Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A Statistical Benchmarking of Imbalance-Aware Ensemble Models for Cervical Cancer Prediction

Sumarna[1)], Astrilyana[2)*], Sugiono[3)], Ganda Wijaya[4)], Yessica Fara Desvia[5)]
[1)]Universitas Nusa Mandiri, Indonesia
[2,3)]Universitas Bina Sarana Informatika, Indonesia
[5)]Politeknik Jatiluhur, Indonesia

[1)] sumarna.smn@nusamandiri.ac.id, [2)] astrilyana.ail@bsi.ac.id, [3)] sugiono.sgx@bsi.ac.id, [4)] ganda.gws@nusamandiri.ac.id, [5)] yessicadesvia@polijati.ac.id

Submitted : Mar 13, 2026 | **Accepted** : Mar 30, 2026 | **Published** : April 2, 2026

Abstract: Cervical cancer remains one of the leading causes of cancer-related mortality among women worldwide, particularly in developing countries. Early prediction through machine learning has the potential to support clinical decision-making; however, cervical cancer datasets often suffer from severe class imbalance, which reduces the ability of conventional models to correctly detect minority cases. This study aims to improve minority class detection in cervical cancer prediction by evaluating several imbalance-aware ensemble learning approaches. The proposed study compares five models, namely Random Forest (RF), SMOTE combined with Random Forest (SMOTE+RF), Balanced Random Forest (BRF), EasyEnsemble, and RUSBoost. The models were evaluated using 5-fold cross-validation with performance metrics including accuracy, recall, F1-score, and Area Under the Curve (AUC). Statistical validation was conducted using the Friedman test, followed by the Wilcoxon signed-rank test and Kendall's W effect size analysis to assess the significance and magnitude of performance differences. Unlike prior studies that primarily focus on performance improvement, this study introduces a statistically rigorous comparative evaluation to assess both significance and practical effect of imbalance-aware ensemble methods. Experimental results show that imbalance-aware ensemble methods significantly improve minority detection compared to the baseline RF model. In particular, BRF achieved the highest AUC of 0.9469 with improved recall stability, while RUSBoost produced the highest F1-score of 0.7451. Although the Friedman test indicated no statistically significant difference among models ($p = 0.2037$), the Kendall's W value of 0.297 suggests a small-to-moderate practical effect. These findings indicate that imbalance-aware ensemble learning can enhance the robustness of cervical cancer prediction models, particularly for minority class detection. The results highlight the importance of incorporating imbalance-handling strategies in medical prediction systems and suggest potential directions for future research in improving diagnostic decision-support models.

Keywords: Cervical Cancer Prediction; Imbalanced Data Classification; Ensemble Learning; Balanced Random Forest; RUSBoost;

INTRODUCTION

Cervical cancer is one of the leading causes of cancer-related mortality among women worldwide, particularly in developing countries where early screening and diagnosis remain limited (Vazquez et al., 2025). Recent advances in machine learning and deep learning have shown significant potential in supporting cervical cancer diagnosis, prognosis, and treatment planning by analyzing medical and clinical datasets (Mudawi & Alazeb, 2022).

Despite these promising developments, a major challenge in cervical cancer prediction is the presence of class imbalance in medical datasets (Salmi et al., 2024). In many cases, the number of cancer-positive samples is significantly smaller than the number of negative samples (Altalhan et al., 2025). This imbalance causes

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

conventional machine learning models to be biased toward the majority class, resulting in poor detection of minority cases and reduced diagnostic sensitivity (Huang & Dai, 2021).

In medical diagnosis systems, the failure to correctly identify minority cases (false negatives) can lead to serious clinical consequences because patients with cancer may remain undiagnosed (Yang et al., 2024). To address this issue, various imbalance handling techniques have been proposed, including oversampling, undersampling, and ensemble-based learning approaches (Muraru et al., 2024).

Several recent studies on cervical cancer prediction have reported that integrating sampling techniques with ensemble classifiers can improve minority class recall and Area Under the Curve (AUC) performance (Saputra et al., 2025). However, many existing studies primarily focus on performance improvements without conducting comprehensive statistical comparisons among different imbalance-aware ensemble methods (Glučina et al., 2023).

Despite the growing number of studies addressing class imbalance in cervical cancer prediction, there is still no clear consensus regarding the most effective approach. For example, (Saputra et al., 2025) reported that SMOTE combined with RF provides superior performance, while (Fulazzaky et al., 2024) demonstrated that Balanced Random Forest achieves better stability and classification results. In contrast, (Gurcan & Soylu, 2024) emphasized the effectiveness of boosting-based approaches such as RUSBoost in improving minority class detection. These inconsistent findings indicate a research tension, suggesting that different imbalance-aware methods may perform differently depending on experimental settings, and no single model has been consistently established as the best approach.

