BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0269–0280
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0269-0280

ENSEMBLE RESAMPLING SUPPORT VECTOR MACHINE, MULTINOMIAL REGRESSION TO MULTICLASS IMBALANCED DATA

Laila Qadrini[1*], Hikmah[2], Elviani Tande[3], Ignasius Presda[4], Aulia Atika Maghfirah[5], Nilawati[6], Handayani[7]

[1,2,3,4,5,6,7]*Statistics Department, Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat*
Jl. Prof. Dr. Baharuddin Lopa, SH, Talumung, Majene, West Sulawesi, 91214, Indonesia

*Corresponding author's e-mail: * **laila.qadrini@unsulbar.ac.id**)*

Article History:

Received: 20th August 2023
Revised: 29th November 2023
Accepted: 31st December 2023

Keywords:

Adasyn;
Bagging;
Multinomial Regression;
SVM.

ABSTRACT

Imbalanced data is a commonly encountered issue in classification analysis. This issue gives rise to prediction errors in the classification process, which in turn affects the sensitivity, particularly in the minority class. Resampling techniques can be employed as a means to mitigate the issue of imbalanced data. Furthermore, ensemble approaches are utilized in the classification procedure to augment the performance of classification. The present study assesses the efficacy of the bagging ensemble approach in conjunction with ADASYN *as a means of addressing the aforementioned issue. The dataset utilized in this work comprises imbalanced Glass Identification data, imbalanced Iris data, and imbalanced synthetic data. The study centres on the utilization of Support Vector Machines (*SVM*) with parameter optimization using repeated cross-validation (k = 10) and the application of multinomial regression. The evaluation of classification outcomes involves a comparison between the ensemble technique and multinomial regression. This comparison is conducted under pre- and post-resampling conditions, with the evaluation metrics being accuracy, sensitivity, and specificity. The analysis of classification outcomes across the three datasets suggests that the ensemble resampling SVM approach and multinomial regression exhibit superior performance compared to the ensemble SVM and multinomial regression approaches when applied to non-resampled data. The resampling of data has been observed to enhance sensitivity, particularly in the minority class.*

How to cite this article:

L. Qadrini, Hikmah, E. Tande, I. Presda, A. A. Maghfirah, Nilawati and Handayani., "ENSEMBLE RESAMPLING SUPPORT VECTOR MACHINE, MULTINOMIAL REGRESSION TO MULTICLASS IMBALANCED DATA," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0269-0280, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

A range of data mining approaches are available depending on the specific aim at hand. The process of classification is widely utilized in data mining [1]. The objective of classification is to systematically arrange novel data or entities into discrete categories, relying on shared attributes within the data. According to the second source [2]. The classification process involves utilizing training data to construct classification models, which are subsequently employed to classify test data [3]. The application of classification extends to several disciplines, leading to the development of multiple classification algorithms throughout history. However, the problem of class imbalance is frequently encountered in the field of classification [4].

Two distinct classification methodologies are determined by the number of existing classes: binary classification and multi-class classification. In recent years, there has been a growing focus within the scientific community on the matter of imbalanced data in multi-class scenarios. Classifying imbalanced multi-class situations is becoming challenging because many existing multi-class classification approaches are primarily geared to handle balanced class distributions. On the other hand, it is more probable that the data in real-life scenarios will exhibit class imbalances [5].

Class disparity refers to the unequal distribution of a dataset, wherein the majority and minority classes display differing degrees of severity. The misclassification of the minority class is common in classification outcomes, primarily attributed to the overwhelming presence of the majority class [6]. This phenomenon leads to challenges in the classification process and results in below-average classification performance. As a result, the accuracy of classifying the majority class is generally higher than that of classifying the minority class, mostly due to this discrepancy [7]. The Support Vector Machine (SVM) Ensemble Resampling is a methodology employed to mitigate the challenge of imbalanced distribution within multi-class datasets. Resampling is a widely employed technique for mitigating the challenges associated with class imbalance. The methodology mentioned above is characterized by its simplicity and efficacy in modifying the distribution of samples across different classes through the process of random or intentional sampling from the initial training dataset [8].

Previous studies on resampling SVM have explored several strategies, such as SMOTE and random forest. These techniques have demonstrated exceptional accuracy, precision, recall, and F1-score performance, achieving values of 99.95%, 81.63%, 90.91%, and 86.02%, respectively. Moreover, a novel ensemble resampling approach has been devised to tackle the challenges associated with classification problems in imbalanced datasets. The proposed methodology employs oversampling techniques for the smaller classes and undersampling techniques for the bigger classes.

