# Comparing Random Forest and Gradient Boosting for Monkeypox Diagnosis

Fahlul Rizki, Widowati, Catur Edi Widodo
Doctoral Program of Information System, Postgraduate School, Universitas Diponegoro, Semarang, Indonesia

***Abstract*—Early and accurate diagnosis of Monkeypox is essential to limit transmission and support effective treatment. This study aims to compare the performance of Random Forest and Gradient Boosting models for classifying Monkeypox cases using clinical symptom data. A synthetic dataset from Kaggle containing 25,000 records with 11 symptom-based features was used to evaluate both models under imbalanced and SMOTE-balanced conditions using stratified 5-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curves, and area under the curve (AUC). The experimental results indicate that both models achieve high recall values on imbalanced data, with Gradient Boosting slightly outperforming Random Forest in discriminative performance (AUC 0.6869 vs. 0.6839). While the application of SMOTE improves precision, it reduces recall and provides only marginal improvements in AUC, indicating a trade-off between sensitivity and precision in symptom-based classification. These findings demonstrate the potential of ensemble learning models for symptom-based Monkeypox classification in synthetic tabular datasets. However, further validation using real-world clinical data is necessary before practical diagnostic deployment.**

***Keywords*—Comparative analysis; Random Forest; Gradient Boosting; clinical symptoms; machine learning**

## I. Introduction

Monkeypox represents the largest outbreak in history, with over 55,000 cases reported in 103 countries [1], [2], highlighting the urgent need for reliable detection systems to control its spread and enable timely treatment [3], [4]. This study compares Random Forest (RF) and Gradient Boosting (GB) for Monkeypox diagnosis using clinical symptom data [5], [6].

RF is an ensemble learning method based on multiple decision trees and has been widely used in medical diagnosis due to its high predictive accuracy and robustness in handling complex clinical data [7], [8], [9].

Similarly, GB has gained substantial attention in disease classification tasks, as it constructs strong predictive models by combining multiple weak learners [10]. Modern implementations such as Extreme Gradient Boosting (XGBoost) further improve scalability and computational efficiency, enabling faster and more accurate disease detection in practical settings [11], [12].

Monkeypox diagnosis requires early detection to support timely treatment and outbreak control, supporting the need for reliable symptom-based prediction systems. Although machine learning has been widely applied to disease detection, research that specifically examines the effectiveness of RF and GB for Monkeypox detection using clinical symptom data remains limited. The presence of class-imbalanced datasets also motivates the use of methods such as Synthetic Minority Oversampling Technique (SMOTE) and stratified k-fold cross-validation to achieve fair and unbiased performance comparison.

This study uses a Kaggle dataset with 25,000 samples and 11 clinical features, labeled as Monkeypox-positive or negative.

Model performance is assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curves, and the area under the curve (AUC), to provide a comprehensive comparison of predictive stability and practical screening relevance. In screening scenarios, false negatives are clinically costly, making recall (sensitivity) more relevant than accuracy. Therefore, recall provides a more meaningful indicator of clinical utility than accuracy, particularly in infectious disease screening scenarios.

Most existing Monkeypox detection studies rely on image-based deep learning models such as convolutional neural networks, particularly for skin lesion classification. However, symptom-based diagnostic modeling using classical machine learning on tabular clinical data remains limited, and comparative evaluations under class imbalance are largely unexplored. To the best of our knowledge, no prior work has systematically compared RF and GB for Monkeypox symptom data under class imbalance conditions, incorporating SMOTE + stratified cross-validation and multi-metric reporting, including clinically relevant recall-based evaluation.

This study contributes by providing a novel comparative assessment of RF and GB for symptom-based Monkeypox prediction using both original imbalanced and SMOTE-balanced datasets, employing stratified cross-validation and clinically relevant performance metrics to systematically evaluate model robustness under class imbalance conditions. However, due to limited access to real-world clinical datasets, publicly available datasets remain important for benchmarking ML-based screening models. However, due to limited access to real-world clinical datasets, publicly available datasets remain important for benchmarking ML-based screening models.

While the dataset is synthetic and does not reflect clinical symptom prevalence or measurement noise, such publicly available resources remain important for benchmarking model behavior before real-world validation. Accordingly, this study

does not intend to provide clinically deployable diagnostic tools, but rather to benchmark classical ensemble methods under controlled synthetic conditions to understand their relative behavior, sensitivity to class imbalance, and metric variability.

The remainder of this study is organized as follows: Section II reviews related works. Section III describes the dataset and methodology. Section IV presents the experimental results. Section V presents the discussion, conclusion, and future work.

## II. Related Work

### A. Monkeypox Outbreaks and Diagnostic Challenges

Monkeypox is a zoonotic viral disease caused by the Monkeypox virus (MPXV), originally endemic to Central and West Africa, but which has experienced significant global spread since 2022, affecting over 50 countries worldwide [13]. Clinical presentation typically begins with non-specific symptoms such as fever and lymphadenopathy, followed by a characteristic but variable rash [14]. The heterogeneous nature of symptoms ranging from rash, headache, and oral lesions to flu-like conditions complicates differential diagnosis, as these features overlap with other infectious diseases [15].

Given the variability in symptom presentation and the possibility of misdiagnosis when relying solely on clinical examination, there has been growing interest in computational methods to support symptom-based classification. In situations where access to real clinical datasets is limited, researchers have turned to publicly available and synthetic tabular data as alternative resources for developing and evaluating machine learning models for Monkeypox symptom classification.

### B. Synthetic Datasets for Infectious Disease Modeling

Synthetic datasets allow researchers to simulate clinical scenarios that are difficult to capture in real-world data due to limited case availability [16], providing a proxy for model development and validation when real clinical data are scarce [17].

For example, synthetic datasets have been used to model clinical characteristics and disease progression in COVID-19 [17] and in rare diseases such as Sickle Cell Disease, Cystic Fibrosis, and Duchenne Muscular Dystrophy [18]. Nevertheless, synthetic healthcare datasets pose challenges related to realism, generalizability, and ethics that must be considered when interpreting model performance [19].

In the case of Monkeypox, publicly released synthetic tabular datasets offer a practical surrogate in regions where clinical datasets are inaccessible, motivating their use in this study as an initial evaluation step for symptom-based MPXV classification.

