Computer Science (CO-SCIENCE)
Volume 6 Issue 2 July 2026
Accreditation **Sinta 3** No. SK : 295/C/C3/KPT/2026

Feature Comparison of RN-SMOTE and DBSCAN-SMOTE for Handling Imbalanced Data

Rias Akmal Sembiring[1], Wawan Gunawan[2*]

[1,2] Informatics, Mercu Buana University Jakarta
Jakarta, Indonesia

e-mail: [1]41523010227@student.mercubuana.ac.id, [2*]wawan.gunawan@mercubuana.ac.id

(*) Corresponding Author

Article Info: Received: 26-02-2026 | Revised : 11-07-2026 | Accepted :16-07-2026

Abstracts - HIV classification using machine learning is often challenged by severe class imbalance, which may reduce predictive reliability and bias classification models toward the majority class. This study aims to compare the effectiveness of two hybrid oversampling techniques, namely SMOTE-DBSCAN and RN-SMOTE, in improving HIV classification performance using an Indonesian HIV dataset. The dataset was obtained from a Non-Governmental Organization (NGO) and initially consisted of 707,389 records with 90 attributes. After data preprocessing, 148,408 validated records with 30 attributes were retained, comprising 135,739 non-reactive and 12,668 reactive HIV cases, indicating a substantial class imbalance. Two feature configurations derived from previous HIV studies were evaluated using SMOTE-DBSCAN and RN-SMOTE. To prevent data leakage, the dataset was partitioned using a stratified 80:20 train-test split before oversampling, while model validation was performed using 10-fold cross-validation on the training data. The balanced datasets were subsequently evaluated using eight classification algorithms. Experimental results demonstrate that RN-SMOTE consistently outperformed SMOTE-DBSCAN across both feature configurations, while Random Forest achieved the best overall predictive performance. Although SMOTE-DBSCAN required less computational time during the balancing process, RN-SMOTE produced more robust classification performance by generating cleaner and more representative minority-class samples. These findings demonstrate the effectiveness of noise-aware hybrid oversampling strategies for improving HIV classification on highly imbalanced datasets and provide practical insights for developing reliable machine learning models to support HIV surveillance and public health decision-making.

Keywords: HIV classification, class imbalance, hybrid oversampling, RN-SMOTE, machine learning.

INTRODUCTION

HIV/AIDS remains a critical public health issue because the infection weakens the immune system and increases the risk of severe complications that may threaten patients' survival. The complexity of HIV transmission patterns requires the application of data-driven analytical approaches to support surveillance systems and more accurate health decision-making (Sui et al., 2025). In Indonesia, variations in social and geographical conditions influence the distribution of HIV cases, making data-based analysis necessary to understand the characteristics of affected populations (Arifin et al., 2023). Machine learning approaches have demonstrated strong potential in improving the accuracy of HIV prediction, particularly when combined with feature selection and data balancing techniques (Rahim et al., 2025).

A major challenge in developing HIV predictive models is data imbalance, where the number of samples in the minority class is significantly smaller than that of the majority class, which may lead to classification bias (Gunawan et al., 2024; Nisa et al., 2023). Such imbalanced data conditions can reduce the model's ability to correctly identify critical cases within the minority class (Kosolwattana et al., 2023). The Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely used methods to address this issue by generating synthetic samples for the minority class (Chawla et al., 2018). However, conventional SMOTE has limitations because it may produce synthetic data containing noise and increase class overlap (Arafa et al., 2022).

Due to the limitations of standard SMOTE, several enhanced oversampling methods have been introduced by integrating clustering and noise reduction approaches to improve the quality of balanced data(Kosolwattana et al., 2023). RN-SMOTE is a method that utilizes the DBSCAN algorithm to detect and remove noise prior to the

oversampling process, resulting in a more representative data distribution (Arafa et al., 2022). Other approaches, such as SMOTE-ENN, which combine oversampling with data cleaning procedures, have also been reported to reduce data complexity and improve classification performance (Riantika et al., 2024). In addition, deep learning–based noise reduction strategies have shown potential in enhancing model stability when dealing with imbalanced datasets (Arafa et al., 2023).

In HIV/AIDS research, selecting appropriate features is a crucial aspect because sociodemographic variables play a significant role in explaining disease distribution patterns (Omrani et al., 2026). Studies in Indonesia reveal disparities in HIV-related knowledge across regions that are closely associated with demographic and geographical factors (Arifin et al., 2023). Predictive HIV models that integrate sociodemographic features with data balancing techniques have been shown to significantly improve classification performance (Rahim et al., 2025). Nevertheless, the effectiveness of each data balancing method may vary depending on dataset characteristics and the epidemiological context of the study (Nisa et al., 2023).

Although numerous SMOTE-based oversampling techniques have been proposed, studies directly comparing RN -SMOTE and DBSCAN - SMOTE remain limited, particularly in the context of HIV classification. Previous research has primarily focused on improving the performance of individual balancing methods or evaluating them within specific epidemiological settings rather than conducting a systematic comparison under identical experimental conditions (Arafa et al., 2022; Kosolwattana et al., 2023; Sandiwarno et al., 2025). Moreover, findings reported in different countries cannot be directly generalized to the Indonesian HIV population because HIV epidemiology is strongly influenced by variations in demographic characteristics, behavioral risk factors, healthcare accessibility, and disease surveillance systems, all of which affect data distribution and feature relevance. Consequently, evaluating balancing techniques using an Indonesian HIV dataset provides a more representative assessment of their effreectiveness within the local epidemiological setting. Building upon this gap, the present study systematically compares RN-SMOTE and DBSCAN-SMOTE using two feature configurations derived from previous HIV studies to examine their effectiveness in improving classification performance. In addition, the proposed evaluation employs multiple classification algorithms to provide a comprehensive assessment of predictive performance and computational efficiency, thereby contributing empirical evidence to the application of hybrid oversampling techniques for HIV classification in Indonesia (Rahim et al., 2025).