The determination of the resampling scale is based on comparing the number of minority and majority classes. The study conducted in this research illustrates the effectiveness of utilizing a combination of different approach types to improve the performance of algorithms [9]. The utilization of the bagging ensemble in research has been shown to improve the performance of SVM classification, resulting in the development of a more accurate predictive model.

The 5-fold cross-validation approach was employed to assess the prediction accuracy rate, which yielded a value of 93.78%. This result was then compared to many state-of-the-art predicting methodologies. The research findings suggest that ensemble approaches can accurately forecast new therapeutic targets and effectively contribute to the advancement of research and drug development endeavours [10]. This work utilizes the Adaptive Synthetic (ADASYN) ensemble and SVM algorithm, a resampling technique commonly used to mitigate class imbalance problems in classification tasks. The ADASYN approach generates synthetic samples for the minority class adaptively, considering the distribution of the surrounding data. This approach enhances the representation of minority classes within the dataset, thereby potentially improving the performance of the classification model in scenarios where there is an imbalance in class distribution [11].

SVM K-Fold is "Support Vector Machine K-Fold Cross-Validation" or "K-Fold Cross-Validation with Support Vector Machine." The evaluation technique employed to assess the performance of an SVM model's performance involves partitioning data into K subsets, commonly referred to as folds. This process is repeated K times, each iteration utilizing one-fold as the test data and the remaining K-1 folds as the training data. The purpose of SVM K-Fold Cross-Validation is to provide a more accurate evaluation of the performance of the SVM model using data that has not been previously observed. This approach enables us to mitigate the issue of overfitting and gain a deeper comprehension of the model's capacity to generalize to unfamiliar data. This approach also facilitates the selection of SVM parameters and enhances the ability of the SVM model to

classify data accurately. This study utilizes multinomial regression, a statistical technique commonly employed in data analysis to examine the association between predictor variables and response variables that include several categories or are polychotomous [12].

2. RESEARCH METHODS

2.1 Research Data

The section on research data will discuss the information collected and analyzed during the research process. This data serves as the foundation for drawing conclusions and making the paper utilizes three datasets for analysis: the unbalanced class Glass Identification (GI), the unbalanced class Iris dataset, which can be obtained from the UCI machine learning repository, and an unbalanced multiclass generated dataset. The specific information pertaining to the data is presented in Table 1.

Table 1. Details of Dataset

Dataset	Number of Observation	Number of Variable	Number of Class
GI	214	10	7
Iris	150	4	3
Generated	1000	4	3

The study of GI involves the classification of various varieties of glass, which is primarily driven by the need for accurate analysis in forensic investigations. Glass fragments found at crime scenes can serve as valuable evidence when accurately identified. The dataset has 214 observations and 10 variables. The set of variables comprises ten elements: RI: refractive index, Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4–10), Mg: Magnesium, Al: Aluminum, Si: Silicon, K: Potassium, Ca: Calcium, Ba: Barium, Fe: Iron,

Type of glass: 1 (building_windows_float_processed), 2 (building_windows_non_float_processed), 3 (vehicle_windows_float_processed), 4 (vehicle_windows_non_float_processed - none in this database), 5 (containers), 6 (tableware), 7 (headlamps). The datasets used in this study were obtained from the USA Forensic Science Service. The predictor variables in this study are of a numeric data type; however, the glass type variable is categorical and can be classified as a factor [9].

The Iris dataset comprises 150 measurements of petal and sepal lengths and widths, with 50 measurements allocated to each of the three species: "setosa", "versicolor", and "virginica". The Iris Dataset is available on R. Moreover, the synthetic unbalanced dataset is generated using a normal distribution. It consists of 1000 data points, with class 1, class 2, and class 3 having class compositions of 10%, 30%, and 60% correspondingly.

2.2 Analysis of the Data

The data analysis was performed using R software version 4.2.3. The first phase of the research entailed exploring data to examine the attributes of the three data clusters. Additionally, bar plots were utilized to visually represent the distribution of classes within each data cluster to evaluate any potential imbalances in the data.

2.3 Steps for Classification Analysis

The process of implementing the Ensemble Resampling SVM involves the following steps:

Step 1: The data analysis process involves data preprocessing

The execution of data preprocessing tasks, such as managing missing values, normalizing, and standardizing data, should be carried out. The dataset should be partitioned into two subsets, namely the training and test data, with proportions of 0.7 and 0.3, respectively. The training data is subjected to modeling.

