

Evaluating the Sensitivity–Specificity Trade-off in Risk Factor-Based Cervical Cancer Classification Using SMOTE and ACO-Optimized XGBoost

Titin Prihatin^{1)*}, Suharjanti²⁾, Resti Lia Andharsaputri³⁾, Hafis Nurdin⁴⁾, Irmawati Carolina⁵⁾

^{1,2,3,4,5)}Universitas Bina Sarana Informatika, Indonesia

¹⁾ titin.tpn@bsi.ac.id, ²⁾ suharjanti.shj@bsi.ac.id, ³⁾ resti.ria@bsi.ac.id, ⁴⁾ hafis.nnr@bsi.ac.id, ⁵⁾ irmawati.imc@bsi.ac.id

Submitted : Aug 26, 2026 | Accepted : Sep 11, 2026 | Published : Oct 1, 2026

Abstract: Class imbalance is a major challenge in risk factor-based cervical cancer classification because positive cases are substantially less frequent than negative cases. This study aims to evaluate the effects of the Synthetic Minority Over-sampling Technique (SMOTE) and Ant Colony Optimization (ACO) on XGBoost performance and to investigate the trade-off between sensitivity, specificity, and overall discriminative ability. The Cervical Cancer (Risk Factors) dataset from the UCI Machine Learning Repository, containing 858 observations, was used with Biopsy as the binary classification target, comprising 803 negative and 55 positive cases. A total of 28 risk-factor features were retained after excluding diagnostic attributes with potential data leakage. Three classification scenarios were evaluated: XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost. Model performance was assessed using Nested Repeated Stratified Cross-Validation, consisting of an outer 5-fold cross-validation repeated five times and an inner 3-fold cross-validation for hyperparameter and decision-threshold optimization. The results showed that SMOTE-XGBoost achieved the highest recall of 0.4182 ± 0.3340 , whereas SMOTE-ACO-XGBoost obtained the highest accuracy of 0.7667 ± 0.1698 and specificity of 0.8042 ± 0.1950 . Baseline XGBoost achieved the highest ROC-AUC and PR-AUC values of 0.6227 ± 0.0663 and 0.1290 ± 0.0739 , respectively. The Friedman test indicated significant differences among the models across all evaluation metrics. Pairwise Wilcoxon signed-rank tests with Holm correction showed that ACO significantly improved accuracy and specificity compared with SMOTE-XGBoost but significantly reduced recall from 0.4182 to 0.2182. These findings demonstrate that class-imbalance handling and hyperparameter optimization do not produce a single model that consistently dominates all performance measures. SMOTE tends to improve sensitivity, whereas ACO optimization shifts the classifier toward higher accuracy and specificity at the cost of reduced sensitivity.

Keywords: Ant Colony Optimization, Cervical Cancer, Class Imbalance, SMOTE, Nested Cross-Validation, XGBoost,

INTRODUCTION

Cervical cancer remains a serious health problem affecting women in many countries. The disease develops in the cervix and is largely associated with persistent infection with high-risk Human Papillomavirus (HPV). Although cervical cancer is largely preventable through HPV vaccination and regular screening, its global burden remains substantial. According to the latest global cancer statistics, approximately 604,000 new cases of cervical cancer and 280,000 deaths were estimated in 2024. Cervical cancer also remains one of the most commonly diagnosed cancers among women worldwide (Sung et al., 2026). These conditions highlight the continued need to improve risk identification and early detection to support cervical cancer prevention and management.

The development of Artificial Intelligence (AI) and machine learning provides opportunities to support cervical cancer risk identification based on patient characteristics. These approaches enable the analysis of demographic factors, behavioral characteristics, and medical histories to identify patterns associated with cervical cancer risk.

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Risk-factor-based studies have increasingly applied various machine learning algorithms to support disease prediction and classification (Ridwansyah et al., 2026). However, a recent systematic review indicates that risk-factor-based cervical cancer prediction research still faces several challenges, particularly in preprocessing, data representation, algorithm selection, and model evaluation (Kondo et al., 2025). This finding highlights the importance of data processing quality and appropriate classification methods in developing reliable predictive models (Ridwansyah et al., 2022).

One of the major challenges in cervical cancer classification is class imbalance, in which negative cases substantially outnumber positive cases. This condition may cause classification algorithms to favor the majority class, resulting in high accuracy but limited ability to correctly identify positive cases. Class imbalance is an important issue in cervical cancer prediction, and previous studies have compared various resampling techniques to improve the ability of machine learning algorithms to recognize minority cases (Muraru et al., 2024). Similar conditions have also been observed in cervical cancer studies using risk-factor datasets, indicating that class balancing should be an important component of classification model development.

One widely used approach for addressing class imbalance is the Synthetic Minority Over-sampling Technique (SMOTE). This technique increases the representation of the minority class by generating synthetic samples based on the characteristics of minority-class observations in the training data. Previous studies have demonstrated that class-imbalance handling can affect the performance of cervical cancer prediction. SMOTE combined with missing-value imputation and ensemble learning, as well as SMOTE-Tomek applied before classification and model optimization, has been used to address the characteristics of imbalanced cervical cancer data (Karamti et al., 2023), (AlMohimeed et al., 2023). These findings emphasize the importance of resampling strategies when developing models for cervical cancer datasets with imbalanced class distributions.

The combination of SMOTE and Extreme Gradient Boosting (XGBoost) has also been investigated for cervical cancer classification. The SMOTE-XGBoost approach has demonstrated that combining class balancing with boosting algorithms can address classification problems in imbalanced data (Anggitasyah & Siregar, 2023). Another study applied missing-value imputation, SMOTE, and hyperparameter optimization using Randomized Search Cross-Validation. The results indicated that the tuned XGBoost model combined with SMOTE provided good performance in predicting cervical cancer risk (Yudha & Rahardi, 2025). These findings further support the potential of combining resampling techniques with XGBoost for cervical cancer classification.

XGBoost is a gradient boosting algorithm that sequentially constructs decision trees to reduce prediction errors from previous iterations. The algorithm incorporates regularization mechanisms and several hyperparameters that influence model complexity and generalization capability. Previous research has shown that XGBoost performance is strongly influenced by hyperparameter configurations, making parameter optimization an important step for improving predictive performance (Boldini et al., 2023). The selection of XGBoost hyperparameters can significantly affect model performance; therefore, an effective optimization strategy is required to identify appropriate parameter configurations (Marinho et al., 2024). Consequently, default parameter settings may not necessarily produce optimal performance for every dataset.