RESEARCH METHOD

1. Research Stages

The research stages describe the systematic workflow adopted to develop and evaluate the proposed HIV classification model, as illustrated in Figure 1. The workflow comprises problem formulation, data collection, exploratory data analysis, data preprocessing, feature configuration, train-test data partitioning, data balancing, model training, validation, and performance evaluation. This study was initiated by identifying the need for a robust classification model capable of predicting HIV status from an imbalanced dataset in Indonesia. The research objective was established to investigate the effectiveness of hybrid oversampling techniques in improving classification performance while maintaining reliable predictive capability. Clearly defining the research objective provides a structured analytical framework for subsequent data processing and model development (Syaifudin et al., 2025).

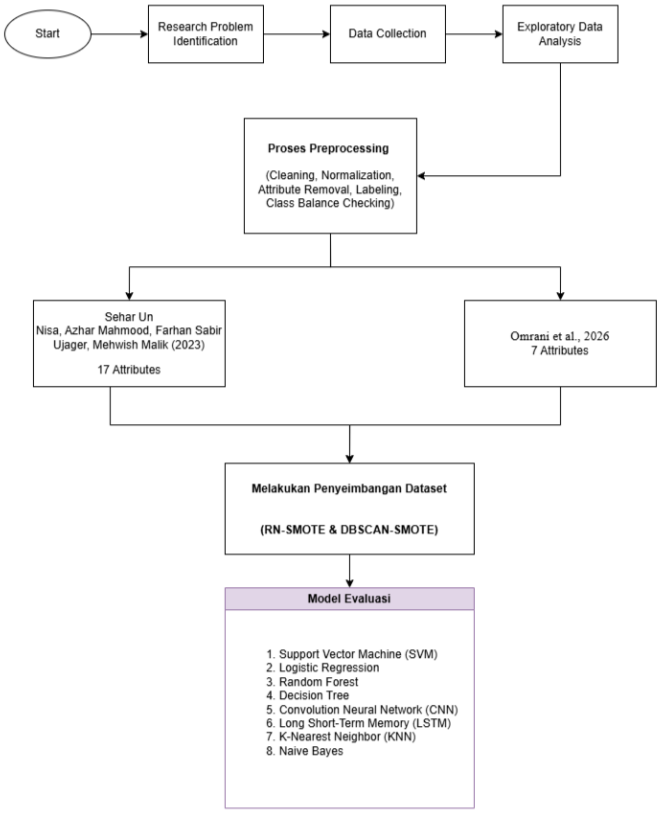
The dataset was obtained from a Non-Governmental Organization (NGO) and comprised HIV screening records collected between 2021 and 2022. The original dataset contained 707,389 records and 90 attributes, representing demographic, behavioral, and clinical information associated with HIV testing. Before model development, Exploratory Data Analysis (EDA) was conducted to examine the dataset characteristics, including data types, attribute distributions, and potential inconsistencies. Univariate and multivariate analyses supported by data visualization were performed to identify variable distributions, relationships among attributes, and potential anomalies that could influence subsequent analytical processes (Syaifudin et al., 2025).

Following EDA, data preprocessing was performed to prepare the dataset for machine learning analysis. Attributes that were not relevant to the objectives of this study were removed according to the variables reported in five previous HIV studies. Subsequently, categorical variables were transformed into numerical representations, incomplete records were eliminated, and the target variable was standardized to ensure data consistency. The distribution of the target classes was then examined to identify the presence of class imbalance before model development. These preprocessing procedures improve data quality and establish a reliable foundation for predictive modeling (Kusnaidi et al., 2022). After preprocessing, the dataset was reduced to 148,408 records with 30 validated attributes, which served as the input for the subsequent feature selection and model development stages.

Following preprocessing, feature selection was conducted based on evidence from previous HIV studies to investigate the influence of different feature configurations on classification performance. Rather than using all validated attributes, the dataset was organized into two independent feature configurations according to variables reported as significant in the literature. The first configuration comprised 17 predictive features adopted from (Nisa et al., 2023), whereas the second configuration consisted of 7 key features identified by (Omrani et al., 2026). Establishing these two feature configurations enabled the proposed balancing methods to be evaluated under

different attribute compositions while maintaining consistency with previous HIV classification studies.

Before the balancing process, each feature configuration was partitioned into training and testing subsets using a stratified 80:20 split to preserve the original class distribution.



Source: Research Result (2026)
Figure 1 Research Stages

Performing data partitioning prior to oversampling prevents information leakage from the training data into the testing data, thereby ensuring that model performance is evaluated under unbiased conditions. The class distribution of the training data confirmed a substantial imbalance between the majority and minority classes, thereby necessitating the application of oversampling techniques before model training.