Step 2: Involves the implementation of SVM K-Fold Cross-Validation

The SVM is a relatively recent method, introduced in 1995, that is used for making predictions in both classification and regression scenarios. In essence, the SVM algorithm endeavours to optimize the margin, which denotes the extent of separation between different classes of data. SVM have the capability to effectively handle datasets with a high number of dimensions by utilizing kernel tricks. The SVM algorithm utilizes a limited number of chosen data points known as support vectors to construct the model employed throughout the classification procedure. In certain cases, SVM can offer a more effective approach for learning nonlinear functions when compared to logistic regression and neural networks.

The objective of optimization logistic regression can be modified to generate SVM. The mathematical representation of the cost function used in logistic regression is as follows:

$$\min_{\theta} \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}(-\log h_\theta(x^{(i)})) + (1-y^{(i)})((-\log(1-h_\theta(x^{(i)})))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

with the parameter λ and the optimization objective function A+λB,

The cost function of SVM is similar to logistic regression, except that the constant "m" is removed because its value does not affect the minimum value of θ. Then, the form $(-\mathbf{log}\,\boldsymbol{h_\theta}(\boldsymbol{x^{(i)}}))$ is change to $\boldsymbol{cost_1}(\boldsymbol{\theta^T x^{(i)}})$ and $((-\mathbf{log}(\mathbf{1}-\boldsymbol{h_\theta}(\boldsymbol{x^{(i)}})))$ is changed to $\boldsymbol{cost_0}(\boldsymbol{\theta^T x^{(i)}})$, and $((-\mathbf{log}(\mathbf{1}-\boldsymbol{h_\theta}(\boldsymbol{x^{(i)}})))$ is changed to $\boldsymbol{cost_0}(\boldsymbol{\theta^T x^{(i)}})$, and multiplied by C=1/λ. It takes the form of:

$$\min_{\theta} C\left[\sum_{i=1}^{m} y^{(i)}\, cost_1(\theta^T x^{(i)}) + (1-y^{(i)})\, cost_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

with the parameter C and the optimization objective function $\boldsymbol{CA}+\boldsymbol{B}$.

The hypothesis of the SVM is $h_{\theta(x)} = 1$ if $\theta^T x \geq 0$ and $h_{\theta(x)} = 0$ if $\theta^T x < 0$

To ascertain the appropriate number of folds (K) for implementation in K-Fold Cross-Validation, it is necessary to do a thorough analysis. The training dataset should be partitioned into K-folds of equal size.

Step 3: Involves the Implementation of K-Fold Iteration

During each iteration of K-Fold Cross-Validation, the training data consists of K-10 folds, while the remaining ten folds are utilized as the test data. Utilize the Bagging-ADASYN-SVM approach on the training data of the current iteration.

Step 4: The Bagging-ADASYN-SVM Method

The Bagging-ADASYN-SVM method is employed in each iteration. Utilize the bagging technique on the training dataset of the present iteration. For every model inside the bagging ensemble: Apply the ADASYN algorithm to the dataset created by the bagging technique. The SVM model should be trained on the dataset generated using the ADASYN sampling technique.

Step 5: Forecasting and Consolidation of Findings

During each iteration of the K-Fold process, the trained ensemble model, specifically the Bagging-ADASYN-SVM model, is utilized to make predictions on the test data of the current iteration. The prediction results obtained from each iteration of the K-Fold cross-validation technique can be combined through aggregation methods such as averaging or majority voting.

Step 6: Evaluating Performance

Assess the ensemble model's performance using K-Fold Cross-Validation, employing appropriate assessment measures for multiclass classification, including accuracy, precision, recall, and confusion matrix.

$$Accuracy = \frac{TP + TNe + TNt}{TP + TNe + TNt + FP + FNe + FNt} \quad (1)$$

$$Sensitivity\ Positive\ Class = \frac{TP}{TP + FNe + FNT} \quad (2)$$

$$Specivicity\ Positive\ Class = \frac{TNe + FNt + FNe + TNt}{FP + TNe + TNt + FNe + TNt} \quad (3)$$

Step 7: Involves the Prediction of a New Class

Following the assessment of the ensemble model using K-Fold Cross-Validation, the trained ensemble model is subsequently employed on the complete training dataset to make predictions on novel classes within previously unobserved data.

Step 8: Parameter Optimization

During the implementation phase, tuning the appropriate parameters for the Bagging, ADASYN, and SVM algorithms is essential. The Utilization of K-Fold Cross-Validation serves the purpose of objectively assessing the performance of a model and mitigating the risk of overfitting on the given dataset **[10]**. The proposed solution integrates Bagging, ADASYN, and SVM approaches to reduce class imbalance and improve classification performance in intricate multiclass scenarios.