Hyperparameter optimization can be performed using conventional search approaches or metaheuristic algorithms. One such metaheuristic approach is Ant Colony Optimization (ACO), an optimization method inspired by the behavior of ant colonies in finding optimal paths through pheromone-based mechanisms. This approach enables the search for solutions within a space consisting of various parameter combinations (Hamid & Ridwansyah, 2024). Previous research has demonstrated the potential of metaheuristic algorithms, including ACO, for hyperparameter optimization in cervical cancer classification (Abdelhay et al., 2026). Although that study employed a deep learning approach and image data, its findings indicate the potential of ACO as a hyperparameter optimization mechanism for cervical cancer classification problems.

Based on previous studies, handling class imbalance and hyperparameter optimization have become two important components in the development of cervical cancer classification models. SMOTE has been used to improve the representation of minority-class samples, while XGBoost has shown potential for analyzing cervical cancer risk factors in imbalanced datasets (Anggitasyah & Siregar, 2023; Karamti et al., 2023). Other studies have also combined resampling techniques with hyperparameter optimization using approaches such as Randomized Search and GridSearchCV (Saputra et al., 2025; Yudha & Rahardi, 2025). Meanwhile, various optimization approaches have been applied to improve XGBoost configurations, indicating that hyperparameter optimization for this algorithm is not a new problem (Marinho et al., 2024). Therefore, the individual use of ACO or XGBoost cannot be considered the primary novelty of this study.

The research gap lies in the limited number of studies that systematically evaluate the effects of class imbalance handling and hyperparameter optimization on XGBoost performance for risk factor-based cervical cancer classification. Previous studies generally applied resampling or hyperparameter tuning separately, while metaheuristic approaches such as ACO have more commonly been applied to image-based data and deep learning models (Abdelhay et al., 2026). These approaches have not clearly demonstrated how sequential class balancing and hyperparameter optimization affect the model's ability to recognize positive and negative classes. To address

*name of corresponding author



this limitation, this study compares XGBoost as the baseline, SMOTE-XGBoost to isolate the effect of class imbalance handling, and SMOTE-ACO-XGBoost to evaluate the additional effect of hyperparameter optimization. Thus, the novelty of this study does not merely lie in combining SMOTE, ACO, and XGBoost, but in providing a controlled evaluation framework to analyze the contribution of each stage to the trade-off between sensitivity, specificity, and overall discriminative performance.

In the context of cervical cancer, sensitivity is important for reducing the risk of unidentified positive cases. However, maximizing sensitivity alone may increase false positives and reduce specificity and precision. Therefore, the optimization objective in this study was not specifically designed to maximize sensitivity, because the primary aim was to evaluate the effect of optimization on the balance among multiple aspects of model performance rather than to develop a cost-sensitive clinical screening model. F1-score was used for decision-threshold optimization as a compromise between precision and recall, while sensitivity, specificity, ROC-AUC, and PR-AUC were evaluated separately. The main contribution of this study is not to claim a clinically superior diagnostic model, but to provide empirical evidence regarding the performance trade-offs produced by SMOTE and ACO optimization. The evaluation employed Nested Repeated Stratified Cross-Validation together with the Friedman Test, Wilcoxon Signed-Rank Test with Holm correction, and Kendall's W to provide a more comprehensive and robust comparison of model performance.

LITERATURE REVIEW

Machine Learning for Cervical Cancer Risk Classification

The application of machine learning in cervical cancer classification has developed as an approach for identifying disease risk based on patients' demographic characteristics, behaviors, and medical histories (Glučina et al., 2023; Vazquez et al., 2025). Risk-factor-based approaches can utilize clinical and epidemiological data that are relatively easy to obtain (Mudawi & Alazeb, 2022). However, previous studies have shown that preprocessing, missing values, class imbalance, and evaluation remain challenges in cervical cancer classification (Kondo et al., 2025).

Cervical cancer risk-factor datasets generally have imbalanced class distributions and contain missing values. These conditions can cause models to become more dominant in learning the majority class, resulting in limited ability to recognize the minority class. Therefore, addressing class imbalance is an important component in developing classification models.

Class Imbalance and Synthetic Minority Over-sampling Technique

Class imbalance refers to a condition in which the distribution of data among classes is disproportionate, potentially causing models to overlook the minority class (Altalhan et al., 2025; Huang & Dai, 2021). In medical data, this condition is a particular concern because the minority class often has greater diagnostic importance (Gurcan & Soylu, 2024; Salmi et al., 2024). One widely used method is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples based on the characteristics of minority-class data (Yang et al., 2024). SMOTE increases the representation of the minority class by generating synthetic samples based on the characteristics of existing minority samples (Carolina et al., 2026).

In cervical cancer classification, the application of SMOTE-Tomek combined with missing-value handling, feature selection, and Bayesian Optimization on the UCI dataset demonstrated that the combination of resampling and optimization can strengthen the classification process (AlMohimeed et al., 2023). Other studies have shown that the combination of imputation, SMOTE, and ensemble learning can also potentially improve the performance of cervical cancer models (Karamti et al., 2023). In addition, oversampling can provide better minority-class detection capabilities than undersampling (Fulazzaky et al., 2024; Muraru et al., 2024; Yang et al., 2024).

Nevertheless, increasing the representation of the minority class does not always improve all evaluation metrics simultaneously (Sumarna et al., 2026). An increase in recall may be accompanied by an increase in false positives, thereby reducing precision or specificity (Çorbacioğlu & Aksel, 2023). Therefore, evaluating the impact of SMOTE should consider multiple metrics simultaneously rather than relying solely on accuracy as an indicator of performance.

Extreme Gradient Boosting for Cervical Cancer Classification

Extreme Gradient Boosting (XGBoost) is an ensemble boosting algorithm that builds decision trees sequentially to improve the prediction errors of previous models (Geron, 2022). The performance of XGBoost is influenced by its hyperparameter configuration; therefore, optimization is required to obtain a model that is appropriate for the characteristics of the data.

Studies on cervical cancer have shown that the combination of SMOTE and XGBoost can be used to address class imbalance (Anggitasyah & Siregar, 2023). Other studies have used the Cervical Cancer Risk Factors dataset from UCI with imputation, SMOTE, and hyperparameter optimization, demonstrating high performance for XGBoost and Random Forest (Saputra et al., 2025; Yudha & Rahardi, 2025). However, the optimization methods

*name of corresponding author



used were still based on Randomized Search, providing an opportunity to employ metaheuristic methods for hyperparameter optimization.

