Comparative Analysis of Weighted and Balanced Random Forests for Imbalanced Data Classification

Student Sizhang Lyu
Student Zhe Huang
Professor Sumanta Basu

Introduction

One of the most critical challenges in machine learning is classifying unbalanced data sets. Imbalanced datasets are characterized by a significant disparity in the number of samples among classes, where one class (the majority class) dominates while others (the minority class) are underrepresented. This imbalance often leads to biased models that disproportionately favor the majority class, resulting in poor performance for the minority class. This is typically of greater importance in applications such as fraud detection, medical diagnosis, and risk prediction.

Random forest, introduced by Breiman (1998), is a widely used ensemble learning method, and it has garnered attention for its robustness, interpretability, and ability to handle high-dimensional data. By combining the predictions of multiple decision trees, random forests inherently reduce overfitting and increase generalization. However, when applied to unbalanced datasets, random forests, like most machine learning algorithms, tend to prioritize the prediction of the majority class. This is because the algorithm optimizes for overall accuracy without considering the class distribution, which can obscure performance metrics for minority classes.

To address these issues, several strategies have emerged, including sampling-based techniques, cost-sensitive adjustments, and algorithmic modifications. Among these innovations, Balanced Random Forest (BRF) and Weighted Random Forest (WRF) Chen et al. discuss two promising variants of the traditional Random Forest framework. BRF balances class representation by constructing each tree on equally sized samples of minority and majority classes, thereby alleviating the imbalance at the data level. However, WRF incorporates class-dependent weights into both the splitting criterion and the predictions of terminal nodes, ensuring that the minority class receives the due emphasis without altering the original data distribution.

This report investigates two key aspects of these methods. First, it examines which circumstances BRF or WRF is more suitable, considering factors such as the degree of imbalance, data complexity, data size, and sparsity. Second, it empirically verifies whether both BRF and WRF can consistently outperform benchmark methods, such as synthetic minority over-sampling technique (SMOTE), feature selection, and meta-cost learning, on simulated datasets characterized by varying levels of imbalance, dataset size, sparsity, and retention. By offering a comparative analysis and a set of guidelines, this study seeks to provide practitioners with a clearer understanding of when and why to implement BRF or WRF, ultimately leading to more robust and equitable predictive models in imbalanced learning contexts.

Method

Benchmark Methods

Three established techniques were used as benchmarks to compare with the proposed methods:

1. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE, or Synthetic Minority Over-sampling Technique, introduced by Chawla et al. (2003), offers a remedy by synthesizing new minority class samples, promoting a more balanced dataset, and improving model fairness. SMOTE operates by randomly selecting a point from the minority class and computing its k-nearest neighbors. The synthetic points are then added between the chosen point and its neighbors. This process helps create a more diverse and representative set of features for the minority class, improving model training and generalization.

Using SMOTE, data scientists can create a more balanced training dataset, often leading to better model performance. This is especially true for metrics relevant to the minority class, such as precision, recall, and the F1 score. SMOTE helps address overfitting to the majority class and enables the model to learn more about the minority class.

- SMOTE was applied to balance the class distribution for each dataset.
- The models were trained on the augmented datasets, and performance metrics were recorded.
- SMOTE has the disadvantage that it assumes that the minority class instances are close in feature space. If the minority class is sparse or the data quality is poor, the synthetic samples created may not be representative.
- However, SMOTE outperforms oversampling and downsampling; we will only use SMOTE as the benchmark method.

The SMOTE algorithm involves the following steps:

1. **Identify Minority Class Instances:** Extract the samples belonging to the minority class from the dataset.
2. **Determine Nearest Neighbors:** For each sample in the minority class, compute the k-nearest neighbors using a distance metric such as Euclidean distance.
3. **Generate Synthetic Samples:** Randomly select one or more neighbors and create synthetic samples by interpolating between the original and chosen neighbors. The interpolation formula is:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{original}} + \lambda \cdot (\mathbf{x}_{\text{neighbor}} - \mathbf{x}_{\text{original}})$$

   where $\lambda \in [0, 1]$ is a random value.
4. **Repeat for Desired Oversampling Rate:** Repeat the process for a specified oversampling percentage to generate the required number of synthetic samples for the minority class.
5. **Combine Synthetic and Original Data:** Add the synthetic samples to the original dataset to create a balanced dataset for training.

2. Feature Selection

Feature selection aims to identify the most relevant features, reduce the dataset's dimensionality, and improve interpretability and model efficiency. In this study, the feature selection process was conducted using three well-established techniques (Zheng et al. 2004):

- **Information Gain:** Measures the reduction in entropy achieved by partitioning the data based on a particular feature, identifying features with the most predictive power.
- **Chi-Square Test:** Evaluates the independence of features and class labels, ranking features based on their statistical significance.
- **Correlation Analysis:** Identifies the linear relationships between features and the target variable, selecting features with strong correlations.

The results from these three methods were compared to identify the ten most significant features they shared. These selected features represent the intersection of the methods, ensuring robust feature selection. Using the reduced feature set, a random forest model was trained to evaluate the impact of feature selection on the performance of imbalanced classification tasks. This approach ensures that the model focuses on the most informative features, potentially improving classification accuracy while reducing computational complexity.

3. Meta-Cost Learning

Meta-cost learning proposed by Domingos (1999) is an algorithm-level technique that integrates cost-sensitive learning into standard classifiers to address class imbalance. Instead of modifying the data distribution, meta-cost assigns higher penalties for misclassifications of the minority class during the training process. This approach ensures the classifier prioritizes the correct classification of minority samples, even in highly imbalanced datasets.

The meta-cost framework operates as follows:

1. **Base Model Selection:** A base classifier, such as a decision tree or a random forest, is the foundation for the meta-cost framework.
2. **Cost Matrix Design:** A cost matrix is constructed to represent the penalties associated with different types of misclassifications. In this study, higher costs were assigned to misclassifying minority class samples compared to majority class samples, reflecting the importance of achieving good performance in the minority class.

3. **Training with Cost Adjustment:** The meta-cost framework iteratively modifies the class labels in the training set based on the cost matrix, effectively re-weighting the importance of specific samples. This process adjusts the classifier's decision boundaries to account for the higher penalties associated with minority misclassifications.

4. **Final Model Training:** The modified training set, incorporating cost adjustments, is used to train the final classifier.

This study applied meta-cost learning using random forests as base classifiers. The cost matrix was designed such that:

- Misclassifying a minority class sample incurred a significantly higher penalty than misclassifying a majority class sample.
- The penalties were proportional to the imbalance ratio, ensuring that the framework adapts to datasets with varying levels of class imbalance.

Meta-cost learning improves the model's sensitivity to minority class predictions without compromising the overall classification performance by dynamically adjusting the classifier to focus on the minority class. This makes it a powerful tool for handling imbalanced datasets in various real-world scenarios, such as fraud detection, medical diagnosis, and anomaly detection.

These benchmark methods represent widely adopted strategies for addressing low prediction accuracy for minority classes in imbalanced datasets and were evaluated on the same datasets to ensure a fair comparison.

Novel Methods

Chen et al. introduced two methods incorporating random forests (Breiman, 1998) in his paper, but the authors did not explain the algorithms in detail. Below is a more nuanced discussion.

