JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 9 (ISSUE 2) DECEMBER 2025: 69-81
https://doi.org/10.21009/JSA.09207

COMPARISON OF OVERSAMPLING, UNDERSAMPLING, AND SMOTE TECHNIQUES FOR MULTICLASS BALANCE DATA HANDLING IN RANDOM FOREST AND MULTINOMIAL LOGISTIC REGRESSION

Fadjryani[1*], Asfar[2], Nazwa[3], Allin Floria Tokandari[4], Tri Andayani Lestari[5], Muhammad Azi Zarir Ghani[6]

[1,2,3,4,5,6]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Tadulako University*
Jl. Soekarno Hatta, Tondo, Mantikulore District, Palu City, Central Sulawesi 94148, Indonesia

Corresponding author's e-mail: ***olahdata.palu@gmail.com***

Article History:
Received: 18, June 2025
Revised: 19, November 2025
Accepted: 29, December 2025
Published: 31, December 2025
Available online.

Keywords:
Random Forest; Multinomial Logistic Regression; Class Imbalance; Oversampling; SMOTE.

ABSTRACT

Class imbalances in multiclass classifications are an important challenge in applied machine learning, particularly in the medical field such as predicting how patients will exit. Although various studies have demonstrated the effectiveness of resampling techniques, the best combination of classification algorithms and balancing methods for highly unbalanced multiclass hospital data is still rarely studied.

This study aims to compare the performance of Random Forest (RF) and Multinomial Logistic Regression (MLR) algorithms in dealing with class imbalances using three resampling techniques: Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Oversampling Technique (SMOTE). The dataset used included 1,032 inpatients with Non-Insulin-Dependent Diabetes Mellitus (NIDDM) at Undata Hospital, Central Sulawesi, for the period January 2021 to December 2023. Data pre-processing includes coding, normalization, and data sharing by stratified sampling (80:20). Feature selection was conducted using Recursive Feature Elimination (RFE), and model evaluation was conducted with 5-fold cross-validation using accuracy, recall, F1-score, and MCC metrics.

The results showed that the combination of RF and ROS provided the best performance with an accuracy of 93.65%, F1-macro of 0.935, and a balanced accuracy of 0.95. This combination has been shown to be able to recognize minority classes well without sacrificing overall accuracy. In contrast, the MLR model shows the lowest performance, especially when using RUSs that cause the loss of important data. Although SMOTE is showing competitive results, it remains below ROS in this context.

This study was limited to structured clinical data and only compared two types of classification models. In the future, deep learning-based approaches or advanced ensembles can be explored. The novelty of this study lies in the thorough evaluation of the combination of balancing techniques and classical classification algorithms for medical predictions with extremely unbalanced multiclass data.

How to cite this article:
Fadjryani, Asfar, Nazwa, A. F. Tokandari, T. A. Lestari, M. A. Z. Ghani, "COMPARISON OF OVERSAMPLING, UNDERSAMPLING, AND SMOTE TECHNIQUES FOR MULTICLASS BALANCE DATA HANDLING IN RANDOM FOREST AND MULTINOMIAL LOGISTIC REGRESSION", Jurnal Statistika dan Aplikasinya, vol. 9, iss. 2, pp. 69 – 81, December 2025

Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Classifying multiclass data with class imbalance is a critical challenge in applied machine learning, especially in the health field such as predicting patient discharge. In this case, the minority class that is often the most clinical and relevant has far fewer agencies than the majority class. As a result, models tend to dismiss minority class patterns as noise, making it difficult to detect critical events and misleading evaluations if they rely solely on biased accuracy metrics [1].

The study confirms that addressing medical data imbalances has been a major focus of the past decade [2]. Resampling techniques such as Random Over Sampling (ROS) and Random Under Sampling (RUS) have significant limitations: ROS risks triggering overfitting through duplication of minority data, while RUS has the potential to eliminate essential information from the majority class. [3]. As an alternative solution, hybrid approaches such as SMOTE (Synthetic Minority Over-sampling Technique) were developed to handle datasets with numerical and categorical mixed features. This approach has been shown to be effective in improving the performance of classification models on complex and unbalanced datasets, particularly on multiclass data with mixed features [4].

In terms of algorithm selection, Multinomial Logistic Regression (MLR) is valued for its high interpretability but tends to be biased toward the majority class when applied to imbalanced data [5]. In contrast, Random Forest (RF) demonstrates greater robustness to class imbalance and a stronger ability to model non-linear relationships [6]. However, the findings indicate that MLR may excel on linear data with limited features of a scenario that has not been adequately explored in the context of post-resampling unbalanced multiclasses [7].