### C. Machine Learning Approaches for Monkeypox Prediction

Machine learning within the healthcare domain utilizes artificial intelligence to autonomously interpret medical data, thereby aiding diagnostic processes and clinical decisions [20], [21]. In the context of infectious diseases, its use increases the accuracy of diagnostic assessments and strengthens preventive healthcare measures [21], [22].

Research has shown that machine learning models can accurately identify infectious diseases from clinical symptom data, such as Monkeypox [20], tuberculosis, COVID-19, HIV, meningitis, malaria, and various bacterial infections [23]. Predictive models may draw on inputs such as radiological imagery, clinical records, and biological markers that hold diagnostic or prognostic value [5], [6], [20].

Recent studies have explored advanced machine learning techniques for symptom-based Monkeypox detection. For example, a stacked ensemble framework combining XGBoost, LightGBM, and Long Short-Term Memory (LSTM) was proposed together with Conditional Tabular GAN (CTGAN) to generate synthetic data for addressing class imbalance, achieving an accuracy of 87.29% and an F1-score of 87.89% in predicting Monkeypox cases based on clinical symptoms [4]. In addition, deep learning approaches have also been widely applied to Monkeypox detection using medical image analysis. A recent study employed GAN-based image augmentation and compared several convolutional neural network architectures, including EfficientNetB3, VGG19, ResNet50, MobileNetV2, VGG16, and InceptionV3, where EfficientNetB3 achieved the best performance with an accuracy of 98.46% for Monkeypox image classification [24]. To provide a clearer overview of recent research on Monkeypox detection, Table I summarizes several previous studies, including the data type, machine learning methods, and reported performance results.

TABLE I. Summary of Recent Monkeypox Detection Studies

| Study | Performance Metrics | | | |
|---|---|---|---|---|
| | ***Data Type*** | ***Method*** | ***Key Technique*** | ***Result*** |
| [4] | Clinical symptom tabular data | Stacking ensemble | XGBoost, LightGBM, LSTM with CTGAN for synthetic data generation | Accuracy 87.29%, F1-score 87.89% |
| [24] | Skin lesion images | Deep learning | GAN-based image augmentation with CNN architectures. | Best accuracy 98.46% using EfficientNetB3 |

Machine learning algorithms are able to uncover intricate symptom patterns that are not easily discernible through manual assessment, which in turn enhances diagnostic accuracy for infectious diseases [23].

### D. Random Forest, Gradient Boosting, and Class Imbalance Handling in Medical Tabular Data

RF and GB are frequently used to process medical tabular data because both are ensemble algorithms capable of handling complex feature interactions. For example, study [25] on cardiovascular disease prediction reported that GB achieved superior evaluation metrics including accuracy, precision, recall, and F1-score while RF produced more

balanced and stable performance. Similarly, study [26] on thyroid disease prediction demonstrated that both GB and RF were highly effective, confirming the potential of ensemble methods to support early diagnosis and clinical decision-making. In Alzheimer's disease classification, study [27] reported strong performance for both models, with GB achieving an evaluation score of 0.95 and RF achieving 0.93. Furthermore, study [28] on coronary heart disease showed that GB handled data complexity more effectively and produced more accurate predictions compared to RF. Collectively, these studies indicate that ensemble learning methods such as RF and GB are consistently applied to medical tabular data due to their stability and high predictive performance.

From a technical perspective, RF relies on ensemble learning to improve accuracy and reduce overfitting, whereas GB incrementally boosts model performance by minimizing residual errors, making both methods suitable for aggregated and complex medical feature sets [25], [26]. However, medical tabular datasets often suffer from class imbalance.

Class imbalance is a critical issue in healthcare data because minority classes are frequently clinically important but underrepresented [29]. When an imbalance is present, machine learning models may become biased toward the majority class and fail to detect minority cases, reducing the overall diagnostic utility [30]. To address this issue, a widely adopted strategy involves the use of SMOTE.

SMOTE generates synthetic samples for the minority class through interpolation, improving minority representation without simple duplication [30]. This technique has been widely used to handle imbalanced medical tabular data, such as in stroke prediction [30], thyroid disease detection [26], and diabetes and breast cancer classification [29]. As a result, SMOTE increases model sensitivity to minority cases, which is crucial in healthcare contexts where rare conditions or high-risk outcomes must be accurately identified [30].

Although RF and GB have been applied across various medical tabular prediction tasks, there is currently no systematic evaluation of their performance for symptom-based Monkeypox diagnosis under class imbalance conditions.

### *E. Summary and Research Gap*

Existing studies on Monkeypox predominantly focus on image-based diagnosis using convolutional neural networks for skin lesion analysis, while research on symptom-based diagnosis using tabular clinical data remains limited. Moreover, for non-Monkeypox diseases, ensemble learning methods such as RF and GB have been widely adopted in medical tabular data due to their strong predictive performance, model robustness, and suitability for complex clinical features. In parallel, class imbalance is a common challenge in medical datasets, and techniques such as SMOTE have been shown to improve the detection of minority cases in various clinical applications.

Despite these advances, several research gaps remain. First, to the best of our knowledge, no prior work has systematically compared RF and GB for symptom-based Monkeypox diagnosis using tabular datasets. Second, existing Monkeypox studies rarely address model performance under class-imbalanced conditions, even though imbalance is common in clinical screening scenarios. Third, recall or sensitivity an important metric in infectious disease screening to avoid false negatives has not been emphasized in previous comparisons. Finally, synthetic tabular Monkeypox datasets have not been utilized to assess method behavior prior to real-world clinical validation.

To address these gaps, this study provides a comparative evaluation of RF and GB for Monkeypox symptom classification under both original imbalanced and SMOTE-balanced conditions, using stratified validation and clinically relevant performance metrics to assess model robustness in early screening contexts.

## III. Dataset and Methodology

This section outlines the methodological framework adopted in this study to compare RF and GB models for symptom-based Monkeypox diagnosis. The methodology encompasses data preprocessing, optional class balancing using SMOTE, model training with hyperparameter optimization, and performance evaluation to ensure a systematic, reproducible, and reliable assessment of the proposed models.

The proposed framework evaluates two ensemble learning algorithms, RF and GB, for symptom-based Monkeypox classification. The workflow consists of data preprocessing, optional class balancing using SMOTE, model training with hyperparameter optimization, and performance evaluation using multiple metrics, including accuracy, precision, recall, F1-score, and AUC. This framework enables a systematic comparison between RF and GB under both imbalanced and balanced data conditions to assess their effectiveness in Monkeypox detection.