Consider a binary classification problem with a training dataset:

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N},$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional feature vector and $y_i \in \{0, 1\}$ is the class label of the i-th observation. Let the minority and majority class subsets be:

$$D_{\text{min}} = \{(\mathbf{x}_i, y_i) \in D : y_i = 1\}, \quad D_{\text{maj}} = \{(\mathbf{x}_i, y_i) \in D : y_i = 0\}.$$

Assume $|D_{\text{min}}| = n_{\text{min}}$ and $|D_{\text{maj}}| = n_{\text{maj}}$ with $n_{\text{min}} \ll n_{\text{maj}}$.

Balanced Random Forest (BRF)

The Balanced Random Forest (BRF) method is designed to mitigate the effects of class imbalance by constructing balanced samples at each iteration of the random forest:

1. **Sampling:** For each iteration $b \in \{1, \ldots, B\}$, where B is the number of trees in the ensemble, draw a balanced bootstrap sample as follows:

$$D_{\text{min}}^{*(b)} \sim D_{\text{min}} \text{ with replacement of size } n_{\text{min}}, \; D_{\text{maj}}^{*(b)} \sim D_{\text{maj}} \text{ with replacement of size } n_{\text{min}},$$

and combine these to form:

$$D^{*(b)} = D_{\text{min}}^{*(b)} \cup D_{\text{maj}}^{*(b)}.$$

2. **Tree Induction:** Let $\mathcal{V} = \{1, \ldots, p\}$ be the set of all feature indices. At each node of the tree, randomly select a subset $\mathcal{V}_{\text{try}} \subset \mathcal{V}$ of size m_{try} uniformly at random:

$$\Pr(\mathcal{V}_{\text{try}} = S) = \frac{1}{\binom{p}{m_{\text{try}}}} \quad \text{for all } S \subseteq \mathcal{V} \text{ with } |S| = m_{\text{try}}.$$

Given the current node's training subset $D_{\text{node}} \subseteq D^{*(b)}$, and a candidate feature $j \in \mathcal{V}_{\text{try}}$ with potential split point s, define:

$$D_{\text{left}}(j, s) = \{(\mathbf{x}_i, y_i) \in D_{\text{node}} : x_{ij} \leq s\}, \quad D_{\text{right}}(j, s) = D_{\text{node}} \setminus D_{\text{left}}(j, s).$$

Let $I(\cdot)$ be an impurity measure (e.g., the Gini index). The impurity reduction for splitting on feature j at point s is:

$$\Delta I(D_{\text{node}}, j, s) = I(D_{\text{node}}) - \left(\frac{|D_{\text{left}}(j,s)|}{|D_{\text{node}}|} I(D_{\text{left}}(j,s)) + \frac{|D_{\text{right}}(j,s)|}{|D_{\text{node}}|} I(D_{\text{right}}(j,s)) \right).$$

Select the feature and split point that yield the most significant impurity reduction:

$$(j^*, s^*) = \arg \max_{j \in \mathcal{V}_{\text{try}}, s} \Delta I(D_{\text{node}}, j, s).$$

Repeat this process recursively to grow the tree to full size without pruning. Denote the resulting tree by T_b.

3. **Ensemble Aggregation:** Repeat the sampling and tree induction steps to grow multiple balanced trees. Aggregate the predictions from all trees (e.g., by majority vote) to form the final, more balanced class prediction. After constructing B trees $\{T_1, T_2, \ldots, T_B\}$, each tree T_b provides a class prediction $\hat{y}_i^{(b)}$ for a new instance $\mathbf{x}_i$. The final prediction of the Balanced Random Forest classifier is given by majority voting:

$$\hat{y}_i = \arg \max_{y \in \{0,1\}} \sum_{b=1}^{B} \mathbb{I}\{\hat{y}_i^{(b)} = y\},$$

where $\mathbb{I}\{\cdot\}$ is the indicator function.

Weighted Random Forest (WRF)

The Weighted Random Forest (WRF) method incorporates cost-sensitive learning to handle class imbalance by assigning differing weights to classes:

1. **Tree Construction:** A Weighted Random Forest consists of B classification trees. For each tree $b \in \{1, \ldots, B\}$:

$$D^{*(b)} \subseteq D$$

is constructed by sampling N observations with replacement from D. Each tree T_b is grown from the bootstrap sample $D^{*(b)}$ using a CART-like procedure, modified to incorporate class weights in both splitting and prediction.

2. **Weighted Node Impurity:** At a given node, let:

$$D_{\text{node}} = \{(\mathbf{x}_i, y_i) : i \in I_{\text{node}}\}$$

be the subset of the bootstrap sample that falls into this node. Define:

$$n_0 = \sum_{i \in I_{\text{node}}} \mathbb{I}(y_i = 0), \quad n_1 = \sum_{i \in I_{\text{node}}} \mathbb{I}(y_i = 1).$$

The total weighted count at the node is:

$$W_{\text{node}} = w_0 n_0 + w_1 n_1.$$

The weighted class proportions are:

$$p_0 = \frac{w_0 n_0}{W_{\text{node}}}, \quad p_1 = \frac{w_1 n_1}{W_{\text{node}}}.$$

The weighted impurity at the node can be measured by a Weighted Gini index:

$$I_w(D_{\text{node}}) = p_0(1 - p_0) + p_1(1 - p_1) = 2p_0 p_1.$$

Consider a candidate split on feature j at threshold s, partitioning the data into:

$$D_{\text{left}}(j,s) = \{(\mathbf{x}_i, y_i) \in D_{\text{node}} : x_{ij} \leq s\}, \quad D_{\text{right}}(j,s) = D_{\text{node}} \setminus D_{\text{left}}(j,s).$$

Denote:

$$n_0^L = \sum_{i \in D_{\text{left}}(j,s)} \mathbb{I}(y_i = 0), \quad n_1^L = \sum_{i \in D_{\text{left}}(j,s)} \mathbb{I}(y_i = 1),$$

$$n_0^R = \sum_{i \in D_{\text{right}}(j,s)} \mathbb{I}(y_i = 0), \quad n_1^R = \sum_{i \in D_{\text{right}}(j,s)} \mathbb{I}(y_i = 1).$$

The weighted totals for the left and right nodes are:

$$W_L = w_0 n_0^L + w_1 n_1^L, \quad W_R = w_0 n_0^R + w_1 n_1^R.$$

The weighted class proportions for the left and right nodes are:

$$p_0^L = \frac{w_0 n_0^L}{W_L}, \quad p_1^L = \frac{w_1 n_1^L}{W_L},$$

$$p_0^R = \frac{w_0 n_0^R}{W_R}, \quad p_1^R = \frac{w_1 n_1^R}{W_R}.$$

The weighted impurity for the left and right nodes is:

$$I_w(D_{\text{left}}(j,s)) = 2 p_0^L p_1^L, \quad I_w(D_{\text{right}}(j,s)) = 2 p_0^R p_1^R.$$

The impurity reduction achieved by the split (j, s) is:

$$\Delta I_w(D_{\text{node}}, j, s) = I_w(D_{\text{node}}) - \frac{W_L}{W_{\text{node}}} I_w(D_{\text{left}}(j,s)) - \frac{W_R}{W_{\text{node}}} I_w(D_{\text{right}}(j,s)).$$

The algorithm selects the feature and split that maximizes ΔI_w. As in standard random forests, only a random subset of m_{try} features is considered at each node.