The effectiveness of resampling techniques is also highly dependent on the characteristics of the dataset. As emphasized, there is no consistently optimal method in all cases; Performance is affected by factors such as data size, imbalance ratio, and classification algorithms [8],[9]. However, comparative studies that simultaneously evaluate multiple resampling techniques using both interpretable and ensemble-based algorithms, particularly in the context of hospital patient discharge status data, remain relatively limited.

Based on this complexity, this study aims to compare the performance of RF and MLR in dealing with multiclass imbalances by applying three resampling techniques: ROS, RUS, and SMOTE. Through a thorough evaluation using precision metrics, minority-class recalls, and F1-scores, the study is expected to provide empirical guidance in the selection of optimal combination of algorithms and data balancing techniques in medical data-driven predictions.

2. METHOD

Data and Materials

Data on Non-Insulin-Dependent Diabetes Mellitus (NIDDM) patients was obtained from electronic medical records of Undata Hospital, Central Sulawesi, Indonesia. This dataset includes data on the output of inpatients throughout the period from January 2020 to December 2022. In total, there were 1,032 patient data with complete information on outcome status, demographic characteristics, and other supporting clinical variables. For analysis purposes, the data is cleaned through a pre-processing process that includes coding categorical variables using numerical representations, and standardization of numerical features. All data has been anonymized to maintain the confidentiality of patient information.

Table 1. Research Variables

Symbol	Variabel	Data Type
Y	How to Discharge Patients	Categorical
X_1	Gender	Categorical
X_2	How to Pay	Categorical
X_3	Age	Numerical
X_4	Length of Stay	Numerical

Research Methods

Research flow

This study was conducted to compare the effectiveness of three data balancing techniques, namely Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Over-sampling Technique (SMOTE), in overcoming the problem of multiclass class imbalance in medical data. The data were then classified using two methods, namely Random Forest and Multinomial Logistic Regression . The research process was carried out in stages as shown in Figure 1.

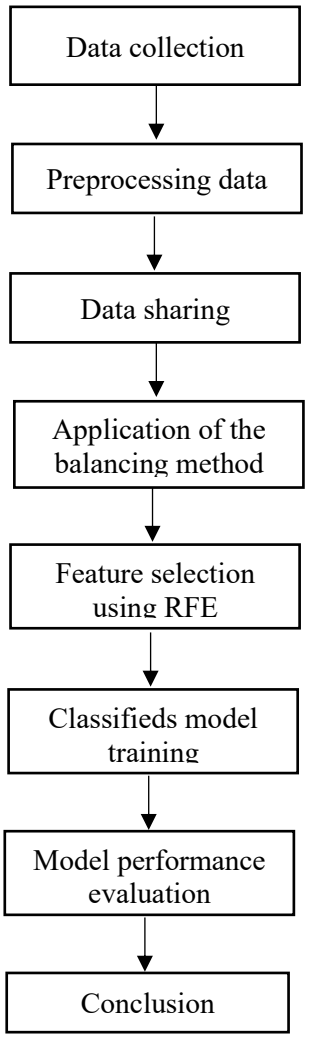

Figure 1. Research Flow

Data Institutions

The preprocessing stage is carried out to prepare data before it is applied into the machine learning model. This process includes the transformation of categorical features using the one-hot encoding technique so that variables can be recognized by classification algorithms, as well as normalization of numerical features to equalize the scale between variable values. Furthermore, the dataset is divided into training data and testing data with an 80:20 ratio using stratified sampling techniques, which aim to maintain a balanced proportion of class distribution in both subsets [10].

Handling Data Imbalance

1. Random Over Sampling (ROS) is the addition of data from minority classes to the training data at random. This addition process is repeated until the amount of minority class data is equal to the number of majority classes [11].
2. Random Under Sampling (RUS) finds a comparable number of minority cases and groups them during the training phase, undersampling randomly removes objects from the class that is the majority [12].
3. This method was first proposed in 2002 by Chawla, where minority classes were oversampled by creating "synthetic training data". The synthetic training data is made based on k-nearest neighbor [11].

Feature Selection

To improve the performance and efficiency of the model, feature selection was carried out using the Recursive Feature Elimination (RFE) technique. RFE works by gradually removing features that are considered least important based on their contribution to model performance. Thus, only the best features are used in the training process, so the risk of overfitting can be minimized [13].