The imbalanced training data were subsequently processed using two hybrid oversampling techniques, namely SMOTE - DBSCAN and RN - SMOTE. In the first approach, the minority class was initially clustered using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, after which the Synthetic Minority Over-sampling Technique (SMOTE) generated synthetic samples within the identified minority clusters. This strategy preserves the local data structure while reducing the influence of isolated observations during oversampling. In the second approach, the Relative Neighborhood (RN) method was first employed to identify and filter noisy minority instances before applying SMOTE to generate synthetic samples from the refined minority class. By integrating neighborhood-based filtering with synthetic oversampling, RN-SMOTE aims to produce a more representative minority class while minimizing the propagation of noisy observations during model learning.

Both balancing methods were independently applied to each feature configuration, resulting in four balanced training datasets for subsequent comparative analysis. After oversampling, the minority class was increased to match the majority class, producing a balanced training set with 108,591 instances in each class. This balanced class distribution provided a more appropriate training dataset for developing robust classification models under imbalanced learning conditions.

The balanced training datasets were subsequently used to train eight classification algorithms, namely Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), K-Nearest Neighbor (KNN), and Naïve Bayes. Model validation was performed using 10-fold cross-validation on the training data to ensure robust performance estimation while reducing the risk of overfitting. Among the evaluated classifiers, the deep learning models were configured using lightweight architectures to ensure a fair comparison with conventional machine learning algorithms. Specifically, CNN employed a single convolutional layer, whereas LSTM consisted of a single recurrent layer with 64 hidden units. Both models were trained using the Adam optimizer with binary cross-entropy

loss for 20 epochs and a batch size of 32. Finally, the trained models were evaluated on the independent testing set using accuracy, precision, recall, and F1-score, while computational runtime was also recorded to compare the predictive effectiveness and computational efficiency of the two hybrid oversampling methods.

2. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a resampling method applied to address class distribution imbalance, a condition in which the proportion between classes in a dataset is unequal and has the potential to cause classification models to be biased toward the majority class (Handoko & Aditya, 2025). Detection of this condition is carried out through the Exploratory Data Analysis (EDA) stage to obtain a comprehensive overview of data patterns and distributions before the modeling process is conducted (Syaifudin et al., 2025). Operationally, SMOTE generates synthetic samples in the minority class by utilizing proximity among instances through a k-nearest neighbors mechanism, so that the data distribution becomes more balanced without directly replicating existing data (Handoko & Aditya, 2025). To ensure that the distance measurement process in the feature space operates consistently, attribute normalization is required to standardize variable scales and avoid bias caused by differences in value ranges (Kusnaidi et al., 2022). This sequence of stages demonstrates that the integration of EDA, normalization, and SMOTE forms a structured methodological procedure for improving the quality of training data in classification scenarios with disproportionate class distributions (Handoko & Aditya, 2025; Kusnaidi et al., 2022; Syaifudin et al., 2025).

The process of generating artificial samples in the SMOTE method uses a linear interpolation formula as shown in Equation 1 (Handoko & Aditya, 2025).

$$\boldsymbol{Xnew = Xi + (Xl - Xi) \,.\, \partial} \qquad \textbf{(1)}$$

Source: Handoko & Aditya (2025)

In the equation, *Xnew* represents the generated synthetic data vector, *Xi* denotes the feature vector of a minority class instance, and *Xl* refers to the feature vector of one of the *k*-nearest neighbors of *Xi*. Meanwhile, ∂ is a random number in the range of 0 to 1 used to determine the interpolation process for generating synthetic data (Handoko & Aditya, 2025).

3. DBSCAN-SMOTE

The DBSCAN–SMOTE method in this study is implemented through two primary stages, namely identifying the density structure of the data using DBSCAN to detect and eliminate noise, followed by applying SMOTE to minority class samples located within sufficiently dense regions. This approach is adopted because oversampling techniques that disregard local distribution characteristics may enlarge class overlap regions and reinforce the presence of outliers in the dataset (Li et al., 2025).

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) is a density-based clustering algorithm that forms clusters according to spatial proximity and the number of neighboring points within a specified radius (Yang et al., 2023). The algorithm operates using two principal parameters, namely the neighborhood radius ε and the minimum number of neighboring points (*MinPts*) required for a point to be classified as a *core point* (Nasaruddin et al., 2025).

The distance between data points is commonly calculated using Euclidean distance (Yang et al., 2023). The ε-neighborhood of a point Xi is defined as:

$$\boldsymbol{N\varepsilon(Xi) = \{Xj \in D \dashv |||Xj - Xi|| \leq \varepsilon\}} \qquad \textbf{(2)}$$

$$\boldsymbol{d(Xi, Xj) = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}} \qquad \textbf{(3)}$$

Source: Yang et al (2023)

In the equation, $N\varepsilon(Xi)$ denotes the set of points within radius$\varepsilon$from *Xi*, where *Xi* is the reference data point, *Xj* is a neighboring data point in the dataset *D*, and $||Xj - Xi||$ represents the distance between them. The parameter *ε\varepsilonε* defines the neighborhood radius, while *d(Xi,Xj)* indicates the Euclidean distance between two points, calculated using *Xi* and x_{jk} as the *k*-th feature values of *Xi* and *Xj*, respectively, with *m* denoting the number of features and *k* the feature index (Yang et al., 2023).