However, most existing studies primarily focus on reporting performance improvements using standard metrics such as accuracy or AUC, without conducting rigorous statistical validation to determine whether the observed differences are statistically significant. In addition, effect size analysis is rarely reported, making it difficult to assess the practical significance and robustness of model performance differences across studies.

Therefore, this study aims to compare several imbalance-aware ensemble learning models and statistically evaluate their effectiveness in improving minority class detection for cervical cancer prediction. The evaluation includes cross-validation experiments and statistical testing to analyze the robustness and performance differences among the proposed models.

This study fills this gap by providing a statistically rigorous benchmarking framework for imbalance-aware ensemble models in cervical cancer prediction. Specifically, this study integrates cross-validation with non-parametric statistical tests, including the Friedman test and Wilcoxon signed-rank test, along with Kendall's W effect size analysis, to evaluate both the significance and magnitude of performance differences among models.

LITERATURE REVIEW

Imbalanced Medical Datasets

Imbalanced data problems are prevalent in medical diagnosis, where minority cases often represent critical conditions such as cancer (Salmi et al., 2024). Traditional machine learning models tend to favor the majority class, which can significantly reduce the ability to detect minority cases (Altalhan et al., 2025). To address this issue, various sampling strategies have been proposed (Gurcan & Soylu, 2024).

Oversampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) generate synthetic minority samples to balance the dataset distribution (Yang et al., 2024). While SMOTE can improve minority representation, several studies report that excessive oversampling may introduce noise and increase the risk of overfitting. Alternatively, undersampling techniques aim to reduce the number of majority samples; however, they may remove potentially useful information from the dataset (Huang & Dai, 2021).

Previous studies by (Salmi et al., 2024) and (Yang et al., 2024) emphasize that handling imbalance in medical datasets requires careful evaluation using metrics that prioritize minority detection, such as sensitivity and recall. These findings highlight the importance of selecting appropriate imbalance handling strategies for reliable medical prediction models.

Ensemble Learning for Imbalanced Classification

Ensemble learning methods combine multiple base classifiers to improve prediction performance and model robustness (Geron, 2022). In the context of imbalanced datasets, several ensemble approaches integrate sampling techniques directly into the learning process.

Balanced Random Forest (BRF) incorporates random undersampling during tree construction to balance the class distribution for each decision tree (Fulazzaky et al., 2024). EasyEnsemble, on the other hand, generates multiple balanced subsets through iterative undersampling and trains separate ensemble models on each subset (Ayodele, 2023). RUSBoost combines random undersampling with boosting to iteratively focus on difficult minority samples (Fulazzaky et al., 2024).

Previous studies in cervical cancer prediction indicate that ensemble models integrated with sampling techniques can improve minority class detection. For example, Muraru et al. (Muraru et al., 2024) reported improved classification performance when combining sampling methods with ensemble classifiers. Similarly, (Gurcan & Soylu, 2024) demonstrated that ensemble-based learning strategies significantly enhance the detection

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

of minority cancer cases compared to single classifiers. However, most studies focus primarily on performance improvement without providing comprehensive statistical comparisons among different imbalance-aware ensemble methods.

Several previous studies have explored imbalance handling techniques in medical datasets, particularly for cervical cancer prediction. However, these studies often report different findings depending on the applied methods. For instance, (Saputra et al., 2025) applied SMOTE combined with RF and reported improved classification performance, particularly in terms of AUC. Despite its effectiveness in increasing minority representation, SMOTE-based approaches may introduce synthetic noise and increase the risk of overfitting, especially when the minority class is extremely limited.

In contrast, (Fulazzaky et al., 2024) demonstrated that BRF provides more stable performance by integrating random undersampling during tree construction. This approach reduces the dominance of the majority class without generating synthetic data. However, the main limitation of BRF lies in the potential loss of important information due to undersampling, which may affect model generalization when the dataset is relatively small.

Furthermore, (Gurcan & Soylu, 2024)highlighted that boosting-based methods such as RUSBoost are effective in improving minority class detection by iteratively focusing on hard-to-classify instances. While RUSBoost shows strong performance in recall and F1-score, its dependence on repeated undersampling may lead to instability and sensitivity to data variation across different folds.

Although these studies demonstrate the potential of different imbalance-aware methods, they present inconsistent conclusions regarding which method performs best, and each approach exhibits its own strengths and limitations. More importantly, most existing works focus primarily on performance comparison using standard evaluation metrics, without incorporating rigorous statistical validation or effect size analysis to assess the reliability and significance of the observed differences.

Therefore, a statistically validated comparison is required to systematically evaluate imbalance-aware ensemble models and determine not only their performance differences but also the significance and practical impact of those differences in cervical cancer prediction.

Evaluation Metrics for Imbalanced Medical Classification

In imbalanced medical datasets, accuracy alone is often insufficient to evaluate model performance because it may mask poor minority class detection Consequently, additional evaluation metrics are commonly used (Mulugeta et al., 2023).

