

Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data

Nur Hanisah Abdul Malek¹, Wan Fairos Wan Yaacob^{1,2}, Yap Bee Wah³, Syerina Azlin Md Nasir¹,
Norshahida Shaadan⁴, Sapto Wahyu Indratno^{5,6}

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Kota Bharu, Malaysia

²Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Shah Alam, Malaysia

³UNITAR Graduate School, UNITAR International University, Petaling Jaya, Malaysia

⁴Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia

⁵Statistics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung, Indonesia

⁶Natural Language Processing and Big Data Analytics (U-CoE AI-VLB), Institut Teknologi Bandung, Bandung, Indonesia

Article Info

Article history:

Received Apr 27, 2022

Revised Sep 28, 2022

Accepted Oct 11, 2022

Keywords:

Bagging

Boosting

Ensemble methods

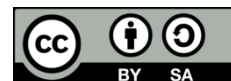
Hybrid sampling

Imbalanced data

ABSTRACT

Training an imbalanced dataset can cause classifiers to overfit the majority class and increase the possibility of information loss for the minority class. Moreover, accuracy may not give a clear picture of the classifier's performance. This paper utilized decision tree (DT), support vector machine (SVM), artificial neural networks (ANN), K-nearest neighbors (KNN) and Naïve Bayes (NB) besides ensemble models like random forest (RF) and gradient boosting (GB), which use bagging and boosting methods, three sampling approaches and seven performance metrics to investigate the effect of class imbalance on water quality data. Based on the results, the best model was gradient boosting without resampling for almost all metrics except balanced accuracy, sensitivity and area under the curve (AUC), followed by random forest model without resampling in term of specificity, precision and AUC. However, in term of balanced accuracy and sensitivity, the highest performance was achieved by random forest with a random under-sampling dataset. Focusing on each performance metric separately, the results showed that for specificity and precision, it is better not to preprocess all the ensemble classifiers. Nevertheless, the results for balanced accuracy and sensitivity showed improvement for both ensemble classifiers when using all the resampled dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Wan Fairos Wan Yaacob

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan

Kampus Kota Bharu, Lembah Sireh, 15050 Kota Bharu, Kelantan, Malaysia

Email: wnfairos@uitm.edu.my

1. INTRODUCTION

One of the main challenges of machine learning is the processing of imbalance data for classification tasks [1]. Recently, the classification of imbalanced data becomes a highly explored issue because when imbalanced data occurred, classifiers have a tendency to produce a biased model with close to zero sensitivity for the minority class. Even not a single minority class sample is classified correctly, the accuracy can reach up to 99% as most majority classes were classified correctly. In other words, accuracy will not give a clear picture of the classifier's performance in an imbalanced dataset. Issues of imbalanced data occurred in many fields such as bankruptcy risk data [2], credit scoring [3], healthcare medical data [4], student performance [5], point cloud data [6], anomalies detection [7] and also water quality data [8]. In real-world applications, the severity of class imbalance may range from mild to severe [9]. The severity of imbalance is said to be mild if

the proportion of minority class is between 20%-40%, moderately imbalance if less than 20% of the data and extreme if less than 1% of the data. A classifier applied without any strategy to process imbalanced data will tend to ignore the minority class and, as a result, will almost inevitably classify it incorrectly.

Basically, there are three approaches to deal with imbalanced data which are data level, algorithm level and ensemble methods [10]. The data-level approach consists of re-sampling the data to reduce class imbalance. There are two basic re-sampling techniques which are under-sampling the majority class and oversampling the minority class. Among oversampling techniques, the most fundamental technique is random oversampling (ROS). Rachburee and Punlumjeak [5] applied adaptive synthetic (ADASYN) method, synthetic minority oversampling technique (SMOTE), SVM-SMOTE and Borderline-SMOTE to predict student performance. They found that Borderline-SMOTE method gave the best prediction result using several classifiers. For under-sampling, random under-sampling (RUS) is the most popular under-sampling technique. Some researchers have opt to combine both oversampling and under-sampling techniques which is called hybrid sampling [6], [11]. These techniques are used to produce a balanced dataset which make the classifiers not biased toward one class or another. Lin and Nguyen [6] used the hybrid sampling technique which involved ROS followed by RUS with a balance loss cost function to resolve imbalanced data. They found that oversampling followed by under-sampling was more effective than under-sampling followed by oversampling. They also found that the proposed method improved performance by 7%. The advantages of ROS-RUS method are that it implies nothing about the data, simple and no heuristic is used [6]. Another study by [12] combined oversampling technique, SMOTE and under-sampling technique to cater the imbalanced issue on 10 datasets. They found that the hybrid sampling had better performance compared to the other technique.