Compared to filter-based feature selection methods, which evaluate features independently of the learning algorithm, RFE directly incorporates model performance into the feature selection process. This allows RFE to capture interactions between features and select a subset that is more relevant to the specific classifier being used. As a result, RFE has been shown to improve classification accuracy and generalization performance, particularly in high-dimensional and noisy datasets, such as medical data [14].

By eliminating irrelevant and redundant features, RFE reduces model complexity and variance, thereby minimizing the risk of overfitting and improving computational efficiency. This is especially beneficial for ensemble models such as Random Forest and linear models like Multinomial Logistic Regression, where feature redundancy can negatively affect stability and interpretability. Previous studies have demonstrated that RFE-based feature selection can enhance model robustness and predictive performance when applied prior to classification [15].

Multinomial Logistics Regression

The general equation of the Multinomialial logistic regression for the three categories of dependent variables is:

$$P(Y = j) = \frac{\exp[g_j(x)]}{\sum_{k=1}^{2} \exp[\, g_k(x)]} = 0; j = 0,1,2; g_0(x) = 0 \quad (1)$$

with: Logit function in the category j logistic regression model.$g_j(x)$

In multinomial logistic regression, parameters are inferred using Maximum Likelihood Estimation (MLE), by maximizing the probability function of the random sample to estimate parameters. The solution to be obtained is by the Newton Raphson iteration method $\hat{\beta}$ [16].

Random Forest

Random Forest is an ensemble learning-based classification algorithm formed from a combination of several decision trees. Random Forest works by building many decision trees during training and issuing classes as a result of the most votes from those trees. This algorithm is known for its stability against overfitting and its ability to handle high-dimensional, nonlinear data [17].

Model Evaluation

The performance of the model with training data was evaluated using 5-fold cross validation, where the training data was divided into five parts and the model was trained and tested alternately on each data fold. Then, the best model is tested using data testing and evaluated using a number of metrics, including accuracy, precision, recall, F1-score, and confusion matrix to provide a comprehensive picture of the quality of the prediction. Here is the formula used in the model evaluation.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \text{x}100\% \tag{2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \text{x}100\% \tag{3}$$

$$\text{F1} - \text{Score} = 2\text{x}\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \text{ x TP}}{2 \text{ x TP} + \text{FP} + \text{FN}} \text{x}100\% \tag{4}$$

$$\text{MacroF1} - \text{Score} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \text{F1}_{\text{i}} \tag{5}$$

$$\text{WeightedF1} - \text{Score} = \frac{\sum_{\text{i}=1}^{\text{N}} \text{w}_{\text{i}} \, . \, \text{F1}_{\text{i}}}{\sum_{\text{i}=1}^{\text{N}} \text{w}_{\text{i}}} \tag{6}$$

3. RESULT

This research includes several stages, namely data exploration and pre-processing, dataset sharing, classification modeling, selection of the best classification model based on evaluation metrics, and classification of test data using the best method for cases of determinants of patient discharge.

Data Exploration

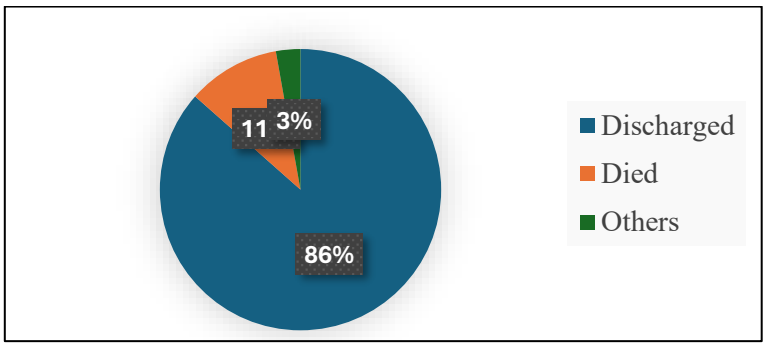

Figure 2. Distribution of Patient Discharge Methods

Figure 2 shows the existence of class imbalances in the dataset. This can be seen from the amount of data in the "Discharged" category, which is shown by the blue pie chart, the most compared to other categories. Such imbalances can cause classification results to be biased towards the dominant class, which ultimately reduces the reliability and representation of the model.

Data Imbalance in Multinomial Regression

Based on the results of data modeling with multinomial without balancing, the matrix confusion and evaluation results are as follows.