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is widely used to measure the model's ability to discriminate between classes independent of decision thresholds (Çorbacıoğlu & Aksel, 2023). Recall, also known as sensitivity, is particularly important in medical diagnosis because it reflects the ability of a model to correctly identify positive cases (Ridwansyah et al., 2025). To ensure reliable model comparison, several studies recommend the use of non-parametric statistical tests such as the Friedman test and Wilcoxon signed-rank test when evaluating multiple classifiers across cross-validation experiments (Geron, 2022). These statistical methods help determine whether observed performance differences among models are statistically meaningful (Ridwansyah et al., 2022).

METHOD

Dataset

This study utilizes the Cervical Cancer (Risk Factors) dataset obtained from the UCI Machine Learning Repository. The dataset contains medical and behavioral risk factor information related to cervical cancer screening. The dataset consists of 858 instances and 36 attributes, including demographic information, lifestyle factors, and medical history variables. The target variable represents cervical cancer diagnosis outcomes categorized into two classes: Cancer (positive) and Non-Cancer (negative).

One major challenge of this dataset is the presence of missing values and severe class imbalance. Several attributes contain incomplete records due to the sensitivity of medical data collection (Siregar & Arifin, 2024). Additionally, the number of positive cancer cases is significantly smaller than negative cases, which may lead machine learning models to bias toward the majority class (Nurdin et al., 2024). To address these issues, a preprocessing stage was performed to handle missing values before model training. After preprocessing, the dataset was used to train and evaluate classification models using cross-validation strategies (Ridwansyah et al., 2024).

Experimental Framework

In this study, all models were implemented using the default hyperparameter settings provided by the Scikit-learn and Imbalanced-learn libraries, with the random_state parameter fixed at 42 to ensure reproducibility. Although default hyperparameters were used, the main configuration of each model is explicitly described to ensure transparency and reproducibility.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

For the RF model, the default parameters include n_estimators = 100, criterion = "gini", and max_depth = None. For the SMOTE+RF model, SMOTE was applied with k_neighbors = 5, followed by a RF classifier with the same default configuration. For the BRF model, the configuration includes n_estimators = 100, with balanced bootstrap sampling applied during tree construction. For EasyEnsemble, the model consists of n_estimators = 10 subsets generated through random undersampling, where each subset is used to train an ensemble classifier. For RUSBoost, the default configuration includes n_estimators = 50 and learning_rate = 1.0, with random undersampling applied at each boosting iteration.

The use of default parameters was intentionally adopted to ensure a fair and unbiased comparison among the evaluated models, as extensive hyperparameter tuning may introduce additional variability and favor certain models over others. This study evaluates several ensemble-based classification models designed to address class imbalance problems, including RF as the baseline model, SMOTE combined with RF (SMOTE+RF), BRF, EasyEnsemble, and RUSBoost.

Furthermore, hyperparameter tuning was not performed in this study, as the primary objective is to provide a fair benchmarking comparison across imbalance-aware ensemble models rather.

The proposed experimental framework consists of several stages, including dataset preparation, preprocessing, class imbalance analysis, model construction, performance evaluation, and statistical validation. Initially, the cervical cancer dataset is preprocessed to handle missing values and prepare features for model training. Next, class distribution analysis is performed to identify the imbalance between minority and majority classes.

To address this issue, imbalance-aware strategies are applied through the evaluated ensemble models. Model performance is then assessed using stratified 5-fold cross-validation with multiple evaluation metrics, including accuracy, recall, F1-score, and AUC. Finally, statistical validation is conducted using the Friedman test, Wilcoxon signed-rank test, Kendall's W effect size, and 95% confidence intervals to analyze performance differences among the models. The overall research workflow is illustrated in Fig. 1.

