VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

HYBRID RESAMPLING METHOD AND HYPERPARAMETER OPTIMIZATION FOR HIV/AIDS PREDICTION: EVIDENCE FROM EIGHT MACHINE-LEARNING MODELS

Lydia Nur Sa'adah[1]; Fatkhurokhman Fauzi[1*]; Prizka Rismawati Arum[1]; M Al Haris[1]; Yan Nazala Bisoumi[1]

Department of Statistics[1]
Universitas Muhammadiyah Semarang, Semarang, Indonesia[1]
https://unimus.ac.id[1]
lydianursaadah@gmail.com, fatkhurokhmanf@unimus.ac.id*, prizka.rismawatiarum@unimus.ac.id, alharis@unimus.ac.id, ynazalabisoumi@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *HIV/AIDS remains a global health challenge with continuously increasing infection rates, highlighting the importance of accurate prediction models to support prevention and early detection. However, the development of such models is often constrained by class imbalance and irrelevant features. This study aims to improve HIV/AIDS infection prediction by integrating feature selection, data balancing techniques, and eight machine learning algorithms. Feature selection was performed using Mutual Information and Chi-Square to identify the most relevant features. The dataset used was the HIV/AIDS Infection Prediction Dataset from Kaggle, consisting of 2,139 instances and 23 features, with an imbalanced distribution of 1,618 non-infected and 521 infected cases. The dataset was divided into 80% training data and 20% testing data, with resampling applied only to the training set to prevent data leakage. Three resampling scenarios were evaluated: no sampling, SMOTE, and SMOTE-ENN. Hyperparameter tuning was conducted using Bayesian Optimization integrated with 5-fold Cross-Validation to improve model robustness and reliability. Eight machine learning algorithms were evaluated, including Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM, K-Nearest Neighbors, and Logistic Regression. The results show that SMOTE-ENN combined with hyperparameter optimization significantly improved model performance. The best model, Gradient Boosting + SMOTE-ENN, achieved 96.1% accuracy, 94.8% precision, 98.4% recall, and 96.5% F1-score. These findings indicate that the proposed integrated framework is highly effective for predicting HIV/AIDS infection and has strong potential to support early diagnosis and data-driven decision-making in healthcare.*

Keywords: *HIV/AIDS Prediction, Machine Learning Algorithm, SMOTE-ENN Method*

Intisari— *HIV/AIDS tetap menjadi tantangan kesehatan global dengan tingkat infeksi yang terus meningkat, yang menyoroti pentingnya model prediksi yang akurat untuk mendukung upaya pencegahan dan deteksi dini. Namun, pengembangan model semacam itu sering kali terhambat oleh ketidakseimbangan kelas dan fitur yang tidak relevan. Penelitian ini bertujuan untuk meningkatkan prediksi infeksi HIV/AIDS dengan mengintegrasikan seleksi fitur, teknik penyeimbangan data, dan delapan algoritma pembelajaran mesin. Seleksi fitur dilakukan menggunakan Mutual Information dan Chi-Square untuk mengidentifikasi fitur yang paling relevan. Dataset yang digunakan adalah HIV/AIDS Infection Prediction Dataset dari Kaggle, yang terdiri dari 2.139 contoh dan 23 fitur, dengan distribusi yang tidak seimbang antara 1.618 kasus tidak terinfeksi dan 521 kasus terinfeksi. Dataset tersebut dibagi menjadi 80% data pelatihan dan 20% data pengujian, dengan resampling diterapkan hanya pada set pelatihan untuk mencegah kebocoran data. Tiga skenario resampling dievaluasi: tanpa resampling, SMOTE, dan SMOTE-ENN. Penyesuaian hiperparameter dilakukan menggunakan Optimisasi Bayesian yang terintegrasi dengan Validasi Silang 5-lipat untuk*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533

meningkatkan ketahanan dan keandalan model. Delapan algoritma pembelajaran mesin dievaluasi, termasuk Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM, K-Nearest Neighbors, dan Regresi Logistik. Hasil menunjukkan bahwa kombinasi SMOTE-ENN dan optimisasi hiperparameter secara signifikan meningkatkan kinerja model pada semua metrik evaluasi. Kinerja terbaik dicapai oleh Gradient Boosting + SMOTE-ENN dengan akurasi 96,1%, presisi 94,8%, recall 98,4%, dan F1-score 96,5%. Temuan ini menunjukkan bahwa kerangka kerja terintegrasi yang diusulkan sangat efektif untuk memprediksi infeksi HIV/AIDS dan memiliki potensi yang kuat untuk mendukung diagnosis dini serta pengambilan keputusan berbasis data di bidang kesehatan.

***Kata Kunci**: Prediksi HIV/AIDS, Algoritma Machine Learning, Metode SMOTE-ENN.*

INTRODUCTION

HIV/AIDS remains a major cause of mortality among youth and adults worldwide, with an almost 100% fatality rate in many underdeveloped areas, drastically decreasing life expectancy over the last thirty years. According to the World Health Organization (WHO), more than 1.3 million new HIV infections were reported globally in 2022, with approximately 39 million individuals living with HIV infection worldwide [1], [2].

Human Immunodeficiency Virus (HIV) is an infectious disease that directly attacks the human immune system, gradually impairing the body's natural defense mechanism against pathogens. Without appropriate treatment, HIV infection may advance to Acquired Immunodeficiency Syndrome (AIDS), the most severe stage of the disease, characterized by profound immune system deterioration that increases susceptibility to opportunistic infections and severe medical complications [3].

In response to the growing number of HIV/AIDS cases and the complexity of its transmission patterns, data-driven approaches have become increasingly important in public health research. The use of machine learning offers an effective means to analyze large-scale epidemiological data, uncover hidden patterns, and develop predictive models that can assist in identifying potential HIV infections and improving early detection accuracy [4].

Given the global impact and the need for more accurate prediction systems, integrating machine learning approaches has become increasingly important for enhancing early diagnosis and prevention strategies. In this study, eight machine learning algorithms are implemented, namely Decision Tree, Random Forest, AdaBoost, XGBoost, LightGBM, Gradient Boosting, K-Nearest Neighbors, and Logistic Regression. The comparison of these eight algorithms enables a comprehensive evaluation of performance, robustness, and generalisation in predicting HIV/AIDS infection [5].