Points that do not meet the minimum density requirement are classified as noise or outliers (Nasaruddin et al., 2025). The elimination of noise prior to oversampling is conducted to prevent the generation of synthetic samples in low-density regions, which may degrade the representational quality of the minority class distribution (Li et al., 2025). After noise points are removed, SMOTE is applied to minority samples located within valid clusters. SMOTE generates synthetic samples through linear interpolation between a minority instance and one of its nearest neighbors (Yang et al., 2023). However, applying SMOTE without considering density structure may

produce synthetic samples in overlapping regions or near outliers, potentially reducing classification performance (Li et al., 2025).

4. RN-SMOTE

RN-SMOTE is an extension of the SMOTE method designed to reduce the influence of noise within the minority class prior to the oversampling process (Arafa et al., 2022). This approach integrates the DBSCAN algorithm to detect and remove outliers, ensuring that synthetic samples are generated only from representative data regions (Arafa et al., 2022). Such a strategy is necessary because conventional SMOTE has the potential to produce synthetic data in noisy areas or regions with class overlap, which may subsequently reduce classification accuracy (Elreedy et al., 2024).

The fundamental process of RN-SMOTE continues to employ the standard SMOTE interpolation mechanism to generate synthetic samples between two minority class instances xi and its neighboring instance xzi.

$$\boldsymbol{z = Xi + \lambda(Xzi - Xi), \lambda \in [0, 1]} \quad \textbf{(4)}$$

Source: Elreedy (2024), Hemmatian et al., (2025)

In the equation, Z represents the synthetic samples, while $\boldsymbol{\lambda}$ denotes a random number used in the synthetic data generation process. Theoretical analysis indicates that interpolation without prior filtering may alter the statistical distribution of the minority class and amplify the influence of noise (Elreedy et al., 2024). Therefore, RN-SMOTE incorporates a density-based filtering stage before performing interpolation (Hemmatian et al., 2025). In the noise reduction stage, DBSCAN clusters the data based on the radius parameter ε and the minimum number of neighboring points (MinPts) (Arafa et al., 2022). A data point is considered part of a valid cluster if it satisfies the following condition:

$$\boldsymbol{|N\varepsilon(x)| \geq MinPts} \quad \textbf{(5)}$$

Source: Arafa et al., (2022)

In the equation, $|N\varepsilon(x)$ denotes the set of neighboring points located within a distance ofε from the reference point x (Arafa et al., 2022). Data points that do not satisfy the density criteria are categorized as noise and removed prior to the oversampling process (Arafa et al., 2022). Following the filtering stage, RN-SMOTE generates synthetic samples exclusively from valid minority clusters, thereby producing a more stable and representative data distribution (Arafa et al., 2022). This strategy mitigates the risk of noise amplification and enhances the model's robustness against overfitting in imbalanced datasets (Elreedy et al., 2024). The integration of density-based filtering and synthetic interpolation renders RN-SMOTE more robust than conventional SMOTE in handling imbalanced datasets containing noise (Hemmatian et al., 2025).

RESULTS AND DISCUSSION

This section presents the results of the feature selection process used to construct the HIV classification model. Feature selection was guided by an integrated HIV prediction framework that combines sociodemographic, behavioral, and biological dimensions, as established in previous studies (Nisa et al., 2023; Omrani et al., 2026). Variables were selected systematically to ensure their relevance to the classification task and their ability to provide meaningful predictive information for HIV status.

Behavioral risk factors and demographic characteristics have consistently been reported as important determinants of HIV infection, particularly among vulnerable populations (Nisa et al., 2023). Accordingly, the first feature configuration adopted 17 predictive attributes derived from the framework proposed (Nisa et al., 2023), emphasizing socio-behavioral indicators associated with injection drug use and high-risk sexual behavior. The selected attributes and their operational definitions are summarized in Table 1.

To complement the socio-behavioral perspective, the second feature configuration incorporated demographic and clinical attributes derived from the HIV surveillance and data mining framework proposed by (Omrani et al., 2026). These variables provide additional epidemiological information that supports a more comprehensive representation of the study population. The corresponding attributes included in this second feature configuration are presented in Table 2.

The selected demographic, behavioral, and healthcare-related attributes were subsequently used as predictor variables, while *rujukan_hasil_hiv* served as the target variable, where a value of 1 indicates reactive HIV status and a value of 2 indicates non-reactive HIV status. These feature configurations provided the basis for the subsequent comparative evaluation of the hybrid oversampling techniques and classification models presented in the following sections.

This section presents the experimental results obtained from the two feature configurations derived from previous HIV studies, as summarized in Tables 1 and 2. Each feature configuration was balanced using two hybrid oversampling techniques, namely SMOTE-DBSCAN and RN-SMOTE. The resulting balanced datasets were

subsequently evaluated using the classification framework described in the Research Stages to compare the effectiveness of both balancing methods in improving HIV classification performance.

Table 1. Feature configuration based on Nisa et al. (2023).article.

