Scientific Journal of Informatics
Vol. 12, No. 2, May 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

A Hybrid Sampling Approach for Handling Data Imbalance in Ensemble Learning Algorithms

Reka Agustia Astari[1*], I Made Sumertajaya[2], Agus Mohamad Soleh[3]
[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia

Abstract.
Purpose: This research aims to address the methodological challenges posed by imbalanced data in classification tasks, where minority classes are severely underrepresented, often leading to biased model performance. It evaluates the effectiveness of hybrid sampling techniques specifically, the Synthetic Minority Oversampling Technique combined with Neighborhood Cleaning Rule (SMOTE-NCL) and with Edited Nearest Neighbors (SMOTE-ENN) in improving the predictive performance of ensemble classifiers, namely Double Random Forest (DRF) and Extremely Randomized Trees (ET), with a focus on enhancing minority class detection.
Methods: A total of eighteen simulated scenarios were developed by varying class imbalance ratios, sample sizes, and feature correlation levels. In addition, empirical data from the 2023 National Socioeconomic Survey (SUSENAS) in Riau Province were employed. The data were partitioned using stratified random sampling (80% training, 20% testing). Models were trained with and without hybrid sampling and optimized through grid search. Their performance was evaluated over 100 iterations using balanced accuracy, sensitivity, and G-mean. Feature importance was interpreted using Shapley Additive Explanations (SHAP).
Results: DRF combined with SMOTE-NCL consistently outperformed all other models, achieving 87.56% balanced accuracy, 82.17% sensitivity, and 86.75% G-mean in the most extreme simulation scenario. On the empirical dataset, the model achieved 76.37% balanced accuracy and 75.49% G-mean.
Novelty: This study introduces a novel integration of hybrid sampling techniques and ensemble learning within an interpretable machine learning framework, providing a robust solution for poverty classification in imbalanced datasets.

Keywords: Data imbalance, Double random forest, Extra trees, Hybrid sampling, Poor household
Received January 2025 / **Revised** June 2025 / **Accepted** June 2025

INTRODUCTION

Classification modelling has become a vital tool for predicting the class of an observation with greater accuracy than conventional methods. One widely adopted approach is ensemble learning, which improves model performance by combining predictions from multiple base learners [1]. This method reduces both bias and variance through techniques such as boosting and bagging. Boosting iteratively enhances accuracy, while bagging aggregates predictions from models trained on different data subsets. Examples include Extra Trees (ET) and Double Random Forest (DRF), both of which have demonstrated effectiveness in classification tasks [2][3].

ET utilizes randomly selected node splits, which increases computational efficiency, whereas DRF bootstraps at each node level during tree construction. In contrast to Random Forest, DRF constructs each tree using the entire dataset at each node split, resulting in deeper trees and improved classification accuracy [4]. Previous studies have shown that ET excels in speed and classification accuracy [5][6]. While DRF yields more consistent results in complex data environments [7].

Based on the results of previous studies, ensemble learning methods, especially DRF and ET, show good results. However, in some cases, the results obtained are less consistent, especially when the data used has an unbalanced distribution. This imbalance causes the model to become biased toward the majority class and limits its ability to accurately predict the minority class [8]. To overcome this problem, various sampling methods, such as oversampling and undersampling, are often used. However, in some cases, these

*Corresponding author.
Email addresses: rekaagustiaastari@apps.ipb.ac.id (Astari), imsjaya@apps.ipb.ac.id (Sumertajaya), agusms@apps.ipb.ac.id (Soleh)
DOI: 10.15294/sji.v12i2.19163

techniques are not effective enough. The application of oversampling can cause overfitting, while under-sampling can cause important information in the dataset to be lost [9].

Hybrid sampling methods have been proposed to address these limitations and have demonstrated superior classification performance, particularly on highly imbalanced datasets [10]. The Synthetic Minority Oversampling Technique (SMOTE) generates synthetic samples to reinforce minority classes, while Neighbourhood Cleaning Rule (NCL) and Edited Nearest Neighbour (ENN) eliminate noisy or overlapping data points, thereby enhancing dataset quality[11]. Techniques such as SMOTE-NCL and SMOTE-ENN have shown better performance in dealing with data imbalance compared to conventional sampling approaches [12]. SMOTE-NCL enhances classification performance by not only generating synthetic samples for the minority class but also cleaning overlapping instances, thereby improving data quality and minority class detection [13]. SMOTE-ENN combines oversampling with noise reduction using edited nearest neighbors, making it more effective in producing balanced and high-quality datasets for accurate classification.

This study uses simulated data to assess model performance across varying levels of class imbalance and feature correlation. Simulated data allows full control over dataset structure, including class ratios, sample size, and correlation patterns. Three imbalance ratios are considered: mild (60:40), moderate (80:20), and extreme (95:5), with low and high correlation levels. The effectiveness of DRF and ET combined with hybrid sampling techniques is evaluated in these scenarios. Additionally, Shapley Additive Explanations (SHAP) are applied to identify key predictive features. Empirical data from poor households in Riau Province are also analyzed to validate the findings. This approach aims to enhance classification performance on severely imbalanced data while offering interpretable results to support real-world decision-making.

METHODS

This study employs both simulated and empirical data to evaluate the performance of hybrid sampling techniques in addressing data imbalance. The simulated data were generated using a multivariate normal distribution with predefined correlation structures and class imbalance ratios. A total of 18 scenarios were constructed, combining three levels of class imbalance (60:40, 80:20, and 95:5) with three correlation structures (low [$\rho = 0{,}1$], high [$\rho = 0{,}9$], mixed [$\rho = 0{,}1$ and $0{,}9$]) [14]. Each scenario was simulated with two different sample sizes, $n = 1000$ and $n = 5000$, as shown in Table 1.