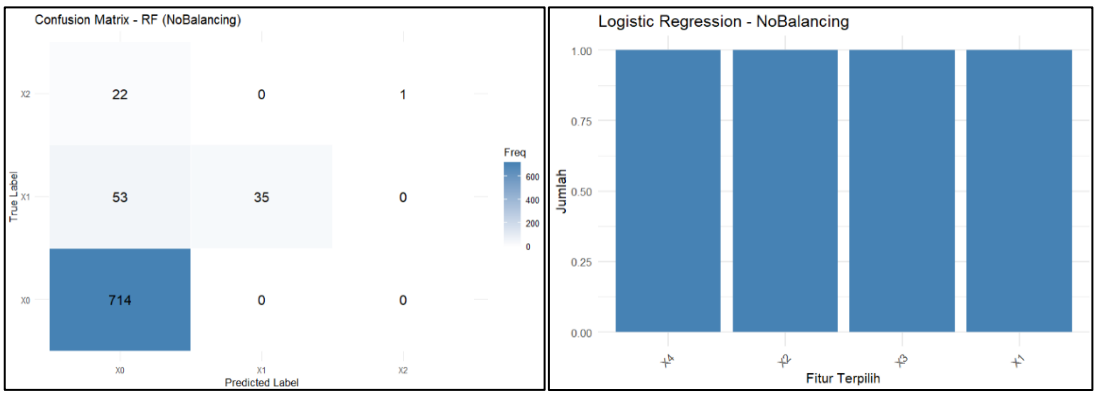

Figure 3. Confusion Matrix and Essential Features for Data Training Before Balancing

In this model, of the four explanatory variables, the variable found to be the most important in classifying the way out of the hospital based on the results shown in Figure 3 is Length of Hospitalization (X_4). These variables are followed by Payment Method (X_2), Age (X_3) and Gender (X_1). These findings suggest that, according to the Multinomial Logistics Regression model, the way a patient discharges has the strongest relationship with the length of the stay and the way of pay.

Table 1. Multinom Model Performance Metrics for Pre-Balancing Training Data

Accuracy	F1 Macro	F1 Weighted	Balanced Accuracy	Recall
0,8630576	0,3088267	0,8018265	0,4990809	NaN

As shown in Table 1, the unbalanced Multinomial model exhibited poor performance, with a f1 macro of 0.31 and a balanced accuracy of 0.50. The presence of a NaN recall for one or more classes indicates that the model was strongly biased toward the majority class.

Data Imbalance in Random Forest

Based on the iterative parameter search process, the best combination of parameters for the Random Forest model used on the unbalanced original training data was found as follows. The model built with a combination of these parameters shows the performance on the training data as shown in Figure 4 and Table 3 below:

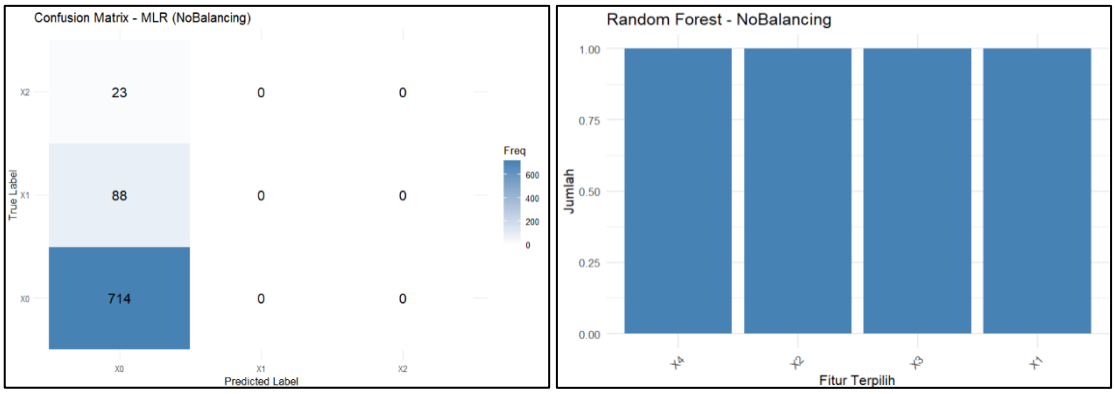

Figure 4. Confusion Matrix and Essential Features For Data Training Before Balancing

Figure 4 shows that Length of Hospitalization (X_4) is the most influential explanatory variable in the Random Forest model for classifying patient discharge outcomes, followed by Payment Method (X_2), Age (X_3), and Gender (X_1). This result indicates that discharge outcomes are more strongly driven by hospitalization duration and payment-related factors than by demographic characteristics. The performance metrics of the Random Forest model trained on the pre-balancing data are subsequently reported in Table 3.