Ant Colony Optimization for Hyperparameter Optimization

Ant Colony Optimization (ACO) is a metaheuristic algorithm that uses a pheromone-based mechanism to search for optimal solutions. In machine learning, ACO can be used to identify hyperparameter combinations based on model performance as the fitness function. Previous research has demonstrated that ACO can be used for hyperparameter optimization in image-based cervical cancer classification; however, it has not been specifically applied to optimize XGBoost hyperparameters for risk-factor-based data (Abdelhay et al., 2026).

Previous Studies and Research Position

Previous studies have employed various resampling strategies, hyperparameter optimization techniques, and classifiers for cervical cancer classification. Anggitasyah and Siregar (2023) applied SMOTE-XGBoost, Karamti et al. (2023) combined SMOTE with ensemble learning, while Muraru et al. (2024) compared several sampling techniques. Other approaches combined resampling with optimization, such as ADASYN-Random Forest with GridSearchCV (Saputra et al., 2025) and SMOTE-XGBoost with Randomized Search CV (Yudha & Rahardi, 2025). Meanwhile, metaheuristic optimization has also been used to automatically search for model configurations; Abdelhay et al. (2026), for example, applied ACO and other metaheuristic algorithms to optimize image-based models. A comparison of the most relevant previous studies is presented in Table 1.

Table 1. Comparison of previous studies and the proposed study

Penelitian	Sampling	Model / Optimasi	Common Metrics	Perbedaan Utama
(Anggitasyah & Siregar, 2023)	SMOTE	XGBoost	AUC	No ACO optimization
(Karamti et al., 2023)	SMOTE	Ensemble	Accuracy, Precision, Recall, F1	No metaheuristic optimization
(Muraru et al., 2024)	Beragam sampling	KNN, LR, RF	Recall / Balanced Accuracy	No XGBoost-ACO
(Saputra et al., 2025)	ADASYN	RF + GridSearchCV	Accuracy, Recall	Conventional optimization
(Yudha & Rahardi, 2025)	SMOTE	XGBoost/RF + Randomized Search	Accuracy, AUC	No metaheuristic optimization
(Abdelhay et al., 2026)	–	Deep Learning + ACO	Accuracy, Recall, Specificity	ACO applied to image data
Penelitian ini	SMOTE	XGBoost + ACO	Accuracy, Recall, Specificity, F1, ROC-AUC, PR-AUC	Stepwise evaluation of the effects of SMOTE and ACO on risk-factor data

Based on this comparison, the contribution of this study does not lie in the novelty of SMOTE, ACO, or XGBoost individually, but rather in the controlled evaluation of XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost. This framework enables the effects of resampling and ACO-based optimization to be analyzed in terms of the trade-off between sensitivity, specificity, and the discriminative ability of the model using Nested Repeated Stratified Cross-Validation.

METHOD

This study employed a quantitative experimental approach to evaluate cervical cancer risk classification using imbalanced data. Three models were compared: XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost. SMOTE was used to address class imbalance, while ACO was used to optimize XGBoost hyperparameters. Evaluation was conducted using Nested Repeated Stratified Cross-Validation to prevent data leakage, with all imputation, SMOTE, hyperparameter optimization, and decision-threshold procedures performed only on the training data within each fold.

Research Dataset

The dataset used in this study is the Cervical Cancer (Risk Factors) dataset from the UCI Machine Learning Repository, consisting of 858 patients and 36 variables (Fernandes et al., 2017). Biopsy was used as the binary target, with 803 negative cases (93.59%) and 55 positive cases (6.41%), indicating a highly imbalanced class distribution. Four diagnostic attributes, namely Dx:Cancer, Dx:CIN, Dx:HPV, and Dx, were excluded because

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

they could potentially cause data leakage. After removing the target variable and these diagnostic attributes, 28 risk factors were retained as predictor features.

Data Preprocessing

The preprocessing stage included converting the value "?" into missing values, converting all features into numerical format, removing diagnostic attributes that could potentially cause data leakage, and handling missing values using median imputation. Imputation was not performed on the entire dataset before validation. In each fold, the median value was calculated only from the training data and subsequently used to transform the validation data.

Handling Class Imbalance Using SMOTE

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic samples for the minority class by interpolating between a minority sample and one of its nearest neighbors. The generation of synthetic samples can be expressed as:

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad \dots (1)$$

where (x_i) represents a minority-class sample, (x_{nn}) represents its nearest minority-class neighbor, and (λ) represents a random value in the interval from 0 to 1. In this study, SMOTE was applied only to the training data. In the nested cross-validation process, SMOTE was independently applied to each inner training fold and outer training fold. Validation data were not oversampled, so the class distribution in the evaluation data remained consistent with the original dataset.

XGBoost Classification

Extreme Gradient Boosting (XGBoost) was used as the primary classification algorithm. XGBoost builds a series of decision trees iteratively, where each subsequent model attempts to correct the prediction errors of the previous model. The XGBoost objective function can be expressed as:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k) \quad \dots (2)$$

dengan (l) represents the loss function between the actual class (y_i) and the predicted result ($\hat{y}_i$), while (Ω) represents the regularization component used to control model complexity.

The baseline XGBoost model used the following configuration: `n_estimators = 250`, `max_depth = 4`, `learning_rate = 0.05`, `subsample = 0.90`, `colsample_bytree = 0.90`, `min_child_weight = 1`, `gamma = 0`, `reg_alpha = 0`, `reg_lambda = 1.0`.

The same configuration was also applied to the SMOTE-XGBoost model. Thus, the difference between the first and second models resulted solely from the application of SMOTE.