Table 1. Simulation scenarios based on correlation and imbalance ratio

Correlation	Data Imbalance			Total n
	60:40	80:20	95:5	
0.1	(0.1,60:40)	(0.1, 80:20)	(0.1, 95:5)	1000
0.1	(0.1,60:40)	(0.1, 80:20)	(0.1, 95:5)	5000
0.9	(0.9,60:40)	(0.9, 80:20)	(0.9, 95:5)	1000
0.9	(0.9,60:40)	(0.9, 80:20)	(0.9, 95:5)	5000
0.1 and 0.9	(0.1,0.9,60:40)	(0.1,0.9, 80:20)	(0.1,0.9, 95:5)	1000
0.1 and 0.9	(0.1,0.9,60:40)	(0.1,0.9, 80:20)	(0.1,0.9, 95:5)	5000

The empirical dataset used in this study was derived from the 2023 National Socioeconomic Survey (SUSENAS), conducted by Badan Pusat Statistik (BPS), comprising data from 8,246 households in Riau Province. It includes comprehensive household-level socioeconomic variables relevant to poverty classification, such as flooring, wall and roof materials, lighting, water sources, cooking fuel, and sanitation facilities. Poverty status was not explicitly provided and was instead constructed based on the official poverty line. Households with a per capita monthly expenditure below Rp 623,910, the regional threshold comprising Rp 173,340 for non-food and Rp 450,570 for food components, were classified as poor; all others were categorized as “non-poor”. The dataset exhibits a pronounced class imbalance, with only 257 households (3.1%) identified as poor, compared to 7.989 (96.9%) non-poor.

To ensure reliable model training and evaluation, both the simulated and empirical datasets were partitioned into training and testing subsets using stratified random sampling with an 80:20 ratio. This approach preserved the original class distribution in each subset, ensuring proportional representation of both majority and minority classes. Given the inherently imbalanced nature of the data, hybrid sampling techniques were employed specifically, SMOTE-NCL and SMOTE-ENN. The SMOTE (Synthetic Minority Oversampling Technique) method was used to generate synthetic samples for the minority class,

while the Neighbourhood Cleaning Rule (NCL) and Edited Nearest Neighbours (ENN) algorithms were applied to remove potentially noisy or misclassified observations from the majority class. This combination of oversampling and data cleaning enhances class separability and reduces the risk of overfitting, particularly in high-dimensional feature spaces.

After resampling, two ensemble learning algorithms were implemented: Double Random Forest (DRF) and Extra Trees (ET). DRF extends the standard Random Forest framework by introducing bootstrapping not only at the dataset level but also at each decision node, thereby increasing model diversity and robustness. In contrast, ET accelerates the model-building process by selecting split thresholds entirely at random, which improves computational efficiency while maintaining competitive accuracy. To optimize the predictive performance of both ensemble models, hyperparameter tuning was conducted using grid search combined with 5-fold cross-validation. The tuning process aimed to identify the optimal combination of model parameters, including the number of features considered at each split (mtry), the number of trees (ntree), the minimum number of samples required at leaf nodes (node size), and the number of random cut points per node (numrandomcut, specific to ET). The full range of hyperparameter values explored is presented in Table 2.

Table 2. Hyperparameter ranges for DRF and ET models

Hyperparameter	Value	Description
mtry	3, 5, 7, 9	Number of features used when searching for the best split
ntree	100, 300, 500, 700	Number of trees
nodesize	1, 3, 5, 7	Minimum number of samples at leaf nodes
numrandomcut	1,2,3,4	Number of random splits at each iteration in decision making

To ensure robustness and reliability of the evaluation process, each modeling procedure was independently repeated 100 times. Performance metrics were then computed based on the confusion matrix, which summarizes the classification outcomes of the predictive models. Table 3 illustrates the structure of the confusion matrix and the definitions of its components.

Table 3. Confusion matrix for binary classification

Actual	Predicted	
	Positive (1)	Negative (0)
Positive (1)	TP (True Positive)	FN (False Negative)
Negative (0)	FP (False Positive)	TN (True Negative)

Descriptions:
TP (True Positive): Number of class 1 (positive) instances correctly predicted as positive.
FN (False Negative): Number of class 1 instances incorrectly predicted as negative.
FP (False Positive): Number of class 0 (negative) instances incorrectly predicted as positive.
TN (True Negative): Number of class 0 instances correctly predicted as negative.

Based on the confusion matrix, the following performance metrics were calculated [15]:

1. Sensitivity (Recall) (Eq.(1): Measures the model's ability to correctly identify positive instances.
$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (1)$$
2. Specificity (Eq.(2): Measures the model's ability to correctly identify negative instances.
$$\text{specificity} = \frac{TN}{\text{TN} + \text{FP}}. \quad (2)$$
3. Balanced Accuracy (Eq.(3): The average of Sensitivity and Specificity, providing an unbiased metric for imbalanced data.
$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}. \quad (3)$$
4. Geometric Mean (G-Mean) (Eq.(4): Reflects the balance between classification performance on both classes.
$$\text{G} - \text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (4)$$

These metrics are particularly suitable for evaluating models trained on imbalanced datasets, where standard accuracy may be misleading. The average scores across the 100 repetitions were used to assess the consistency and generalizability of each model.

The complete research workflow for data analysis is systematically illustrated in Figure 1.