Table 2. Random Forest Model Performance Metrics for Pre-Balancing Training Data

Accuracy	F1 Macro	F1 Weighted	Balanced Accuracy	Recall
0.8643623	0.3747732	0.8226571	0.5318350	NaN

Based on the results reported in Table 3, the Random Forest model without data balancing achieved a relatively high accuracy of 86.44%. However, the f1 macro and recall values were low, at 0.37 and NaN, respectively. This indicates that although the model appears to perform well in terms of overall accuracy, it is strongly biased toward the majority class and has limited ability to correctly identify the minority class.

Data Balancing Method

By eliminating variables using Recursive Feature Elimination (RFE), the model was evaluated using 5-fold cross-validation. The performance evaluation was conducted based on five metrics, namely accuracy, f1 macro, f1 weighted, balanced accuracy, and recall. The detailed performance results of the multinomial model on the balanced training data are summarized in Table 3.

Table 3. Multinom Model Performance Metricss for Train Data After Balancing

Model	Accuracy	F1 Macro	F1 Weighted	Balanced Accuracy	Recall
Multinomial + Random Over Sampling	0.4472015	0.4296848	0.4296848	0.5854342	0.3652968
Multinomial + Random Under Sampling	0.5320879	0.5317949	0.5317949	0.6521739	0.5517241
Multinomial + SMOTE	0.5840078	0.3974998	0.5650638	0.5727281	NaN

The addition of balancing techniques such as ROS, RUS, and SMOTE to the multinomial model improves performance, although the improvement is not statistically significant. Among the evaluated methods, the multinomial model with RUS balancing showed the best results compared to other techniques, achieving an accuracy of 53% and a balanced accuracy of 0.65. These performance metrics, as summarized in Table 3 , indicate that the multinomial model with the RUS balancing

method provides the overall best performance. Meanwhile, the performance results of the Random Forest model under the same balancing strategies are presented in Table 4 for comparison.

Table 4. Random Forest Model Performance Metrics for Train Data After Balancing

Model	Accuracy	F1 Macro	F1 Weighted	Balanced Accuracy	Recall
Random Forest + Random Over Sampling	0.9351055	0.9335055	0.9335055	0.9513305	1.0000000
Random Forest + Random Under Sampling	0.5382418	0.5383759	0.5383759	0.6521739	0.5384615
Random Forest + SMOTE	0.8920829	0.6620310	0.8756429	0.7954493	0.4705882

Based on the model evaluation results presented in Table 4 , it can be observed that the Random Forest (RF) model with the Random Over Sampling (ROS) balancing technique provides the best performance. This model achieves an accuracy of 93.65%, along with F1-macro, F1-weighted, and balanced accuracy values that are all above 0.93. These results indicate that the model is not only highly accurate overall but also effective in handling class imbalance.

Model Performance On The Test Data

This section presents the performance of the models on the test data, both before and after applying data balancing techniques. The comparison aims to evaluate the impact of balancing methods on the model's ability to generalize to unseen data. The Model Performance Metrics for Test Data Without Balancing are reported in Table 5 below.

Table 5 Model Performance Metrics for Test Data Without Balancing

Model	Accuracy	Precision	Recall	F1	Balanced Accuracy	MCC
LMR	0.8647	0.6859	0.3599	0.5403	0.5234	0.1536
RF	0.8647	0.8647	0.3333	0.9275	0.5000	NA

Without data balancing, both models (LMR and RF) achieved the same accuracy (86.47%), but showed poor performance in handling class imbalance, as indicated by low recall and balanced accuracy. LMR had moderate F1 and MCC scores, while RF showed a high F1 score but extremely low recall and balanced accuracy, with MCC not available, likely due to class prediction issues.

Figure 5. Multinomial Matrix Confusion and Essential Features for Test Data

In this model, through variable selection (RFE) on the data that has been balanced, the variables found to be the most important in classifying the way out based on the results shown in Figure 5 are How to Pay (X_2), Length of Stay (X_4) and Gender (X_1).