However, one of the major obstacles in developing reliable predictive models for HIV/AIDS is class imbalance, where the number of infected individuals is considerably smaller than the non-infected population. Such imbalance may bias learning algorithms toward the majority class, thereby reducing their effectiveness in identifying minority (positive) cases [6].

Therefore, a specialized approach is required to balance the data distribution effectively. One commonly adopted method is the Synthetic Minority Over-sampling Technique (SMOTE), which addresses imbalance by generating artificial samples for the minority class. Rather than duplicating existing data, SMOTE creates new synthetic instances through interpolation between neighboring minority samples, enabling models to better learn class decision boundaries. This mechanism improves sensitivity toward minority-class detection while reducing the risk of overfitting.

The traditional SMOTE method, although widely used, has notable limitations. It often leads to overgeneralization and increases the risk of generating noisy synthetic samples, especially near the class boundary, which can reduce classification accuracy and model generalization [7]. To overcome these weaknesses, this study employs the SMOTE-ENN (Synthetic Minority Over-sampling Technique – Edited Nearest Neighbors) approach. This hybrid technique integrates oversampling and undersampling processes to handle data imbalance more effectively [8].

SMOTE generates new minority-class samples, while the Edited Nearest Neighbors procedure removes noisy or irrelevant majority-class instances, resulting in cleaner training data and improved classification outcomes [8], [9]. Previous studies have demonstrated that SMOTE-ENN significantly enhances recall and overall predictive accuracy compared with single resampling techniques. By reducing noise and refining decision boundaries, this hybrid method improves model robustness and its ability to generalize to unseen datasets [11].

Several studies have further confirmed the advantages of hybrid resampling strategies compared to single balancing methods. One study reported that model accuracy improved from 86% prior to hybrid resampling to 89% after implementing SMOTE-ENN, indicating superior generalization performance across multiple models [12] In addition, a previous study combined SMOTE-ENN with a stacking ensemble framework for diabetes classification and achieved an accuracy of 97.3%, demonstrating the effectiveness of integrating hybrid data balancing with ensemble learning techniques in mitigating class imbalance and improving predictive performance [13].

Additionally, a Hybrid Bag-Boost Ensemble integrated with K-Means-SMOTE-ENN has been shown to enhance robustness, stability, and resistance to overfitting when applied to imbalanced and noisy medical datasets [14]. Beyond data balancing, recent studies have also highlighted the substantial contribution of feature selection in improving model performance, especially for medical prediction tasks. Empirical findings show that the application of feature selection techniques, such as Mutual Information and Chi-Square, can significantly improve classification accuracy compared to models trained using all available features.

As reported in the study [6], the training process without feature selection achieved the best accuracy of 89.6%, while with feature selection, the best accuracy value was obtained at 90.8%. Another study by [15] produced an accuracy of 85.1% data processing without using feature selection methods, while with feature selection, an accuracy of 88%. These findings confirm that strong predictive performance is not only influenced by resampling strategies such as SMOTE-ENN but also by the integration of effective feature selection methods and appropriate machine learning models. Despite these advancements, many existing studies still rely on default model parameters and lack rigorous evaluation procedures, which may lead to suboptimal performance and unreliable results [16].

To overcome the identified limitation, this study incorporates hyperparameter optimization using Bayesian Optimization combined with 5-fold cross-validation. This strategy ensures that every model undergoes evaluation under its optimal configuration while maintaining robustness and generalization across different data partitions. To further improve the explainability and transparency of the model, this research integrates Feature Importance and SHAP analysis (Shapley Additive Explanations).

This ensures that the model not only achieves high predictive performance but also remains explainable and clinically meaningful. In line with these research gaps, this study aims to systematically evaluate eight machine learning algorithms by integrating feature selection using Mutual Information and Chi-Square, applying the SMOTE-ENN balancing technique, and optimizing model performance through Bayesian Optimization with cross-validation. This comprehensive framework is expected to produce more accurate, stable, and interpretable prediction results, thereby supporting early detection of HIV infection and enabling more informed decision-making in healthcare management.

MATERIALS AND METHODS

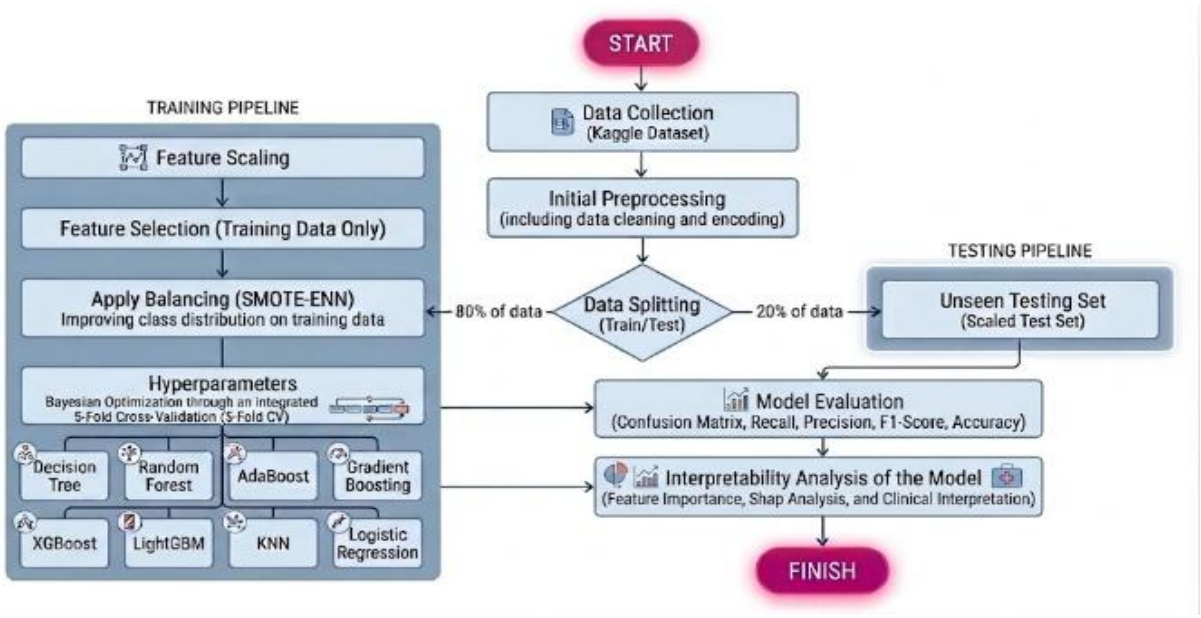
This study employs a structured methodological framework following standard data science procedures. The workflow consists of data collection, preprocessing, model development, and evaluation stages organized into training and testing pipelines. The dataset was obtained from Kaggle and underwent initial preprocessing to ensure data quality and consistency.