3. **Weighted Terminal Node Prediction:** When a node can no longer be split (e.g., it is pure or a stopping criterion is met), it becomes a terminal node. Let the terminal node contain n_0 observations from class 0 and n_1 observations from class 1. The weighted votes for each class are:

$$V_0 = w_0 n_0, \quad V_1 = w_1 n_1.$$

The predicted class at this terminal node is:

$$\hat{y}_{\text{node}} = \arg \max_{y \in \{0,1\}} V_y.$$

4. **Ensemble Aggregation:** After growing B weighted trees $\{T_1, \ldots, T_B\}$, each tree provides weighted votes $(V_0^{(b)}(\mathbf{x}), V_1^{(b)}(\mathbf{x}))$ for a new instance $\mathbf{x}$. The final predicted class is based on the averaged weighted votes across all trees:

$$\bar{V}_0(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} V_0^{(b)}(\mathbf{x}), \quad \bar{V}_1(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} V_1^{(b)}(\mathbf{x}).$$

The final prediction is:

$$\hat{y}(\mathbf{x}) = \arg \max_{y \in \{0,1\}} \bar{V}_y(\mathbf{x}).$$

This weighting ensures that misclassification costs are explicitly considered both in the splitting process and at the leaf predictions, thus achieving a cost-sensitive ensemble classifier.

These novel methods, BRF and WRF, extend the standard random forest approach to better handle imbalanced datasets. They ultimately improve the predictive fairness and robustness of the resulting models in practical applications. We implemented both methods as specified in the BRFWRF package, which can be downloaded from https://github.com/Googer315/BRFWRF.

Experimental Setup

Synthetic Data Generation

This study evaluates the performance of Balanced Random Forest (BRF) and Weighted Random Forest (WRF) through an extensive simulation experiment. Each data set consists of 20 variables, and all synthetic datasets are constructed to systematically vary four key factors: total sample size, class imbalance ratio, data sparsity, and complexity of the underlying class structure. By considering all factorial combinations of these factors, we generated 36 distinct datasets, each reflecting a unique configuration of the experimental conditions.

Let each dataset be denoted by $\mathcal{D}(n, \alpha, s, c)$, where n is the total sample size, α is the minority class proportion, $s \in \{0, 1\}$ is the sparsity indicator (with $s = 1$ indicating sparse data), and $c \in \{0, 1\}$ is the complexity indicator (with $c = 1$ indicating heterogeneous class structure). The complete set of configurations is given by the product $\{500, 1000, 5000\} \times \{0.05, 0.20, 0.40\} \times \{0, 1\} \times \{0, 1\}$, yielding $3 \times 3 \times 2 \times 2 = 36$ datasets.

- **Dataset Size** (n): We consider three sample sizes, $n \in \{500, 1000, 5000\}$. These sample sizes allow us to examine how the estimators scale from small to relatively large datasets.
- **Minority Ratio** (α): The minority class proportion $\alpha \in \{0.05, 0.20, 0.40\}$. This ensures that we examine performance under severe imbalance ($\alpha = 0.05$), moderate imbalance ($\alpha = 0.20$), and relatively mild imbalance ($\alpha = 0.40$).
- **Sparsity** (s): Two scenarios are considered: $s = 0$ (non-sparse) and $s = 1$ (sparse). For the sparse scenario, features are drawn from a Poisson distribution with parameter $\lambda = 0.2$, leading to a high incidence of zeros. For the non-sparse scenario, features are drawn from a normal distribution $\mathcal{N}(\mu, 1)$ $\exists \mu$, representing continuous, non-zero-centered data.
- **Complexity** (c): Two configurations are assessed: standard two-class data generation ($c = 0$) and heterogeneous class structure ($c = 1$). In the complex configuration, the majority class comprises three distinct subpopulations (clusters) to reflect internal heterogeneity.

Each generated dataset is recorded in a structured list (`all_datasets`), along with descriptive identifiers reflecting the specific values of n, α, s, and c.

Scenario 3 (Sparsity)

To model sparse conditions, each feature X_j is generated from a Poisson distribution with $\lambda = 0.2$:

$$X_j \sim \text{Poisson}(0.2),$$

yielding a high proportion of zero-valued observations. Such data generation mimics real-world scenarios with inherently sparse signals, such as document-term matrices in text analytics or high-dimensional omics data with low expression levels in most samples.

Scenario 4 (Complexity)

To introduce more complex class structures, we employ a mixture model for the majority class. Let $Y \in \{0, 1\}$ be the class label, with $Y = 0$ denoting the majority class and $Y = 1$ the minority class. For the standard scenario ($c = 0$), both classes are drawn from simple normal distributions. For the complex scenario ($c = 1$), the minority class remains a single Gaussian distribution:

$$X \mid Y = 1 \sim \mathcal{N}(3, 1),$$

while the majority class is generated from a mixture of three normal distributions:

$$X \mid Y = 0 \sim 0.8\mathcal{N}(0, 1) + 0.1\mathcal{N}(-3, 1) + 0.1\mathcal{N}(3, 1).$$

This construction captures a heterogeneous majority class that includes one dominant cluster around zero and two smaller clusters at means -3 and 3, respectively. Such a data-generating process more closely approximates real-world conditions in which within-class heterogeneity and overlapping distributions complicate classification tasks. By systematically varying these factors and incorporating complexity and sparsity, we create a controlled environment for comparing the efficacy of BRF and WRF under a broad spectrum of practical conditions.

Metrics for Imbalanced Data

In the presence of imbalanced data, overall classification accuracy is not always a reliable indicator of the model's performance. Accuracy can overemphasize the majority class and undervalue the minority class. As a result, it is often more informative to consider metrics derived from the confusion matrix, which provide a more nuanced understanding of the model's predictive capabilities across both classes.

Confusion Matrix Based Measures

- **Specificity:** Measures the true negative rate. It indicates how well the model identifies negatives.
- **Sensitivity (Recall):** Measures the true positive rate, focusing on how well the model captures the minority class instances.
- **AUC (Area Under the ROC Curve):** Captures the model's ability to rank positives above negatives, independent of classification thresholds.

These metrics allow for a more comprehensive evaluation, mainly when dealing with imbalanced data. They ensure that minority and majority classes are appropriately represented in the model's performance assessment.

The models were trained and tested on the same synthetic datasets under standardized cross-validation procedures. Results from benchmark methods (SMOTE, feature selection, and meta-cost learning) were compared against those from the novel methods (BRF and WRF) to:

- Assess the efficacy of BRF and WRF in addressing class imbalance compared to benchmark methods.
- Understand the trade-offs in performance across different dataset characteristics (e.g., sparsity, imbalance ratios, and complexity).
- Identify scenarios where each novel methods outperform each other.

This comparative analysis provides a comprehensive understanding of the strengths and limitations of each method for imbalanced classification tasks.

Results

Comparison with Benchmark Methods

When compared against a selection of benchmark methods—namely, Feature Selection, MetaCost, and SMOTE—both BRF and WRF demonstrated superior overall performance, particularly in terms of their ability to maintain a favorable balance between sensitivity and specificity, as well as a high area under the ROC curve (AUC). As summarized in Table 1, BRF and WRF achieved comparable AUC values of 0.871, surpassing all benchmark methods. For instance, Feature Selection, despite its exceptionally high sensitivity (0.976), was limited by markedly reduced specificity (0.491) and a lower AUC (0.734). This indicates that the emphasis on identifying minority class instances through Feature Selection was accompanied by a notable increase in false positives.