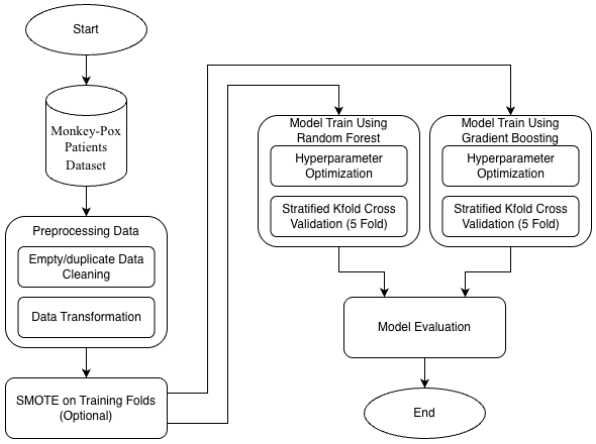


Fig. 1. Research methodology framework.

As illustrated in Fig. 1, each stage of the workflow is described in detail to facilitate a clear understanding of the procedures applied throughout this research.

### *A. Dataset*

This study employs the Monkey-Pox PATIENTS Dataset, published on Kaggle in 2023 [31]. The dataset contains 25,000 synthetic tabular records, of which 15,909 (63.64%) represent confirmed Monkeypox cases and 9,091 (36.36%) represent

non-cases. Each record corresponds to a single patient and comprises 11 features, including one identifier (Patient_ID), one categorical feature (Systemic Illness with categories such as None, Fever, Swollen Lymph Nodes, and Muscle Aches and Pain), and eight binary clinical features (Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection, and Sexually Transmitted Infection). The binary target variable, MonkeyPox, indicates disease presence (1) or absence (0).

The dataset is synthetic and was generated based on clinical findings reported by Adler et al. in The BMJ (2022), titled "Clinical features and novel presentations of human Monkeypox in a central London center during the 2022 outbreak: descriptive case series". It was developed to support research and benchmarking of machine learning models for Monkeypox diagnosis using symptom-based data.

It is important to emphasize that the dataset used in this study is synthetic and was originally generated for research and benchmarking purposes. While synthetic datasets enable controlled experimentation and reproducibility, they may not fully represent real-world clinical distributions, measurement noise, or complex symptom correlations observed in real patient data. Consequently, the results of this study should be interpreted primarily as methodological performance comparisons rather than clinically validated diagnostic outcomes. Future studies should validate the proposed approaches using real-world clinical datasets to ensure practical applicability.

### B. Preprocessing and Class Balancing

Data preprocessing and class balancing were conducted to ensure reliable and unbiased model training. Boolean clinical features were converted into binary numerical values (0 and 1), while the categorical feature Systemic Illness was encoded using one-hot encoding. The target label MonkeyPox was transformed into a binary format, and the patient identifier (Patient_ID) was removed to prevent potential data leakage.

To preserve the original class distribution during evaluation, a stratified splitting strategy was applied within a 5-fold cross-validation framework. Class imbalance observed after preprocessing was addressed using SMOTE, which generates synthetic samples for the minority class to reduce bias toward the majority class. Although the dataset imbalance (approximately 63:37) is moderate rather than extreme, such an imbalance may still influence model predictions by biasing classifiers toward the majority class. In medical screening scenarios, even a moderate imbalance can reduce sensitivity for minority disease cases, motivating the evaluation of resampling techniques such as SMOTE.

To avoid data leakage, SMOTE was applied exclusively to the training data within each cross-validation fold, while validation data remained unchanged. Class balancing is particularly important in medical diagnosis tasks, as imbalanced data may lead to misleading performance estimates and reduced sensitivity to minority-class cases. By integrating SMOTE within the training folds, the proposed approach enables a fair and robust comparison of RF and GB Boosting models for Monkeypox diagnosis.

### C. Models

This study evaluates two tree-based machine learning classifiers, RF and GB, for monkeypox diagnosis using tabular clinical symptom data. RF constructs an ensemble of decision trees using bootstrap sampling and feature randomness to reduce variance, while GB builds trees sequentially, with each model correcting the errors of its predecessor to improve predictive performance.

Model optimization was performed using a randomized hyperparameter search strategy implemented through manual random sampling of parameter combinations. In this procedure, candidate hyperparameters were randomly selected from a predefined search space and evaluated across multiple iterations. A total of 10 random search iterations were conducted to explore different model configurations while maintaining computational efficiency. Each configuration was evaluated using stratified 5-fold cross-validation to ensure robust and unbiased performance estimation. In addition to mean performance values, the standard deviation across cross-validation folds was also recorded to provide a more reliable estimation of model stability and performance variability.

The hyperparameter ranges were selected based on commonly used configurations in ensemble learning literature and preliminary experiments to balance predictive performance and computational efficiency. Table II presents the hyperparameter search space used for RF and GB during the optimization process.

TABLE II. HYPERPARAMETER SEARCH SPACE FOR RF AND GB

| Parameter | Hyperparameter Search |
|---|---|
| | *Candidate Values* |
| n_estimators | 100, 200, 300, 500 |
| max_depth | 5, 10, 20, 30, None |
| min_samples_split | 2, 5, 10 |
| min_samples_leaf | 1, 2, 4 |
| max_features | sqrt, log2 |
| class_weight | None, balanced, balanced_subsample |
| criterion (RF) | entropy, log_loss |
| learning_rate (GB) | 0.01, 0.05, 0.1 |

Each model configuration was evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, ROC curve, and AUC. These metrics were averaged across cross-validation folds to obtain stable performance estimates.

To ensure a fair comparison, both RF and GB were trained and evaluated under identical experimental settings using both the original imbalanced dataset and the SMOTE-balanced dataset. These models were selected due to their robustness, interpretability, and proven effectiveness in handling heterogeneous tabular clinical data, making them well-suited for symptom-based disease classification tasks such as Monkeypox diagnosis.

The best-performing model was selected based on cross-validation performance, with primary emphasis on recall to reflect its clinical importance in screening scenarios, while accuracy and other metrics were used as supporting indicators. A detailed multi-metric evaluation is presented in the Evaluation Metrics section.