Second, in algorithm level approach, machine learning model is modified to adapt the imbalanced data. Next, the third approach is ensemble method. Ensemble method combines several base learners' decision to produce more precise prediction than each base learner's decision [13]. There are two commonly used ensemble families in machine learning which are bagging and boosting. Bootstrap aggregating or bagging is a method that learns multiple base classifiers in parallel. The advantage of this bagging method is that it can lower the variance while retaining low bias of the base classifiers. This is done by averaging outputs from base classifiers [11]. Boosting method also works by combining multiple base learners. However, it trains the multiple learners in sequential way [14]. The weights are allocated to the instances by each learner and then the weighted instances are utilized by the next learner. The weights of instances which are incorrectly classified are increased, while the instances' weights that are correctly classified are decreased. Both bagging and boosting methods provide higher stability to the classifiers and are good in reducing variance. In a previous study, they compared the performance of a single model and modified ensemble bagging model by using banking financial ratios data. The results showed that the modified ensemble bagging model was always more accurate compared to the single model [2]. This is supported by another study [15] which found that ensemble bagging model increased the performance of decision trees C4.5 and CART model. Evangelista and Sy [16] used four ensemble models which are homogeneous ensembles (boosting and bagging) and heterogeneous ensembles (stacking and voting) to enhance different single classifier's performance. The results in the study revealed that voting ensemble model performed slightly better than boosting and bagging models. Meanwhile, Priasni and Oswari [17] applied three ensemble learning models which are voting, Adaboost and bagging to the Naïve Bayes, decision tree and support vector machine classifiers. They found that Adaboost model using decision tree as base classifier had the highest accuracy and precision while bagging model using support vector machine as base classifier had the highest f-measure, area under the curve (AUC) and recall.

However, ensemble methods which employs resampling techniques are expected to work better in handling imbalanced data. This was proven by [11] when they found that their hybrid sampling combined with bagging model (RSYNBagging) had the best classification performance based on the AUC-ROC plot. This study demonstrated the advantage of combining oversampling and under-sampling techniques with ensemble model to cater imbalanced class issue. Lu *et al.* [18] also used hybrid sampling with bagging (HSBagging) which adopted random under-sampling technique and SMOTE integrated with bagging algorithm. The study found that HSBagging outperformed the other related UnderBagging and SMOTEBagging methods. Many researchers used machine learning models such as random forest (RF) [19], [20], extreme gradient boosting (XGBoost) [3], ensemble models [21], [22], hybridization of random forest and extreme gradient boosting [23], gradient boosting (GB) and conventional machine learning model such as decision tree (DT), artificial neural network (ANN), k-nearest neighbors (KNN), support vector machine (SVM) and Naïve Bayes (NB) [8] for classification. However, there is still no conclusive evidence as to which is the best approach. The aim of this study is therefore to investigate the predictive performance of five conventional machine learning models which are SVM, NB, KNN, ANN, DT and two popular ensemble models which are random forest and gradient boosting using three resampling techniques such as random oversampling (ROS), random under-sampling (RUS) and hybrid sampling of ROS-RUS method on the imbalanced water quality classification (WQC) dataset. This paper is organized as follows: section 2 describes the methodology for evaluating the machine

learning models with application of sampling techniques (ROS, RUS and ROS-RUS). The results are presents and discussed in section 3 and the conclusion is given in section 4.

2. METHOD

2.1. Water quality data

This study used secondary data on various parameters of water quality which were obtained from Department of Environment (DOE) Malaysia. DOE performs regular water quality monitoring of Kelantan River for 4, 5 or 6 times per year based on the stations. Kelantan River is one of the main rivers in Malaysia which is located in the north-east of peninsular Malaysia. The data are for 2005 to 2020. In 2005 until 2015, the data were from 8 stations situated along Kelantan River, namely Jambatan Kusia, Jambatan Sultan Yahya Petra, Kota Bahru, Tangga Kerai, Bandar Kuala Kerai, Jambatan bandar Rantau Panjang-Golok, Kampung Kuala Sat, Jeli, Kampung Bukit Bunga, Kampung Lubok Setol and Kampung Jeram Perdah. Later, in 2016, data from a new station at Loji Air Lemal, Pasir Mas was included. In 2018, three new monitoring stations were added in Kelantan River which are Sg. Relai, Loji Ayer Lanas and Skim Bekalan Air Merbau Chondong. Hence, giving the total observations in this study is 685 observations measured 4, 5 or 6 times per year for 16 years at 12 locations. The dataset consists of the target variable which is the water quality classification (WQC) and 13 physicochemical parameters which are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solid (TSS), pH, Ammoniacal Nitrogen (NH₃-N), temperature, conductivity, salinity, turbidity, nitrogen (NO₃), phosphorus (PO₄) and *Escherichia coli* (E-coli). WQC are constructed based on the water quality index value range as shown in Table 1. The water quality is classified as clean if the WQI value range between 81 to 100 and slightly polluted if range between 60 to 80 [8].