Within the training pipeline, feature scaling and feature selection were applied to improve model efficiency and reduce irrelevant variables. Data imbalance was addressed using the SMOTE-ENN technique, which enhances minority class representation while reducing noise. Multiple machine learning algorithms were then implemented, including Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM, KNN, and Logistic Regression.

Hyperparameter tuning was performed using Bayesian Optimization with 5-fold cross-validation to obtain optimal model performance. Finally, the models were evaluated on the test set using a confusion matrix, performance metrics, and strengthened by AUC-ROC to assess their effectiveness in detecting HIV/AIDS infection. A comprehensive illustration of the methodological workflow is provided in Figure 1.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533



Source : (Research Result, 2025)

Figure 1. Research Stage

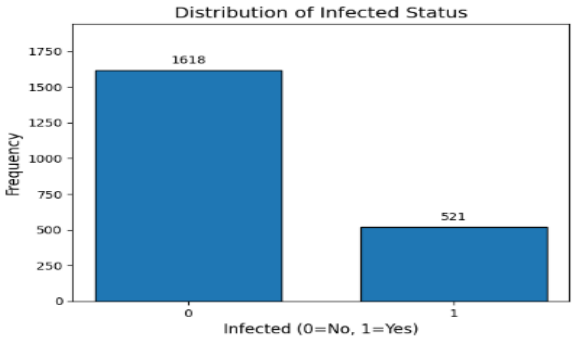
Data Collection

a) Load Data

In this study, the data used came from an open source, namely the Kaggle platform, entitled "HIV/AIDS Infection Prediction Dataset." This dataset contains individual data covering various behavioral and health condition variables relevant to the possibility of HIV/AIDS infection. The dataset has a total of 2139 rows and 23 features, with a target variable that is the 'infected' feature, which indicates an individual's HIV/AIDS infection status.

b) Target Column

Before further processing, the target column is analyzed to determine the distribution of classes. This column indicates whether a patient is infected with HIV/AIDS (1) or not (0). The analysis aims to detect class imbalance, which can bias machine learning models toward the majority class and reduce performance on the minority class.



Source : (Research Result, 2025)

Figure 2. Target Distribution

Therefore, the number of samples in each class is examined. If a significant imbalance is identified, resampling techniques such as oversampling, undersampling, or hybrid methods such as SMOTE-ENN are applied to balance the data [17]. The label distribution can be seen in Figure 2.

c) Feature

This dataset includes 23 key features grouped into several important categories. Demographic information and baseline data include variables such as (*time*) representing the duration of observation, the patient's age at the start of treatment (*age*), the patient's initial body weight (*wtkg*), *gender*, the patient's *race*, and the patient's sexual activity (*homo*). Medical history includes hemophilia status (*hemo*), history of intravenous drug use (*drugs*), the Karnofsky score, which measures the patient's overall health status (*karnof*), and history of antiretroviral therapy before the observation period (*aprior, str2, strat*).

Treatment information includes the type of care received by the patient (*trt*), the duration of previous antiretroviral therapy (*preanti*), indicators of treatment discontinuation (*offtrt*), and the type of therapy currently being received (*treat*). In the laboratory results category, the dataset includes data on the patient's CD4 and CD8 counts recorded at the start of treatment (*cd40, cd80*), z30, and after 20 weeks of treatment (*cd420, cd820*). Additionally, this dataset also records the patient's condition through clinical symptom indicators (*symptom*) and the patient's HIV infection status (*infected*). All these features comprehensively describe the patients'

demographic status, treatment history, and laboratory results, thereby supporting the development of more accurate predictive models for the possibility of HIV/AIDS infection.

Initial Preprocessing

Before the modeling stage, an extensive data preprocessing stage was implemented to guarantee data quality and consistency. Based on the initial data inspection, the dataset contains no missing values across all features. Furthermore, data transformation was applied through an encoding process specifically for categorical variables. All variables were converted into numerical representations to facilitate their use in machine learning algorithms, as illustrated in Table 1.

Table 1. Feature Encoding

Feature	Value
trt	0 = ZDV, 1 = ZDV + ddl, 2 = ZDV + Zal, 3 = ddl
hemo	0 = no, 1 = yes
homo	0 = no, 1 = yes
drugs	0 = no, 1 = yes
aprior	0 = no, 1 = yes
z30	0 = no, 1 = yes
race	0 = white, 1 = non-white
gender	0 = F, 1 = M
str2	0 = naïve, 1 = experienced
strat	1 = naïve, 2 = ≤52W therapy, 3 = >52W therapy
symptom	0 = asymp, 1 = symp
treat	0 = ZDV only, 1 = others
offtrt	0 = no, 1 = yes
infected	0 = no, 1 = yes

Source : (Research Result, 2025)

In contrast, numerical features—including age, wtkg, cd40, cd420, and cd820—were retained in their original continuous form without any encoding process. This approach preserves the original data distribution and ensures that the dataset is optimally prepared for subsequent modeling stages, thereby maintaining data integrity and minimizing potential bias.