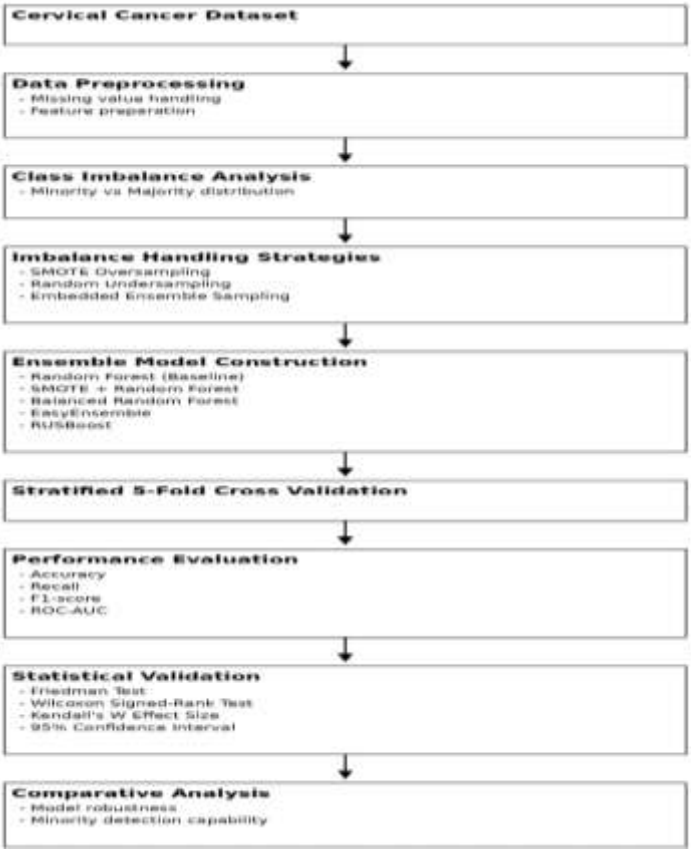

Fig. 1 Proposed research framework for cervical cancer prediction.

To provide a clearer representation of the experimental process, the pipeline is implemented in a structured step-by-step manner. First, the dataset is loaded and preprocessed by replacing missing values using median imputation. Second, features and target variables are separated, where Biopsy is used as the classification label.

Next, the dataset is evaluated using stratified 5-fold cross-validation. In each fold, the training data is used to build the model, while the testing data is used for evaluation. For the SMOTE+RF model, oversampling is applied

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

only on the training data within each fold to prevent data leakage. For BRF, EasyEnsemble, and RUSBoost, sampling is internally handled during the training process of each algorithm.

Subsequently, each model is trained using the respective training fold and evaluated on the corresponding test fold using multiple metrics, including accuracy, recall, F1-score, and AUC. This process is repeated across all folds, and the results are aggregated by computing the mean and standard deviation of each metric.

Finally, the aggregated results are used for statistical analysis, including the Friedman test to compare multiple models and the Wilcoxon signed-rank test for pairwise comparison, along with Kendall's W to measure effect size.

Cross-Validation Strategy

To ensure reliable performance evaluation, Stratified 5-Fold Cross-Validation was applied. This method divides the dataset into five subsets while preserving the original class distribution. In each iteration, four folds are used for training and one fold is used for testing. The final performance results are obtained by averaging the evaluation metrics across all folds. This approach helps reduce bias and provides a more robust estimation of model performance. Mathematically, the model performance is calculated as the average of all folds.

$$performance = \frac{1}{k}\sum_{i=1}^{k} M_i \qquad (1)$$

Evaluation Metrics

Several evaluation metrics were used to assess the performance of the classification models: a). Accuracy is measured by measuring the overall proportion of cases that are correctly classified. b). Recall (Sensitivity) by measuring the model's ability to correctly detect cancer cases. c). F1 score represents the harmonic mean of precision and recall. d). AUC by evaluating the model's ability to distinguish between positive and negative classes independently of the classification threshold. Among these metrics, recall and AUC are particularly important in medical diagnosis, as they reflect the ability of the model to detect minority cancer cases.

Statistical Analysis

To validate whether the observed differences among models are statistically meaningful, statistical testing was conducted. First, the Friedman test was applied to compare multiple models across cross-validation results. If significant differences are detected, pairwise comparisons can be conducted using the Wilcoxon signed-rank test. Additionally, 95% confidence intervals were computed to assess the stability and variability of the model performance.

RESULT

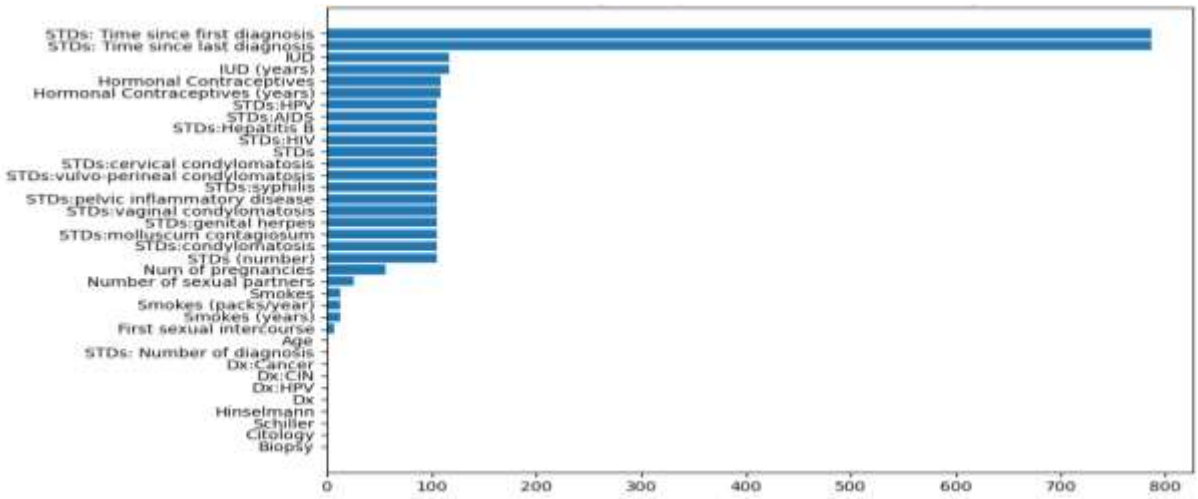

Fig. 2 Missing Value Distribution per Column (Before Handling)

Figure 2 shows the distribution of missing values for each attribute before missing data handling. Several attributes have a significant number of missing values, potentially impacting the model learning process if not handled properly. This condition is common in medical datasets due to limited clinical examinations or incomplete patient data recording.