Research Dataset Features	Variables in Nisa (2023)	Variable Categories
penjangkauan_jumlah_kondom	Condom usage / Sexual behavior	Sexual behavior
penjangkauan_jumlah_jarum	Syringe acquired from	Injection-related behavior
klien_umur	Age	
klien_jenis_kelamin	Gender	Sosiodemografis
klien_status_pernikahan	Marital status	
klien_pasangan_tetap	Sexual relation with any one	
klien_seksual_anal	Sexual practices	
klien_seksual_vaginal	Sexual practices	Sexual behavior
klien_kondom_selalu	Condom category / AVG	
klien_kondom_terakhir	Condom used last intercourse	
klien_suntik_jumlah	Syringe out all	
klien_suntik_frekuensi	Frequency of daily injecting	Injection-related behavior
klien_jarum_berbagi	Needle sharing	
klien_jarum_sendiri	Syringe used by self	
rujukan_hasil_ims	Treated for STI	Biologis
rujukan_hasil_tb	Biological condition	Biologis
rujukan_hasil_hiv	Predictive status	Outcome

Source: (Nisa et al., 2023)

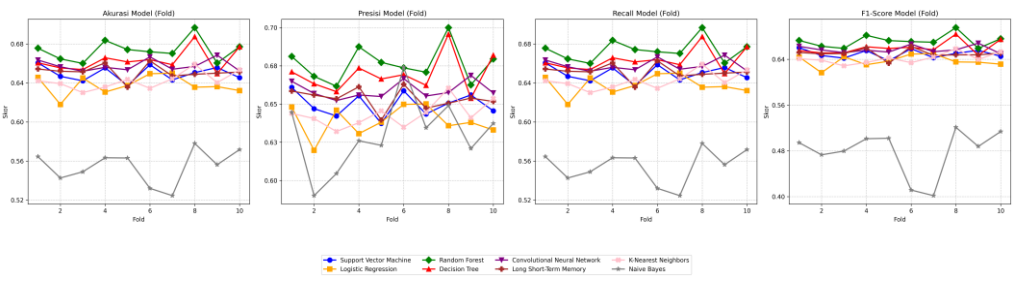
Table 2. Feature configuration based on Omrani (2026) article.

Research Dataset Features	Variables in Omrani (2026)
penjangkauan_jenis_kontak	Type of client interaction with the healthcare service system
penjangkauan_status_kontak	Continuity or follow-up status within the surveillance process
penjangkauan_jenis_kegiatan	Engagement in healthcare services or HIV prevention programs
klien_umur	Age at diagnosis
klien_jenis_kelamin	Gender
rujukan_hasil_vct	Proses diagnosis awal HIV
rujukan_hasil_hiv	Outcome / status HIV

Source: (Omrani et al., 2026)

Tables 1 and 2 present the two feature configurations adopted in this study based on previous HIV prediction research. The first configuration, derived from (Nisa et al., 2023), consists of a broader set of demographic, behavioral, biological, and clinical variables, whereas the second configuration, adapted from (Omrani et al., 2026), includes a more compact set of key predictive attributes. The use of these two feature configurations enables a comparative evaluation of the proposed balancing methods under different levels of feature complexity. This experimental design facilitates the assessment of whether the effectiveness of DBSCAN - SMOTE and RN - SMOTE remains consistent across different feature compositions prior to classification model development.

1. Dataframe Based on Nisa (2023)

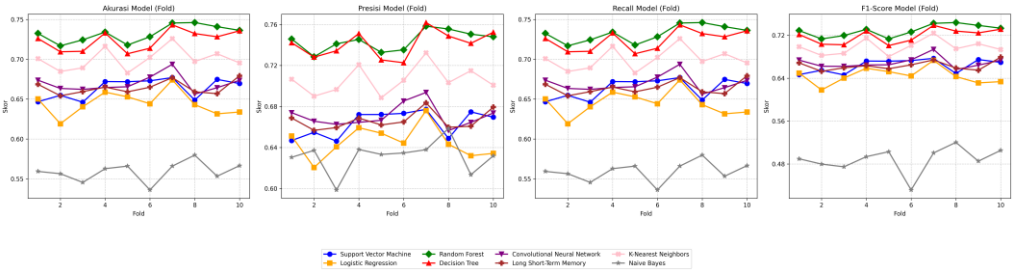


Source: Research Result (2026)

Figure 2 Performance Evaluation Results of SMOTE-DBSCAN on the Nisa (2023) Feature Set

Based on the evaluation results using the SMOTE - DBSCAN balancing approach in (Nisa et al., 2023), the balancing process required an execution time of 2 minutes and 10 seconds. Subsequently, the evaluation of

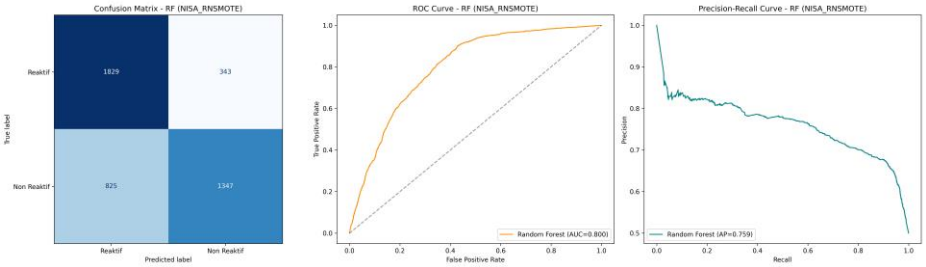
eight classification algorithms was completed within 43 minutes and 55 seconds. Among all evaluated algorithms, Random Forest (RF) demonstrated the best performance, achieving an accuracy of 0.6846, precision of 0.6872, recall of 0.6846, and an F1-score of 0.6835. These findings indicate that Random Forest exhibits a more optimal capability in modeling data patterns following the application of density-based oversampling. The consistency across all evaluation metrics further suggests that the ensemble-based learning mechanism employed by Random Forest is more effective in addressing class imbalance in the HIV dataset compared to the other algorithms evaluated within the SMOTE–DBSCAN framework.