### D. Experimental Setup

In this study, RF and GB classifiers were trained and evaluated using an identical experimental protocol to ensure a fair comparison. Model performance was assessed using stratified 5-fold cross-validation, which preserves the original class distribution in each fold and reduces bias caused by data imbalance. For each configuration, the model was trained and validated across the five folds, and the averaged cross-validation results were used to determine the optimal model.

Experiments were performed on both the original imbalanced dataset and a SMOTE-balanced version to analyze model behavior under different class distribution scenarios. The same training, validation, and evaluation procedures were applied consistently across both datasets.

To ensure reproducibility, a fixed random seed was applied during data splitting, resampling, and model training. All experiments were implemented in Python using the scikit-learn library, with imbalanced-learn for SMOTE-based resampling and pandas and NumPy for data preprocessing. Detailed performance metrics are reported separately in the Evaluation Metrics section.

### E. Evaluation Metrics

Model performance was evaluated to assess the predictive capability of RF and GB for monkeypox diagnosis based on clinical symptom data. Evaluation was conducted under two data scenarios: the original imbalanced dataset and the SMOTE-balanced dataset, allowing analysis of model robustness under class imbalance conditions.

Multiple performance metrics were employed, including accuracy, precision, recall, F1-score, ROC curve, and AUC values. These metrics provide complementary perspectives on classification performance and are commonly used in medical decision-support systems.

Accuracy represents the ratio of correctly predicted instances to the overall number of samples evaluated [32]. The accuracy calculation formula is shown in Formula (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

In this formula, TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

The precision metric reflects the percentage of correctly identified positive cases among all predicted positives [32]. The formula for calculating precision is presented in Formula (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

In this formula, TP indicates the number of true positives, whereas FP indicates the number of false positives.

The recall metric indicates the fraction of real positive cases that are successfully recognized by the model [32]. The calculation formula for recall is presented in Formula (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In this formula, TP indicates the number of true positives, whereas FN indicates the number of false negatives.

F1-score serves as a combined indicator of precision and recall, derived from their harmonic mean [32]. The calculation formula for F1-score is presented in Formula (4):

$$\mathrm{F1} = 2\,x\,\frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

In this formula, the F1-score is obtained using the calculated values of precision and recall.

The ROC curve evaluates the model's ability to classify positive and negative outcomes by analyzing the interaction between TPR and FPR under changing thresholds [33]. The formulas for calculating TPR and FPR are presented in the following Formula (5) and Formula (6):

$$\mathrm{TPR} = \frac{TP}{TP+FN} \quad (5)$$

$$\mathrm{FPR} = \frac{FP}{FP+TN} \quad (6)$$

Based on these formulas, the ROC curve can be created using the values of TPR and FPR.

The AUC metric measures the total area beneath the ROC curve, indicating the model's capability to correctly distinguish positive cases from negative ones [33]. An AUC score approaching 1 signifies better model performance, whereas a score closer to 0 reflects poorer predictive ability [33]. The formula for calculating AUC is shown in Formula (7):

$$\mathrm{AUC} = \sum_{i=1}^{n-1}(FPR_{i+1} + FPR_i)\,x\,\frac{(TPR_{i+1}+TPR_i)}{2} \quad (7)$$

Using this formula, the ROC metric indicates how effectively the model separates positive from negative classes.

This multi-metric evaluation framework enables a comprehensive and clinically meaningful comparison of RF and GB models for symptom-based Monkeypox screening.

## IV. Experimental Results

This section presents the experimental results obtained from evaluating RF and GB models for Monkeypox diagnosis using symptom-based clinical features. The experiments were conducted under two dataset configurations, namely the original imbalanced dataset and a SMOTE-balanced dataset, in order to assess the impact of class distribution on model behavior and evaluation metrics. Performance was analyzed across multiple dimensions, including cross-validation metrics (accuracy, precision, recall, and F1-score) as well as threshold-independent measures based on ROC and AUC. The results aim to provide a comprehensive understanding of classifier behavior under different class distribution scenarios and to identify performance trade-offs relevant to early screening and diagnostic support applications.

### A. Cross-Validation Performance (Imbalanced Dataset)

This subsection presents the cross-validation performance of RF and GB models when trained on the original imbalanced dataset. The results reflect the baseline behavior of both models under real class distribution conditions, prior to the application of resampling techniques.

TABLE III. CROSS-VALIDATION PERFORMANCE METRICS ON AN IMBALANCED DATASET.

| Model | Performance Metrics | | | |
|---|---|---|---|---|
| | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
| RF | 0.7083 | 0.7142 | 0.9414 | 0.8122 |
| GB | 0.7102 | 0.7200 | 0.9286 | 0.8111 |

Based on Table III, RF trained on the imbalanced dataset achieved the highest recall value (0.9414), indicating a strong ability to correctly identify Monkeypox-positive cases. This is particularly relevant in screening scenarios, where minimizing false negatives is clinically important. However, the model exhibits moderate precision (0.7142), implying a higher tendency to misclassify negative cases as positive.

Similarly, GB obtained a recall of 0.9286 and a precision of 0.7200, demonstrating comparable screening capability with a slightly more balanced trade-off between recall and precision. Both models achieved similar accuracy (RF: 0.7083, GB: 0.7102) and F1-scores (RF: 0.8122, GB: 0.8111), suggesting no substantial performance difference under imbalanced conditions.

These baseline results indicate that both ensemble models are sensitive to the minority class (positive cases), but precision-related errors remain present due to class imbalance, motivating the use of class-balancing techniques in subsequent experiments.

### B. Cross-Validation Performance (SMOTE-Balanced Dataset)

This subsection reports the cross-validation performance of both classifiers when trained on the SMOTE-balanced dataset. The results allow us to examine how class balancing affects predictive behavior compared to the original imbalanced scenario, particularly with respect to minority-class sensitivity and metric stability.

TABLE IV. CROSS-VALIDATION PERFORMANCE METRICS ON SMOTE-BALANCED DATASET.

| Model | Performance Metrics | | | |
|---|---|---|---|---|
| | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
| RF | 0.6556 | 0.7639 | 0.7033 | 0.7323 |
| GB | 0.6609 | 0.7620 | 0.7182 | 0.7395 |

Based on the results in Table IV, data balancing via SMOTE leads to notable changes in model behavior compared to the original imbalanced scenario. For both classifiers, accuracy decreases slightly after balancing (from 0.7083 to 0.6556 for RF, and from 0.7102 to 0.6609 for GB), indicating that oversampling reduces the strong bias toward the majority class that previously inflated accuracy. In contrast, precision increases for both models (0.7639 for RF and 0.7620 for GB), suggesting that balanced training improves the reliability of positive predictions by reducing false positives.