Ant Colony Optimization for XGBoost Hyperparameter Optimization

In the SMOTE-ACO-XGBoost model, Ant Colony Optimization (ACO) was used to optimize the XGBoost hyperparameters. Each ant represented a hyperparameter combination selected based on pheromone probabilities. ACO was configured with 8 ants for 10 iterations, an evaporation rate of 0.25, 3 elite ants, and an initial pheromone value of 1.0 for each parameter candidate. The initial random state was set to 42 to support reproducibility. The search space consisted of `n_estimators = {100, 200, 300, 400, 500}`, `max_depth = {2, 3, 4, 5, 6}`, `learning_rate = {0.01, 0.03, 0.05, 0.10, 0.20}`, `subsample = {0.6, 0.7, 0.8, 0.9, 1.0}`, `colsample_bytree = {0.6, 0.7, 0.8, 0.9, 1.0}`, `min_child_weight = {1, 2, 3, 5, 7}`, `gamma = {0, 0.1, 0.3, 0.5, 1.0}`, `reg_alpha = {0, 0.01, 0.1, 0.5, 1.0}`, and `reg_lambda = {0.5, 1.0, 2.0, 5.0, 10.0}`. PR-AUC was used as the ACO fitness function because the dataset was highly imbalanced and PR-AUC directly evaluates the precision-recall relationship for the positive class. The fitness of each candidate was calculated using out-of-fold probabilities from inner 3-fold stratified cross-validation, with median imputation and SMOTE applied only to the inner training folds to prevent data leakage. At each iteration, all solutions were ranked according to their PR-AUC values, and the three best solutions were selected as elite solutions. The pheromone values first underwent evaporation at a rate of 0.25, after which the pheromone associated with candidates selected by the elite ants was increased according to the fitness value of each solution. The pheromone update mechanism was expressed as:

$$\tau_{ij}^{t+1} = (1 - \rho)\tau_{ij}^t + \Delta\tau_{ij} \quad \dots (3)$$

where $\rho = 0.25$, τ_{ij} is the pheromone value for the $-j$ candidate of the $-i$ for hyperparameter, and $\Delta\tau_{ij}$ represents the pheromone deposit based on the PR-AUC of the elite solution. After 10 iterations, the hyperparameter

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

combination with the highest PR-AUC was selected as the XGBoost configuration for the corresponding outer training fold.

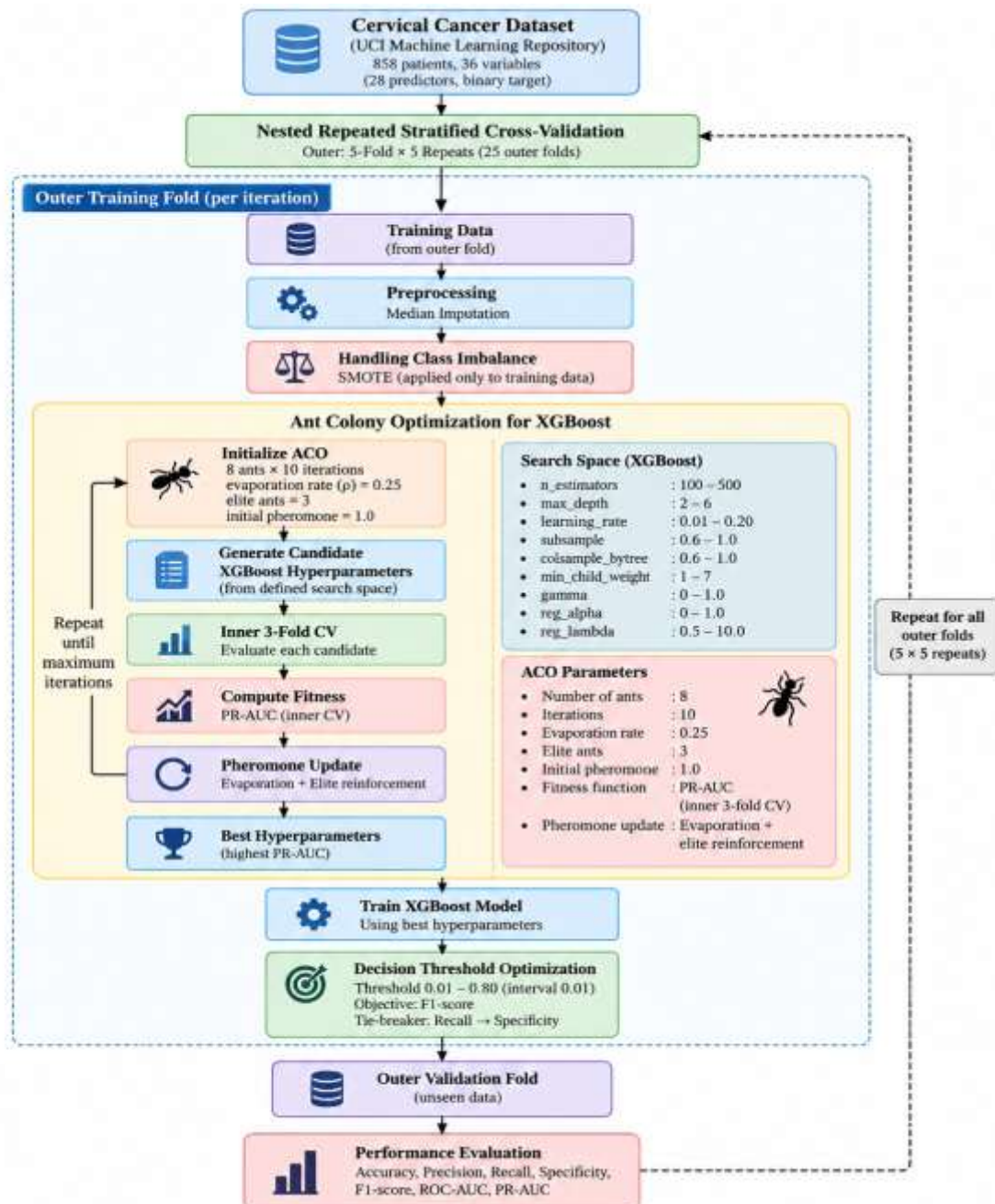


Fig. 1. Workflow of the proposed SMOTE-ACO-XGBoost framework with nested repeated stratified cross-validation

Fig. 1 illustrates the overall workflow of the proposed method, including preprocessing, SMOTE-based class balancing, ACO-based hyperparameter optimization for XGBoost, decision threshold optimization, and outer-fold performance evaluation

Nested Repeated Stratified Cross-Validation

Evaluation was conducted using 5-fold stratified cross-validation with 5 repeats in the outer layer and 3-fold stratified cross-validation in the inner layer. The inner cross-validation was used for ACO hyperparameter optimization and decision-threshold optimization, whereas the outer validation was used solely to obtain the final performance. Thus, imputation, SMOTE, hyperparameter optimization, and threshold optimization did not use information from the outer validation data.