Figure 5 shows the confusion matrix of the Multinomial Logistic Regression (MLR) model with Random Under Sampling (RUS). In the training data, the model performance is quite low with large classification errors, especially in the X0 and X2 classes. In contrast, in the testing data, accuracy increased with the largest correct predictions in classes X_0 (80), X_2 (91), and X_1 (8), indicating an increase in model generalization. The quantitative performance metrics of this best multinomial model on the test data are summarized in Table 6 .

Table 6 Best Multinom Model Performance Metrics for Test Data

Accuracy	Precision	Recall	F1	Balanced Accuracy	MCC
0.4202899	0.3802226	0.4166526	0.2791058	0.5364537	0.02005848

Even so, as shown in Table 7, the overall performance of the model remains low, with an accuracy of 42%, precision and recall values 0,42, and an F1-score of only 0.28. The balanced accuracy was recorded at 0.54, while the MCC was very low (0.02), indicating the model's very limited predictive capabilities.

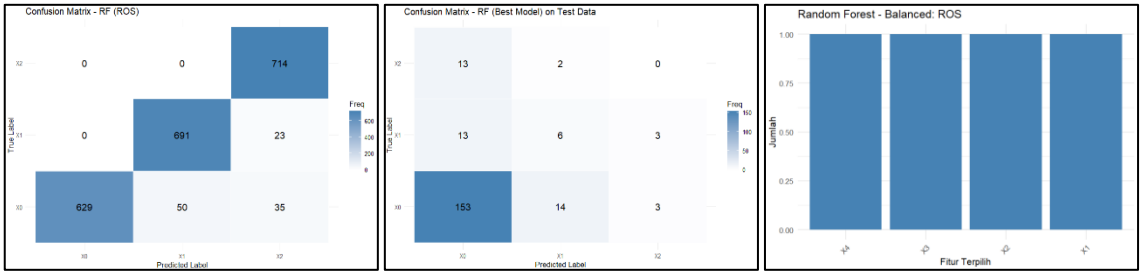

Figure 6. Confusion Matrix Random Forest and Essential Features for Test Data

Figure 6 shows the confusion matrix of the Random Forest (RF) model trained using the Random Over Sampling (ROS) technique. In the training data, the model showed high performance with the largest correct predictions in classes X_2 (714), X_1 (691), and X_0 (629).

In the testing data, the model still showed good performance, with correct predictions in the X0 class as many as 153, X2 as many as 13, and X1 as many as 6. This suggests that the model is also able to recognize minority classes better. The most influential variable in the classification was Length of Stay (X_4), followed by Payment Method (X_2), Age (X_3), and Gender (X_1). The Best Random Forest Model Performance Metrics for the test data are summarized in Table 7 .

Table 7. Best Random Forest Model Performance Metrics for Test Data

Accuracy	Precision	Recall	F1	Balanced Accuracy	MCC
0.7681159	0.3909091	0.3758253	0.5747591	0.5598699	0.16356729

The use of the Random Over Sampling (ROS) technique in Random Forest actually increases the fairness of the model. Although the accuracy drops to 76.81%, the balanced accuracy increases to 0.56, and the MCC rises to 0.16. This suggests that the model becomes more balanced in recognizing both classes, including minority classes, with more stable recall and precision.

4. DISCUSSION

Based on the results of the classification model evaluation with data testing, Random Forest and Logistik Multinomial Regression without balancing techniques produced the same accuracy (86.44%). However, other important metrics at random forest such as recall (0.33), and balanced accuracy (0.52) suggest that these models tend to biased towards the majority class and are less able to recognize the minority class. This is common in unbalanced data, where the model appears to be accurate overall but fails in an even classification performance [18].

The application of Random Oversampling (ROS) in Random Forest resulted in a decrease in accuracy to 76.81%, but accompanied by an increase in balanced accuracy (0.56) and MCC value (0,16). This improvement indicates that the model has become more sensitive and fair in recognizing both classes, especially minority classes. This model is able to overcome distribution inequality without sacrificing predictive performance significantly. In contrast, the application of Random Undersampling (RUS) to Multinomial Logistic Regression showed very low performance, with an accuracy of only 42.02% and other metrics close to zero. This is most likely due to the loss of important data from the majority class due to aggressive undersampling, which negatively impacts the model's ability to study data patterns.

In this study, the model used was limited to Random Forest and Multinomial Logistic Regression as they represent two distinct but commonly used approaches in ensemble learning classification and generalized linear regression. Random Forest was chosen for its superior ability to handle non-linear features and high data complexity, as well as its ability to perform internal feature selection and reduce overfitting. Meanwhile, Multinomial Logistic Regression is used as a comparator because it is a simple and interpretable linear classification model, but tends to be incapable of handling complex relationships between features.