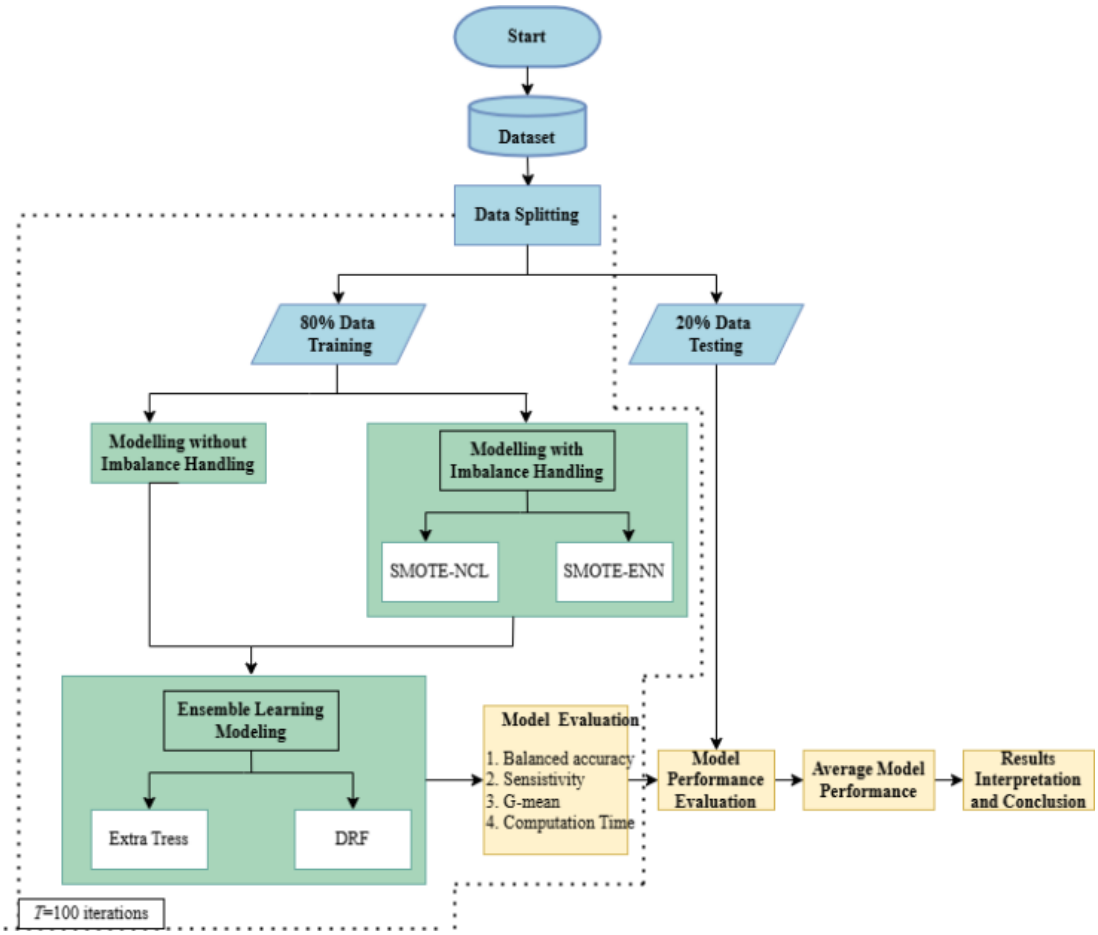

Figure 1. Systematic workflow of the data analysis process

Ensemble Model

The ensemble method is a set of classification models whose individual decisions are combined to classify new samples by using majority voting to combine various classification models [16]. By combining several different learning models, ensemble learning can overcome the problem of uncertainty and weakness in one model, resulting in more accurate predictions [17]. Ensemble learning refers to a method of learning by combining several basic models to build a single model that is larger but stronger than its components. Ensemble learning can also reduce the risk of overfitting. This is due to the combination of weak classifiers with relatively low precision to train classifiers [18]. In this study, ensemble learning is implemented through tree-based bagging methods using decision tree classifiers as base learners. The aggregated outputs from multiple learners are combined via majority voting to determine the final classification label. This strategy allows for better handling of high-dimensional and imbalanced datasets, especially when integrated with resampling techniques such as SMOTE-NCL and SMOTE-ENN. Two ensemble algorithms are utilized, namely Double Random Forest (DRF) and Extra Trees (ET). While both are based on decision trees, they introduce diversity in distinct ways: DRF by applying bootstrapped sampling at both the dataset and node level, and ET by randomly selecting cut points without bootstrapping, resulting in enhanced computational efficiency and greater model variance.

Extremely Randomized Trees (ET)

ET is an ensemble model based on the decision tree algorithm, merging by majority vote for classification cases and arithmetic mean for regression cases [19]. When compared to decision trees, ET uses a slightly different approach to tree formation. ET tends to generate additional decision trees that help prevent overfitting [20] and reduce variance and can minimize bias [21]. ET is similar to random forest. However,

there is a key difference in building the decision tree. When building a decision tree, ET randomly assigns each feature to each tree node regardless of which variable is most informative. So, when trained, the process becomes faster because it eliminates the need to find the best feature for each split node.
In simple terms, according to [3], the ET algorithm procedure can be explained as follows:

a. Determine the number of trees to be formed based on the parameter M (the number of ensemble trees to be formed) using all samples. The greater the value of M, the better the prediction accuracy. However, it will affect the performance of the computation.
b. In each ensemble tree formed, it is determined by the parameter k, which is the number of random attributes used in each node.
c. The parameter n_{min}, the minimum number of samples used in the separation process is determined for each ensemble tree formed. The greater the value of n_{min}, the smaller the ensemble tree and vice versa, if the value of n_{min} Is small, the size of the ensemble tree that will be formed will be larger. Therefore, the n_{min} parameter determines how large the tree will be.
d. The next step is to perform the final prediction, using the majority vote for the classification case and the arithmetic mean for the regression case.

Double Random Forest (DRF)

The DRF model was developed by Han [2]. The development of this model is an alternative solution to improve the performance produced by random forests that experience underfitting. In contrast to random forests that build trees using bootstrap results, DRF uses bootstrapping at each node during the tree-building process, not just once, bootstrapping at the root node as in random forests, but at each node when determining the best sorting criteria during the decision tree building process. This causes DRF to have a more diverse set of trees, so that it has the possibility of more accurate prediction results.
In simple terms, the formation of the DRF algorithm can be explained as follows:

a. Form a tree using all D training data.
b. Select the best splitting using the following criteria:

- For each node t, draw a random sample of size $n_t^*$with bootstrap recovery, if $n_t > n \times 0.1$, if not qualified, no bootstrap is performed.
- Randomly select $m \approx \sqrt{p}$or $m \approx {}^{p}/_{3}$ explanatory variables.
- Repeating steps (a) to (c) until a stopping criterion is reached, resulting in the prediction of a single tree.
  c. Repeating steps (1) and (2) until b classification trees are obtained.
  d. Combining the prediction results obtained from each tree.