Table 1. Water quality classification

Parameter	Water quality classification	
	Slightly Polluted	Clean
Water quality index	60-80	81-100

2.2. Data pre-processing

Data pre-processing is a vital step to prepare the data before developing water quality predictive models using machine learning classifier. It involves a number of important steps, such as data clean-up, data transformation and feature selection. Data clean-up and transformation are methods used to remove outliers and standardize data to have similar units. This study used z-score method to standardize the data and Mahalanobis distance to detect outliers. Based on the Mahalanobis distance, 27 outliers were detected in the dataset and removed from the dataset. The number of remaining samples is 658. Next, for missing values analysis, only 3 variables which are turbidity, phosphorus and E-coli has missing values with the missing percentage of 1.0%, 1.8% and 1.0% respectively. The missing values were imputed using expectation maximization (EM) method. This study used R programming software to analyse the data.

2.3. Conventional machine learning models

2.3.1. K-nearest neighbours

This K-nearest neighbours (KNN) algorithm classifies the samples by discovering the given points nearest neighbours and assigns the class of majority of K neighbours to it. In the event of a draw, different techniques could be used to solve it. However, KNN is not suggested for large data set since all processing occurs during the testing, and it iterates through all the training data and calculates the nearest neighbours each time [24]. This study used K = 10 configuration for the KNN model.

2.3.2. Support vector machines

Support vector machine (SVM) is one of the classifying methods based on the theory of statistical learning. SVM uses the structural risk minimization principle to address overfitting problem in machine learning by reducing the model's complexity and fitting the training data successfully. Minimization of risk can enhance the generalization of the SVM model [25]. Estimates of the SVM model are created based on small sub-set of training data which is known as support vector. The capability to interpret support vector machine decisions can be improved by recognizing vectors that are chosen as support vector [26]. SVM maps the initial data in a high-dimension feature space in which an optimal separating plane is created by using suitable kernel function. For classification, the optimal separating plane is the line that dividing the plane into two parts and each class is placed into different side. Along each part of the separating plane, 2 parallel

hyperplanes could be built to separate the training data. The hyperplane is optimal if the margin between closest training vector and the hyperplane is maximal. This study used complexity constant, $C=5$ to set the misclassification tolerance. Large value of C can lead to overfitting problem while small value may cause over generalization. This study used the polynomial kernel since it is suitable for the case where all training data are normalized.

2.3.3. Artificial neural network

Artificial neural network (ANN) works like a human brain's nervous system which comprises of interconnected neurons that work together in parallel [8]. It is widely used in many fields because of its advantages such as self-organizing, self-learning and self-adapting abilities. Neural network's structure is composed of 3 layers which are the input, middle and output layer. Input variables are entered into the algorithm in the input layer. In the middle layer, the input variables are multiplied by weights before they are summed by a constant value. Then, an activation function is added to the sum of the weighted inputs. Activation function are needed to transform the input signals into output signals. Recent artificial neural network algorithms employ activation functions that are non-linear [27]. This is because non-linear activation functions allow backpropagation and multi-layer neurons stacking to produce complex mapping between input and output networks which are needed to study complex dataset. Most popular activation functions are Gaussian, Sigmoid and Tansig. In the output layer, the prediction is obtained from the parallel computation in the middle layer. The mathematical formula of neuron computation is given by $I_j = f(\sum_i w_{ij} \alpha_i + \theta_j)$ where w_{ij} are the weights, α_i are the input variables and θ_j are the biases. This study used the default hidden layer which consist of one hidden layer with Sigmoid activation function and size equal to (number of attributes + number of classes)/2+1.

2.3.4. Decision tree

Decision tree (DT) is a simple and explicit algorithm that makes decisions based on values from all relevant input parameters. DT uses entropy to select the root variable and based on that, it looks to the values of the other parameters. It has all the parameter decisions organized in a tree from top to bottom and plans the decision based on different values of different parameters [28]. Decision tree models frequently found in previous studies to perform well on imbalanced data. However, decision tree-based ensembles models including random forests (RF) and gradient boosting (GB) almost always outperform the single decision tree. The advantages of decision tree-based model are not sensitive to missing values, ability to manage both regular attributes and data and highly efficient.