In comparison, MetaCost and SMOTE, two widely recognized techniques designed to address class imbalance, offered moderate improvements over Feature Selection but could not match the consistently robust performance of BRF and WRF. While MetaCost achieved the highest sensitivity (0.981) and SMOTE maintained strong sensitivity (0.975), both methods exhibited relatively lower specificity and AUC. Specifically, MetaCost's specificity (0.613) and AUC (0.797), together with SMOTE's specificity (0.627) and AUC (0.801), remained substantially below those of BRF and WRF. Thus, although MetaCost and SMOTE effectively enhanced minority class detection, their gains were not as well-rounded, ultimately failing to produce the same balanced performance observed with BRF and WRF.

In sum, the comparison highlights BRF and WRF's superior capacity to achieve consistently strong classification metrics. Both models offered a more favorable trade-off, maintaining high specificity without compromising the accurate detection of minority instances or overall discriminative ability. This underscores the relative advantage of BRF and WRF over these particular benchmark methods when seeking robust, balanced results.

Table 1: Average Metrics

Method	Specificity	Sensitivity	AUC
BRF	0.782	0.959	0.871
WRF	0.774	0.967	0.871
Feature Selection	0.491	0.976	0.734
MetaCost	0.613	0.981	0.797
SMOTE	0.627	0.975	0.801

Compasion of BRF and WRF

Both BRF and WRF were fitted with default hyperparameters, with node size 1 and 100 trees in total. For WRF, the only hyperparameter in addition to the standard RF is the weight, we assigned the default weight to be proportional to the inverse frequency. For BRF, the only hyperparameter in addition to the standard RF is the number of features m_{try} to be selected to fit into each decision tree, and we assigned the default to be the square root of the total number of features. The code and output of BRF and WRF on all 36 data sets are recorded in the appendix section.

Overall, both Balanced Random Forest (BRF) and Weighted Random Forest (WRF) exhibited near-perfect performance in scenarios free from complex structures in the dataset and feature sparsity. Specifically, when data were balanced or only mildly imbalanced and featured continuous, normally distributed predictors, both models achieved accuracy, sensitivity, specificity, and area under the ROC curve (AUC) values at or near 1. These results confirm that under favorable conditions, both BRF and WRF perform at an essentially optimal level.

However, model performance diverged significantly under more challenging conditions. In datasets characterized by high feature sparsity, BRF generally maintained greater robustness than WRF. While both methods experienced declines in performance, WRF was notably more sensitive to extreme sparsity, often exhibiting substantially lower AUC and failing to consistently identify minority class instances. BRF, by contrast, sustained comparatively higher balanced accuracy and sensitivity, highlighting its greater resilience when a large proportion of features carried zero or near-zero values.

In extremely imbalanced datasets, particularly those with as little as 5% minority class representation, WRF frequently struggled to detect any minority instances at all, resulting in sensitivity values of zero. BRF, although not immune to the difficulties posed by severe imbalance, performed more reliably in identifying minority class members. Furthermore, when the majority class structure was heterogeneous—incorporating multiple subclusters—WRF demonstrated slightly better performance relative to BRF. In these complex majority-class scenarios, WRF's higher specificity scores and improved balanced accuracy indicated its ability to capture subtle distinctions within the data distribution when downsampling could indicate loss of important information.

Increasing dataset size generally improved performance for both models, as larger sample sizes helps mitigate the negative effects of imbalance and sparsity. Nonetheless, the detrimental impact of extreme sparsity on WRF persisted even as datasets grew in size, while BRF's overall performance remained comparatively stable. These findings suggest that data characteristics, rather than sample size alone, play a central role in determining each model's effectiveness.

In sum, BRF exhibited particular strengths in sparse, highly imbalanced, and heterogeneous scenarios, showing more reliable detection of minority classes and demonstrating greater versatility across diverse data conditions. WRF, although highly effective in moderately imbalanced and non-sparse datasets, proved less adaptable in the face of adverse conditions. These results underscore the importance of considering the specific properties of a dataset—such as the degree of sparsity, class imbalance, and complexity—when selecting a classification method. Preprocessing strategies to reduce sparsity or rebalance classes are likely to further enhance the performance of both models, ensuring a better match between the learning algorithm and the problem's intrinsic characteristics. Also, a tradeoff in hyperparameters exists for both models: increasing the number of variables used to fit each tree m_{try} could improve the performance but experiences slower computation and is more prone to overfitting, and assigning much higher weights to the minority class than default could dramatically improve specificity but could also drag sensitivity down.

Conclusion and Further Directions

This simulation study demonstrates that both Balanced Random Forest (BRF) and Weighted Random Forest (WRF) are capable of near-optimal performance in balanced, non-sparse settings. However, performance diverges significantly under more challenging conditions. BRF emerges as the more robust and versatile method, maintaining higher sensitivity and balanced accuracy in scenarios marked by severe sparsity, extreme class imbalance, and

heterogeneous majority class distributions. In contrast, WRF tends to favor the majority class in difficult conditions, resulting in weaker detection of minority instances.

These findings highlight that the choice between BRF and WRF should be guided by the specific characteristics and constraints of the data set. Employing WRF may be preferable when class imbalance is present but not extreme, when it is critical to retain all data points (e.g., when each sample carries unique and valuable information), and when the data set is not sufficiently large to accommodate extensive downsampling. On the other hand, BRF is the more suitable option when the data set is sparse, highly imbalanced, or when the majority class is generated by a coherent, homogeneous process, as it more effectively addresses these complex and adverse scenarios.

Benchmark comparisons with methods like Feature Selection, MetaCost, and SMOTE further underscore the advantages of BRF and WRF. The latter two demonstrated a more favorable overall performance profile, confirming the conclusion of Chen et al. These results reinforce the importance of selecting appropriate methods based on class distribution and other data attributes such as sparsity and complexity.

Future research could explore several promising avenues. First, extending the investigation to even more complex data scenarios, such as higher-dimensional feature spaces or multi-class settings, may yield additional insights into the scalability and generalizability of BRF and WRF. Second, hybrid methods that combine BRF or WRF with feature extraction or a voting ensemble classifier between the two could be developed to further improve their robustness. Finally, applying these insights and models to domain-specific problems—such as clinical diagnostics, fraud detection, or text classification—would help validate the practical utility and efficacy of these approaches in addressing real-world challenges.

References

- Chen, C., Liaw, A., & Breiman, L. (n.d.). Using random forest to learn imbalanced data. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf
- Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. (2003). SMOTEBoost: Improving prediction of the minority class in ... Data Science Association. http://www.datascienceassn.org/sites/default/files/SMOTEBoost Improving Prediction of the Minority Class in Boosting .pdf
- Breiman, L. (1998). Classification and regression trees. Chapman and Hall/CRC.
- Colantonio, S., Little, S., Salvetti, O., & Perner, P. (2010). Prototype-based classification in unbalanced biomedical problems. Studies in Computational Intelligence, 143–163. https://doi.org/10.1007/978-3-642-14078-5_7
- Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. 2008 Fourth International Conference on Natural Computation, 192–201. https://doi.org/10.1109/icnc.2008.871
- Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on Imbalanced Data. ACM SIGKDD Explorations Newsletter, 6(1), 80–89. https://doi.org/10.1145/1007730.1007741
- Domingos, P. (1999). Metacost. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/312129.312220