Hybrid Sampling

Methods for handling data imbalance to improve algorithms and increase data accuracy have combined oversampling and undersampling, called hybrid sampling. One strategy is to perform data sampling to reduce noise and class imbalance and apply thresholding to reduce bias towards the majority group. Some techniques for combining hybrid sampling methods are as follows:

a. Neighborhood Cleaning Rule (NCL)
   NCL is a technique that can be used to overcome data imbalance by reducing the data caused by cleaning [22]. When a minority class dataset is in the region of the majority class generated from SMOTE, NCL removes it from the dataset. This is done by calculating the ratio between the number of minority class samples and the number of majority class samples in their immediate neighborhood, or community. If the ratio exceeds a certain threshold, the minority class samples are considered as noisy or outliers and removed from the dataset. This is done to prevent overfitting and improve data quality after oversampling.
b. Edited Nearest Neighbor (ENN)
   The ENN method is one of the techniques performed by removing samples from the majority class that are considered to have differences from the majority class based on the nearest neighbor rule, this nearest neighbor rule states that the nearest neighbor sample of each majority sample is found based on the distance between the two samples and is identified as a noise sample or not by assessing the consistency of the nearest label [23].
c. Synthetic Minority Over-sampling Technique (SMOTE)
   SMOTE was proposed by Chawla [24] as an effective method to overcome data imbalance by resampling the relevant data. The main idea behind SMOTE is to increase the number of minority classes compared to majority classes by using synthetic data based on k-nearest neighbors.

Determination of the closest distance for numerical variables is calculated with the Euclidean distance, shown in Eq. (5).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5}$$

The closest distance calculation for categorical variables is by using the value difference metric (VDM) distance, shown in Eq.(6).

$$\Delta(X,Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r \tag{6}$$

Both equations (5) and (6) are used within the SMOTE algorithm to generate new samples that preserve the structure of the minority class while reducing the bias introduced by class imbalance.

Shapley Addictive Explanation (SHAP)

To interpret the predictive behavior of machine learning models, this study employed the Spectral SHAP method, an extension of the Shapley Additive Explanations (SHAP) framework. SHAP is grounded in cooperative game theory, where each feature is treated as a player and the model prediction as the payout, aiming to fairly attribute the contribution of each feature to the final output [25]. This method was chosen due to its consistency, local accuracy, and its ability to provide both global and instance-level explanations. The Shapley value ϕ_j for feature *j* is computed as shown in Eq.(7).

$$\phi_j = \sum_{s \subseteq M(j)} \frac{(|S|!\,(M - |S| - 1)!}{M!}\big(v(S \cup \{j\}) - v(S)\big) \tag{7}$$

Where,

ϕ_j	: Shapley value, representing the contribution of the *j feature*,
S	: A coalition, which is a subset of features excluding feature *j*,
M	: Total number of features,
$v(S \cup \{j\})$	: The prediction obtained from the coalition *S* with the inclusion of feature *j*,
$v(S)$	: The prediction obtained from the coalition *S* without feature *j*

A high positive ϕ_{ij} indicates a synergistic interaction, while a negative value suggests redundancy. The full matrix of SHAP interaction values across all feature pairs forms a symmetric M x M matrix, offering a comprehensive view of inter-feature dependencies within the model.

RESULTS AND DISCUSSIONS

To systematically evaluate classification performance under imbalanced conditions, a simulation framework was developed by combining three levels of class imbalance (low, moderate, high) with three levels of inter-feature correlation (low, moderate, high), resulting in 18 distinct scenarios. Each scenario was modelled using six algorithmic configurations: DRF, DRF SMOTE-NCL, DRF SMOTE-ENN, ET, ET SMOTE-NCL, and ET SMOTE-ENN. These configurations represent hybrid combinations of ensemble learning algorithms and oversampling techniques, designed to assess model robustness under varying data complexities and imbalance structures. Model performance was assessed using metrics derived from the confusion matrix, namely balanced accuracy, sensitivity, and G-mean. Balanced accuracy evaluates the model's ability to classify both classes equitably, while sensitivity measures the effectiveness in detecting the minority class. G-mean, calculated as the square root of the product of sensitivity and specificity, provides a comprehensive view of performance across both classes. Hyperparameter tuning was conducted for each model using a systematic grid search to determine the optimal parameter configurations. The best-performing hyperparameters for each model are summarized in Table 4.

Table 4. Best hyperparameter results

Method	mtry	ntree	nodesize	numrandomcut
DRF	9	100	1	3
DRF SMOTE NCL	5	700	1	3
DRF SMOTE ENN	7	700	1	4
ET	9	100	3	-
ET SMOTE NCL	3	100	1	-
ET SMOTE ENN	7	100	5	-

These results indicate that the DRF-SMOTE variants required higher numbers of trees (ntree = 700) to ensure model stability in the presence of complex synthetic data, while the ET models achieved sufficient performance with fewer trees (ntree = 100). Smaller mtry values in SMOTE-NCL variants suggest tighter feature selection to control overfitting, whereas higher nodesize values in ET-SMOTE-ENN indicate a design to reduce model complexity and enhance generalization. Model evaluations were performed across training and testing sets using the aforementioned metrics. Performance results were visualized and categorized based on model type and imbalance severity. The detailed outcomes for all 18 simulation scenarios are shown in Table 5.