The steps involved in doing a Multinomial Regression (MR) analysis are as follows:

Step 1: In the data analysis process involves data preprocessing

The execution of data preprocessing tasks, such as managing missing values, normalization, and data standards, is required. The data should be partitioned into distinct training and testing sets.

Step 2: Involves implementing resampling techniques or strategies for handling the minority class

To address the issue of class imbalance, it is advisable to employ methodologies such as oversampling or undersampling on the minority class. These strategies can help develop a more balanced dataset regarding class distribution. The present study employed ADASYN (Adaptive Synthetic Sampling) to generate synthetic samples from the minority class.

Step 3: Involves conducting an MR analysis

The resampled dataset was subjected to multinomial regression analysis. This approach enables the user to construct a model that captures the association between many predictor factors and a target variable that exhibits more than two distinct classes.

Step 4: In the process is model training

The MR model should be trained on the resampled dataset. To achieve optimal performance, it is necessary to adjust both the model parameters and hyperparameters.

Step 5: The research process involves model validation

Employ cross-validation methodologies to assess the performance of the model impartially. This approach aids in mitigating overfitting and offers a more accurate evaluation of performance on data that has not been encountered before.

Step 6: The evaluation of performance

Assess the multinomial regression model by employing suitable assessment measures for multiclass classification, such as accuracy, precision, recall, F1-score, or confusion matrix.

Step 7: The prediction of a new class

Following the model evaluation process, the trained MR model can be used to make predictions on novel data instances, assigning them to appropriate classes.

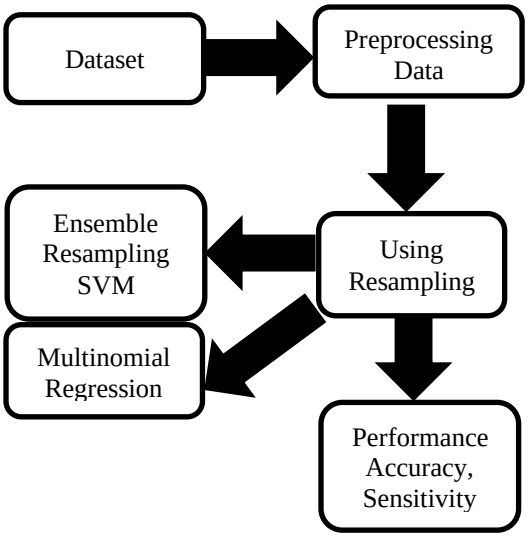


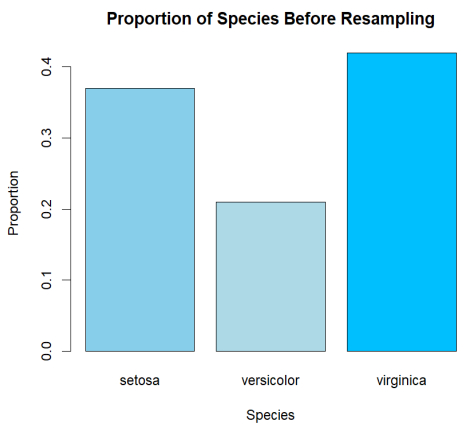
Figure 1. The Process Research

Figure 1 presents the Research Steps for the ERSVM and the MR methods used to the three datasets used in this study.

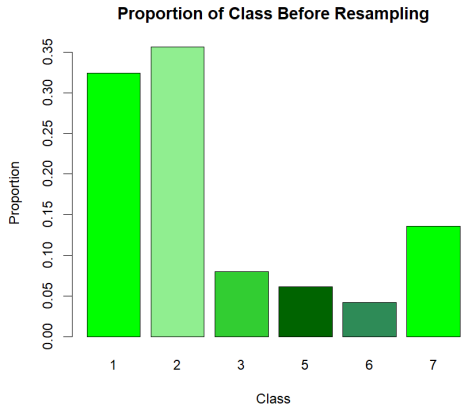
3. RESULTS AND DISCUSSION

3.1 Data Exploration

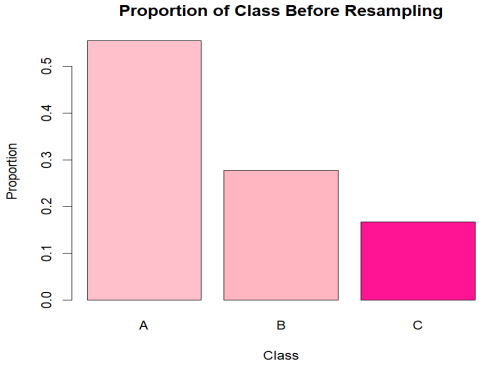
Data exploration is conducted to observe the distribution of class proportions in the imbalanced Iris, imbalanced GI, and imbalanced synthetic datasets. The distribution of class proportions is visually shown by a bar plot. The bar plot in **Figure 1** visually represents the class imbalance seen in the three datasets.