Appendix

The code and original output file can also be downloaded from https://github.com/Googer315/BRFWRF.

```
Confusion Matrix for brf model on dataset: Data_n500_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n500_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

AUC for brf model on dataset: Data_n500_imbalance5perc : 1
AUC for wrf model on dataset: Data_n500_imbalance5perc : 1
Confusion Matrix for brf model on dataset: Data_n500_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n500_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n500_imbalance20perc : 1
AUC for wrf model on dataset: Data_n500_imbalance20perc : 1
Confusion Matrix for brf model on dataset: Data_n500_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n500_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n500_imbalance40perc : 1
AUC for wrf model on dataset: Data_n500_imbalance40perc : 1
Confusion Matrix for brf model on dataset: Data_n1000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n1000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

AUC for brf model on dataset: Data_n1000_imbalance5perc : 1
AUC for wrf model on dataset: Data_n1000_imbalance5perc : 1
Confusion Matrix for brf model on dataset: Data_n1000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n1000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n1000_imbalance20perc : 1
AUC for wrf model on dataset: Data_n1000_imbalance20perc : 1
Confusion Matrix for brf model on dataset: Data_n1000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n1000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n1000_imbalance40perc : 1
AUC for wrf model on dataset: Data_n1000_imbalance40perc : 1
Confusion Matrix for brf model on dataset: Data_n5000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n5000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   0
         1   0  10

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 3.505e-05

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00
            Specificity : 1.00
         Pos Pred Value : 1.00
         Neg Pred Value : 1.00
             Prevalence : 0.05
         Detection Rate : 0.05
   Detection Prevalence : 0.05
      Balanced Accuracy : 1.00

       'Positive' Class : 1

AUC for brf model on dataset: Data_n5000_imbalance5perc : 1
AUC for wrf model on dataset: Data_n5000_imbalance5perc : 1
Confusion Matrix for brf model on dataset: Data_n5000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n5000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160   0
         1   0  40

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.2
         Detection Rate : 0.2
   Detection Prevalence : 0.2
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n5000_imbalance20perc : 1
AUC for wrf model on dataset: Data_n5000_imbalance20perc : 1
Confusion Matrix for brf model on dataset: Data_n5000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: Data_n5000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140   0
         1   0  60

               Accuracy : 1
                 95% CI : (0.9817, 1)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0
            Specificity : 1.0
         Pos Pred Value : 1.0
         Neg Pred Value : 1.0
             Prevalence : 0.3
         Detection Rate : 0.3
   Detection Prevalence : 0.3
      Balanced Accuracy : 1.0

       'Positive' Class : 1

AUC for brf model on dataset: Data_n5000_imbalance40perc : 1
AUC for wrf model on dataset: Data_n5000_imbalance40perc : 1
Confusion Matrix for brf model on dataset: SparseData_n500_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 176   8
         1  14   2

               Accuracy : 0.89
                 95% CI : (0.8382, 0.9298)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9998

                  Kappa : 0.0984

 Mcnemar's Test P-Value : 0.2864

            Sensitivity : 0.2000
            Specificity : 0.9263
         Pos Pred Value : 0.1250
         Neg Pred Value : 0.9565
             Prevalence : 0.0500
         Detection Rate : 0.0100
   Detection Prevalence : 0.0800
      Balanced Accuracy : 0.5632

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n500_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n500_imbalance5perc : 0.5631579
AUC for wrf model on dataset: SparseData_n500_imbalance5perc : 0.5
Confusion Matrix for brf model on dataset: SparseData_n500_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 152   9
         1   8  31

               Accuracy : 0.915
                 95% CI : (0.8674, 0.9497)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 6.879e-06

                  Kappa : 0.7319

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7750
            Specificity : 0.9500
         Pos Pred Value : 0.7949
         Neg Pred Value : 0.9441
             Prevalence : 0.2000
         Detection Rate : 0.1550
   Detection Prevalence : 0.1950
      Balanced Accuracy : 0.8625

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n500_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 160  26
         1   0  14

               Accuracy : 0.87
                 95% CI : (0.8153, 0.9133)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.006433

                  Kappa : 0.4628

 Mcnemar's Test P-Value : 9.443e-07

            Sensitivity : 0.3500
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8602
             Prevalence : 0.2000
         Detection Rate : 0.0700
   Detection Prevalence : 0.0700
      Balanced Accuracy : 0.6750

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n500_imbalance20perc : 0.8625
AUC for wrf model on dataset: SparseData_n500_imbalance20perc : 0.675
Confusion Matrix for brf model on dataset: SparseData_n500_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 127  15
         1  13  45

               Accuracy : 0.86
                 95% CI : (0.8041, 0.9049)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 9.991e-08

                  Kappa : 0.6635

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.7500
            Specificity : 0.9071
         Pos Pred Value : 0.7759
         Neg Pred Value : 0.8944
             Prevalence : 0.3000
         Detection Rate : 0.2250
   Detection Prevalence : 0.2900
      Balanced Accuracy : 0.8286

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n500_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 137  19
         1   3  41

               Accuracy : 0.89
                 95% CI : (0.8382, 0.9298)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.315e-10

                  Kappa : 0.7165

 Mcnemar's Test P-Value : 0.001384

            Sensitivity : 0.6833
            Specificity : 0.9786
         Pos Pred Value : 0.9318
         Neg Pred Value : 0.8782
             Prevalence : 0.3000
         Detection Rate : 0.2050
   Detection Prevalence : 0.2200
      Balanced Accuracy : 0.8310

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n500_imbalance40perc : 0.8285714
AUC for wrf model on dataset: SparseData_n500_imbalance40perc : 0.8309524
Confusion Matrix for brf model on dataset: SparseData_n1000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 177   4
         1  13   6

               Accuracy : 0.915
                 95% CI : (0.8674, 0.9497)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.98791

                  Kappa : 0.3727

 Mcnemar's Test P-Value : 0.05235

            Sensitivity : 0.6000
            Specificity : 0.9316
         Pos Pred Value : 0.3158
         Neg Pred Value : 0.9779
             Prevalence : 0.0500
         Detection Rate : 0.0300
   Detection Prevalence : 0.0950
      Balanced Accuracy : 0.7658

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n1000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n1000_imbalance5perc : 0.7657895
AUC for wrf model on dataset: SparseData_n1000_imbalance5perc : 0.5
Confusion Matrix for brf model on dataset: SparseData_n1000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 152  13
         1   8  27

               Accuracy : 0.895
                 95% CI : (0.844, 0.9338)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0002364

                  Kappa : 0.6557

 Mcnemar's Test P-Value : 0.3827331

            Sensitivity : 0.6750
            Specificity : 0.9500
         Pos Pred Value : 0.7714
         Neg Pred Value : 0.9212
             Prevalence : 0.2000
         Detection Rate : 0.1350
   Detection Prevalence : 0.1750
      Balanced Accuracy : 0.8125

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n1000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158  19
         1   2  21

               Accuracy : 0.895
                 95% CI : (0.844, 0.9338)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0002364

                  Kappa : 0.6097

 Mcnemar's Test P-Value : 0.0004803

            Sensitivity : 0.5250
            Specificity : 0.9875
         Pos Pred Value : 0.9130
         Neg Pred Value : 0.8927
             Prevalence : 0.2000
         Detection Rate : 0.1050
   Detection Prevalence : 0.1150
      Balanced Accuracy : 0.7563

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n1000_imbalance20perc : 0.8125