Table 5.Average model performance on simulated data

Level of Imbalance	Metrics	Modeling					
		DRF	DRF SMOTE-NCL	DRF SMOTE-ENN	ET	ET SMOTE-NCL	ET SMOTE-ENN
Low (60:40)	*Balanced Accuracy*	0,500	0,830	0,511	0,502	0,501	0,503
	Sensitivity	0,189	0,771	0,571	0,133	0,247	0,463
	G-mean	0,385	0,716	0,505	0,378	0,417	0,500
Moderate (80:20)	*Balanced Accuracy*	0,500	0,782	0,516	0,499	0,511	0,506
	Sensitivity	0,010	0,857	0,224	0,008	0,358	0,179
	G-mean	0,067	0,777	0,411	0,331	0,484	0,372
High (95:5)	*Balanced Accuracy*	0,504	0,721	0,489	0,500	0,511	0,510
	Sensitivity	0,006	0,814	0,150	0,001	0,199	0,151
	G-mean	0,002	0,828	0,336	0,184	0,380	0,324

Table 5 presents the average performance of classification models under three simulated class imbalance scenarios: low (60:40), moderate (80:20), and high (95:5). The results demonstrate that the level of data imbalance substantially impacts model performance, particularly in the sensitivity and G-mean metrics, which are critical for identifying minority class instances.

Among all evaluated methods, Double Random Forest (DRF) combined with SMOTE-NCL consistently outperformed other models across all imbalance levels. In the low imbalance scenario (60:40), DRF SMOTE-NCL achieved a balanced accuracy of 83.00%, sensitivity of 77.10%, and G-mean of 71.60%, while its performance remained robust under extreme imbalance (95:5), with balanced accuracy of 72.10%, sensitivity of 81.40%, and G-mean of 82.80%. These results highlight the model's strong ability to detect minority classes while maintaining overall classification balance.

In contrast, models without resampling, such as standard DRF and Extremely Randomized Trees (ET), showed near-zero sensitivity under moderate and high imbalance, indicating a severe inability to classify minority instances. For example, standard DRF yielded sensitivity values as low as 1% under 80:20 and only 0.6% under 95:5 imbalance. Although SMOTE-ENN improved performance over baseline models, its effectiveness was consistently lower than that of SMOTE-NCL across all scenarios. Notably, ET combined with SMOTE-NCL also showed improved metrics (e.g., sensitivity of 19.90% and G-mean of 38% under 95:5), but did not surpass the performance of DRF SMOTE-NCL.

These findings confirm the superiority of SMOTE-NCL, particularly when integrated with ensemble learning methods like DRF, in addressing class imbalance. The high G-mean values further indicate a better balance between sensitivity and specificity, reinforcing the recommendation to adopt DRF + SMOTE-NCL in scenarios with severe class imbalance.

Subsequently, the results of each simulated dataset were visualized and grouped according to the applied model and the level of data imbalance. The following figures present the model performance visualization across varying degrees of data imbalance. Figures 2,3, and 4 illustrate the performance distributions of six classification models under three levels of data imbalance. Regarding balanced accuracy, the Double Random Forest (DRF) combined with SMOTE-NCL consistently achieved the highest performance, with median values of approximately 0,85 under high imbalance, 0.84 under moderate imbalance, and around 0,78 under low imbalance. The DRF SMOTE-ENN model also performed well, recording median values ranging from 0.8 to 0.87. In contrast, models without resampling, particularly Extremely Randomized Trees (ET), exhibited lower and more unstable performance, especially under high imbalance scenarios, with median balanced accuracy values hovering around 0.50.

In terms of sensitivity and G-mean, both DRF SMOTE-NCL and DRF SMOTE-ENN again outperformed the other models. Under high imbalance, DRF SMOTE-NCL achieved a median sensitivity of 0.85, while DRF SMOTE-ENN reached 0.83 both exceeded 0.84 under moderate imbalance. ET-based models, even when combined with resampling, produced lower sensitivity, especially in the absence of any balancing method. G-mean results further confirmed the superiority of DRF-based hybrid models, with median scores of 0.84 for DRF SMOTE-NCL and 0.82 for DRF SMOTE-ENN under high imbalance, compared to only 0.70 for the baseline ET model. These findings underscore the critical role of hybrid resampling techniques, particularly when integrated with ensemble classifiers like DRF, in improving model performance for real-world imbalanced classification problems.