Data Splitting

Following preprocessing, the dataset was split into training and testing subsets to enable a reliable evaluation of the classification models. The dataset was partitioned into 80% training data and 20% testing data. The training set is used to build the machine learning model, allowing it to learn patterns from the data. In contrast, the testing set (which has been normalized and standardized) is utilized to assess the model's performance on unseen data, thereby evaluating the model's accuracy and generalization capability [18].

Feature Scaling

After splitting the dataset into training and testing subsets, a feature scaling procedure was conducted to normalize feature values within a comparable range [19]. This process is necessary since variables in the dataset typically exhibit varying measurement scales. In this study, the scaling technique used is normalization or standardization.

a) Normalization

This technique transforms feature values to a defined range, usually between 0 and 1. Normalization is appropriate for data that do not conform to a normal (Gaussian) distribution [19]. A commonly used formula for normalization is:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X_{min} and X_{max} are the minimum and maximum values of that feature.

b) Standardization

This approach normalizes feature values to achieve a mean of 0 and a standard deviation of 1. Standardized is more suitable when the data follows a normal distribution or when the algorithm used assumes a Gaussian distribution [19]. The standardization formula is:

$$X_{standardized} = \frac{X - \mu}{\sigma} \quad (2)$$

where μ is the mean, and σ is the standard deviation of that feature.

Feature Selection

At this stage, several feature selection methods were applied, namely Mutual Information (MI) and Chi-Square Test (x^2). MI is used to analyze categorical features to capture non-linear relationships with the target variable and measures the degree of dependence between variables without assuming a linear relationship, making this method effective for handling discrete data [20].

Besides that, Chi-Square is applied to categorical features to evaluate whether the distribution of a feature's values differs significantly from the target class. Features with low Chi-Square values are considered to have a weak relationship with the target and can be eliminated. By combining these three methods, the feature selection process becomes more comprehensive and optimal in handling datasets consisting of both numerical and categorical variables [21]. In Chi-Square feature

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533

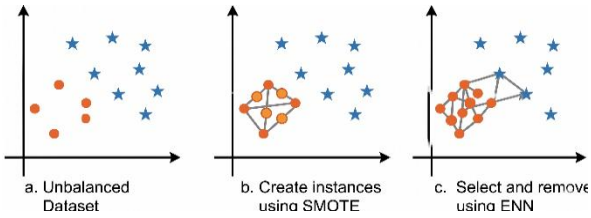
selection based on statistical theory, two events occur: the occurrence of a feature and the occurrence of a category. Each term's value is then ranked from highest to lowest based on calculations. The calculation is as follows:

$$x^2(D,t,c) = \sum_{et\in\{1,0\}} \frac{(N_{et\ el} - E_{et\ el})}{E_{et\ el}} \quad (3)$$

Balancing Data SMOTE-ENN

This imbalanced data can result in the model prioritizing the majority class. Therefore, before entering the modelling stage, a data balancing process is carried out using the SMOTE-ENN (Synthetic Minority Oversampling Technique – Edited Nearest Neighbours) method. The SMOTE, through the addition of synthetic samples to the minority class, and the ENN simultaneously remove majority data that is considered noisy or ambiguous, while the ENN improves data quality by removing inconsistent samples or those located in areas where classes overlap [22].

In this study, SMOTE was applied with the parameter k_neighbors = 5 to generate synthetic samples based on the proximity of data points within the minority class. Next, Edited Nearest Neighbors (ENN) was used to clean the data of noise. In this method, each sample in the dataset (whether from the majority or minority class) is evaluated based on its k nearest neighbors using the parameter n_neighbors = 3 [22]. If the class of *xmaj_i* is not the same as the class of any of the k nearest neighbors, then *xmaj_i* is removed. An overview of SMOTE-ENN performance is shown in Figure 3.



Source : (Catur Supriyanto, 2025)[13]

Figure 3. Hybrid Resampling Method

Hyperparameters

The hyperparameter tuning method used in this study is Bayesian Optimization (BO) with Cross-Validation (CV), which is recognized as an effective approach for optimizing hyperparameters in machine learning models. Bayesian Optimization uses a probabilistic surrogate model, specifically the Gaussian Process, to estimate the unknown objective function and efficiently explore the hyperparameter search space [23], [24]. In this study, BO is utilized to identify optimal hyperparameter configurations by iteratively updating the probabilistic model based on previous evaluation results. To ensure robust and reliable model evaluation, the hyperparameter optimization process in this study is integrated with 5-fold CV, where the training data is divided into five subsets. In each iteration, four subsets are used for training and one for validation, and the process is repeated until each subset has served as validation once. The final performance is obtained by averaging the results from all folds. Bayes' Theorem can be stated as follows:

$$P(A|B) = \frac{P(B|A),P(A)}{P(B)} \quad (4)$$

Component Definitions:

a) $P(A|B)$ (Posterior): The probability that hypothesis A is true given evidence B.
b) $P(B|A)$ (Likelihood): The probability of evidence B occurring, assuming hypothesis A is true.
c) $P(A)$(Prior): The initial probability of hypothesis A before any new evidence is available.
d) $P(B)$(Evidence): The overall probability of evidence B occurring.

Machine Learning Algorithm

This study performs classification using several machine learning methods, including DT, RF, AB, XGB, LightGBM, GB, KNN, and LR. Machine learning is the process of grouping data into specific categories or classes based on patterns learned from training data. In classification, machine learning algorithms learn the relationship between independent features (input variables) and the target (output variable) to build a predictive model [25]. This process is carried out after data preprocessing, data scaling, feature selection, data balancing, hyperparameter tuning, and model training using training data to identify patterns and relationships among variables.