Recall values show a more substantial shift. For RF, recall decreases from 0.9414 to 0.7033, while GB decreases from 0.9286 to 0.7182 after balancing. This pattern is expected, as imbalanced training tends to inflate recall by encouraging the model to classify a large proportion of samples as positive. After balancing, predictions become more conservative, yielding a more realistic trade-off between precision and recall in screening scenarios. The F1-scores for both models (0.7323 for RF and 0.7395 for GB) reflect this improved balance between sensitivity and specificity under the SMOTE condition.

Overall, these results indicate that class balancing mitigates overfitting tendencies observed under the imbalanced condition, producing more stable performance across metrics. While both models exhibit reduced recall post-balancing, the improvement in precision and the more balanced F1-scores suggest enhanced generalization capability rather than deteriorating performance. Among the two models, GB shows slightly higher recall and F1-score on the balanced dataset, implying better adaptation to the adjusted class distribution. These findings highlight that SMOTE contributes to fairer evaluation and more clinically meaningful performance estimation for symptom-based Monkeypox prediction tasks.

### C. Comparative Performance Across Metrics

This subsection provides a comparative analysis of RF and GB models across multiple evaluation metrics under both imbalanced and SMOTE-balanced conditions. The comparison highlights performance trade-offs across accuracy, precision, recall, and F1-score, enabling a more comprehensive interpretation of each model's suitability for symptom-based Monkeypox diagnosis.

Table III and Table IV collectively reveal the performance trade-offs between RF and GB across both imbalanced and SMOTE-balanced conditions. Under the imbalanced scenario, both models achieve similar accuracy (~0.71) and F1-scores (~0.81), while recall values are notably high for both classifiers (RF: 0.9414, GB: 0.9286). This indicates a strong emphasis on identifying Monkeypox-positive cases, which is desirable for early screening; however, the relatively moderate precision (RF: 0.7142, GB: 0.7200) suggests a higher proportion of false positives, reflecting limited specificity under skewed class distribution. The relatively small performance differences between RF and GB suggest that both models exhibit comparable predictive capability on the evaluated dataset.

Following class balancing, changes occur across all metrics. Accuracy decreases for both models (a reduction of approximately 5% for RF and 4% for GB), suggesting that imbalanced data inflated accuracy by biasing predictions toward the majority class. Conversely, precision increases for both models (RF: from 0.7142 to 0.7639; GB: from 0.7200 to 0.7620), indicating improved reliability of positive predictions. Meanwhile, recall drops substantially for both

models (RF: −23.8%, GB: −21.6%), reflecting a shift from over-sensitive screening to more conservative prediction behavior.

When comparing classifiers, GB consistently achieves slightly higher precision and F1-scores under both scenarios, while RF exhibits higher recall in the imbalanced condition but becomes less sensitive once balanced. This suggests that GB better adapts to balanced data and maintains a more favorable sensitivity-precision trade-off, whereas RF benefits more from class imbalance at the cost of specificity.

From a clinical perspective, high recall is advantageous in screening to minimize missed infections, while precision and F1-score become more relevant in confirmatory settings. Therefore, imbalanced training favors early detection workflows, whereas balanced training provides a fairer evaluation of generalization capability for downstream diagnostic scenarios.

### *D. ROC and AUC Analysis*

This subsection analyzes the ROC curves and AUC scores for both RF and GB models. ROC-AUC provides threshold-independent insight into the discriminative ability of each classifier, allowing a more robust comparison beyond fixed operating points and class distribution effects.

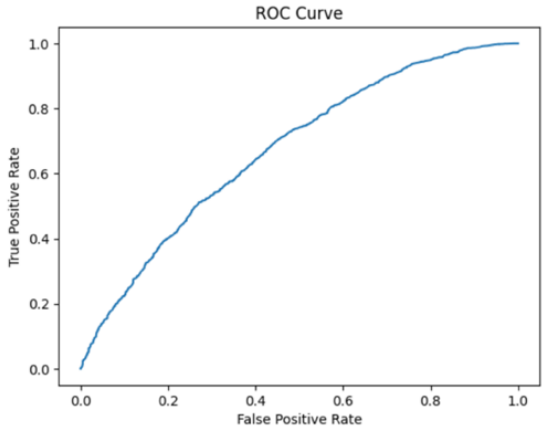


Fig. 2. ROC curve for RF (Imbalanced dataset).

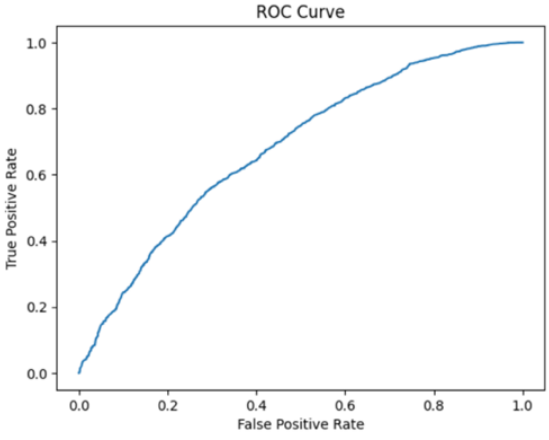


Fig. 3. ROC curve for GB (Imbalanced dataset).

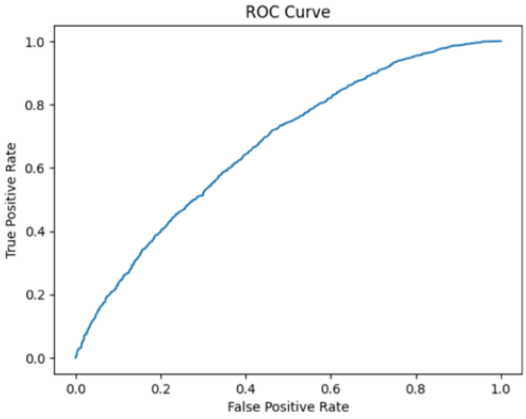


Fig. 4. ROC curve for RF (Balanced dataset).