Taking into account all evaluation metrics, the best model in this study is Random Forest with Random Oversampling, as it is able to provide a better balance between accuracy and the ability to detect minority classes. These findings are consistent with results from and that suggest that simple oversampling techniques such as ROS can significantly improve model performance on unbalanced data, especially when used in conjunction with ensemble models such as Random Forest [19],[20].

In addition, the results of the study show that random oversampling provides the best performance in datasets with moderate class imbalances, especially when used with the Random Forest model. In the study, ROS was able to significantly improve model performance compared to other balancing methods, while for extreme imbalances, the hybrid resampling method was considered more suitable. Thus, the selection of balancing methods needs to be adjusted to the level of imbalance and complexity of the data, in order to be able to optimize the performance of the classification model used [21].

The use of Random Over Sampling (ROS) in the Random Forest model improves the model's ability to recognize minority classes by ensuring that minority samples are adequately represented in the bootstrap samples used to construct individual trees [22]. Although ROS duplicates minority observations, Random Forest is relatively robust to overfitting due to its ensemble nature and random feature selection. As a result, the application of ROS leads to improved balanced accuracy and MCC, indicating a fairer classification performance across classes, despite a decrease in overall accuracy

Class imbalance substantially affects model predictions by biasing the learning process toward the majority class, making accuracy a misleading performance metric [23]. This phenomenon is evident in the unbalanced Random Forest model, which achieved high accuracy but low recall and balanced accuracy, indicating poor detection of minority outcomes. In medical datasets, such bias is particularly problematic, as minority classes often represent clinically critical outcomes.

The poor performance of Random Under Sampling (RUS), particularly when applied to Multinomial Logistic Regression, can be attributed to the loss of valuable information from the majority class. Aggressive under sampling reduces the effective sample size and eliminates important variability in the data, leading to unstable parameter estimation and degraded predictive performance [22].

Although SMOTE is widely used to address class imbalance, its performance in this study was inferior to ROS. The generation of synthetic samples through interpolation may introduce noise and produce clinically unrealistic combinations of features, especially in complex medical data [24], [25]. In contrast, ROS preserves the original data structure by replicating real observations, making it more suitable for datasets with moderate imbalance.

From a practical perspective, the improved sensitivity of the Random Forest model with Random Oversampling (ROS) has important implications for hospital decision-making. This finding aligns with previous research [26], which highlights that a major challenge in patient discharge prediction lies in addressing class imbalance, where failure to identify minority-class (high-risk) patients may

lead to serious clinical consequences. A more balanced prediction enables earlier identification of high-risk patients, supports safer discharge planning, facilitates better allocation of medical resources such as intensive care, and reduces the risk of overlooking critical cases, ultimately contributing to improved patient outcomes.

5. CONCLUSION

This study investigated the performance of Random Forest and Multinomial Logistic Regression models for classifying patient discharge outcomes under multiclass imbalanced conditions using several resampling techniques. The results demonstrate that although Random Forest without data balancing achieved high overall accuracy, the integration of Random Over Sampling (ROS) significantly improved the model's ability to detect minority classes, resulting in a more balanced and clinically meaningful classification performance. Overall, Random Forest combined with ROS emerged as the most effective approach, offering an optimal trade-off between accuracy and sensitivity to clinically relevant minority classes.

From a practical perspective, the improved sensitivity of the Random Forest model with ROS has important implications for hospital decision-making. This enhancement highlights that the main challenge in patient discharge prediction is addressing class imbalance, where the model's failure to detect the minority class (high-risk patients) can lead to very serious consequences. This enhanced predictive capability directly enables clinical teams to formulate safer discharge plans, and proactively supports the allocation of medical resources such as intensive care in the management of high-risk patients. Thus, the model significantly reduces the risk of overlooking critical cases, which is essential for improving overall patient outcomes.

Despite these findings, this study has several limitations. The analysis was conducted using data from a single hospital, which may limit the generalizability of the results to other healthcare settings with different patient populations and clinical practices. Furthermore, only two classification models and three resampling techniques were evaluated, and potential biases related to class distribution and data collection may have influenced the observed performance. Therefore, future research is advised to validate these findings using multi-center datasets to assess their robustness across diverse healthcare contexts. Further studies may also explore more advanced machine learning approaches, such as gradient boosting or deep learning models, as well as cost-sensitive learning and hybrid resampling strategies, to further improve minority-class detection and classification performance in imbalanced medical datasets.