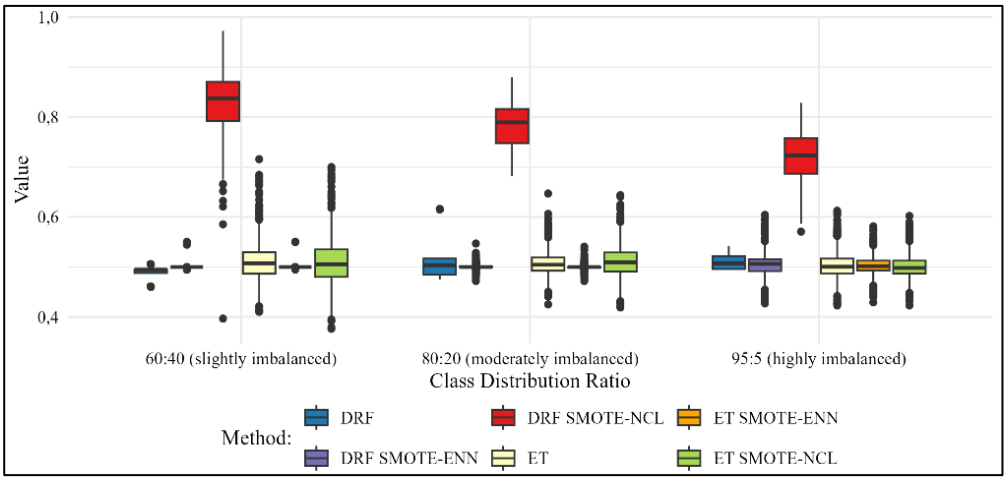

Figure 2. Balanced accuracy comparison across classification methods

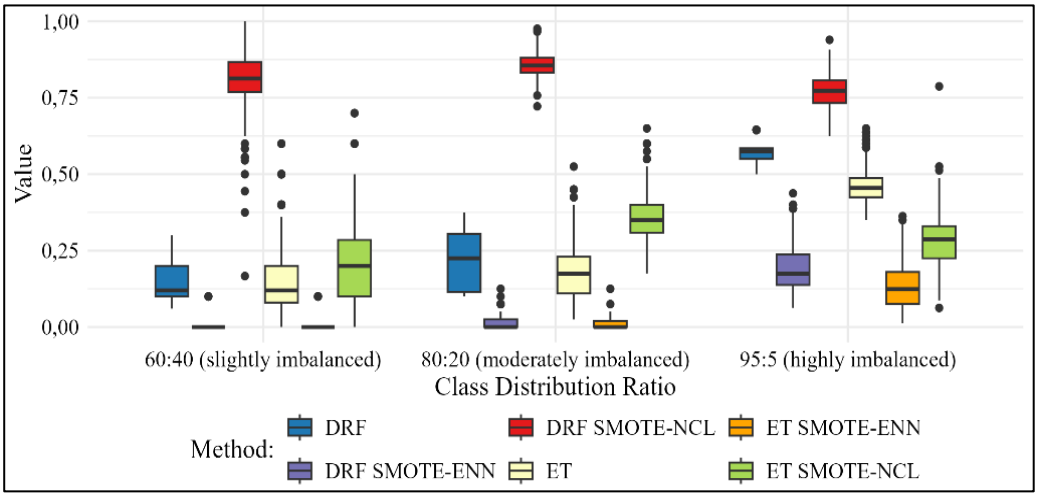

Figure 3. Sensitivity comparison across classification methods

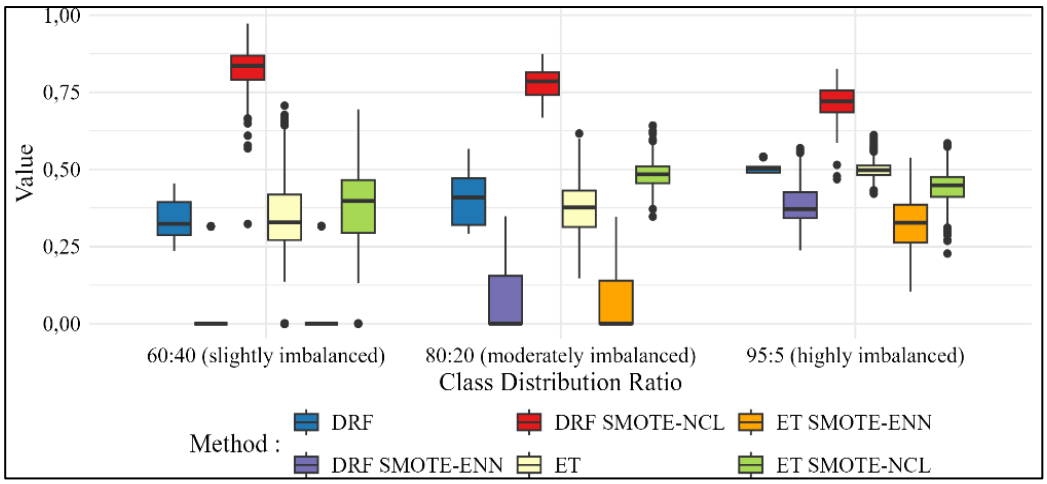

Figure 4. G-mean comparison across classification methods

Model performance was evaluated using metrics specifically relevant to imbalanced classification tasks, namely, balanced accuracy, sensitivity, and G-mean, while also considering computational efficiency in terms of model execution time. Based on the results from 100 repetitions per model-method combination (Table 6), the Double Random Forest (DRF) model integrated with SMOTE-NCL consistently outperformed other approaches, as indicated by its comparatively higher scores across all evaluation metrics.

These findings underscore the critical importance of selecting appropriate imbalance-handling strategies to improve classification quality, particularly in the context of poverty mapping in Riau Province. The superior sensitivity and G-mean values further demonstrate the model's robustness in detecting minority (i.e., poor household) instances, reinforcing the value of hybrid resampling techniques for real-world, imbalanced socioeconomic datasets.

Table 6. Comparison of average model performance on empirical data

Method	Balanced accuracy	Sensitivity	G-Mean	Computation time
DRF	0.518043	0.039216	0.197719	24. 089
DRF SMOTE NCL	0.763779	0.648627	0.754961	457.897
DRF SMOTE ENN	0.726528	0.571373	0.709702	371. 838
ET	0.514685	0.031961	0.157673	0. 734
ET SMOTE NCL	0.703357	0.549412	0.685564	96. 492
ET SMOTE ENN	0.741349	0.654118	0.736007	79. 611

Table 6 and Figure 5 provide a comparative evaluation of the DRF and ET models, both with and without hybrid sampling techniques (SMOTE-NCL and SMOTE-ENN), using empirical poverty data from Riau Province. The baseline models exhibited the weakest performance, with DRF achieving only 51.80% balanced accuracy, 3.92% sensitivity, and a G-Mean of 19.77%. Similarly, ET yielded 51.47% balanced accuracy, 3.20% sensitivity, and a G-Mean of 15.77%, although it required the shortest computation time. The application of hybrid sampling techniques led to substantial performance improvements. Among DRF-based models, DRF SMOTE-NCL recorded the highest overall performance, with 76.38% balanced accuracy, 64.86% sensitivity, and a G-Mean of 75.50%, albeit with a relatively long computation time of 457.90 seconds. In contrast, ET SMOTE-ENN achieved the best overall results, with 74.13% balanced accuracy, 65.41% sensitivity, and a G-Mean of 73.60%, while maintaining a lower computation time of 79.61 seconds.