*name of corresponding author



Experimental Setup and Scenarios

All experiments were implemented in Python using the Google Colab environment with NumPy, Pandas, Scikit-learn, Imbalanced-learn, XGBoost, SciPy, and Statsmodels libraries. XGBoost was executed using `tree_method='hist'` without explicit GPU configuration. Google Colab was accessed through a computer running Windows 11 Pro 64-bit with a 12th Gen Intel Core i5-12400 2.50 GHz processor and 8 GB RAM. Because the computations were performed using a cloud-based Google Colab runtime, the local computer specifications are reported as the experimental access environment. Three scenarios were compared: (1) XGBoost using 28 risk factors without SMOTE and ACO; (2) SMOTE-XGBoost using SMOTE without ACO optimization; and (3) SMOTE-ACO-XGBoost using SMOTE with ACO-based hyperparameter optimization. The comparison between the first and second scenarios was used to evaluate the effect of SMOTE, while the comparison between the second and third scenarios was used to evaluate the effect of ACO optimization.

Decision Threshold Optimization

Decision threshold optimization was performed separately from hyperparameter optimization. After the hyperparameter configuration was obtained, out-of-fold probabilities from the inner 3-fold stratified cross-validation were used to evaluate threshold values ranging from 0.01 to 0.80 with an interval of 0.01. The threshold was selected based on the highest F1-score, while recall and specificity were used sequentially as tie-breaking criteria. Thus, PR-AUC was used as the fitness function for ACO-based hyperparameter optimization, whereas F1-score was used as the objective for decision threshold optimization. The entire threshold optimization process was performed only on the training data to prevent data leakage.

Performance Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall/sensitivity, specificity, F1-score, ROC-AUC, and PR-AUC. PR-AUC was included because it focuses more specifically on the precision-recall relationship for the positive class and is relevant for datasets with substantial class imbalance. The results were reported based on 25 outer validation folds to assess both average performance and performance stability.

Statistical Analysis

In addition to comparing the average metrics, statistical analysis was conducted to determine whether differences in model performance were consistent across the same outer validation folds. First, the Friedman Test was used as a non-parametric test to simultaneously compare the three models. The hypotheses were defined as follows: H_0 There is no significant difference in performance among the models. H_1 At least one model has significantly different performance. The significance level was set at $\alpha = 0.05$ if $p < 0.05$ H_0 was rejected. The magnitude of the overall effect was analyzed using Kendall's W, calculated as:

$$W = \frac{\chi_F^2}{N(k-1)} \dots (4)$$

With:

(χ_F^2) = Friedman test statistic;

(N) = number of paired outer-validation scores; and

(k) = number of models.

The interpretation of Kendall's W used in this study is presented in Table 1.

Table 2. Interpretation of Kendall's W

Kendall's W	Interpretation
< 0.10	Very small
0.10–0.29	Small
0.30–0.49	Moderate
≥ 0.50	Large

If the Friedman Test indicated a significant overall difference, the analysis was continued using the Pairwise Wilcoxon Signed-Rank Test to compare the following pairs: XGBoost vs. SMOTE-XGBoost, XGBoost vs. SMOTE-ACO-XGBoost, and SMOTE-XGBoost vs. SMOTE-ACO-XGBoost. Because three pairwise comparisons were conducted for each metric, Holm correction was applied to the Wilcoxon p-values to control the family-wise error rate. A difference was considered statistically significant when the adjusted p-value after Holm correction was less than 0.05. Statistical analyses were conducted separately for accuracy, precision, recall/sensitivity, specificity, F1-score, ROC-AUC, and PR-AUC.

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

RESULT

This study evaluated three classification scenarios, namely XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost, using Nested Repeated Stratified Cross-Validation. The outer validation layer employed 5-fold cross-validation repeated five times, resulting in 25 evaluation results for each model. All results are reported as the mean and standard deviation.

Classification Performance

The evaluation results of the three models are presented in Table 3.

Table 3. Performance Comparison of the Classification Models

Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC	PR-AUC
XGBoost	0.7665 ± 0.1150	0.0972 ± 0.0544	0.3164 ± 0.2229	0.7973 ± 0.1352	0.1367 ± 0.0718	0.6227 ± 0.0663	0.1290 ± 0.0739
SMOTE-XGBoost	0.6276 ± 0.2551	0.0759 ± 0.0576	0.4182 ± 0.3340	0.6419 ± 0.2933	0.1147 ± 0.0763	0.5673 ± 0.0874	0.1012 ± 0.0649
SMOTE-ACO-XGBoost	0.7667 ± 0.1698	0.0992 ± 0.1946	0.2182 ± 0.2180	0.8042 ± 0.1950	0.0901 ± 0.0676	0.5481 ± 0.0882	0.0984 ± 0.0508

Based on Table 3, SMOTE-XGBoost achieved the highest recall of 0.4182 ± 0.3340 , compared with 0.3164 ± 0.2229 for XGBoost and 0.2182 ± 0.2180 for SMOTE-ACO-XGBoost. In contrast, SMOTE-ACO-XGBoost achieved the highest accuracy and specificity, reaching 0.7667 ± 0.1698 and 0.8042 ± 0.1950 , respectively. The baseline XGBoost model obtained the highest F1-score, ROC-AUC, and PR-AUC, with values of 0.1367 ± 0.0718 , 0.6227 ± 0.0663 , and 0.1290 ± 0.0739 , respectively. These results indicate a trade-off among the models rather than the dominance of a single method across all evaluation metrics. The main performance comparison is visualized in Fig. 2.