(a)



(b)



(c)

Figure 2. Bar plot. (a) Iris Imbalanced Class, (b) GI Imbalanced Class, (c) Generated data Imbalanced Class

Figure 2 illustrates the presence of uneven class distributions across all three datasets. The GI dataset has the most pronounced data imbalance. There is a lack of available data for Class 4 (vehicle_windows_non_float_processed), necessitating the need for preprocessing prior to undertaking any classification study. Class 4, which pertains to non-float processed car windows, has been eliminated from the GI dataset, leaving a total of six surviving classes. Following that, the process of data preparation is carried out with ADASYN in order to tackle the imbalances present in all three datasets. The utilization of ADASYN has been shown to yield advantageous outcomes in the generation of enhanced and more efficient data for the purpose of mitigating data imbalances **[11]**.

3.2 Resampling Technique for Data Balancing

The utilization of resampling techniques is employed in the research investigation to achieve data balance. The act of achieving balance within an imbalanced dataset presents numerous notable benefits: Enhanced Model Performance: Imbalanced datasets often pose challenges for models in effectively capturing patterns from minority classes. By achieving a balance in the dataset, the model will allocate greater attention to the minority class, perhaps resulting in enhanced performance in the identification and prediction of said class.

3.2.1 Mitigation of Bias

Machine learning models that are trained on datasets with uneven class distributions often demonstrate a tendency to favor the dominant class. The act of balancing the data can effectively decrease the probability of the model disregarding the minority class and result in more equitable conclusions.

3.2.2 Mitigation of Majority Predictions

In the absence of data balancing techniques, models may exhibit a bias towards predicting the majority class consistently, owing to its higher prevalence. This phenomenon can lead to a situation where the classification model achieves a high level of accuracy overall but has a relatively low ability to accurately anticipate instances belonging to the minority class.

3.2.3 Enhanced Sensitivity (Recall)

In situations where accurately identifying the minority class is of utmost importance, data balancing techniques can improve the model's sensitivity (recall) for that particular class. This phenomenon proves advantageous in scenarios when the occurrence of false negatives entails significant repercussions.

3.2.4 Overfitting Reduction

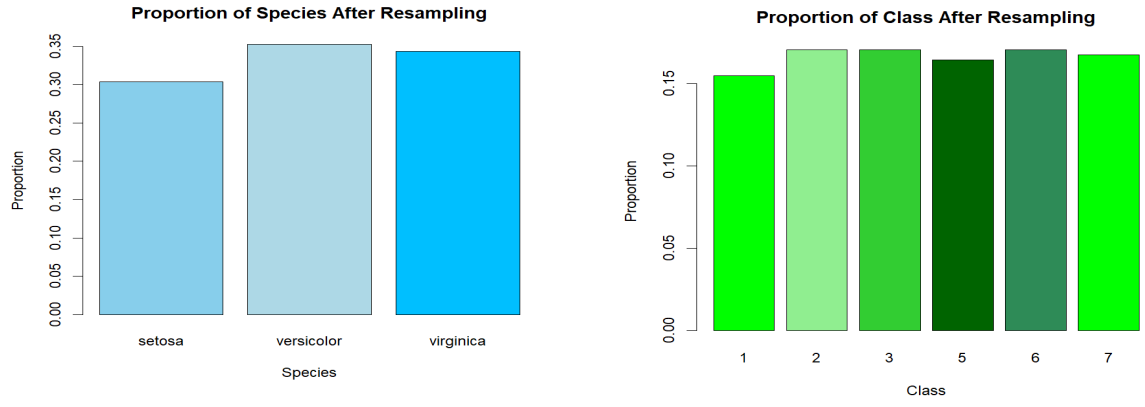
The reduction of overfitting can be achieved by balancing the data, as this prevents the model from excessively learning noise from the majority of the data.

3.2.5 Enhanced Realism

The act of balancing the data can replicate real-world scenarios in which classes exhibit a more equitable distribution. **Figure 3** displays the outcomes of data balancing for the three datasets subsequent to ADASYN resampling.