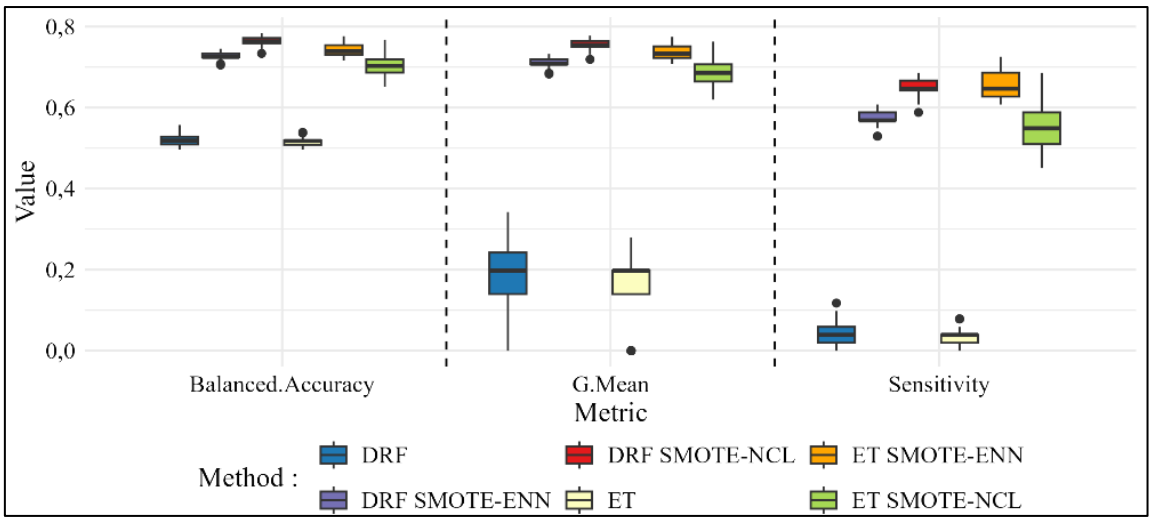

Figure 5. Boxplot of the average confusion matrix of empirical data

As illustrated in Figure 5, models incorporating hybrid sampling, particularly ET SMOTE-ENN, DRF SMOTE-NCL, and DRF SMOTE-ENN, consistently exhibited higher and more stable median values for both balanced accuracy and G-Mean across 100 iterations. Although sensitivity remained relatively low across most models, balanced accuracy and G-Mean proved to be more reliable metrics for evaluating classifier performance under imbalanced data conditions.

The SHAP analysis was employed to provide a transparent and interpretable understanding of each variable's contribution to the model's predictions. This approach enables the best-performing model to serve not only as a predictive tool but also as a means of gaining insights into the most influential factors in determining poverty status. Based on the model evaluation results, the DRF model combined with SMOTE-NCL outperformed the other models. Consequently, SHAP was applied to interpret the DRF SMOTE-NCL model, as illustrated in Figure 6.

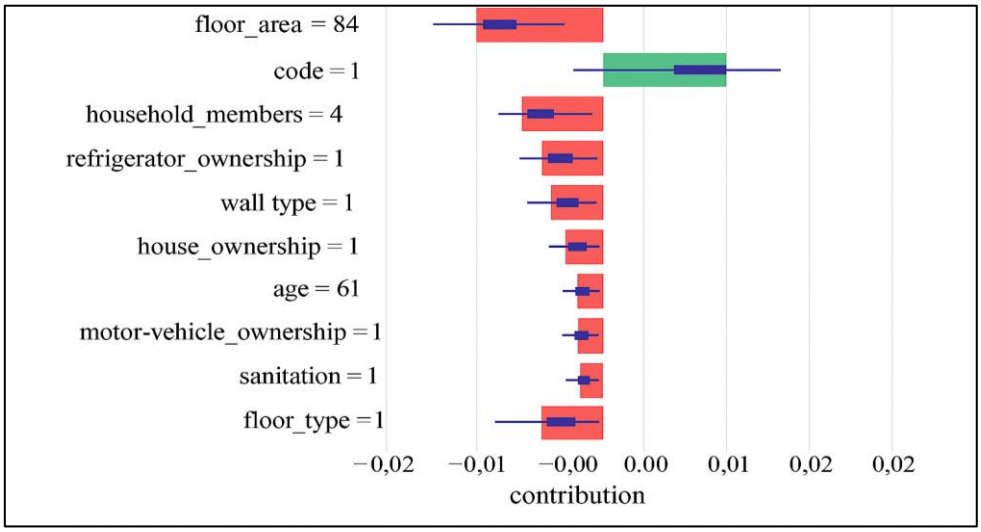

Figure 6. SHAP Plot: the largest contribution in the model

Figure 6 illustrates the most influential factors affecting the likelihood of a household being classified as poor in Riau Province, based on the DRF model analysis. The plot reveals that features such as refrigerator ownership, floor area, and regional code are the most impactful in the model. Other important variables include wall quality, type of flooring, access to sanitation, household size (number of household members), and motorcycle ownership. Households that own a refrigerator are generally less likely to be classified as poor, whereas those with smaller floor areas face a higher risk of poverty. Moreover, larger household sizes

are associated with increased vulnerability to poverty. This analysis highlights that poverty is influenced by a combination of physical housing conditions, asset ownership, and geographic location. Therefore, poverty alleviation programs should adopt a comprehensive approach that considers these multidimensional factors.

CONCLUSION

This study comprehensively evaluated the effectiveness of hybrid sampling methods for addressing class imbalance in poverty classification, using both simulation scenarios and empirical household data from Riau Province. Among all tested approaches, the Double Random Forest (DRF) combined with SMOTE-NCL consistently outperformed other models in terms of balanced accuracy, sensitivity, and G-Mean, particularly under conditions of high class imbalance and low to moderate feature correlation. On empirical data, where the proportion of poor households was extremely low (3.1%), models without imbalance handling performed poorly, especially in detecting the minority class. The DRF with SMOTE-NCL effectively improved classification performance, whereas Extra Trees (ET) yielded less stable results, although it still benefited from hybrid sampling techniques. These findings establish DRF with SMOTE-NCL as a robust approach for highly imbalanced classification tasks. Additionally, SHAP-based interpretability analysis revealed key features influencing poverty prediction, including floor area, geographic code, household size, refrigerator ownership, and housing quality indicators. These results emphasize that poverty is a multidimensional issue, shaped not only by physical household conditions but also by social and geographic factors. Therefore, poverty alleviation strategies should be data-driven and holistic, incorporating structural determinants to ensure effective and sustainable policy interventions.