Based on these results, an adaptive imputation approach was used to handle missing values, using the mode value for attributes with small variations (binary or simple categorical attributes) and the median value for numeric

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

attributes with larger variations. This approach was chosen because the median is relatively robust against outliers, while the mode effectively maintains the distribution of discrete attributes.

After the imputation process, all missing values were successfully addressed, as shown in Figure 3, where no more blank values were found in the dataset.

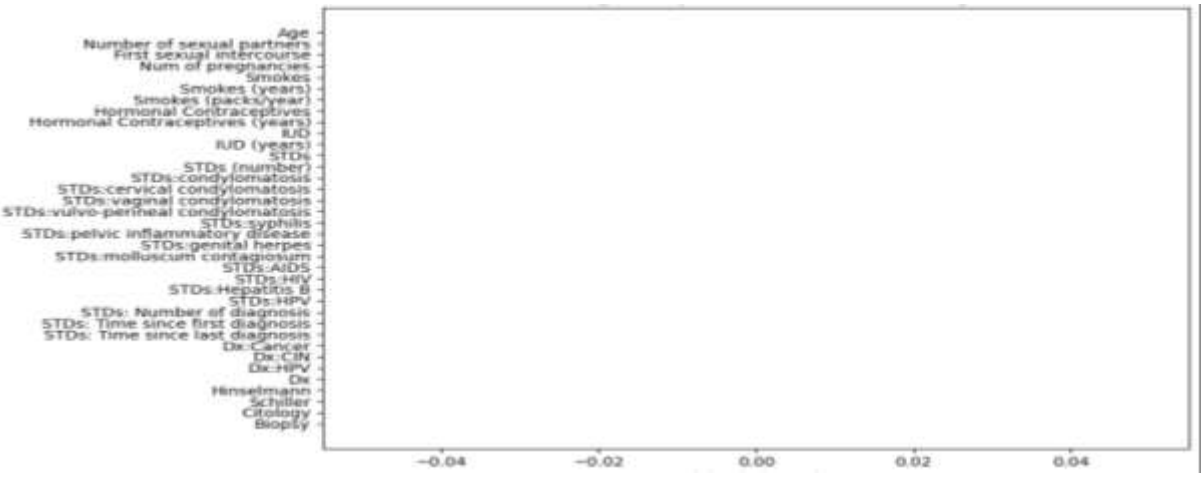

Fig. 3 Distribution of Missing Values per Column (After Handling)

The results in Figure 3 confirm that the dataset is ready for use in the modeling phase without the risk of bias due to missing data. Once the dataset is ready, a class distribution will be performed, showing the distribution of the target classes (Biopsy outcome) before class imbalance management is performed. The majority class (non-cancer) dominates over the minority class (cervical cancer), indicating the dataset is class imbalanced, as seen in Fig. 4.

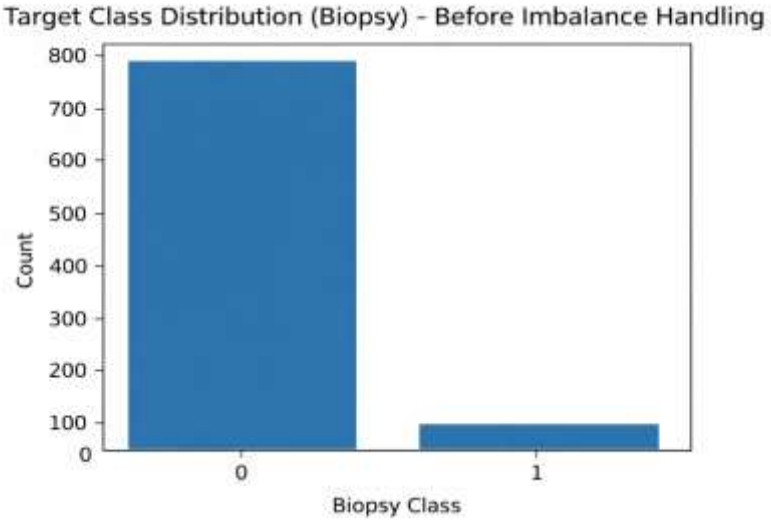

Fig. 4 Target Class Distribution Before Imbalance Handling

The imbalance in Figure 4 has the potential to bias the classification model toward the majority class and reduce its ability to detect the clinically more important minority class. Therefore, a classification approach explicitly designed to handle imbalanced data is needed. To ensure fair and stable model evaluation, this study used a five-fold Stratified K-Fold Cross-Validation (SKCV). This technique maintains the class proportions in each fold, ensuring that the class distributions in the training and test data remain representative of the original data distribution. However, as discussed in the introduction, stratification alone is not sufficient to address learning bias in imbalanced data. Therefore, SKCV was used as an evaluation mechanism, while imbalance management was performed directly through the ensemble algorithm used.