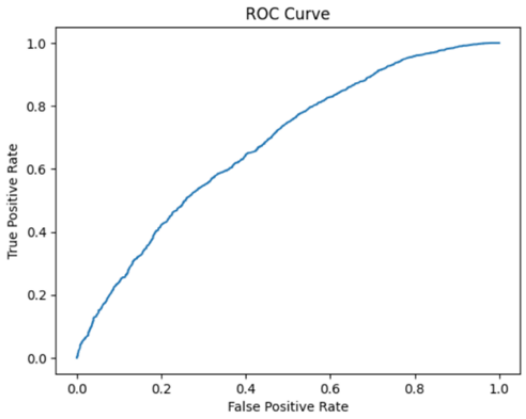


Fig. 5. ROC curve for GB (Balanced dataset).

Fig. 2 to Fig. 5 illustrate the ROC curves for RF and GB under both imbalanced and SMOTE-balanced conditions. Across all scenarios, the ROC curves lie above the diagonal reference line, indicating performance better than random classification. However, notable differences in curve shape and steepness are observed, reflecting how class distribution affects classifier sensitivity-specificity trade-offs.

For the imbalanced dataset, the ROC curves of RF (Fig. 2) and GB (Fig. 3) exhibit moderately steep initial slopes, indicating the ability to achieve relatively high true positive rates (TPR) at low false positive rates (FPR). Between the two classifiers, GB demonstrates a slightly smoother curve, which aligns with its more stable performance across the imbalanced evaluation metrics reported earlier. The overall shape suggests that the imbalanced condition inflates sensitivity (recall), consistent with the high recall values observed in Table III.

After balancing with SMOTE, the ROC curves of RF (Fig. 4) and GB (Fig. 5) become less steep in the early region and more uniformly distributed toward the diagonal-upper region of the plot. This indicates that balancing reduces aggressive positive classification behavior by distributing decision boundaries more evenly across classes. The resulting curves suggest more conservative decision thresholds, consistent with the reduction in recall and improvement in precision reported in Table IV.

When comparing the two classifiers across both conditions, GB consistently yields slightly smoother ROC curves, indicating better probability ranking ability across thresholds. Meanwhile, RF shows a sharper sensitivity shift under imbalance followed by noticeable stabilization after SMOTE balancing, which aligns with its larger recall reduction reported previously.

To provide a quantitative assessment, Table V compares the AUC scores across all configurations. The AUC values show minimal variation between imbalanced and balanced settings, confirming that SMOTE does not substantially degrade overall discriminative capacity, but instead mitigates sensitivity-driven inflation observed in the imbalanced scenario.

TABLE V. AUC SCORES UNDER IMBALANCED AND BALANCED CONDITIONS.

| Model | Performance Metrics |
|---|---|
| | *AUC Score* |
| RF (Imbalance) | 0.6839 |
| GB (Imbalance) | 0.6831 |
| RF (Balanced) | 0.6869 |
| GB (Balanced) | 0.6859 |

As shown in Table V, GB achieves slightly higher AUC values than RF under both imbalanced and SMOTE-balanced conditions, indicating superior probability ranking performance. The difference in AUC scores between imbalanced and balanced configurations is very small for both models (a reduction of 0.0008 for RF and 0.0010 for GB), confirming that SMOTE predominantly affects threshold-dependent metrics such as precision and recall rather than threshold-independent metrics like AUC. Overall, these findings demonstrate that imbalanced training favors sensitivity at the cost of specificity, whereas SMOTE-balanced training promotes more stable and clinically meaningful discrimination. This supports earlier metric-based observations that GB adapts better to balanced class distributions, while RF benefits more from skewed prevalence in early screening scenarios.

### E. Feature Importance Analysis

To improve interpretability and understand how clinical symptoms influence model predictions, feature importance analysis was conducted using SHAP (SHapley Additive exPlanations) on the best-performing model. SHAP provides a unified framework for explaining machine learning predictions by quantifying the contribution of each feature to the model output [9].

Fig. 6 presents the SHAP summary plot for the Gradient Boosting model trained on the SMOTE-balanced dataset. The plot illustrates the relative importance of clinical symptoms in determining the predicted probability of Monkeypox infection. Features appearing at the top of the plot have the greatest influence on the model's predictions. In this study, symptoms related to systemic illness and comorbidity indicators, such as muscle aches and pain, HIV infection, and sexually transmitted infections, exhibit relatively higher contributions to the classification outcome.

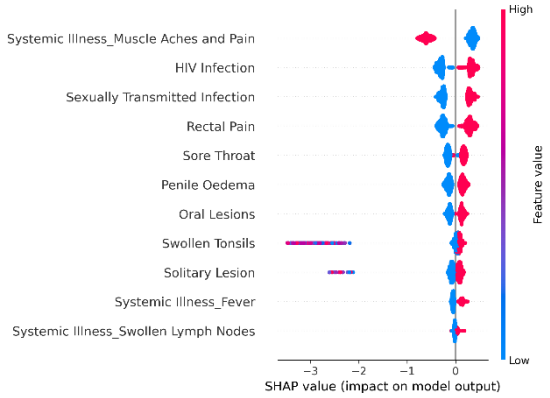


Fig. 6. SHAP summary plot for the Gradient Boosting Model.

The analysis indicates that several symptoms contribute more strongly to the prediction process, suggesting that the model captures meaningful symptom patterns within the dataset. Since the dataset used in this study is synthetic, these importance scores should be interpreted as model-level insights rather than clinically validated associations.

## V. DISCUSSION, CONCLUSION, AND FUTURE WORK

This section synthesizes the experimental findings, summarizes the key conclusions, and outlines directions for future research. The discussion interprets the results in the context of class imbalance effects on ensemble learning performance for Monkeypox detection, while the conclusion highlights the principal contributions and implications of this study. Finally, the future work subsection identifies research extensions aimed at improving model robustness, clinical relevance, and real-world applicability.

### A. Discussion

The experimental results provide insight into the influence of class imbalance on the performance of ensemble learning models for symptom-based Monkeypox detection. Under the original imbalanced dataset, both RF and GB achieved relatively high recall and F1-scores, indicating strong sensitivity toward minority-class (positive) samples. This behavior is desirable in early screening contexts, where minimizing false negatives is a priority. However, the moderate precision observed in both classifiers reflects a tendency to produce false positives due to skewed class distribution, highlighting limited specificity in the absence of resampling.