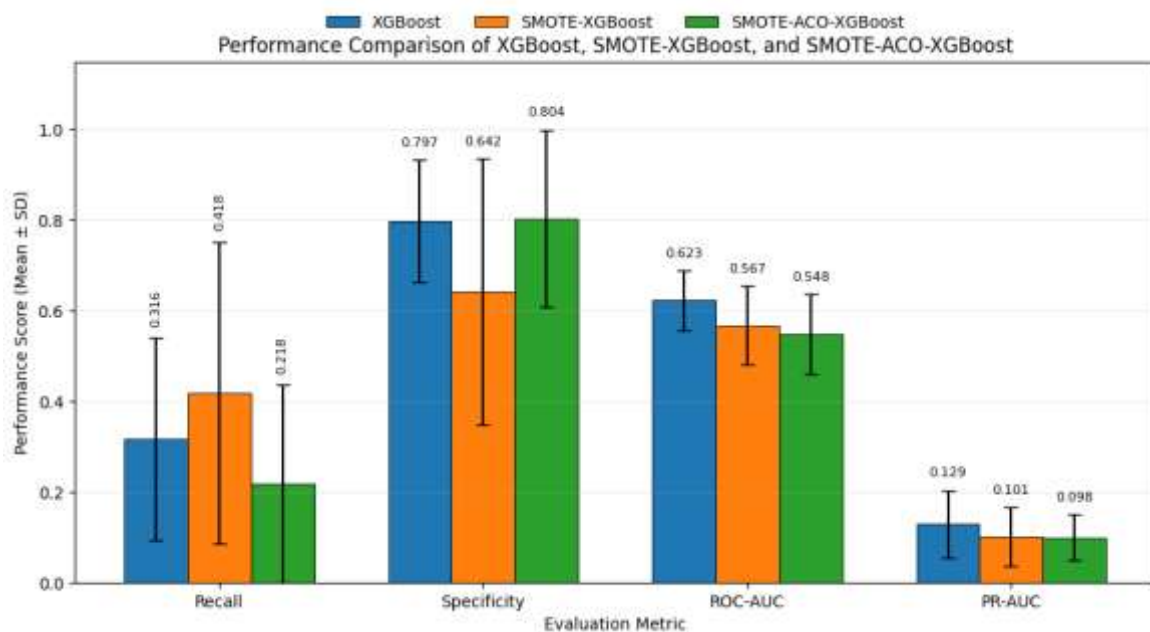


Fig. 2. Performance comparison of XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost across selected evaluation metrics. Error bars indicate standard deviation across 25 outer validation folds.

The standard deviation of recall was relatively high, particularly for SMOTE-XGBoost (0.4182 ± 0.3340). The dataset contained only 55 positive cases, resulting in approximately 11 positive cases in each outer validation fold. Consequently, misclassification of only a few cases could lead to substantial changes in recall. Therefore, the performance results, particularly sensitivity, should be interpreted with caution and should not be considered evidence of strong predictive stability.

Statistical Comparison

Differences in the performance of the three models were analyzed using the Friedman Test. The magnitude of the differences was measured using Kendall's W. The test results are presented in Table 4.

*name of corresponding author



Table 4. Friedman Test and Kendall's W results

Metric	Friedman p-value	Kendall's W	Effect Size	Significance
Accuracy	0.0171	0.163	Small	Significant
Precision	0.0212	0.154	Small	Significant
Recall	0.0020	0.248	Small	Significant
Specificity	0.0261	0.146	Small	Significant
F1-Score	0.0160	0.165	Small	Significant
ROC-AUC	0.000134	0.357	Moderate	Significant
PR-AUC	0.000020	0.434	Moderate	Significant

The Friedman Test indicated significant differences across all metrics ($p < 0.05$). Kendall's W showed a small effect size for accuracy, precision, recall, specificity, and F1-score, whereas ROC-AUC and PR-AUC showed moderate effect sizes. These findings indicate that differences between the models were more consistent in terms of discriminative ability than in most threshold-based metrics.

Pairwise Comparison

Because the Friedman Test indicated significant differences, the analysis was continued using the Pairwise Wilcoxon Signed-Rank Test with Holm correction. The main results of the pairwise comparisons are presented in Table 5.

Table 5. Selected pairwise Wilcoxon-Holm comparison

Metric	Comparison	Mean A	Mean B	Holm-adjusted p-value	Result
Recall	XGBoost vs SMOTE-XGBoost	0.3164	0.4182	0.1975	Not significant
Recall	XGBoost vs SMOTE-ACO-XGBoost	0.3164	0.2182	0.0372	Significant
Recall	SMOTE-XGBoost vs SMOTE-ACO-XGBoost	0.4182	0.2182	0.0115	Significant
Accuracy	SMOTE-XGBoost vs SMOTE-ACO-XGBoost	0.6276	0.7667	0.0142	Significant
Specificity	SMOTE-XGBoost vs SMOTE-ACO-XGBoost	0.6419	0.8042	0.0093	Significant

The application of SMOTE increased recall from 0.3164 to 0.4182 compared with XGBoost; however, the difference was not statistically significant after Holm correction ($p = 0.1975$). In contrast, ACO optimization in the SMOTE-XGBoost model significantly reduced recall from 0.4182 to 0.2182 ($p = 0.0115$). The recall of SMOTE-ACO-XGBoost was also significantly lower than that of the baseline XGBoost ($p = 0.0372$). However, ACO increased accuracy from 0.6276 to 0.7667 ($p = 0.0142$) and specificity from 0.6419 to 0.8042 ($p = 0.0093$) compared with SMOTE-XGBoost. These results demonstrate a shift in the classification characteristics following hyperparameter optimization using ACO.

ACO Hyperparameter Selection

ACO was performed independently within each outer training fold. Therefore, the selected hyperparameter combinations were not necessarily identical across all folds. The most frequently selected parameter values are presented in Table 6.

Table 6. Most frequently selected hyperparameters by ACO

Hyperparameter	Most Frequent Value	Frequency	Selection Rate
n_estimators	100	10/25	40%
max_depth	6	13/25	52%
learning_rate	0.01	7/25	28%

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

subsample	0.8	7/25	28%
colsample_bytree	0.9	7/25	28%
min_child_weight	1	9/25	36%
gamma	0.5	8/25	32%
reg_alpha	1.0	7/25	28%
reg_lambda	2.0	9/25	36%

The percentages in Table 6 represent the selection frequency of the most frequently occurring hyperparameter values (modal values) across the 25 outer training folds and should not be interpreted as the success rate or robustness of ACO. For example, `max_depth = 6` was selected in 13 of the 25 folds (52%), whereas the most frequently selected values for the other hyperparameters occurred in only 28–40% of the folds. These relatively low modal frequencies indicate that ACO did not converge to a single dominant hyperparameter configuration across the training folds. Instead, the selected configurations varied depending on the composition of each outer training set. Therefore, Table 6 reflects hyperparameter variability rather than optimizer robustness. The effect of ACO is more appropriately assessed based on its out-of-sample performance relative to the SMOTE-XGBoost baseline. In this case, ACO significantly improved accuracy and specificity but significantly reduced recall. Since other metaheuristic optimizers were not evaluated under the same experimental protocol, the results of this study cannot establish the superiority of ACO over alternative optimization algorithms.

DISCUSSIONS

The results indicate that class imbalance handling and hyperparameter optimization had different effects on the predictive characteristics of XGBoost. The main finding of this study is not that a single model dominated across all metrics, but rather that there was a trade-off among sensitivity, specificity, and overall discriminative ability.