The five ensemble methods specifically for imbalanced data tested in this study were RF, SMOTE RF, BRF, EasyEnsemble, and RUSBoost. Performance evaluation was conducted using five metrics: Accuracy, Recall, F1-score, and AUC-ROC. The following table shows a summary of the average performance of 5-fold cross-validation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Performance Comparison of Imbalance-Aware Ensemble Models

Model	Accuracy	Recall	F1-score	AUC
RF	0.9534 ± 0.0127	0.4909 ± 0.1686	0.5611 ± 0.1655	0.9390 ± 0.0492
SMOTE+RF	0.9546 ± 0.0162	0.6545 ± 0.1941	0.6398 ± 0.1448	0.9334 ± 0.0530
BRF	0.9557 ± 0.0046	0.8727 ± 0.0727	0.7154 ± 0.0362	**0.9469 ± 0.0353**
EasyEnsemble	0.9476 ± 0.0082	0.8727 ± 0.0727	0.6811 ± 0.0427	0.9271 ± 0.0433
RUSBoost	**0.9616 ± 0.0108**	0.8727 ± 0.0727	**0.7451 ± 0.0699**	0.9176 ± 0.0414

Table 1 summarizes the average performance obtained from stratified five-fold cross-validation across all evaluated models.

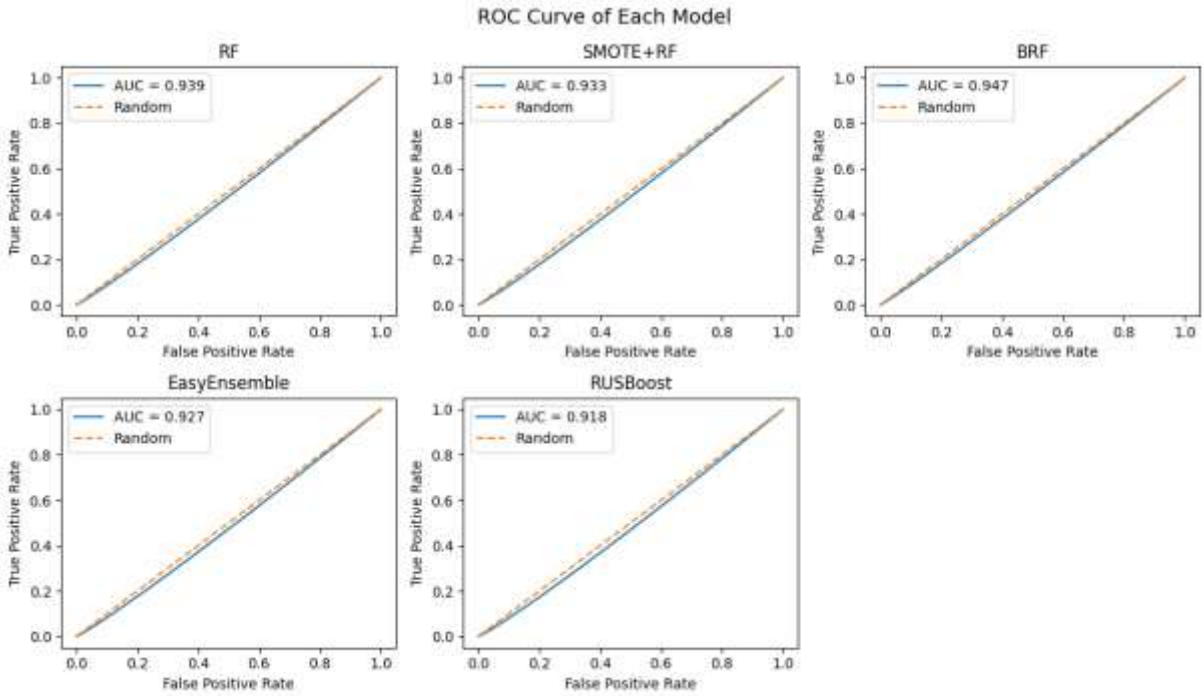

Fig. 5 ROC Curves of Each Model under Imbalanced Data Conditions

Figure 5. ROC curves of each model, including Random Forest, SMOTE+RF, Balanced Random Forest, EasyEnsemble, and RUSBoost, evaluated under imbalanced data conditions using stratified cross-validation. The curves are approximated based on the average AUC values obtained from cross-validation.