After applying SMOTE balancing, meaningful shifts occurred across all evaluation metrics. Accuracy decreased for both models, suggesting that the imbalanced setting initially inflated accuracy by biasing predictions toward the majority class. In contrast, precision increased while recall decreased, indicating that SMOTE reduced aggressive positive classification and resulted in more conservative and balanced decision boundaries. This trade-off manifests in a more realistic precision-recall balance, which may be preferable in diagnostic follow-up scenarios rather than broad screening.

ROC-AUC results complemented these findings by revealing minimal variations between imbalanced and balanced configurations. Since AUC is threshold-independent, its stability indicates that class balancing primarily influences threshold-dependent metrics rather than the fundamental discriminative capacity of the classifiers. Across both conditions, GB demonstrated slightly higher AUC values and smoother ROC curves, suggesting better ranking ability across decision thresholds and stronger adaptation to balanced distributions. Meanwhile, RF and GB exhibited higher recall under imbalance but experienced a more pronounced reduction after balancing, indicating sensitivity to class distribution shifts.

Overall, these findings demonstrate that the suitability of a given classifier depends on the operational context: imbalanced conditions may be more suitable for early-stage screening pipelines where sensitivity is prioritized, while balanced conditions provide a fairer assessment for downstream diagnostic settings where both sensitivity and specificity are relevant. An advantage of this study lies in its systematic comparison of two widely used ensemble learning algorithms under both imbalanced and SMOTE-balanced conditions using stratified cross-validation and multiple evaluation metrics. This approach provides a more comprehensive understanding of model behavior in symptom-based Monkeypox prediction. However, several limitations should be acknowledged. First, the dataset used in this study is synthetic and may not fully represent real-world clinical variability or symptom prevalence. Second, the study focuses on classical ensemble methods and does not include deep learning models or multimodal medical data such as medical images. Therefore, future work should evaluate these models on real clinical datasets and explore hybrid or deep learning approaches to improve predictive performance and clinical applicability.

### *B. Conclusion*

This study compared RF and GB for Monkeypox diagnosis using a synthetic clinical symptom dataset under imbalanced and SMOTE-balanced conditions. The results show that class imbalance significantly affects classifier behavior, particularly in threshold-dependent metrics. Both models achieved high recall under the imbalanced condition but exhibited lower precision, indicating limited specificity. After balancing, precision improved and recall decreased, yielding more stable and clinically interpretable outcomes. GB consistently produced slightly higher precision, F1-scores, and AUC values across both conditions, indicating better probability ranking performance and adaptability to balanced class distributions. In contrast, RF benefited more from class imbalance due to its higher recall in the imbalanced scenario. These results highlight the importance of considering class distribution and clinical objectives when deploying symptom-based machine learning models for infectious disease detection.

### *C. Future Work*

While this study demonstrates the effect of SMOTE-based balancing on ensemble model performance, several directions remain for future research. First, the dataset used in this study is synthetic and symptom-based; therefore, validation using real-world clinical datasets is necessary to assess generalizability and practical applicability. Second, only one resampling technique (SMOTE) was evaluated. Future studies may investigate hybrid imbalance handling methods such as SMOTE-ENN, SMOTE-Tomek, and Borderline-SMOTE, which combine oversampling and noise reduction to improve class boundary learning. Third, extending the model comparison to additional ensemble approaches such as XGBoost or LightGBM may further enhance predictive performance. Finally, integrating explainability frameworks such as SHAP or LIME could provide deeper insight into symptom contributions and support interpretable clinical decision systems.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the research funding provided by the Foundation of Universitas Aisyah Pringsewu (Yayasan Aisyah Lampung).

## REFERENCES

[1] B. Cabanillas et al., “A compilation answering 50 questions on monkeypox virus and the current monkeypox outbreak,” Allergy, vol. 78, no. 3, pp. 639–662, Mar. 2023, doi: 10.1111/all.15633.

[2] N. L. Bragazzi et al., “Knowing the unknown: The underestimation of monkeypox cases. Insights and implications from an integrative review of the literature,” Front. Microbiol., vol. 13, pp. 1–12, Sep. 2022, doi: 10.3389/fmicb.2022.1011049.

[3] A. S. Sathwik, B. Naseeba, J. C. Kiran, K. Lokesh, V. S. D. Ch, and N. P. Challa, “Early detection of monkeypox skin disease using patch based DL model and transfer learning techniques,” EAI Endorsed Trans. Pervasive Health Technol., vol. 9, pp. 1–9, May 2023, doi: 10.4108/eetpht.9.4313.

[4] S. Nagro, “A stacked ensemble approach for symptom-based monkeypox diagnosis,” Comput. Biol. Med., vol. 191, p. 110140, Jun. 2025, doi: 10.1016/j.compbiomed.2025.110140.

[5] S. Paul, P. Ranjan, S. Kumar, and A. Kumar, “Disease Predictor Using Random Forest Classifier,” in 2022 International Conference for Advancement in Technology, ICONAT 2022, Goa: Institute of Electrical and Electronics Engineers Inc., Jan. 2022, pp. 1–4. doi: 10.1109/ICONAT53423.2022.9726023.

[6] M. Maindola et al., “Utilizing Random Forests for High-Accuracy Classification in Medical Diagnostics,” in Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024, Greater Noida: Institute of Electrical and Electronics Engineers Inc., Sep. 2024, pp. 1679–1685. doi: 10.1109/IC3I61595.2024.10828609.

[7] I. Altaf, M. A. Butt, and M. Zaman, “Systematic Consequence of Different Splitting Indices on the Classification Performance of Random Decision Forest,” in 2022 2nd International Conference on Intelligent Technologies, CONIT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/CONIT55038.2022.9848372.

[8] S. Georganos et al., “Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling,” Geocarto Int., vol. 36, no. 2, pp. 121–136, 2021, doi: 10.1080/10106049.2019.1595177.

[9] D. Fania, I. Waspada, and H. A. Wibawa, “Addressing Data Limitations in Cardiovascular Disease Prediction: Integration of Public Databases and Clinical Records,” Institute of Electrical and Electronics Engineers (IEEE), Jan. 2026, pp. 293–298. doi: 10.1109/icicos68590.2025.11329869.

[10] A. Es-Smairi et al., “Rare earth (Tm, Y, Gd, and Eu) doped ZnS monolayer: a comparative first-principles study,” Electronic Structure, vol. 6, Jan. 2024, doi: 10.1088/2516-1075/ad17d5.