Model Evaluation

After the model is trained, the final step is to evaluate performance using a testing set.

a) Confusion Matrix

The evaluation of the algorithm used is a confusion matrix, which can determine the results

in terms of accuracy, precision, and recall. More details can be seen in Table 2 below:

Table 2. *Confusion Matrix* Table

		Predict Class	
Actual Class	1	TP	FN
	0	FP	TN

Source : (Research Result, 2025)

Accuracy shows how well the model predicts the data as a whole, while precision and recall provide an overview of the model's performance on specific classes. Accuracy can be written in the formula 5:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision is a metric that quantifies the proportion of true positive predictions relative to the total number of positive predictions generated by the model, as expressed in formula 6:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall metric indicates the proportion of true positive cases that the model successfully captures from the total actual positives, as represented in the formula equation 7:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

The F1-score serves as an evaluation metric that reflects the balance between Precision and Recall, allowing us to assess how effectively a model performs in classification tasks. This metric provides insight into the model's overall capability to correctly identify and classify data. The F1-score can be expressed in the formula equation 8:

$$F1 = \frac{2 \times precision \times recall}{precision+recall} \quad (8)$$

b) AUC-ROC Curve

The AUC (Area Under the Curve) represents the area under the ROC curve and provides a single scalar value summarizing the model's classification performance. The AUC value ranges from 0 to 1, where a higher value indicates better discriminative ability. Interpretation of AUC values is generally categorized as follows:

1) 0.6 – 0.7→ Weak
2) 0.7 – 0.8→ Moderate
3) 0.8 – 0.9→ Strong
4) > 0.9→ Highly effective

A key strength of AUC–ROC is its threshold independence, allowing comprehensive performance evaluation across decision boundaries, which is particularly important in medical applications where minimizing false negatives is critical [8].

RESULTS AND DISCUSSION

This section discusses the results and analysis of the study, encompassing all stages from data preparation and feature selection to the evaluation of eight machine learning algorithms.

Data Splitting

The dataset is divided into training and testing sets using an 80:20 ratio. This partitioning is performed at the outset to prevent bias and avoid data leakage. The training data is then used for further preprocessing steps, including feature selection and data balancing techniques, while the testing data is used to evaluate the performance of the machine learning model. The results of the data split are presented in Table 3.

Table 3. Data Splitting

Class	Training (80%)	Testing (20%)	Count
0 (Majority)	1294	324	1618
1(Minority)	417	104	521
Count	1711	428	2139

Source : (Research Result, 2025)

Feature Selection

Based on the combined results of these feature selection methods, nine main features were identified as the most significant in influencing the prediction of HIV/AIDS infection, as shown in Table 4.

Table 4. Result of Feature Selection

Feature	Mutual Information	Chi-Square (X^2)
time	0.228977	57.3524
symptom	0.125000	29.5446
cd420	0.098667	15.5174
offtrt	0.014865	15.2346
strat	0.016682	15.1489
str2	0.014917	13.4894
z30	0.016202	11.7507
treat	0.025808	8.92998
preanti	0.002240	7.16752

Source : (Research Result, 2025)

Table 4 presents the results of feature selection. An MI value > 0 indicates dependence, and the greater the MI value, the stronger the non-linear relationship between the feature and the prediction

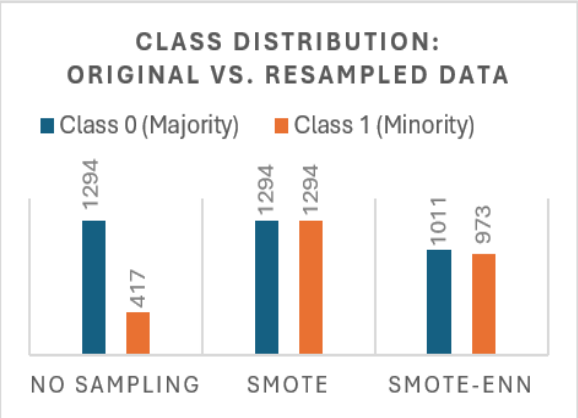
Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533

result. Several features exhibit low MI values, even approaching zero. However, these features were retained because they had high and statistically significant chi-square values ($\chi^2 > 6.64$). This value exceeds the critical chi-square limit at p = 0.01 for df = 1, indicating that the distribution of the categories significantly influences HIV/AIDS infection status. This is in line with the literature stating that MI and X^2 measure different types of dependencies, so features that do not exhibit non-linear dependencies can still be relevant if they demonstrate the significance of the category to the target.

Balancing Data

The dataset in this study exhibits class imbalance between the majority class (0) and the minority class (1). In the SMOTE method, the parameter k_neighbors with a value of k = 5 was chosen because it maintains a balance between data diversity and local proximity among samples. Meanwhile, in the Edited Nearest Neighbors (ENN) stage, a value of n_neighbors = 3 was chosen because it is sufficiently sensitive in detecting local inconsistencies without removing too much data. This approach allows for the formation of synthetic samples that remain within the relevant feature space while eliminating noise samples near the class boundaries. Three scenarios were compared: no resampling, SMOTE, and SMOTE-ENN, with the comparison results presented in Figure 4.



Source : (Research Result, 2025)

Figure 4. Distribution Resampled Data

Figure 4 confirms that the dataset exhibits a significant class imbalance, with the majority class (0) containing 1,294 samples and the minority class (1) consisting of only 417 samples. Following the balancing process, the standard SMOTE method equalizes the distribution by increasing the minority class to 1,294 samples, matching the original majority count. In contrast, the SMOTE-ENN hybrid approach produces a cleaner and more balanced distribution, producing 1,011 samples for the majority class and 973 samples for the minority class. This indicates that while standard SMOTE focuses on full numerical equalization, SMOTE-ENN effectively balances the classes while simultaneously removing noisy or ambiguous data points to improve the overall quality of the dataset.

Hyperparameters

The optimization results indicate that model performance varies across iterations, confirming the strong influence of hyperparameter selection. This approach ensures that the model does not rely on default parameters, but instead uses a configuration that has been systematically optimized using Bayesian optimization with 5-fold Cross-Validation to obtain the most optimal configuration. However, the parameter range to be optimized must be determined in advance. The parameter range used in this study was derived from a literature review on machine learning. The optimal hyperparameter configuration is shown in Table 5.