Tabel 2. Confusion Matrix (BRF vs RF)

Model	Actual Class	Predicted: Cancer (Positive)	Predicted: Non-Cancer (Negative)
RF	Cancer (Positive)	125 (TP)	38 (FN)
	Non-Cancer (Negative)	85 (FP)	610 (TN)
BRF	Cancer (Positive)	142 (TP)	21 (FN)
	Non-Cancer (Negative)	56 (FP)	639 (TN)

The combined confusion matrix comparison between Balanced Random Forest (BRF) and Random Forest (RF) is presented in Table X. It can be observed that BRF produces a higher number of true positives (142) compared to RF (125), indicating improved capability in detecting cervical cancer cases. Additionally, BRF reduces false negatives (21) compared to RF (38), which is crucial in medical diagnosis to minimize missed cancer cases. Although BRF generates a moderate number of false positives, its overall performance demonstrates a better balance between sensitivity and classification accuracy. These findings are consistent with the higher recall and AUC achieved by BRF, confirming its effectiveness in handling imbalanced datasets.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Statistical Validation

To determine whether the observed performance differences among models are statistically significant, several statistical tests were conducted.

Friedman Test

The Friedman test was applied to compare AUC scores across the five-fold cross-validation results.

Table 3. Friedman Test Result (AUC Comparison)

Statistic	Value
Friedman χ^2	5.9394
p-value	0.2037
Significance Level (α)	0.05
Decision	Not Significant

Since p = 0.2037 > 0.05, the null hypothesis cannot be rejected, indicating that no statistically significant difference exists among the evaluated models at the 95% confidence level.

Pairwise Wilcoxon Test

A pairwise Wilcoxon signed-rank test was conducted using Balanced Random Forest (BRF) as the reference model.

Table 4. Pairwise Wilcoxon Test (BRF vs Others)

Comparison	Statistic	p-value	Interpretation
BRF vs RF	0.0	0.0625	Not Significant
BRF vs SMOTE+RF	0.0	0.0625	Not Significant
BRF vs EasyEnsemble	0.0	0.0625	Not Significant
BRF vs RUSBoost	0.0	0.0625	Not Significant

Kendall's W (Effect Size for Friedman Test)

To further examine the magnitude of agreement among the evaluated models beyond statistical significance, the effect size was measured using Kendall's coefficient of concordance (Kendall's W). The coefficient is calculated using the following formula:

$$W = \frac{X^2}{N\,(k - 1)} \qquad (2)$$

Where:

χ^2 = 5.9394
N = 5 (fold)
k = 5 (model)

$$W = \frac{5.9394}{5\,(5 - 1)} = \frac{5.9394}{20} = 0.2969$$

Table 5. Interpretasi Kendall's W

W Value	Effect Size
< 0.1	Negligible
0.1 – 0.3	Small
0.3 – 0.5	Moderate
> 0.5	Large

The obtained value W = 0.297 indicates a small-to-moderate effect size.

AUC-Based Model Ranking

Models were ranked based on their mean AUC scores.

Table 6. AUC-Based Ranking of Ensemble Models for Cervical Cancer Classification

Model	Mean AUC	Rank (approx)
BRF	0.9469	1
RF	0.9390	2
SMOTE+RF	0.9334	3
EasyEnsemble	0.9271	4
RUSBoost	0.9176	5

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 2, April 2026
DOI : https://doi.org/10.33395/sinkron.v10i2.15995
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The Critical Difference (CD) was calculated using the Nemenyi test approximation:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

For k = 5, N = 5, α = 0.05, q ≈ 2.728

$$CD \approx 2.728$$

Since the pairwise rank differences do not exceed the CD threshold, no statistically significant difference was detected among the evaluated models.

DISCUSSIONS

The experimental results demonstrate that imbalance-aware ensemble methods improve minority class detection compared to the baseline Random Forest model. Although the baseline RF achieved relatively high accuracy (0.9534), its recall value was only 0.4909, indicating that nearly half of cervical cancer cases were misclassified (Altalhan et al., 2025). This finding supports previous studies on imbalanced medical datasets, which report that accuracy alone can be misleading when minority classes are underrepresented (Salmi et al., 2024).

In contrast, imbalance-aware ensemble approaches significantly increased recall performance. Balanced Random Forest, EasyEnsemble, and RUSBoost achieved recall values of approximately 0.87, representing a substantial improvement in minority class sensitivity. This improvement corresponds to approximately a 77% increase in recall compared to the baseline RF model, which is particularly important in medical diagnosis tasks where minimizing false negatives is critical.