3.2.6 Improved Model Stability

By utilizing a dataset that is more evenly distributed, it is expected that the model will exhibit more stability and reliability in its ability to generate predictions for the minority class.



(a)

(b)

(c)

Figure 3. Bar plot. (a) Iris Imbalanced Class, (b) GI Imbalanced Class, (c) Generated data Imbalanced Class

3.3 Modeling and Evaluation of Classification Results

In this section, we will discuss the process of modeling and evaluating classification results. This involves the creation of a model that can accurately classify data into different categories, as well as the assessment of the model's performance. The first step in modeling classification results is to select an appropriate algorithm. The classification modeling is performed on the training data in two scenarios: the imbalanced data scenario and the balanced data scenario, utilizing ADASYN. The classifier utilized in this study is the ERSVM, more precisely the Bagging-ADASYN-SVM. **Table 2** presents the ideal parameters for all four approaches.

Table 2. Optimize Parameter

Dataset	Condition	Ensemble Bagging SVM	Multinomial Regression
GI	Without Resampling	Repeated cv = 10	Default
	With Resampling	Repeated cv = 10	Default
IRIS	Without Resampling	Repeated cv = 10	Default
	With Resampling	Repeated cv = 10	Default
Generated	Without Resampling	Repeated cv = 10	Default
	With Resampling	Repeated cv = 10	Default

Following the completion of the modeling procedure, cross-validation is conducted utilizing the test data, subsequently leading to the computation of accuracy, sensitivity, and specificity metrics. **Table 3** provides an elucidation of the metrics pertaining to accuracy, sensitivity, and specificity.

Table 3. Confusion Matrix for Three-Class Classification

Class	Positively Classified	Negatively Classified	Neutrally Classified
Positive	True Positive (TP)	False Negative (FN)	False Netral (FNt)
Negative	False Positive (FP)	True Negative (TN)	False Netral (FNt)
Neutral	False Positive (FP)	False Negative (FN)	True Netral (TNt)

Nevertheless, the assessment predominantly focuses on comparing sensitivity values, specifically for the minority class. Accuracy refers to the level of precision exhibited by predicted outcomes. Sensitivity, often known as the true positive proportion, refers to the accuracy with which the intended class is accurately identified. In contrast, specificity, which refers to the genuine negative proportion, accurately identifies instances belonging to the unwanted class. The presence of data imbalance has been found to be associated with decreased sensitivity values **[12]**. The level of accuracy achieved by both techniques demonstrates a high degree of effectiveness when applied to the GI dataset, as illustrated in **Table 4**.

Table 4. Comparison of Accuracy and Sensitivity of The GI Data Imbalanced Class

Condition	Method's Code	Accuracy	Sensitivity					
			"1"	"2"	"3"	"5"	"6"	"7"
Without Resampling	ESVM	0,66	0,9	0,63	0	0,66	0,5	0,62
	MR	0,53	0,65	0,4	0	1	0,5	0,75
With Resampling	ERSVM	0,63	0,85	0,5	0	1	0,5	0,75
	MR	0,53	0,85	0,4	0	0	0	0,75

The ESVM technique demonstrates superior accuracy in the context of the GI dataset without the need for resampling. Conversely, the MR method yields comparable accuracy regardless of the inclusion or exclusion of the GI dataset. Both the ERSVM and MR classification algorithms have the highest sensitivity for class "1" when resampling is employed. The Utilization of resampling techniques enhances the sensitivity of class "7" in both the ERSVM and MR approaches. The condition that yields the highest accuracy is the one without resampling, and the accuracy number tends to approach 1. Nevertheless, an excessively high level of accuracy may suggest the occurrence of overfitting.

Overfitting refers to a situation in which the model has mostly memorized the irrelevant details or noise included in the training data rather than grasping the fundamental patterns. Consequently, this can result in inadequate generalization when applied to novel, unknown data **[13]**. Resampling approaches, such as oversampling (the addition of samples) or undersampling (the reduction of instances), have the potential to incorporate novel models into the dataset. If the newly acquired samples fail to accurately reflect the minority or majority class, they have the potential to add noise to the dataset, hence impeding the model's capacity to appropriately generalize.

Excessive implementation of undersampling may result in the loss of crucial information from the initial dataset, hence diminishing the model's capacity to comprehend pre-existing patterns within the data. The process of resampling may not consistently produce fresh samples that effectively represent the minority class. When the minority data exhibits a high degree of complexity or diversity, the accuracy of the resampling outcomes may be compromised. Following the process of resampling, it may be necessary to make adjustments to the utilized model in order to accommodate the alterations in the distribution of data. If executed appropriately, the performance of the model can remain unchanged. The associations between characteristics in the data can be influenced by changes in data distribution resulting from resampling techniques.

