun dan mengevaluasi dua model klasifikasi machine learning, yaitu Logistic Regression dan Support Vector Machine (SVM), untuk memprediksi wilayah WHO suatu negara berdasarkan data HIV/AIDS global. Berdasarkan seluruh proses eksperimen yang telah dilakukan, dapat ditarik beberapa kesimpulan utama. Pertama, kedua model menghasilkan akurasi data uji yang identik sebesar 39,29%, yang mencerminkan kompleksitas inherent dari dataset akibat ketidakseimbangan kelas dan tumpang tindih distribusi fitur antar wilayah. Kedua, Logistic Regression terbukti lebih unggul dan stabil dibandingkan SVM dasar, dengan rata-rata akurasi validasi silang 5-fold sebesar 0,4405 berbanding 0,3968. Ketiga, meskipun hyperparameter tuning berhasil meningkatkan akurasi validasi silang SVM menjadi 0,4905, model tersebut menunjukkan indikasi overfitting dengan akurasi data uji yang justru lebih rendah (0,3571). Keempat, fitur ART_coverage_% merupakan fitur yang paling berpengaruh dalam membedakan wilayah WHO, khususnya untuk membedakan wilayah Europe dari kawasan lainnya. Kelima, kelas minoritas seperti South-East Asia dan Western Pacific sama sekali tidak berhasil diprediksi dengan benar oleh kedua model, menegaskan bahwa penanganan ketidakseimbangan kelas merupakan prioritas utama untuk penelitian selanjutnya. Untuk peningkatan performa di masa mendatang, sangat direkomendasikan untuk menerapkan teknik SMOTE (*Synthetic Minority Oversampling Technique*) guna mengatasi ketidakseimbangan kelas, mengeksplorasi algoritma yang lebih kompleks seperti Random Forest atau Gradient Boosting, serta mempertimbangkan penambahan fitur dari sumber data eksternal lainnya agar model dapat belajar pola yang lebih representatif dari setiap wilayah WHO.

REFERENCES

[1] WHO, "HIV data and statistics," World Health Organization, 2025. [Online]. Available: https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics

[2] UNAIDS/WHO, "HIV statistics, globally and by WHO region, 2025," WHO Information Sheet, 2025. [Online]. Available: https://cdn.who.int/media/docs/default-source/hq-hiv-hepatitis-and-stis-library/who-ias-hiv-statistics_2025-new.pdf

[3] S. Rasheed et al., "Heart Disease Prediction Using GridSearchCV and Random Forest," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 10, Mar. 2024. https://doi.org/10.4108/eetpht.10.5523

[4] A. Khalid et al., "Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease," *Processes*, vol. 11, no. 3, p. 734, 2023. https://doi.org/10.3390/pr11030734

[5] S. Domínguez-Rodríguez et al., "Scalable and robust machine learning framework for HIV classification using clinical and laboratory data," *Scientific Reports*, 2025. https://doi.org/10.1038/s41598-025-00085-4

[6] B. Olatosi et al., "Machine Learning Approaches to Identify Communities with High HIV Prevalence in Resource-Limited Settings," *medRxiv*, 2025. https://doi.org/10.1101/2025.11.10.25339949

[7] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, p. 54, 2023. https://doi.org/10.3390/info14010054

[8] Z. Zhang et al., "Multi-Class Imbalanced Learning with Support Vector Machines via Differential Evolution," *arXiv*, 2025. https://arxiv.org/abs/2502.14597

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 908-914
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1140

[9] J. Hemmatian et al., "Addressing imbalanced data classification with Cluster-Based Reduced Noise SMOTE," *PLOS ONE*, 2025. https://doi.org/10.1371/journal.pone.0317396

[10] N. U. Maulidevi and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, 2022. https://doi.org/10.1016/j.jksuci.2020.02.004

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953

[12] T. Wongvorachan, "An Investigation of SMOTE Based Methods for Imbalanced Datasets With Data Complexity Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, 2023. https://doi.org/10.1109/TKDE.2022.3179381

[13] H. Abid et al., "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *Journal of Big Data*, Springer, 2024. https://doi.org/10.1186/s40537-024-00943-4

[14] P. M. Nyarige et al., "The successful scaling-up of antiretroviral therapy globally has many lessons for advancing universal health coverage," *BMC Global and Public Health*, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12860156/

[15] Global Burden of Disease 2017 HIV Collaborators, "Global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2017, and forecasts to 2030," *The Lancet HIV*, vol. 6, no. 12, pp. e831–e859, 2019. https://doi.org/10.1016/S2352-3018(19)30196-1