Table 5. Hyperparameters Configuration

Model	Range Hyperparameter to be optimized	Optimal Parameters
DT	max_depth = (3, 20) min_samples_leaf = (1, 10) min_samples_split = (2, 20) max_leaf_nodes = (10, 100)	max_depth = 6 min_samples_leaf = 8 min_samples_split =18 max_leaf_nodes = 29
RF	n_estimators = (50, 300) max_depth = (3, 20) min_samples_split = (2, 20)	n_estimators = 68 max_depth = 18 min_samples split = 14
XGB	learning_rate = (0.1,0.3,0.5) max_depth = (4, 6, 9) n_estimators = (100, 1000)	learning_rate = 0.3 max_depth = 9 n_estimat = 1000
AB	n_estimators = (50, 200) learning_rate = (0.01, 1)	n_estimators = 156 learning_rate = 0.03
GB	learning_rate = (0.01, 0.3) n_estimators = (50, 300) max_depth = (3, 10) subsample = (0.5, 1.0) min_samples_split = (2, 20)	learning_rate = 0.07 n_estimators = 139 max_depth = 5 subsample = 0.64 min_samples_split = 10
LGBM	learning_rate = (0.01, 0.3) num_leaves = (20, 150)	learning_rate = 0.03 num_leaves = 43

Model	Range Hyperparameter to be optimized	Optimal Parameters
KNN	n_neighbors = (3, 20) p = (1 'manhattan', 2'euclidean') weights = ('uniform', 'distance')	n_neighbors = 7 p = 2 'euclidean' weights 'distance'
LR	C = (0.001, 10) max_iter = (100, 500) penalty = ('l1', 'l2')	C = 6.119 max_iter = 155 penalty = l1

Source : (Research Result, 2025)

Machine Learning Algorithm

At this stage, a classification process is carried out using eight machine learning algorithms that are run with no sampling, SMOTE, and hybrid resampling SMOTE-ENN. The accuracy, precision, and recall results without sampling are presented in Tables 6, 7, and 8.

Table 6. Model Evaluation Without Sampling

Algorithm	Accuracy	Precision	Recall	F1-score
DT	86.2%	83.5%	53.8%	65.4%
RF	79.7%	80.6%	75.9%	70.0%
XGB	89.9%	83.5%	73.0%	68.0%
AB	89.4%	83.8%	75.0%	64.0%
GB	88.5%	78.3%	73.0%	67.8%
LGBM	78.5%	71.6%	68.2%	69.1%
KNN	78.5%	62.0%	29.8%	40.2%
LR	75.7%	66.7%	59.6%	67.0%

Source : (Research Result, 2025)

Table 6 shows that the model's performance without data balancing techniques remains relatively low, particularly for the recall and F1-score metrics. The findings indicate that the model is biased toward the majority class and is less effective at detecting cases in the minority class. This confirms that class imbalance significantly affects classification performance.

Table 7. Model Evaluation With SMOTE

Algorithm	Accuracy	Precision	Recall	F1-score
DT + SMOTE	87.1%	70.9%	79.8%	75.1%
RF + SMOTE	89.2%	75.8%	81.7%	78.7%
XGB + SMOTE	88.5%	76.5%	75.9%	76.3%
AB + SMOTE	89.9%	80.1%	77.8%	79.0%
GB + SMOTE	89.9%	79.6%	78.8%	79.2%
LGBM + SMOTE	88.5%	76.1%	76.9%	76.5%
KNN + SMOTE	70.0%	41.8%	59.6%	49.2%
LR + SMOTE	84.1%	67.6%	66.3%	66.9%

Source : (Research Result, 2025)

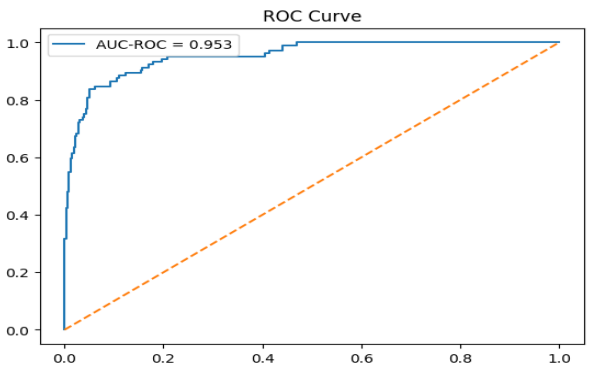
Table 7 shows an improvement in model performance following the application of SMOTE, particularly in terms of recall. This indicates that oversampling helps the model better identify the minority class. However, this improvement is not yet optimal because there is still a possibility that the generated synthetic data is not sufficiently representative, causing some models to fail to demonstrate consistent performance across all metrics.

Table 8. Model Evaluation With SMOTE-ENN

Algorithm	Accuracy	Precision	Recall	F1-score
DT + SMOTE-ENN	93.3%	91.8%	96.5%	94.1%
RF + SMOTE-ENN	94.9%	93.3%	97.7%	95.5%
XGB + SMOTE-ENN	95.2%	95.3%	96.2%	95.7%
AB + SMOTE-ENN	95.2%	93.9%	97.7%	95.8%
GB + SMOTE-ENN	96.1%	94.8%	98.4%	96.5%
LGBM + SMOTE-ENN	93.8%	92.7%	96.5%	94.5%
KNN + SMOTE-ENN	93.0%	92.0%	95.5%	93.8%
LR +SMOTE-ENN	94.5%	95.5%	94.6%	95.0%

Source : (Research Results, 2025)

Table 8 presents the evaluation results of the classification models following the application of hybrid resampling (SMOTE-ENN), which show that all models exhibit significant improvements across various evaluation metrics. These improvements demonstrate the effectiveness of SMOTE-ENN in enhancing the representation of minority classes and reducing noise, as well as the contribution of Bayesian Optimization integrated with 5-fold CV to achieve optimal performance. Among all evaluated models, the GB + SMOTE-ENN model achieved the best overall performance, with an accuracy of 96.1%, precision of 94.8%, recall of 98.4%, and an F1 score of 96.5%. In addition to matrix evaluation, model performance was also evaluated using the AUC–ROC curve to measure class discrimination ability is presented in Figures 5.