AUC for wrf model on dataset: SparseData_n1000_imbalance20perc : 0.75625
Confusion Matrix for brf model on dataset: SparseData_n1000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 127  12
         1  13  48

               Accuracy : 0.875
                 95% CI : (0.821, 0.9174)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 4.424e-09

                  Kappa : 0.7038

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8000
            Specificity : 0.9071
         Pos Pred Value : 0.7869
         Neg Pred Value : 0.9137
             Prevalence : 0.3000
         Detection Rate : 0.2400
   Detection Prevalence : 0.3050
      Balanced Accuracy : 0.8536

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n1000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 131  13
         1   9  47

               Accuracy : 0.89
                 95% CI : (0.8382, 0.9298)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.315e-10

                  Kappa : 0.733

 Mcnemar's Test P-Value : 0.5224

            Sensitivity : 0.7833
            Specificity : 0.9357
         Pos Pred Value : 0.8393
         Neg Pred Value : 0.9097
             Prevalence : 0.3000
         Detection Rate : 0.2350
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.8595

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n1000_imbalance40perc : 0.8535714
AUC for wrf model on dataset: SparseData_n1000_imbalance40perc : 0.8595238
Confusion Matrix for brf model on dataset: SparseData_n5000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 179   5
         1  11   5

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9762

                  Kappa : 0.3443

 Mcnemar's Test P-Value : 0.2113

            Sensitivity : 0.5000
            Specificity : 0.9421
         Pos Pred Value : 0.3125
         Neg Pred Value : 0.9728
             Prevalence : 0.0500
         Detection Rate : 0.0250
   Detection Prevalence : 0.0800
      Balanced Accuracy : 0.7211

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n5000_imbalance5perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n5000_imbalance5perc : 0.7210526
AUC for wrf model on dataset: SparseData_n5000_imbalance5perc : 0.5
Confusion Matrix for brf model on dataset: SparseData_n5000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 142   7
         1  18  33

               Accuracy : 0.875
                 95% CI : (0.821, 0.9174)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.003629

                  Kappa : 0.6459

 Mcnemar's Test P-Value : 0.045500

            Sensitivity : 0.8250
            Specificity : 0.8875
         Pos Pred Value : 0.6471
         Neg Pred Value : 0.9530
             Prevalence : 0.2000
         Detection Rate : 0.1650
   Detection Prevalence : 0.2550
      Balanced Accuracy : 0.8562

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n5000_imbalance20perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 156  22
         1   4  18

               Accuracy : 0.87
                 95% CI : (0.8153, 0.9133)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0064329

                  Kappa : 0.5113

 Mcnemar's Test P-Value : 0.0008561

            Sensitivity : 0.4500
            Specificity : 0.9750
         Pos Pred Value : 0.8182
         Neg Pred Value : 0.8764
             Prevalence : 0.2000
         Detection Rate : 0.0900
   Detection Prevalence : 0.1100
      Balanced Accuracy : 0.7125

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n5000_imbalance20perc : 0.85625
AUC for wrf model on dataset: SparseData_n5000_imbalance20perc : 0.7125
Confusion Matrix for brf model on dataset: SparseData_n5000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 129  17
         1  11  43

               Accuracy : 0.86
                 95% CI : (0.8041, 0.9049)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 9.991e-08

                  Kappa : 0.6569

 Mcnemar's Test P-Value : 0.3447

            Sensitivity : 0.7167
            Specificity : 0.9214
         Pos Pred Value : 0.7963
         Neg Pred Value : 0.8836
             Prevalence : 0.3000
         Detection Rate : 0.2150
   Detection Prevalence : 0.2700
      Balanced Accuracy : 0.8190

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseData_n5000_imbalance40perc
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 134  15
         1   6  45

               Accuracy : 0.895
                 95% CI : (0.844, 0.9338)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 3.698e-11

                  Kappa : 0.7388

 Mcnemar's Test P-Value : 0.08086

            Sensitivity : 0.7500
            Specificity : 0.9571
         Pos Pred Value : 0.8824
         Neg Pred Value : 0.8993
             Prevalence : 0.3000
         Detection Rate : 0.2250
   Detection Prevalence : 0.2550
      Balanced Accuracy : 0.8536

       'Positive' Class : 1

AUC for brf model on dataset: SparseData_n5000_imbalance40perc : 0.8190476
AUC for wrf model on dataset: SparseData_n5000_imbalance40perc : 0.8535714
Confusion Matrix for brf model on dataset: RetentionData_n500_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 171   0
         1  19  10

               Accuracy : 0.905
                 95% CI : (0.8556, 0.9418)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9973

                  Kappa : 0.4737

 Mcnemar's Test P-Value : 3.636e-05

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.3448
         Neg Pred Value : 1.0000
             Prevalence : 0.0500
         Detection Rate : 0.0500
   Detection Prevalence : 0.1450
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n500_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 189  10
         1   1   0

               Accuracy : 0.945
                 95% CI : (0.9037, 0.9722)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.69976

                  Kappa : -0.0092

 Mcnemar's Test P-Value : 0.01586

            Sensitivity : 0.0000
            Specificity : 0.9947
         Pos Pred Value : 0.0000
         Neg Pred Value : 0.9497
             Prevalence : 0.0500
         Detection Rate : 0.0000
   Detection Prevalence : 0.0050
      Balanced Accuracy : 0.4974

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n500_imbalance5perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n500_imbalance5perc_clusters3 : 0.4973684
Confusion Matrix for brf model on dataset: RetentionData_n500_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   0
         1  16  40

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.457e-06

                  Kappa : 0.7826

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.7143
         Neg Pred Value : 1.0000
             Prevalence : 0.2000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n500_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   0
         1  16  40

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.457e-06

                  Kappa : 0.7826

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.7143
         Neg Pred Value : 1.0000
             Prevalence : 0.2000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n500_imbalance20perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n500_imbalance20perc_clusters3 : 0.95
Confusion Matrix for brf model on dataset: RetentionData_n500_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   0
         1  14  60

               Accuracy : 0.93
                 95% CI : (0.8853, 0.9612)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.051e-15

                  Kappa : 0.8438

 Mcnemar's Test P-Value : 0.000512

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.8108
         Neg Pred Value : 1.0000
             Prevalence : 0.3000
         Detection Rate : 0.3000
   Detection Prevalence : 0.3700
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n500_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   0
         1  14  60

               Accuracy : 0.93
                 95% CI : (0.8853, 0.9612)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.051e-15

                  Kappa : 0.8438

 Mcnemar's Test P-Value : 0.000512

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.8108
         Neg Pred Value : 1.0000
             Prevalence : 0.3000
         Detection Rate : 0.3000
   Detection Prevalence : 0.3700
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n500_imbalance40perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n500_imbalance40perc_clusters3 : 0.95
Confusion Matrix for brf model on dataset: RetentionData_n1000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 171   0
         1  19  10

               Accuracy : 0.905
                 95% CI : (0.8556, 0.9418)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9973

                  Kappa : 0.4737

 Mcnemar's Test P-Value : 3.636e-05

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.3448
         Neg Pred Value : 1.0000
             Prevalence : 0.0500
         Detection Rate : 0.0500
   Detection Prevalence : 0.1450
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n1000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n1000_imbalance5perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n1000_imbalance5perc_clusters3 : 0.5