6. ACKNOWLEDGMENTS

This research would not have been realized without the support and contributions of various parties. We would like to thank all members of the research team who have worked hard and collaborated in every stage of this process. We would like to express our special gratitude to the supervisors, Fadjryani S.T., M.Si and Asfar S.Pd., M.Si, who have provided guidance, valuable input, and moral support during this research.

We also appreciate the institutions that have provided the necessary facilities and resources to carry out this research. Finally, we thank all respondents who have participated in the data collection, which allowed us to gain in-depth insights into patient discharge. The support and contribution of all parties is very meaningful for the success of this research.

7. REFERENCES

[1] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, “Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods,” *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 36, 2022, doi: 10.1186/s12911-022-01775-z.

[2] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif Intell Rev*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10884-2.

[3] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," 2024, *Frontiers Media SA*. doi: 10.3389/fdgth.2024.1430245.

[4] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. 2018. doi: 10.1007/978-3-319-98074-4.

[5] S. Mauludiah, Y. Arif, M. Faisal, and D. Putra, "Struggling Models: An Analysis of Logistic Regression and Random Forest in Predicting Repeat Buyers with Imbalanced Performance Metrics," *Applied Information System and Management (AISM)*, vol. 7, Sep. 2024, doi: 10.15408/aism.v7i2.39326.

[6] J. Dong and Q. Qian, "A Density-Based Random Forest for Imbalanced Data Classification," *Future Internet*, vol. 14, no. 3, Mar. 2022, doi: 10.3390/fi14030090.

[7] J. J. Levy, J. J. Levy, J. J. Levy, A. J. O'Malley, and A. J. O'Malley, "Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning," *BMC Med Res Methodol*, vol. 20, no. 1, Jun. 2020, doi: 10.1186/s12874-020-01046-3.

[8] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, "The effect of rebalancing techniques on the classification performance in cyberbullying datasets," Jan. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-023-09084-w.

[9] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[10] Y. Ye, Q. Wu, J. Z. Huang, M. K. P. Ng, and X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognit.*, vol. 46, pp. 769–787, 2013, [Online]. Available: https://api.semanticscholar.org/CorpusID:18377491

[11] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[12] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *International Journal of Computer Science Issues(IJCSI)*, vol. 9, Sep. 2012.

[13] A. B. Pillay, D. Pathmanathan, A. Abu, and H. Omar, "RFE-Based Feature Selection to Improve Classification Accuracy for Morphometric Analysis of Craniodental Characters of House Rats," *Sains Malays*, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:261922533

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach Learn*, vol. 46, no. 1, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[15] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Stat Comput*, vol. 27, no. 3, pp. 659–678, 2017, doi: 10.1007/s11222-016-9646-1.

[16] David W. Hosmer and Stanley Lemeshow, *Applied Logistic Regression*. 2000.

[17] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[18] C. Hand and E. Fitkov-Norris, "Not seeing the wood for the trees: Influences on random forest accuracy," *International Journal of Market Research*, vol. 66, pp. 559–566, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:269904925

[19] T. Fulazzaky, A. Saefuddin, and A. M. Soleh, "Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning: A Comparative Analysis of Balanced Random Forest, SMOTE-RF, SMOTEBoost, and RUSBoost," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 969–980, Dec. 2024, doi: 10.15294/sji.v11i4.15937.

[20] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," *Jurnal Matematika, Statistika dan Komputasi*, vol. 21, no. 1, pp. 88–102, Sep. 2024, doi: 10.20956/j.v21i1.35552.

[21] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[22] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J Big Data*, vol. 11, no. 1, p. 87, 2024, doi: 10.1186/s40537-024-00943-4.

[23] M. Altalhan, A. Algarni, and T. Monia, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. PP, p. 1, Jan. 2025, doi: 10.1109/ACCESS.2025.3531662.

[24] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRAS: An oversampling approach for imbalanced datasets," *Mach Learn*, vol. 110, no. 2, pp. 279–301, 2021.

[25] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, 2024, doi: https://doi.org/10.1016/j.mlwa.2024.100597.

[26] Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu, "Application of machine learning in predicting hospital readmissions: a scoping review of the literature," *BMC Med Res Methodol*, vol. 21, no. 1, p. 96, 2021.