[11] N. Gunasekara, B. Pfahringer, H. Gomes, and A. Bifet, “Gradient boosted trees for evolving data streams,” Mach. Learn., vol. 113, no. 5, pp. 3325–3352, May 2024, doi: 10.1007/s10994-024-06517-y.

[12] A. Shrivastava et al., "Leveraging XGBoost for Predictive Analytics in Healthcare: Enhancing Disease Diagnosis," in Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 1666–1672. doi: 10.1109/IC3I61595.2024.10829136.

[13] D. K. Bonilla-Aldana and A. J. Rodriguez-Morales, "Is monkeypox another reemerging viral zoonosis with many animal hosts yet to be defined?," Veterinary Quarterly, vol. 42, no. 1, pp. 148–150, Jun. 2022, doi: 10.1080/01652176.2022.2088881.

[14] J. Kaler, A. Hussain, G. Flores, S. Kheiri, and D. Desrosiers, "Monkeypox: A Comprehensive Review of Transmission, Pathogenesis, and Manifestation," Cureus, Jul. 2022, doi: 10.7759/cureus.26531.

[15] V. Jaiswal et al., "Symptomatology, prognosis, and clinical findings of Monkeypox infected patients during COVID-19 era: A systematic-review," Immun. Inflamm. Dis., vol. 10, no. 11, Nov. 2022, doi: 10.1002/iid3.722.

[16] R. Kong et al., "Simulated Infectious Diseases Datasets with Controlled Data Bias," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2025, pp. 5551–5559. doi: 10.1145/3711896.3737401.

[17] G. Isaac H, P. Daniel M, J. Sunny, and M. Volodymyr M, "Semiparametric inference of effective reproduction number dynamics from wastewater pathogen surveillance data," International Biometric Society, vol. 80, no. 3, pp. 1–9, Sep. 2024.

[18] J. M. Mendes, A. Barbar, and M. Refaie, "Synthetic data generation: a privacy-preserving approach to accelerate rare disease research," Front. Digit. Health, vol. 7, 2025, doi: 10.3389/fdgth.2025.1563991.

[19] E. Karoulla, N. Matragkas, S. Ul Islam, and G. Epiphaniou, "Challenges and Risks of Using Synthetic Data for AI-Driven Healthcare Applications," in Studies in Health Technology and Informatics, IOS Press BV, Jun. 2025, pp. 81–85. doi: 10.3233/SHTI250677.

[20] S. Rampogu, "A review on the use of machine learning techniques in monkeypox disease prediction," Science in One Health, vol. 2, p. 100040, Jan. 2023, doi: 10.1016/j.soh.2023.100040.

[21] R. Dhanalakshmi, N. Vijayaraghavan, S. Narasimhan, and R. Rajesh, "A Demonstrative Study on Prediction of Disease Using Bayesian Machine Learning Techniques," in 2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023, PUDUCHERRY: Institute of Electrical and Electronics Engineers Inc., Nov. 2023, p. 10395684. doi: 10.1109/ICSCAN58655.2023.10395684.

[22] C. Xu, L. Y. Zhao, C. S. Ye, K. C. Xu, and K. Y. Xu, "The application of machine learning in clinical microbiology and infectious diseases," Front. Cell. Infect. Microbiol., vol. 15, May 2025, doi: 10.3389/fcimb.2025.1545646.

[23] J. Aalam, S. N. Ahmad Shah, and R. Parveen, "An extensive review on infectious disease diagnosis using machine learning techniques and next generation sequencing: State-of-the-art and perspectives," Comput. Biol. Med., vol. 189, p. 109962, Mar. 2025, doi: 10.1016/j.compbiomed.2025.109962.

[24] D. Sehgal and I. Saini, "GAN-Based Image Augmentation and Comparative Analysis of Various CNN Models for Monkeypox Detection," in Proceedings - 1st International Conference on Electronics, Communication and Signal Processing, ICECSP 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICECSP61809.2024.10697992.

[25] E. Logananthan, M. Naveenkumar, J. K. Kanimozhi, N. Prakash, M. Venkatesan, and A. Yogananthan, "Evaluating the Impact of Gradient Boosting and Random Forest on Accuracy Improvement in Cardiovascular Risk Prediction," in Proc. Int. Conf. Smart Technol. Commun. Robot. (STCR), Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/STCR62650.2025.11019833.

[26] J. Shanthalakshmi Revathy, K. S. Santhoshini, and T. K. Sri Swetha, "Predictive Modeling For Thyroid Disease Using Gradient Boosting And Random Forest Algorithms," in 2nd International Conference on Signal Processing, Communication, Power and Embedded Systems, SCOPES 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/SCOPES64467.2024.10990454.

[27] M. H. Shakir et al., "Optimizing Alzheimer's Disease Diagnosis Using Ensemble Machine Learning Techniques: A Comparative Study," Institute of Electrical and Electronics Engineers (IEEE), Nov. 2025, pp. 1–8. doi: 10.1109/ijcnn64981.2025.11228086.

[28] V. M. Moorthy, B. N. Dharamsoth, V. Muthukaruppan, A. Elango, and K. Ganesan, "Predicting Coronary Heart Disease Using Data Mining and Machine Learning Solutions," An. Acad. Bras. Cienc., vol. 97, no. 3, 2025, doi: 10.1590/0001-3765202520240811.

[29] V. Khullar, M. Angurala, K. D. Singh, P. Prasant, V. Pabbi, and M. R. M. Veeramanickam, "Exploring Methods for Dealing with Class Imbalances in Supervised Machine Learning Structured Datasets," in ACCESS 2023 - 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 209–214. doi: 10.1109/ACCESS57397.2023.10199296.

[30] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Handling Class Imbalance in Health Datasets: A Comparative Study of SMOTE and SMOTEENN with TabNet," in 2025 8th International Conference on Informatics and Computational Sciences (ICICoS), IEEE, Oct. 2025, pp. 305–310. doi: 10.1109/ICICoS68590.2025.11329876.

[31] M. Ahmed, "Monkey-Pox Patients Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset

[32] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Pneumothorax detection system in thoracic radiography images using CNN method," Scientific Journal of Informatics, vol. 11, no. 4, pp. 981–990, Jan. 2025, doi: 10.15294/sji.v11i4.16635.

[33] Y. Yao, Q. Lin, and T. Yang, "Large-scale Optimization of Partial AUC in a Range of False Positive Rates," 2022.