Source: Research Result (2026)

Figure 3 Performance Evaluation Results of SMOTE-RN on the Nisa (2023) Feature Set

Based on the evaluation results of (Nisa et al., 2023) using the RN - SMOTE balancing approach, the Random Forest algorithm achieved the best overall classification performance among all evaluated models. The RF model obtained an accuracy of 0.7325, precision of 0.7441, recall of 0.7325, and an F1-score of 0.7293, indicating a balanced and consistent improvement across all evaluation metrics in distinguishing HIV status classes. These results demonstrate that the integration of the RN - SMOTE technique effectively enhances class distribution quality, allowing the ensemble learning mechanism of Random Forest to better capture complex data patterns within imbalanced HIV datasets. The data balancing process using RN-SMOTE required an execution time of 2 minutes 18 seconds, The subsequent model evaluation phase required 44 minutes 47 seconds, reflecting the computational cost of improved classification performance. Because the combination of RN-SMOTE and Random Forest under the (Nisa et al., 2023) feature configuration achieved the highest predictive performance among all experimental settings, additional visual evaluation was conducted using the confusion matrix, ROC curve, and Precision–Recall curve, as presented in Figure 4.



Source: Research Result (2026)

Figure 4 Confusion matrix, ROC curve, and Precision–Recall curve of the best-performing model (Random Forest with RN-SMOTE) under the Nisa et al. (2023) feature configuration.

Figure 4 presents the confusion matrix, ROC curve, and Precision-Recall curve for the best-performing model obtained using Random Forest with RN-SMOTE under the (Nisa et al., 2023) feature configuration. The confusion matrix shows that most reactive and non-reactive HIV cases were correctly classified, although a small number of misclassifications remained between the two classes. Furthermore, the ROC curve achieved an Area Under the Curve (AUC) of 0.800, indicating good discriminative capability, while the Precision–Recall curve obtained an Average Precision (AP) score of 0.759, demonstrating stable predictive performance on the minority class. These visual evaluation results are consistent with the quantitative performance metrics presented previously and further confirm the effectiveness of RN-SMOTE in improving HIV classification on imbalanced HIV datasets.

2. Dataframe Based on Omrani (2026)

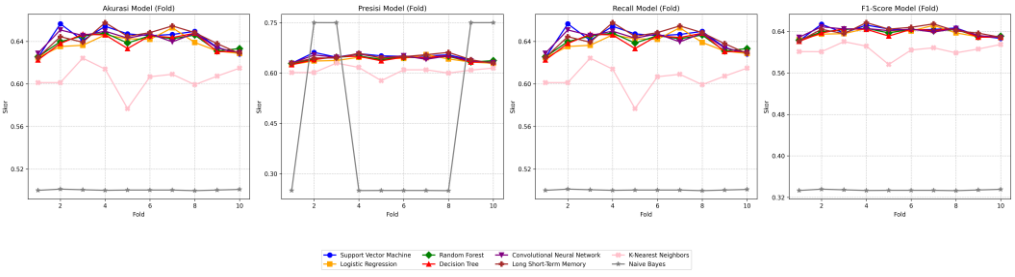
Based on the evaluation results referring to (Omrani et al., 2026) under the SMOTE - DBSCAN balancing approach, the Support Vector Machine (SVM) algorithm demonstrated the best overall classification performance among the evaluated models. The SVM model achieved an accuracy of 0.6471, precision of 0.6526, recall of 0.6471, and an F1-score of 0.6439, indicating stable and balanced predictive capability across evaluation metrics in distinguishing HIV status classes. These findings suggest that the SMOTE–DBSCAN method effectively

improves class distribution by refining synthetic sample generation and noise handling, enabling SVM to construct clearer decision boundaries within the imbalanced dataset. The data balancing process using SMOTE–DBSCAN

Table 3. Evaluation Results Table of Two Dataframes Using Two SMOTE Methods

Literature Review	Algorithm Model	SMOTE - DBSCAN				SMOTE - RN			
		Acc	*Pre*	*Rec*	*F1*	*Acc*	*Pre*	*Rec*	*F1*
Nisa et al, 2023	*SVM*	0.6565	0.6568	0.6565	0.6564	0.6678	0.6679	0.6678	0.6678
	LR	0.6425	0.6433	0.6425	0.6420	0.6351	0.6353	0.6351	0.6350
	RF	0.6846	0.6872	0.6846	0.6835	0.7325	0.7441	0.7325	0.7293
	DT	0.6683	0.6737	0.6683	0.6657	0.7231	0.7403	0.7231	0.7180
	CNN	0.6669	0.6729	0.6669	0.6640	0.6763	0.6766	0.6763	0.6762
	LSTM	0.6614	0.6620	0.6614	0.6610	0.6761	0.6763	0.6761	0.6760
	KNN	0.6390	0.6395	0.6390	0.6387	0.6945	0.6999	0.6945	0.6925
	NB	0.5711	0.6358	0.5711	0.5132	0.5730	0.6489	0.5730	0.5106
Omrani et al, 2026	*SVM*	0.6471	0.6526	0.6471	0.6439	0.6565	0.6667	0.6565	0.6512
	LR	0.6446	0.6456	0.6446	0.6439	0.6763	0.6786	0.6763	0.6753
	RF	0.6432	0.6501	0.6432	0.6390	0.7169	0.7379	0.7169	0.7104
	DT	0.6379	0.6421	0.6379	0.6352	0.7159	0.7388	0.7159	0.7090
	CNN	0.6448	0.6484	0.6448	0.6426	0.6786	0.6812	0.6786	0.6775
	LSTM	0.6430	0.6431	0.6430	0.6429	0.6577	0.6604	0.6577	0.6562
	KNN	0.6103	0.6103	0.6103	0.6102	0.6591	0.6595	0.6591	0.6588
	NB	0.5005	0.7501	0.5005	0.3344	0.5002	0.7501	0.5002	0.3338