REFERENCES

[1] M. A. Ganaie, M. Tanveer, I. Beheshti, N. Ahmad, and P. N. Suganthan, "Heterogeneous Oblique Double Random Forest," 2023.

[2] S. Han, H. Kim, and Y. S. Lee, "Double random forest," *Machine Learning*, vol. 109, no. 8, pp. 1569–1586, 2020, doi: 10.1007/s10994-020-05889-1.

[3] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[4] Han et al 2020. Double random forest. Mach Learn. 109(8):15691586. doi:10.1007/s10994-020-05889-1., "double random forest," *Machine Learning*, vol. 109(8):156, pp. 1569–1586, 2020.

[5] L. Yin and Y. Chen, "An Intrusion Detection Model Based on Extra Trees Algorithm with Dimensionality Reduction and Oversampling," *Journal of Computing and Information Technology*, vol. 31, no. 4, pp. 219–231, 2023, doi: 10.20532/cit.2023.1005771.

[6] K. Bluwstein, B. Marcus, J. Andreas, K. Sujit, and Ş. Özgür, "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach," *ECB Working Paper*, vol. 49, no. 11, pp. 8–20, 2021, doi: doi.org/10.2139/ssrn.3969562.

[7] K. A. N. Aldania, Annisarahmi Nur Aini., Agus Mohamad Soleh ., "Jurnal Resti," *Resti*, vol. 1, no. 1, pp. 19–25, 2023.

[8] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artificial Intelligence in Medicine*, vol. 101, p. 101723, 2019, doi: 10.1016/j.artmed.2019.101723.

[9] G. Gumelar and H. Al Fatta, "Kombinasi Algoritma Klasifikasi Dengan Algoritma Oversampling Untuk Menangani Ketidakseimbangan Kelas Pada Level Data," vol. 10, no. 2, pp. 29–39, 2023.

[10] A. Newaz and F. S. Haq, "A Novel Hybrid Sampling Framework for Imbalanced Learning," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4200131.

[11] P. A. R. Mukhlashin, A. Fitrianto, A. M. Soleh, and W. Z. A. Wan Muhamad, "Ensemble learning with imbalanced data handling in the early detection of capital markets," *Journal of Accounting and Investment*, vol. 24, no. 2, pp. 600–617, 2023, doi: 10.18196/jai.v24i2.17970.

[12] M. Fauzi, M. Aria, and R. Pohan, "Algoritma Informed -RRT * Menggunakan Hybrid Sampling Untuk Menemukan Solusi akhir jalur yang Cepat Informed-RRT * Using Hybrid Sampling to Finding Fast Final Path Solution," vol. 9, no. 2, 2021.

[13] M. Dewi, T. H. Saragih, and R. Herteno, "Penerapan SMOTE-NCL untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Jantung Koroner," *Jurnal Informatika Polinema*, vol. 10, no. 1, pp. 27–34, 2023, doi: 10.33795/jip.v10i1.1394.

[14] F. Jabnabillah and N. Margina, "Analisis Korelasi Pearson Dalam Menentukan Hubungan Antara Motivasi Belajar Dengan Kemandirian Belajar Pada Pembelajaran Daring," *Jurnal Sintak*, vol. 1, no. 1, pp. 14–18, 2022.

[15] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," pp. 3125–3128, 2010, doi: 10.1109/ICPR.2010.764.

[16] B. Sunarko *et al.*, "Penerapan Stacking Ensemble Learning untuk Klasifikasi Efek Kesehatan Akibat Pencemaran Udara," *Edu Komputika Journal*, vol. 10, no. 1, pp. 55–63, 2023, doi: 10.15294/edukomputika.v10i1.72080.

[17] H. Ke, H; Gong, S; He, J; Zhang, L; Cui, B.Wang, Y; Mo, J; Zhou, Y; & Zhang, "Development and application of an automated air quality forecasting system based on machine learning.," *Science of The Total Environment*, 2022.

[18] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–14, 2022, doi: 10.1186/s12911-022-02075-2.

[19] D. Al Mahkya, K. A. Notodiputro, and B. Sartono, "Extra Trees Method for Stock Price Forecasting With Rolling Origin Accuracy Evaluation," *Media Statistika*, vol. 15, no. 1, pp. 36–47, 2022, doi: 10.14710/medstat.15.1.36-47.

[20] V. R. S. C. S. K. A. Aina, Listya Nur; Nastiti, "Implementasi Extra Trees Classifier dengan Optimasi Grid Search CV pada Prediksi Tingkat Adaptasi," *Multimedia Artificial Intelligent Networking Database*, vol. 9, no. 1, pp. 78–88, 2024.

[21] A. Aminifar, M. Shokri, F. Rabbi, V. K. I. Pun, and Y. Lamo, "Extremely Randomized Trees with Privacy Preservation for Distributed Structured Health Data," *IEEE Access*, vol. 10, pp. 6010–6027, 2022, doi: 10.1109/ACCESS.2022.3141709.

[22] K. Agustianto and P. Destarianto, "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling," *Proceedings - 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2019*, vol. 1, no. October 2019, pp. 86–89, 2019, doi: 10.1109/ICOMITEE.2019.8921159.

[23] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *Journal of Biomedical Informatics*, vol. 107, no. June, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[24] N. V. Chawla, W. Bowyer, Kevin, L. O. H. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1002/eap.2043.

[25] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. November, pp. 4766–4775, 2017.