This phenomenon can provide difficulties for the model in terms of achieving successful generalization. Resampling methods may only partially address class imbalance issues in certain cases. If the number of samples in the majority class remains very large even after resampling, it may be necessary to further address the issue of class imbalance. The sensitivity for the class "setosa" achieves a maximum value. The ERSVM method demonstrates a commendable level of accuracy, albeit with modest variations in comparison to the MR method. In the case of the "Virginica" class, it can be observed from **Table 5** that sensitivity exhibits the lowest values across all settings.

Table 5. Comparison of Accuracy and Sensitivity of The Iris Data Imbalanced Class

Condition	Method's Code	Accuracy	Sensitivity		
			"Setosa"	"Versicolor"	"Virginica"
Without Resampling	ESVM	0,90	0,92	0,85	0,46
	MR	0,90	1	0,85	0,73
With Resampling	ERSVM	0,80	1	0,92	0,64
	MR	0,68	0,76	0,85	0,53

Based on the data obtained, it is evident that the class dataset exhibits an imbalance, wherein one class contains a substantially larger number of instances compared to the others. This imbalance has the potential to introduce bias in the model, favouring the majority class and consequently resulting in reduced accuracy. The performance of the SVM model can be influenced by the complexity of the model or the hyperparameters employed in its construction. In the event that the model exhibits excessive complexity or inadequate tuning, its ability to effectively generalize to novel, unobserved data may be compromised. The performance of the model can also be influenced by the quality of the input characteristics and the procedures taken in data preprocessing.

The presence of noise or irrelevant elements within the data can result in unsatisfactory outcomes. In the event that the dataset is of insufficient size, it is plausible that the model may require a greater number of examples in order to get a more comprehensive understanding of the underlying patterns. The concept of randomness is employed in the technique known as bagging, which entails the generation of many bootstrap samples from the dataset and subsequently training models on each of these samples. The introduction of randomness through the bootstrapping process and the random selection of features might result in fluctuating accuracy levels across multiple iterations. Additional elements: It is possible that there are other elements unique to the dataset and the training process of the model that may impact its accuracy. The classification outcomes for the generated data are presented in **Table 6**.

Table 6. Comparison of Accuracy and Sensitivity of The Generated Data Imbalanced Class

Condition	Method's Code	Accuracy	Sensitivity		
			"1"	"2"	"3"
Without Resampling	ERSVM	0,60	0	0	1
	MR	0,60	0	0	1
With Resampling	ERSVM	0,33	0,33	0,32	0,33
	MR	0,21	0,66	0,24	0,13

The accuracy of the resampling method exhibits a general decline across all three datasets. However, an accuracy level above 80% is considered satisfactory **[14]**. This suggests that accuracy alone is insufficient for evaluating the quality of classification results. Evaluating the efficiency of classification methods for Imbalanced data scenarios requires more than just accuracy **[15]**. Sensitivity can also be employed for the evaluation of classification performance. The precision metric affected by the presence of imbalanced data.

The data sets consistently exhibit a diminished level of sensitivity in classes with a smaller number of participants. In certain classifiers, the ADASYN process is preceded by the occurrence of the lowest sensitivity value, which is zero. This observation indicates that the classification process has a bias towards the majority class. The sensitivity results of the Imbalanced Class GI data indicate that the sensitivity for kinds "1" and "2" is zero, but the sensitivity for class "3" is one in both the ERSVM and MR classification outcomes.

A sensitivity score of 0 indicates that the model is unable to accurately identify positive classes, resulting in inaccurate classifications. This suggests that the model fails to recognize and classify positive cases, instead mislabeling them as negative. The current condition is deemed undesirable as it indicates a full failure of the model to identify positive cases, which may result in significant implications depending on the specific application employed. A sensitivity rating of 1 indicates that the model is capable of accurately detecting all instances of positive cases. The dataset incorporates instances that have positive outcomes. In this hypothetical case, the model demonstrates optimal performance in accurately identifying positive classifications.

Nevertheless, attaining a sensitivity rating of 1 is a formidable task due to the typical existence of a trade-off between sensitivity and specificity. A high level of sensitivity may indicate that the model has a greater tendency to produce a larger number of false positive results, wherein negative instances are incorrectly classified as positive, in order to minimize the risk of missing positive cases. Hence, it is imperative to achieve a judicious evaluation that strikes a balance between sensitivity and specificity, taking into account the unique demands and circumstances of the particular application.