Source: Research Result (2026)

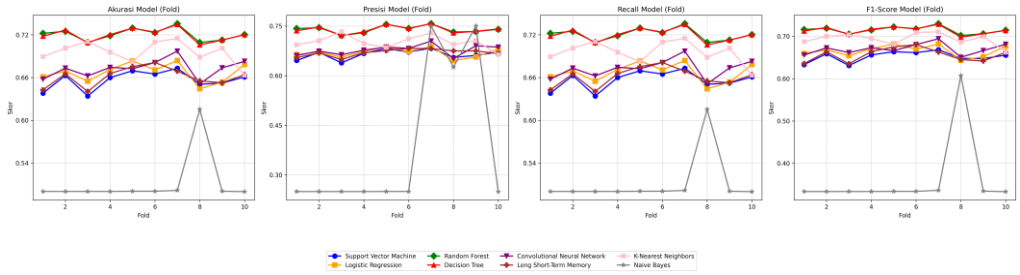


Source: Research Result (2026)

Figure 5 Performance Evaluation Results of SMOTE-DBSCAN on the Omrani (2026) Feature Set

required an execution time of 1 minute 59 seconds, Meanwhile, the model evaluation phase required 34 minutes 20 seconds to obtain the final performance metrics, demonstrating computational efficiency alongside improved classification performance.

Based on the evaluation results referring to (Omrani et al., 2026) using the SMOTE - RN balancing method, the Random Forest (RF) algorithm achieved the best classification performance among all evaluated models. The RF model obtained an accuracy of 0.7169, precision of 0.7379, recall of 0.7169, and an F1-score of 0.7104, demonstrating superior capability in capturing class patterns and maintaining balanced predictive performance across all evaluation metrics. These results indicate that RN-SMOTE effectively reduces noise while generating representative synthetic samples, thereby enhancing the learning process of ensemble-based models such as Random Forest in handling imbalanced HIV datasets.



Source: Research Result (2026)

Figure 6 Performance Evaluation Results of SMOTE-RN on the Omrani (2026) Feature Set

The data balancing process using RN-SMOTE required an execution time of 2 minutes 05 seconds, while the overall model evaluation process required 40 minutes 12 seconds to produce the final performance results, reflecting the computational cost associated with improved classification robustness.

Overall, the evaluation results on the feature sets derived from (Nisa et al., 2023) and (Omrani et al., 2026) indicate that the application of SMOTE-based data balancing methods significantly improves classification performance on imbalanced HIV datasets. Among the evaluated approaches, SMOTE - RN generally produced superior performance compared to SMOTE - DBSCAN, particularly when combined with the Random Forest algorithm, which consistently achieved the most stable accuracy, precision, recall, and F1-score values. Although SMOTE-RN required slightly longer computational time, it was able to generate a more representative class distribution, thereby enhancing the overall predictive quality and robustness of the classification models.

CONCLUSION

This study demonstrates that hybrid data balancing techniques play a significant role in improving the reliability of HIV classification under imbalanced data conditions. Both DBSCAN - SMOTE and RN - SMOTE effectively reduced class imbalance, enabling classification models to learn more representative decision boundaries and produce more balanced predictive performance. The incorporation of demographic, behavioral, and clinical variables further strengthened the predictive capability of the developed models.

Comparative evaluation across two feature configurations and multiple classification algorithms indicates that RN - SMOTE consistently provides more robust and stable classification performance than DBSCAN - SMOTE, particularly when combined with ensemble learning methods. Although DBSCAN-SMOTE offers greater computational efficiency, RN - SMOTE demonstrates superior capability in reducing noise and improving minority-class representation, resulting in more reliable HIV prediction. These findings highlight that the effectiveness of a balancing method depends not only on its ability to generate synthetic samples but also on its capacity to preserve meaningful class boundaries during the learning process.

This study contributes to the existing literature by providing a systematic comparison of two hybrid oversampling techniques using an Indonesian HIV dataset and two feature configurations derived from previous studies. The findings provide empirical evidence that can support the selection of appropriate data balancing strategies for HIV classification and other healthcare prediction tasks involving class-imbalanced datasets. Future research may extend this work by investigating additional hybrid balancing approaches and validating their performance using datasets from different healthcare institutions and epidemiological settings.