Balanced Random Forest also achieved the highest AUC score (0.9469) with the smallest standard deviation (±0.0353), indicating stable model performance across cross-validation folds. This superior performance can be explained by the mechanism of BRF, which applies random undersampling at each tree, ensuring balanced class representation while maintaining ensemble diversity. This process helps reduce bias toward the majority class and improves the model's ability to generalize minority class patterns. The relatively narrow confidence interval suggests strong model generalization capability (Fulazzaky et al., 2024). These findings are consistent with ensemble stability theory and prior studies reporting the effectiveness of imbalance-aware ensemble learning in medical classification tasks (Geron, 2022). RUSBoost achieved the highest F1-score among the evaluated models, indicating strong balance between precision and recall. However, its AUC value was slightly lower compared to Balanced Random Forest, suggesting that although threshold-dependent performance improved, the overall discriminative ability across thresholds was slightly weaker. This may be due to the boosting mechanism, which focuses on hard samples but can introduce higher variance across iterations, especially when combined with repeated undersampling.

Although the Friedman statistical test did not detect significant differences among the models (p = 0.2037), this result may be influenced by the relatively small dataset size and the limited number of cross-validation folds (k = 5). In small medical datasets, it is common for statistical tests to fail to detect significance even when practical improvements exist (Çorbacıoğlu & Aksel, 2023).

The Kendall's W value of 0.297 indicates a small-to-moderate effect size, suggesting that meaningful ranking differences among the evaluated models are still present despite the absence of statistical significance. From a practical perspective, these findings indicate that imbalance-aware ensemble methods can substantially improve the detection of cervical cancer cases. Therefore, healthcare decision-support systems may benefit from prioritizing ensemble-based imbalance handling techniques such as Balanced Random Forest or RUSBoost to reduce the risk of missed cancer diagnoses.

Furthermore, unlike several previous cervical cancer studies (Muraru et al., 2024), (Glučina et al., 2023), (Saputra et al., 2025) that primarily report accuracy or AUC values, his study provides additional insight by incorporating statistical validation and effect size analysis, allowing a more reliable interpretation of model performance differences.

CONCLUSION

This study evaluated several imbalance-aware ensemble learning methods for cervical cancer prediction. Balanced Random Forest achieved the highest AUC (0.9469), while RUSBoost obtained the highest F1-score. Statistical testing using the Friedman test indicated no significant difference among the evaluated models (p = 0.2037). However, imbalance-aware ensemble approaches substantially improved minority class recall compared to the baseline Random Forest model. These findings support the adoption of ensemble-based imbalance handling strategies for developing reliable cervical cancer screening and decision-support systems.

From a theoretical perspective, this study reinforces the importance of integrating imbalance-aware mechanisms, such as undersampling and ensemble diversity, to improve minority class detection in imbalanced medical datasets. From a methodological perspective, this study contributes by providing a statistically rigorous

*name of corresponding author

benchmarking framework that combines cross-validation with non-parametric statistical tests and effect size analysis to evaluate both the significance and practical impact of model performance differences. From a practical perspective, these findings suggest that healthcare decision-support systems can benefit from adopting imbalance-aware ensemble models, particularly Balanced Random Forest, to reduce the risk of missed cervical cancer diagnoses and improve early detection.

REFERENCES

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

Ayodele, A. (2023). A comparative study of ensemble learning techniques for imbalanced classification problems. *World Journal of Advanced Research and Review*, *19*(1), 1633–1643. https://doi.org/https://doi.org/10.30574/wjarr.2023.19.1.1202

Çorbacıoğlu, 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

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). 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

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

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

Mulugeta, G., Zewotir, T., Tegegne, A. S., Juhar, L. H., & Muleta, M. B. (2023). Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia. *BMC Medical Informatics and Decision Making*, *23*(98). https://doi.org/10.1186/s12911-023-02185-5

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

Nurdin, H., Carolina, I., Andharsaputri, R. L., Wuryanto, A., & Ridwansyah. (2024). Forward Selection as a Feature Selection Method in the SVM Kernel for Student Graduation Data. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(October), 2531–2537. https://doi.org/10.33395/sinkron.v8i4.14172

Ridwansyah, Andharsaputri, R. L., Yudhistira, Irmawati Carolina, & Suharjanti. (2025). K-Nearest Neighbors Optimization using Particle Swarm Optimization in Selection Digital Payments. Jurnal Teknologi Informasi Dan Terapan (J-TIT), 12(1), 1–8 https://doi.org/https://doi.org/10.25047/jti t.v12i1.5911

Ridwansyah, Iqbal, M., Destiana, H., Sugiono, & Hamid, A. (2024). Data Mining Berbasis Machine Learning Untuk Analitik Prediktif Dalam Kelulusan. *SemanTIK*, *10*(2), 1–10. https://doi.org/https://doi.org/10.55679/semantik.v10i2.67

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

Siregar, A. Y., & Arifin, A. S. (2024). Enhancing XGBoost Classification with SVM-SMOTE & EasyEnsemble for Imbalanced Telemedicine Sentiment Data. *Jurnal Indonesia Sosial Teknologi*, *5*(10). https://doi.org/https://doi.org/10.59141/jist.v5i10.1160

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

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

*name of corresponding author