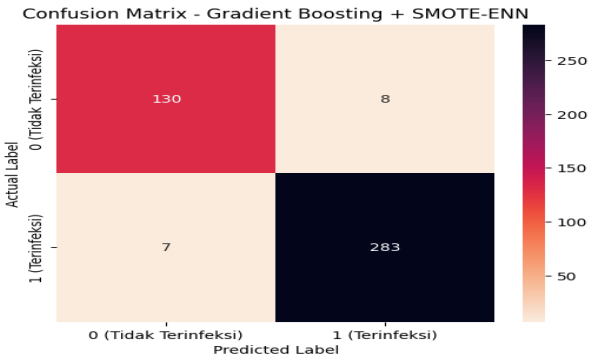
Source : (Research Result, 2025)

Figure 5. ROC curve

The best model (Gradient Boosting + SMOTE-ENN) achieved an AUC of 0.953, demonstrating excellent classification performance and a strong capacity to separate HIV-positive from HIV-negative cases. The Confusion Matrix is shown in Figure 6.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533



Source : (Research Result, 2025)

Figure 6. Confusion Matrix for Best Model

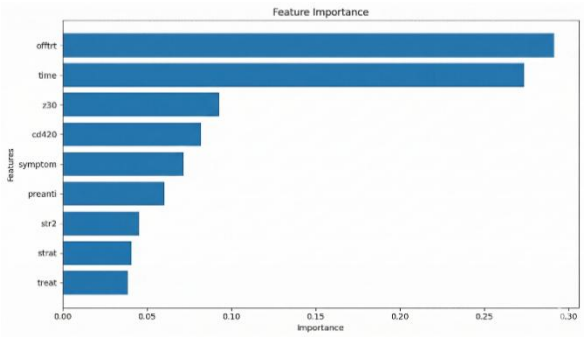
Figure 6 demonstrates that the Gradient Boosting model optimized using the SMOTE-ENN method is capable of classifying both classes very well. Based on the Confusion Matrix, for class 0 (uninfected), the model correctly classified 130 data points as True Negatives, and only 8 data points were incorrectly predicted as False Positives. Meanwhile, for class 1 (infected), the model was able to correctly identify 283 data points as True Positives, with a very low error rate only 7 data points were misclassified as False Negatives, demonstrating excellent detection capability for positive cases. Overall, the combination of high recall, good specificity, and a strong AUC value indicates that this model has robust performance and is reliable for use in classification tasks, particularly in the medical field.

Interpretability Analysis of the Model

In medical applications, model interpretability is a critical requirement to ensure that predictive outcomes can be understood, validated, and trusted by healthcare stakeholders.

a) Feature Importance Analysis

The feature importance analysis highlighted the most significant variables affecting HIV/AIDS infection prediction, as illustrated in Figure 7.



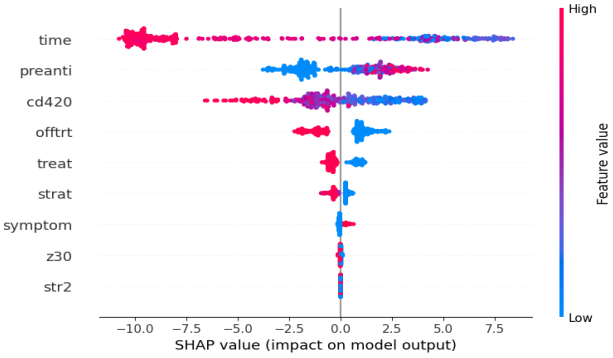
Source : (Research Result, 2025)

Figure 7. Feature Importance

Figure 7, the feature importance results indicate that treatment-related variables, particularly offtrt (treatment discontinuation) and time (duration of observation), are the most influential predictors in the model. This finding suggests that both the continuity of treatment and the length of observation play a critical role in determining patient outcomes. In addition, laboratory variables, such as cd420, show moderate importance, highlighting the relevance of immune system indicators in assessing disease progression. Similarly, preanti (duration of prior antiretroviral therapy) and z30, indicating that prior treatment exposure also influences prediction outcomes. Conversely, variables related to demographic characteristics (e.g., age, gender, race) and certain aspects of medical history (e.g., hemophilia status, drug use history) demonstrate relatively lower importance.

b) SHAP Analysis

In addition to global feature importance, SHAP (SHapley Additive exPlanations) was employed to provide a more detailed and consistent interpretation of model predictions. SHAP visualization is shown below in Figure 8.



Source : (Research Result, 2025)

Figure 8. SHAP Visualization

Figure 8 shows a SHAP summary plot illustrating the contribution of each feature to the prediction of HIV/AIDS infection. Each point represents a patient, where a positive SHAP value indicates a prediction of infection (1), while a negative value indicates no infection (0). The visualization results show that the features time, preanti, and cd420 are the most influential variables. The time variable indicates that a longer monitoring duration improves the accuracy of determining infection status, while cd420 (CD4 cell count) serves as a primary indicator of immune status, where low values are associated with an increased risk of infection.

Additionally, treatment-related variables such as preanti, treat, and offtrt indicate that the history and status of antiretroviral therapy significantly influence the prediction results. Other variables such as symptom, strat, and z30 have a smaller contribution but still support the classification process.

c) Clinical Interpretation

The integration of feature importance and SHAP analysis enables meaningful clinical interpretation:

1) Time demonstrates a strong influence, indicating that patient monitoring duration significantly affects HIV/AIDS progression.
2) Preanti highlights the importance of prior antiretroviral therapy history in risk prediction.
3) Cd420 remains a key immunological indicator of disease progression.
4) Offtrt and treat emphasize the critical role of treatment status and adherence in determining patient outcomes.
5) Strat and symptom provide additional clinical context with moderate contribution. Also, Z30 and str2 show relatively low importance.

From a clinical perspective, the dominance of treatment-related variables, particularly offtrt and preanti, underscores the importance of therapy adherence and treatment history in predicting HIV/AIDS outcomes. The significant role of time reflects the progressive nature of the disease. This suggests that, within this model, clinical and treatment-related factors contribute more substantially to the model's predictive performance than baseline demographic characteristics. Overall, the integration of feature importance and SHAP analysis provides a comprehensive understanding of applicability in clinical decision-making and early detection of HIV/AIDS.