REFERENCES

Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2022). RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 5059–5074. https://doi.org/10.1016/j.jksuci.2022.06.005

Arafa, A., El-Fishawy, N., Badawy, M., & Radad, M. (2023). RN-Autoencoder: Reduced Noise Autoencoder for classifying imbalanced cancer genomic data. *Journal of Biological Engineering*, *17*(1), 7. https://doi.org/10.1186/s13036-022-00319-3

Arifin, B., Rokhman, M. R., Zulkarnain, Z., Perwitasari, D. A., Mangau, M., Rauf, S., Noor, R., Padmawati, R. S., Massi, M. N., Schans, J. van der, & Postma, M. J. (2023). The knowledge mapping of HIV/AIDS in Indonesians living on six major islands using the Indonesian version of the HIV-KQ-18 instrument. *PLOS ONE*, *18*(11), e0293876. https://doi.org/10.1371/journal.pone.0293876

Chawla, N. V., Herra, F., Garcia, S., & Fernandez, A. (2018). *SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary*. https://doi.org/https://doi.org/10.1613/jair.1.11192

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Gunawan, W., Purnomo, H., Dewi, C., Iriani, A., & Sembiring, I. (2024). Systematic Literature Review: Disease Classification Modeling Using Deep Learning Algorithms. *2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA)*, 1–6. https://doi.org/10.1109/ICTIIA61827.2024.10761789

Handoko, C. B., & Aditya, C. S. K. (2025). Penerapan Teknik SMOTE Dalam Mengatasi Imbalance Data Penyakit Diabetes Menggunakan Algoritma ANN. *Smart Comp: Jurnalnya Orang Pintar Komputer*, *14*(1). https://doi.org/10.30591/smartcomp.v14i1.7045

Hemmatian, J., Hajizadeh, R., & Nazari, F. (2025). Addressing Imbalanced data Classification With Cluster-Based Reduced Noise SMOTE. *Journal.Plos.Org*. https://doi.org/doi.org/10.1371/journal.pone.0317396

Kosolwattana, T., Liu, C., Hu, R., Han, S., Chen, H., & Lin, Y. (2023). A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Mining*, *16*(1), 15. https://doi.org/10.1186/s13040-023-00330-4

Kusnaidi, M. R., Gulo, T., & Aripin, S. (2022). Penerapan Normalisasi Data Dalam Mengelompokkan Data Mahasiswa Dengan Menggunakan Metode K-Means Untuk Menentukan Prioritas Bantuan Uang Kuliah Tunggal. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 330–338. https://doi.org/10.47065/josyc.v3i4.2112

Li, Y., Yang, Y., Song, P., Duan, L., & Ren, R. (2025). An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. *Scientific Reports*, *15*(1), 23521. https://doi.org/10.1038/s41598-025-09506-w

Nasaruddin, N., Masseran, N., Idris, W. M. R., & Ul-Saufie, A. Z. (2025). SMOTE-PCADBSCAN: A Novel Approach for Addressing Class Imbalance in Water Quality Prediction. *Sains Malaysiana*, *54*(6), 1629–1639. https://doi.org/10.17576/jsm-2025-5406-17

Nisa, S. U., Mahmood, A., Ujager, F. S., & Malik, M. (2023). HIV/AIDS predictive model using random forest based on socio-demographical, biological and behavioral data. *Egyptian Informatics Journal*, *24*(1), 107–115. https://doi.org/10.1016/j.eij.2022.12.005

Omrani, A. S., Almohaizeie, A., Al Hammadi, A., Wali, G., Al Salman, J., Chaponda, M., & Albaksami, O. (2026). Consensus-based expert recommendations on the management of heavily treatment-experienced people living with HIV in the Middle East. *Journal of Infection and Public Health*, *19*(3), 103110. https://doi.org/10.1016/j.jiph.2025.103110

Rahim, A. M. A., Hartato, B. P., Ridwan, A., & Asharudin, F. (2025). Machine Learning-Based Approach for HIV/AIDS Prediction: Feature Selection and Data Balancing Strategy. *Machine Learning*, *9*(2), 338~347.

Riantika, I., Sartono, B., & Anwar Notodiputro, K. (2024). Effectiveness of SMOTE-ENN to Reduce Complexity in Classification Model. *Indonesian Journal of Statistics and Its Applications*, *8*(1), 70–82. https://doi.org/10.29244/ijsa.v8i1p70-82

Sandiwarno, S., Niu, Z., & Nyamawe, A. S. (2025). SES-Net: A Novel Multi-Task Deep Neural Network Model for Analyzing E-learning Users' Satisfaction via Sentiment, Emotion, and Semantic. *International Journal of Human–Computer Interaction*, *41*(8), 4910–4933. https://doi.org/10.1080/10447318.2024.2356356

Sui, Q., Li, G., Peng, Y., Zhang, J., Zhang, Y., & Zhao, R. (2025). Scalable and robust machine learning framework for HIV classification using clinical and laboratory data. *Scientific Reports*, *15*(1), 18727. https://doi.org/10.1038/s41598-025-00085-4

Syaifudin, A., Hapsoro, H. W., & Pratama, I. W. (2025). *IMPLEMENTASI EXPLORATORY DATA ANALYSIS UNTUK ANALISIS DATA LEMAK TUBUH*. *Vol.20 No.1*. https://doi.org/https://doi.org/10.47775/ictech.v20i1.328

Yang, Y., Akbarzadeh Khorshidi, H., & Aickelin, U. (2023). A Diversity-Based Synthetic Oversampling Using Clustering for Handling Extreme Imbalance. *SN Computer Science*, *4*(6), 848. https://doi.org/10.1007/s42979-023-02249-3