Effect of SMOTE on Minority-Class Detection

The application of SMOTE increased recall from 0.3164 to 0.4182. This increase indicates that oversampling provided greater representation of the minority class and encouraged the model to identify positive cases more frequently. However, this increase was not statistically significant after Holm correction ($p = 0.1975$) and was accompanied by decreases in accuracy and specificity. This finding demonstrates that SMOTE does not automatically improve all aspects of model performance. In a dataset containing only 55 positive cases, an increase in sensitivity may be accompanied by an increase in false positives. Therefore, model evaluation should consider recall together with precision, specificity, and other discriminative metrics.

Effect of ACO Hyperparameter Optimization

Compared with SMOTE-XGBoost, SMOTE-ACO-XGBoost increased accuracy from 0.6276 to 0.7667 and specificity from 0.6419 to 0.8042. Both increases were statistically significant. However, recall significantly decreased from 0.4182 to 0.2182. This finding indicates that ACO did not improve sensitivity but instead changed the model characteristics toward a more conservative pattern in assigning positive predictions. Therefore, the contribution of ACO in this study is more appropriately interpreted as an optimization that produced higher accuracy and specificity at the cost of reduced sensitivity. The difference between inner and outer validation performance also highlights the importance of nested cross-validation in obtaining a more objective estimate of model generalization.

Discriminative Performance and Variability

The baseline XGBoost model achieved the highest ROC-AUC and PR-AUC values, at 0.6227 and 0.1290, respectively, indicating better discriminative performance than the SMOTE-based models. The Friedman Test also showed significant differences in both metrics, with Kendall's W values of 0.357 for ROC-AUC and 0.434 for PR-AUC. However, the relatively low ROC-AUC and PR-AUC values indicate that the absolute discriminative ability of the models remains limited. Therefore, the results should not be interpreted as evidence that the models are ready for clinical diagnosis. Performance variability across folds was also relatively high, particularly for the recall of SMOTE-XGBoost (0.4182 ± 0.3340). With only 55 positive cases, each outer validation fold contained approximately 11 positive cases, meaning that misclassification of only a few cases could substantially affect recall. The variation in ACO-selected hyperparameters across folds also indicates that the optimal configuration was not fully stable.

As summarized in Table 1, previous studies employed various resampling techniques, classifiers, and optimization strategies and often reported higher performance. However, direct numerical comparisons should be interpreted cautiously because of differences in dataset preparation, target definition, sampling strategies, and

*name of corresponding author



validation protocols. In this study, SMOTE tended to improve sensitivity, whereas ACO optimization increased accuracy and specificity at the expense of sensitivity. Therefore, the main contribution of this study lies in identifying the methodological trade-offs between resampling and optimization strategies rather than demonstrating superior predictive performance. These findings also highlight the importance of using multiple evaluation metrics and validation protocols that minimize data leakage.

Implications and Limitations

The findings indicate that model selection should be adjusted according to the intended application. SMOTE-XGBoost may be more appropriate when sensitivity to positive cases is prioritized, whereas SMOTE-ACO-XGBoost provides better characteristics in terms of accuracy and specificity. The baseline XGBoost demonstrated the highest discriminative ability based on ROC-AUC and PR-AUC.

This study has several limitations, including the small number of positive cases (55), the use of a single public dataset, and the restriction of features to risk factors. Although nested cross-validation reduces the risk of bias resulting from optimization, it does not replace external validation. Future studies should use external datasets with larger numbers of positive cases, compare additional imbalance-handling techniques, and evaluate ACO fitness functions that explicitly consider sensitivity or a combination of sensitivity and specificity.

CONCLUSION

This study evaluated XGBoost, SMOTE-XGBoost, and SMOTE-ACO-XGBoost for risk-factor-based cervical cancer classification using a highly imbalanced dataset. The results showed that no single model outperformed the others across all evaluation metrics. SMOTE-XGBoost achieved the highest recall of 0.4182 ± 0.3340 , but its improvement over XGBoost was not statistically significant. In contrast, SMOTE-ACO-XGBoost achieved the highest accuracy and specificity, at 0.7667 ± 0.1698 and 0.8042 ± 0.1950 , respectively, but its recall significantly decreased to 0.2182 ± 0.2180 .

The baseline XGBoost achieved the highest ROC-AUC and PR-AUC, at 0.6227 ± 0.0663 and 0.1290 ± 0.0739 , respectively, indicating the best discriminative ability among the three models. The Friedman Test showed significant differences across all metrics, while Kendall's W indicated moderate effect sizes for ROC-AUC and PR-AUC.

Overall, SMOTE tended to improve sensitivity, whereas ACO, under the configuration used in this study, shifted the model toward higher accuracy and specificity at the expense of sensitivity. Given the limited number of positive cases and the relatively low discriminative ability of the models, these models should not be considered substitutes for clinical diagnosis. External validation using larger datasets and optimization strategies that explicitly account for sensitivity are recommended for future research.

REFERENCES

- Abdelhay, E. H., Elgamily, K. M., & Walaa Omar El-Farouk Badr. (2026). Metaheuristic optimization of deep CNNs for multi-class diagnosis of cervical cancer and lymphoma. *Scientific Reports*, 16. <https://doi.org/https://doi.org/10.1038/s41598-026-51619-3>
- AlMohimeed, A., Saleh, H., Mostafa, S., Redhwan M. A. Saad, & Talaat, A. S. (2023). Cervical Cancer Diagnosis Using Stacked Ensemble Model and Optimized Feature Selection: An Explainable Artificial Intelligence Approach. *Computers*, 12(10). <https://doi.org/https://doi.org/10.3390/computers12100200>
- Altalhan, M., Algarni, A., & Monia, T. H. A. (2025). Imbalanced Data Problem in Machine Learning: A Review. *Turki Hadj Alouane Monia*, 11. <https://doi.org/10.1109/ACCESS.2025.3531662>
- Anggitasyah, D., & Siregar, M. A. P. (2023). Penerapan Metode Smote Extreme Gradient Boosting Untuk Klasifikasi Penyakit Kanker Serviks Di Kota Medan. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, 12(2). <https://doi.org/https://doi.org/10.35889/jutisi.v12i2.1479>
- Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, 15(1), 73. <https://doi.org/10.1186/s13321-023-00743-7>
- Carolina, I., Andharsaputri, R. L., Suharjanti, S., Prihatin, T., & Nurdin, H. (2026). Comparative Analysis of Multi-Class Classifier Models with Resampling Techniques for Imbalanced Student Graduation Prediction. *Paradigma*, 28(1). <https://doi.org/https://doi.org/10.31294/p.v28i1.11976>
- Çorbacioğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies. *Turkish Journal of Emergency Medicine*, 23(4). https://doi.org/10.4103/tjem.tjem_182_23
- Fernandes, K., Cardoso, J. S., & Fernandes, J. C. (2017). *Cervical Cancer (Risk Factors)*. Springer. UCI Machine Learning Repository. https://doi.org/10.1007/978-3-319-58838-4_27
- Fulazzaky, T., Saefuddin, A., & Soleh, A. M. (2024). Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning: A Comparative Analysis of Balanced Random Forest, SMOTE-RF, SMOTEBoost, and RUSBoost. *Scientific Journal of Informatics*, 11(4).