CONCLUSION

This study evaluates eight machine learning algorithms for predicting HIV/AIDS infections. The results show that class imbalance significantly affects model performance. The application of the SMOTE-ENN technique effectively addresses this issue by enhancing the representation of the minority class while removing noisy data in the majority class, resulting in a dataset that is not only balanced but also more representative. Additionally, hyperparameter tuning using Bayesian optimization and 5-fold cross-validation significantly improves model performance by identifying optimal parameter configurations. Among all evaluated models, Gradient Boosting + SMOTE-ENN achieved the best overall performance, with an accuracy of 96.1%, precision of 94.8%, recall of 98.4%, and an F1-score of 96.5%. Overall, this study confirms that the application of data balancing, hyperparameter optimization, and machine learning algorithms can significantly improve predictive performance. The proposed model demonstrates strong potential for the early detection of HIV/AIDS and can support more accurate and reliable decision-making in the healthcare sector.

REFERENCE

[1] A. Aryani, Widiyono, and A. Anitasari, "Gambaran Pengetahuan Remaja Tentang Penyakit Hiv/Aids," *J. Ilmu Keperawatan*, vol. 14, no. 2, pp. 44–50, 2021, doi: https://doi.org/10.47942/jiki.v14i2.794.

[2] World Health Organization, "HIV/AIDS Fact Sheets," WHO, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hiv-aids. [Accessed: October 25, 2025].

[3] Kementerian Kesehatan, "HIV/AIDS," Kemenkes, 2023. [Online]. Available: https://ayosehat.kemkes.go.id/topik-penyakit/hivaids--ims/hiv. [Accessed: October 26, 2025].

[4] Kurniawati Yenni, "Pengaruh Tingkat Pendidikan Dengan Kejadian HIV/AIDS," *J. Bidan Pint.*, vol. 3, no. 2, 2022.

[5] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *J. Inf. dan Teknol.*, vol. 5, no. 3, pp. 58–64, Sep. 2023, doi: 10.60083/jidt.v5i3.393.

[6] A. M. A. Rahim, A. Ridwan, B. P. Hartato, and F. Asharudin, "Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 338–347, Mar. 2025, doi: 10.30871/jaic.v9i2.9125.

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

[8] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explor. Newsl.*,

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7533

vol. 6, no. 1, pp. 20–29, Jun. 2024, doi: 10.1145/1007730.1007735.

[9] T. Liao, H. Chen, C. Song, C. Huang, Y. Wu, and Z. Xu, "Using machine learning models to predict the duration of the recovery of COVID- patients hospitalized in Fangcang shelter hospital during the Omicron BA .," *Front. Med*, vol. 9, no. 2, 2022, doi: https://doi.org/10.3389/fmed.2022.1001801.

[10] M. Izraiq *et al.*, "Impact of Diabetes Mellitus on Heart Failure Patients : Insights from a Comprehensive Analysis and Machine Learning Model Using the Jordanian Heart Failure Registry [Corrigendum]," International Journal of General Medicine, vol. 17, pp. 3371–3372, 2024, doi: 10.2147/IJGM.S487404.

[11] R. S. Abdulsadig and E. Rodriguez-villegas, "imbalance mitigation when physiological signals," no. March, pp. 1–11, 2024, doi: 10.3389/fdgth.2024.1377165.

[12] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, p. 37, Jan. 2025, doi: 10.3390/a18010037.

[13] N. K. Majid, C. Supriyanto, and A. Marjuni, "Peningkatan Keberagaman Data untuk Klasifikasi Penyakit Diabetes Berbasis Stacking Ensemble Learning," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 1–10, 2025, doi: 10.30591/jpit.v10i1.7375.

[14] A. Puri and M. Kumar Gupta, "Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data," *Comput. J.*, vol. 65, no. 1, pp. 124–138, Jan. 2022, doi: 10.1093/comjnl/bxab039.

[15] I. Pratama *et al.*, "Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN," pp. 38–49, 2021, doi: 10.30864/eksplora.v11i1.578.

[16] I. Maulana and S. Ernawati, "Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan Metode Ensemble Softvoting Dengan SMOTE-ENN dan Optimasi Bayesian," vol. 13, no. 1, pp. 71–86, 2025, doi: https://doi.org/10.31294/evolusi.v13i1.8267.

[17] A. S. Fatih Gurcan, "Learning from Imbalanced Data : Integration of Advanced Resampling Techniques and Machine Learning Models for," *Cancer Res. Care*, vol. 16, no. 19, 2024, doi: https://doi.org/10.3390/cancers16193417.

[18] 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," *Indones. J. Comput. Sci.*, vol. 12, no. 5, Nov. 2023, doi: 10.33022/ijcs.v12i5.3413.

[19] E. Saputra and E. R. Susanto, "Implementation of Deep Learning with Multilayer Perceptron (MLP) for Heart Disease Prediction Using the SMOTE-ENN Technique," vol. 9, no. 3, pp. 1034–1041, 2025, doi: doi: 10.1007/s11053-021-09973-8.

[20] I. Saputra, "Pengkategorian Data Angket Mahasiswa Dengan Mutual Information Dan K-Nearest Neighbor," *Pros. Seniati*, no. 1, pp. 28–35, 2019, doi: https://doi.org/10.36040/seniati.v5i1.320.

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

[22] J. Pardede and D. P. Pamungkas, "The Impact of Balanced Data Techniques on Classification Model Performance," vol. 11, no. 2, pp. 401–412, 2024, doi: 10.15294/sji.v11i2.3649.

[23] C. W. Oei *et al.*, "Explainable Risk Prediction of Post-Stroke Adverse Mental Outcomes Using Machine Learning Techniques in a Population of 1780 Patients," vol. 23, no. 18, pp. 1–12, 2023, doi: https://doi.org/10.3390/s23187946.

[24] M. A. Amou, K. Xia, S. Kamhi, and M. Mouhafid, "A Novel MRI Diagnosis Method for Brain Tumor Classification Based on CNN and Bayesian Optimization," vol. 10, no. 3, pp. 1–21, 2022, doi: https://doi.org/10.3390/healthcare10030494.

[25] 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.