Confusion Matrix for brf model on dataset: RetentionData_n1000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   0
         1  16  40

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.457e-06

                  Kappa : 0.7826

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.7143
         Neg Pred Value : 1.0000
             Prevalence : 0.2000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n1000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   0
         1  16  40

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.457e-06

                  Kappa : 0.7826

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.7143
         Neg Pred Value : 1.0000
             Prevalence : 0.2000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n1000_imbalance20perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n1000_imbalance20perc_clusters3 : 0.95
Confusion Matrix for brf model on dataset: RetentionData_n1000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   1
         1  14  59

               Accuracy : 0.925
                 95% CI : (0.8793, 0.9574)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 5.677e-15

                  Kappa : 0.8318

 Mcnemar's Test P-Value : 0.001946

            Sensitivity : 0.9833
            Specificity : 0.9000
         Pos Pred Value : 0.8082
         Neg Pred Value : 0.9921
             Prevalence : 0.3000
         Detection Rate : 0.2950
   Detection Prevalence : 0.3650
      Balanced Accuracy : 0.9417

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n1000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   0
         1  14  60

               Accuracy : 0.93
                 95% CI : (0.8853, 0.9612)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.051e-15

                  Kappa : 0.8438

 Mcnemar's Test P-Value : 0.000512

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.8108
         Neg Pred Value : 1.0000
             Prevalence : 0.3000
         Detection Rate : 0.3000
   Detection Prevalence : 0.3700
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n1000_imbalance40perc_clusters3 : 0.9416667
AUC for wrf model on dataset: RetentionData_n1000_imbalance40perc_clusters3 : 0.95
Confusion Matrix for brf model on dataset: RetentionData_n5000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 171   0
         1  19  10

               Accuracy : 0.905
                 95% CI : (0.8556, 0.9418)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9973

                  Kappa : 0.4737

 Mcnemar's Test P-Value : 3.636e-05

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.3448
         Neg Pred Value : 1.0000
             Prevalence : 0.0500
         Detection Rate : 0.0500
   Detection Prevalence : 0.1450
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n5000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   8
         1   0   2

               Accuracy : 0.96
                 95% CI : (0.9227, 0.9826)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.32702

                  Kappa : 0.322

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 0.2000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9596
             Prevalence : 0.0500
         Detection Rate : 0.0100
   Detection Prevalence : 0.0100
      Balanced Accuracy : 0.6000

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n5000_imbalance5perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n5000_imbalance5perc_clusters3 : 0.6
Confusion Matrix for brf model on dataset: RetentionData_n5000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   0
         1  16  40

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 2.457e-06

                  Kappa : 0.7826

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.7143
         Neg Pred Value : 1.0000
             Prevalence : 0.2000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n5000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   1
         1  16  39

               Accuracy : 0.915
                 95% CI : (0.8674, 0.9497)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 6.879e-06

                  Kappa : 0.7671

 Mcnemar's Test P-Value : 0.000685

            Sensitivity : 0.9750
            Specificity : 0.9000
         Pos Pred Value : 0.7091
         Neg Pred Value : 0.9931
             Prevalence : 0.2000
         Detection Rate : 0.1950
   Detection Prevalence : 0.2750
      Balanced Accuracy : 0.9375

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n5000_imbalance20perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n5000_imbalance20perc_clusters3 : 0.9375
Confusion Matrix for brf model on dataset: RetentionData_n5000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   0
         1  14  60

               Accuracy : 0.93
                 95% CI : (0.8853, 0.9612)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.051e-15

                  Kappa : 0.8438

 Mcnemar's Test P-Value : 0.000512

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.8108
         Neg Pred Value : 1.0000
             Prevalence : 0.3000
         Detection Rate : 0.3000
   Detection Prevalence : 0.3700
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: RetentionData_n5000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126   0
         1  14  60

               Accuracy : 0.93
                 95% CI : (0.8853, 0.9612)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.051e-15

                  Kappa : 0.8438

 Mcnemar's Test P-Value : 0.000512

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.8108
         Neg Pred Value : 1.0000
             Prevalence : 0.3000
         Detection Rate : 0.3000
   Detection Prevalence : 0.3700
      Balanced Accuracy : 0.9500

       'Positive' Class : 1

AUC for brf model on dataset: RetentionData_n5000_imbalance40perc_clusters3 : 0.95
AUC for wrf model on dataset: RetentionData_n5000_imbalance40perc_clusters3 : 0.95
Confusion Matrix for brf model on dataset: SparseRetentionData_n500_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 180   8
         1  10   2

               Accuracy : 0.91
                 95% CI : (0.8615, 0.9458)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9942

                  Kappa : 0.1346

 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.2000
            Specificity : 0.9474
         Pos Pred Value : 0.1667
         Neg Pred Value : 0.9574
             Prevalence : 0.0500
         Detection Rate : 0.0100
   Detection Prevalence : 0.0600
      Balanced Accuracy : 0.5737

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n500_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n500_imbalance5perc_clusters3 : 0.5736842
AUC for wrf model on dataset: SparseRetentionData_n500_imbalance5perc_clusters3 : 0.5
Confusion Matrix for brf model on dataset: SparseRetentionData_n500_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 144   9
         1  16  31

               Accuracy : 0.875
                 95% CI : (0.821, 0.9174)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.003629

                  Kappa : 0.6334

 Mcnemar's Test P-Value : 0.230139

            Sensitivity : 0.7750
            Specificity : 0.9000
         Pos Pred Value : 0.6596
         Neg Pred Value : 0.9412
             Prevalence : 0.2000
         Detection Rate : 0.1550
   Detection Prevalence : 0.2350
      Balanced Accuracy : 0.8375

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n500_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 156  25
         1   4  15

               Accuracy : 0.855
                 95% CI : (0.7984, 0.9007)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0282766

                  Kappa : 0.4358

 Mcnemar's Test P-Value : 0.0002041

            Sensitivity : 0.3750
            Specificity : 0.9750
         Pos Pred Value : 0.7895
         Neg Pred Value : 0.8619
             Prevalence : 0.2000
         Detection Rate : 0.0750
   Detection Prevalence : 0.0950
      Balanced Accuracy : 0.6750

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n500_imbalance20perc_clusters3 : 0.8375
AUC for wrf model on dataset: SparseRetentionData_n500_imbalance20perc_clusters3 : 0.675
Confusion Matrix for brf model on dataset: SparseRetentionData_n500_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 124  16
         1  16  44

               Accuracy : 0.84
                 95% CI : (0.7817, 0.8879)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 3.655e-06

                  Kappa : 0.619

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7333
            Specificity : 0.8857
         Pos Pred Value : 0.7333
         Neg Pred Value : 0.8857
             Prevalence : 0.3000
         Detection Rate : 0.2200
   Detection Prevalence : 0.3000
      Balanced Accuracy : 0.8095

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n500_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 129  20
         1  11  40

               Accuracy : 0.845
                 95% CI : (0.7873, 0.8922)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.572e-06