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

- <https://doi.org/https://doi.org/10.15294/sji.v11i4.15937>
- Geron, A. (2022). *Hands on Machine learning with Scikit Learn Keras and Tensor Flow Concepts, Tools, and Techniques to Build Intelligent Systems* (3rd ed.). O'Reilly Media, Inc.
- Glučina, M., Ariana Lorencin, Nikola Anđelić, & Ivan Lorencin. (2023). Cervical Cancer Diagnostics Using Machine Learning Algorithms and Class Balancing Techniques. *Applied Sciences*, 13(12). <https://doi.org/https://doi.org/10.3390/app13021061>
- Gurcan, F., & Soylu, A. (2024). Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers*, 16(19). <https://doi.org/https://doi.org/10.3390/cancers16193417>
- Hamid, A., & Ridwansyah. (2024). Optimizing Heart Failure Detection : A Comparison between Naive Bayes and Particle Swarm Optimization. *Paradigma*, 26(1), 30–36. <https://doi.org/https://doi.org/10.31294/p.v26i1.3284>
- Huang, C. Y., & Dai, H. L. (2021). Learning from class-imbalanced data: review of data driven methods and algorithm driven methods. *Data Science in Finance and Economics*, 1(1), 21–36. <https://doi.org/10.3934/DSFE.2021002>
- Karamti, H., Alharthi, R., Anizi, A. Al, Reemah M Alhebshi, Eshmawi, A. A., Alsubai, S., & Muhammad Umer. (2023). Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach. *Cancers*, 15(17). <https://doi.org/10.3390/cancers15174412>
- Kondo, T. S., Daniel Ngondya, & Hamim Rusheke. (2025). Application of artificial intelligence in cervical cancer diagnosis using risk factors: A systematic review. *Telematics and Informatics Reports*, 20. <https://doi.org/https://doi.org/10.1016/j.teler.2025.100250>
- Marinho, T. L., Diego Carvalho do Nascimento, & Bruno Almeida Pimentel. (2024). Optimization on selecting XGBoost hyperparameters using meta-learning. *Expert Systems*, 41(9). <https://doi.org/https://doi.org/10.1111/exsy.13611>
- Mudawi, N. Al, & Alazeb, A. (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors*, 22(11). <https://doi.org/https://doi.org/10.3390/s22114132>
- Muraru, M. M., Simó, Z., & Iantovics, L. B. (2024). Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods. *Applied Sciences*, 14(22). <https://doi.org/https://doi.org/10.3390/app142210085>
- Ridwansyah, R., Rahayu, S., Purnama, J. J., Riyanto, V., & Hamid, A. (2026). Pengoptimalan Seleksi Fitur Berbasis Particle Swarm Optimization pada Prediksi Gagal Jantung dengan Random Tree. *JASIEK (Jurnal Aplikasi Sains, Informasi, Elektronika Dan Komputer)*, 8(1). <https://doi.org/https://doi.org/10.26905/jasiek.v8i1.16595>
- Ridwansyah, R., Riyanto, V., Hamid, A., Rahayu, S., & Purnama, J. J. (2022). Grouping Data in Predicting Infant Mortality Using K-Means and Decision Tree. *Paradigma*, 24(2), 168–174. <https://doi.org/10.31294/paradigma.v24i2.1399>
- Salmi, M., Atif, D., Oliva, D., Abraham, A., & Sebastian Ventura. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review (Springer Nature)*, 57(10). <https://doi.org/10.1007/s10462-024-10884-2>
- Saputra, R. M., Alzami, F., Pramudi, Y. T. C., Erawan, L., Megantara, R. A., Ricardus Anggi Pramunendar, & Yusuf, M. (2025). Improving Cervical Cancer Classification Using ADASYN and Random Forest with GridSearchCV Optimization. *Informatics, Electrical Engineering, and Mechanical Engineering*, 16(1). <https://doi.org/https://doi.org/10.35970/infotekmesin.v16i1.2552>
- Sumarna, Astrilyana, Sugiono, Wijaya, G., & Yessica Fara Desvia. (2026). Improving Minority Class Detection in Cervical Cancer Prediction Using Imbalance-Aware Ensemble Learning. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, 10(2). <https://doi.org/10.33395/sinkron.v10i2.15995>
- Sung, H., Filho, A. M., Laversanne, M., Ferlay, J., Rebecca L Siegel, Soerjomataram, I., Ahmedin Jemal, & Bray, F. (2026). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA: A Cancer Journal for Clinicians*, 76(4). <https://doi.org/10.3322/caac.70090>
- Vazquez, B., Rojas-García, M., Rodríguez-Esquivel, J. I., Marquez-Acosta, J., Aranda-Flores, C. E., Cetina-Pérez, L. del C., Soto-López, S., Estévez-García, J. A., Bahena-Román, M., Madrid-Marina, V., & Torres-Poveda, K. (2025). Machine and Deep Learning for the Diagnosis, Prognosis, and Treatment of Cervical Cancer: A Scoping Review. *Diagnostics*, 15(12).
- Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. *Frontiers in Digital Health*, 6(1430245). <https://doi.org/10.3389/fdgth.2024.1430245>
- Yudha, M. A. R., & Rahardi, M. (2025). Comparative Analysis of Random Forest and XGBoost Models for Cervical Cancer Risk Prediction using SHAP-based Explainable AI. *Journal of Applied Informatics and Computing*, 9(6). <https://doi.org/https://doi.org/10.30871/jaic.v9i6.10357>

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.