4. CONCLUSIONS

The accuracy of the ERSVM and MR classifiers on the unbalanced class Iris dataset is satisfactory, whereas the accuracy achieved by the ESVM and MR classifiers is superior. The performance of the ERSVM and MR classifiers on the imbalanced class GI dataset is marginally inferior to the accuracy achieved by the ESVM and MR classifiers. The accuracy achieved by the ERSVM and MR classifiers on the imbalanced synthetic dataset is comparatively lower than that of the ESVM and MR classifiers. The utilization of resampling can be a complex task due to the inherent uncertainty associated with it. Hence, it is imperative to know the influence of resampling on model performance by gaining a comprehensive understanding of data features, running experiments with different resampling approaches, and conducting rigorous evaluations. Moreover, the inclusion of evaluation criteria other than accuracy, namely sensitivity, specificity, and F1-score, might offer a more thorough understanding of the performance of a model when dealing with imbalanced data.

REFERENCES

[1] Jumairah&Mulyadi, "Analisis Perbandingan Klasifikasi Algoritma CART dengan Algoritma C 4 . 5 Pada Kasus Penderita Kanker Payudara," *J. Tekno Kompak*, vol. 17, no. 1, pp. 171–183, 2017.

[2] F. Marisa, S. Kom, A. L. Maukar, T. M. Akhriza, and P. D. MMSI, *Data mining konsep dan penerapannya*. Deepublish, 2021.

[3] A. Irma Prianti, "Pebandingan Metode K-Nearest Neighbor dan Adaptive Boosting pada Kasus Klasifikasi Multi Kelas," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 13, no. 1, pp. 39–47, 2020, doi: 10.36456/jstat.vol13.no1.a3269.

[4] L. Qadrini, H. Hikmah, and M. Megasari, "Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 386–391, 2022, doi: 10.47065/josyc.v3i4.2154.

[5] Q. Meidianingsih, D. E. Wardani, E. Salsabila, and A. N. Mutia, "Perbandingan Performa Metode Berbasis Support Vector Machine untuk Penanganan Klasifikasi Multi Kelas Tidak Seimbang," vol. 23, no. 1, pp. 8–18, 2023.

[6] S. Li, W. Song, H. Qin, and A. Hao, "Deep variance network: An iterative, improved CNN framework for unbalanced training datasets," *Pattern Recognit.*, vol. 81, pp. 294–308, 2018.

[7] F. Arofah and A. Sofro, "Penerapan Regresi Logistik Multinomial untuk Analisis Model Tingkat Depresi pada Lansia," *MATHunesa J. Ilm. Mat.*, vol. 10, no. 1, pp. 84–93, 2022, doi: 10.26740/mathunesa.v10n1.p84-93.

[8] L. Qadrini, "Handling Unbalanced Data With Smote Adaboost," *J. Mantik*, vol. 6, no. 2, pp. 2332–2336, 2022.

[9] S. Goswami and E. J. Wegman, "Comparison of different classification methods on glass identification for forensic research," *J. Stat. Sci. App*, vol. 4, pp. 65–84, 2016.

[10] J. He and P. Chalise, "Nested and repeated cross validation for classification model with high-dimensional data," *Rev. Colomb. Estadística*, vol. 43, no. 1, pp. 103–125, 2020.

[11] J. Beinecke and D. Heider, "Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making," *BioData Min.*, vol. 14, pp. 1–11, 2021.

[12] R. D. Permatasari, S. W. Rizki, and N. N. Debataraja, "PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE DALAM MENGATASI DATA TIDAK SEIMBANG PADA METODE CLASSIFICATION AND REGRESSION TREE," *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 9, no. 1, 2020.

[13] L. Qadrini, "Undersampling dan K-Fold Random Forest Untuk Klasifikasi Kelas Tidak Seimbang," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1967–1974, 2023, doi: 10.47065/bits.v4i4.3141.

[14] A. Herdiansah, R. I. Borman, D. Nurnaningsih, A. A. J. Sinlae, and R. R. Al Hakim, "Klasifikasi Citra Daun Herbal dengan Menggunakan Backpropagation Neural Networks Berdasarkan Ekstraksi Ciri Bentuk," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, pp. 388–395, 2022.

[15] A. A. Arifiyanti and E. D. Wahyuni, "SMOTE: Metode penyeimbang kelas pada klasifikasi data mining," *Scan J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, pp. 34–39, 2020.