                  Kappa : 0.6144

 Mcnemar's Test P-Value : 0.1508

            Sensitivity : 0.6667
            Specificity : 0.9214
         Pos Pred Value : 0.7843
         Neg Pred Value : 0.8658
             Prevalence : 0.3000
         Detection Rate : 0.2000
   Detection Prevalence : 0.2550
      Balanced Accuracy : 0.7940

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n500_imbalance40perc_clusters3 : 0.8095238
AUC for wrf model on dataset: SparseRetentionData_n500_imbalance40perc_clusters3 : 0.7940476
Confusion Matrix for brf model on dataset: SparseRetentionData_n1000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 179   5
         1  11   5

               Accuracy : 0.92
                 95% CI : (0.8733, 0.9536)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9762

                  Kappa : 0.3443

 Mcnemar's Test P-Value : 0.2113

            Sensitivity : 0.5000
            Specificity : 0.9421
         Pos Pred Value : 0.3125
         Neg Pred Value : 0.9728
             Prevalence : 0.0500
         Detection Rate : 0.0250
   Detection Prevalence : 0.0800
      Balanced Accuracy : 0.7211

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n1000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190   9
         1   0   1

               Accuracy : 0.955
                 95% CI : (0.9163, 0.9792)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.454710

                  Kappa : 0.1743

 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 0.1000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9548
             Prevalence : 0.0500
         Detection Rate : 0.0050
   Detection Prevalence : 0.0050
      Balanced Accuracy : 0.5500

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n1000_imbalance5perc_clusters3 : 0.7210526
AUC for wrf model on dataset: SparseRetentionData_n1000_imbalance5perc_clusters3 : 0.55
Confusion Matrix for brf model on dataset: SparseRetentionData_n1000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 148  10
         1  12  30

               Accuracy : 0.89
                 95% CI : (0.8382, 0.9298)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0005019

                  Kappa : 0.6626

 Mcnemar's Test P-Value : 0.8311704

            Sensitivity : 0.7500
            Specificity : 0.9250
         Pos Pred Value : 0.7143
         Neg Pred Value : 0.9367
             Prevalence : 0.2000
         Detection Rate : 0.1500
   Detection Prevalence : 0.2100
      Balanced Accuracy : 0.8375

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n1000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  20
         1   3  20

               Accuracy : 0.885
                 95% CI : (0.8325, 0.9257)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0010154

                  Kappa : 0.5725

 Mcnemar's Test P-Value : 0.0008492

            Sensitivity : 0.5000
            Specificity : 0.9812
         Pos Pred Value : 0.8696
         Neg Pred Value : 0.8870
             Prevalence : 0.2000
         Detection Rate : 0.1000
   Detection Prevalence : 0.1150
      Balanced Accuracy : 0.7406

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n1000_imbalance20perc_clusters3 : 0.8375
AUC for wrf model on dataset: SparseRetentionData_n1000_imbalance20perc_clusters3 : 0.740625
Confusion Matrix for brf model on dataset: SparseRetentionData_n1000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126  13
         1  14  47

               Accuracy : 0.865
                 95% CI : (0.8097, 0.9091)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 3.686e-08

                  Kappa : 0.6801

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7833
            Specificity : 0.9000
         Pos Pred Value : 0.7705
         Neg Pred Value : 0.9065
             Prevalence : 0.3000
         Detection Rate : 0.2350
   Detection Prevalence : 0.3050
      Balanced Accuracy : 0.8417

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n1000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 126  15
         1  14  45

               Accuracy : 0.855
                 95% CI : (0.7984, 0.9007)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 2.602e-07

                  Kappa : 0.6531

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7500
            Specificity : 0.9000
         Pos Pred Value : 0.7627
         Neg Pred Value : 0.8936
             Prevalence : 0.3000
         Detection Rate : 0.2250
   Detection Prevalence : 0.2950
      Balanced Accuracy : 0.8250

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n1000_imbalance40perc_clusters3 : 0.8416667
AUC for wrf model on dataset: SparseRetentionData_n1000_imbalance40perc_clusters3 : 0.825
Confusion Matrix for brf model on dataset: SparseRetentionData_n5000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 178   6
         1  12   4

               Accuracy : 0.91
                 95% CI : (0.8615, 0.9458)
```

```
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.9942

                  Kappa : 0.2623

 Mcnemar's Test P-Value : 0.2386

            Sensitivity : 0.4000
            Specificity : 0.9368
         Pos Pred Value : 0.2500
         Neg Pred Value : 0.9674
             Prevalence : 0.0500
         Detection Rate : 0.0200
   Detection Prevalence : 0.0800
      Balanced Accuracy : 0.6684

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n5000_imbalance5perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 190  10
         1   0   0

               Accuracy : 0.95
                 95% CI : (0.91, 0.9758)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.583067

                  Kappa : 0

 Mcnemar's Test P-Value : 0.004427

            Sensitivity : 0.00
            Specificity : 1.00
         Pos Pred Value :  NaN
         Neg Pred Value : 0.95
             Prevalence : 0.05
         Detection Rate : 0.00
   Detection Prevalence : 0.00
      Balanced Accuracy : 0.50

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n5000_imbalance5perc_clusters3 : 0.6684211
AUC for wrf model on dataset: SparseRetentionData_n5000_imbalance5perc_clusters3 : 0.5
Confusion Matrix for brf model on dataset: SparseRetentionData_n5000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 146   8
         1  14  32

               Accuracy : 0.89
                 95% CI : (0.8382, 0.9298)
```

```
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0005019

                  Kappa : 0.6746

 Mcnemar's Test P-Value : 0.2864220

            Sensitivity : 0.8000
            Specificity : 0.9125
         Pos Pred Value : 0.6957
         Neg Pred Value : 0.9481
             Prevalence : 0.2000
         Detection Rate : 0.1600
   Detection Prevalence : 0.2300
      Balanced Accuracy : 0.8562

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n5000_imbalance20perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 159  20
         1   1  20

               Accuracy : 0.895
                 95% CI : (0.844, 0.9338)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 0.0002364

                  Kappa : 0.6008

 Mcnemar's Test P-Value : 8.568e-05

            Sensitivity : 0.5000
            Specificity : 0.9938
         Pos Pred Value : 0.9524
         Neg Pred Value : 0.8883
             Prevalence : 0.2000
         Detection Rate : 0.1000
   Detection Prevalence : 0.1050
      Balanced Accuracy : 0.7469

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n5000_imbalance20perc_clusters3 : 0.85625
AUC for wrf model on dataset: SparseRetentionData_n5000_imbalance20perc_clusters3 : 0.746875
Confusion Matrix for brf model on dataset: SparseRetentionData_n5000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 127  15
         1  13  45

               Accuracy : 0.86
                 95% CI : (0.8041, 0.9049)
```

```
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 9.991e-08

                  Kappa : 0.6635

 Mcnemar's Test P-Value : 0.8501

            Sensitivity : 0.7500
            Specificity : 0.9071
         Pos Pred Value : 0.7759
         Neg Pred Value : 0.8944
             Prevalence : 0.3000
         Detection Rate : 0.2250
   Detection Prevalence : 0.2900
      Balanced Accuracy : 0.8286

       'Positive' Class : 1

Confusion Matrix for wrf model on dataset: SparseRetentionData_n5000_imbalance40perc_clusters3
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 130  16
         1  10  44

               Accuracy : 0.87
                 95% CI : (0.8153, 0.9133)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 1.305e-08

                  Kappa : 0.6814

 Mcnemar's Test P-Value : 0.3268

            Sensitivity : 0.7333
            Specificity : 0.9286
         Pos Pred Value : 0.8148
         Neg Pred Value : 0.8904
             Prevalence : 0.3000
         Detection Rate : 0.2200
   Detection Prevalence : 0.2700
      Balanced Accuracy : 0.8310

       'Positive' Class : 1

AUC for brf model on dataset: SparseRetentionData_n5000_imbalance40perc_clusters3 : 0.8285714
AUC for wrf model on dataset: SparseRetentionData_n5000_imbalance40perc_clusters3 : 0.8309524
```