2.3.5. Naïve Bayes

Bayes approach employs probability statistics knowledge to classify the data and estimate the outcome. The Bayes model uses prior and posterior probabilities in order to prevent overfitting problem and bias from using only sample information [29]. A classification technique that uses Bayes theorem and the independent conditions assumption is known as Naïve Bayes (NB). When the target value is specified, the attributes are meant to be conditionally independent from each other [29]. This technique makes the complexity of the Bayes model much simpler. The probability of event A occurs given that event B occurred is different from the probability of event B occurs given that event A occurred. Assume that A_1, A_2, \dots, A_n are the event vectors and B is the dataset class, hence the Naïve Bayes formula may be written as shown in (1):

$$P(B|A_1, A_2, \dots, A_n) = \frac{P(B) \times P(A_1, A_2, \dots, A_n|B)}{P(A_1, A_2, \dots, A_n)} = \frac{P(B) \times \prod_{j=1}^n P(A_j|B)}{P(A_1, A_2, \dots, A_n)} \quad (1)$$

where the $P(A)$ is a prior probability that represents the event vectors and $P(A_j|B)$ is the dataset class prior probability. This study used default values for this algorithm.

2.4. Ensemble methods for imbalanced problem

2.4.1. Bagging ensemble method for machine learning

Random forest is a classification model that uses multiple base models typically decision trees, on a given subset of data independently and makes decisions based on all models [30]. It uses feature randomness and bagging when building each individual decision tree to produce independent forest of trees. RF is a method of calculating the mean of several deep decision trees formed in different parts of the same training set, with the aim of reducing the variance. The prediction by this committee is more accurate than that of any individual tree and robust against overfitting. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node [31]. RF algorithm works by creating n_{tree} bootstrap sub-samples of original dataset with replacement first. Then, for each bootstrap samples, train a decision tree model. The new data are predicted by aggregating the prediction of the n_{tree} models (majority votes for classification).

2.4.2. Boosting ensemble method for machine learning

Gradient boosting is a boosting-based machine learning algorithm which trains multiple weak classifiers typically decision tree to create a robust classifier for regression and classification problems [32]. It assembles the model in a stage-wise way similar to the what the other boosting techniques do and it generalizes them by optimizing a suitable cost function. In the GB algorithm, incorrectly classified cases for a step are given increased weight during the next step. The advantages of GB are that it has exceptional accuracy in predicting and fast process.

2.5. Sampling techniques

This section outlines three sampling techniques utilized in this study to address the issue of imbalanced data. Random under-sampling, random oversampling and hybrid sampling ROS-RUS are among the approaches used. The details about each sampling technique are discussed briefly below.

2.5.1. Random under-sampling

Random under-sampling (RUS) method works by randomly removing the instances of the majority class until a certain desired majority-to-minority ratio is achieved. However, the drawback of this method is it may delete useful data which cause information loss [18]. This random deletion may also modify the majority class distribution and therefore modify their representative features. When this occurs, a large number of majority cases will be misclassified. However, despite these drawbacks, RUS generally works better than other under-sampling methods [11].

2.5.2. Random oversampling

Among oversampling techniques, the most fundamental technique is random oversampling. In random oversampling (ROS), minority class samples are randomly selected and duplicated till the data become balanced [11]. Nevertheless, this approach has led to overfitting problem where the classifiers become biased to the duplicated samples. Consequently, the classifiers are not able to classify new instances correctly.

2.5.3. Hybrid sampling

Interesting results can be obtained by combining random oversampling with random under-sampling. The classifier's performance could be enhanced to a greater extent. In the Hybrid sampling (ROS-RUS) method, the minority class data is mixed with the majority class data after oversampling and then all data are down sampled, so that they are matched with the input of network. The imbalanced ratio of the data set generated is also random, resulting in additional diversity from which the ensemble can also benefit [33]. Given a dataset TR with N samples $\{x_i, y_i\}, i = 1, 2, \dots, N$, where x_i is the sample in the m dimension feature space and the label of the class $y_i \in C = \{Y_0, Y_1\}$. x_i are a random vector attributes x defined on R^d , with unknown probability density function $f(x)$. Let N_j be the number of samples belonging to class Y_j . First, random oversampling procedure chooses $y^* = Y_j$ with probability π_j . Then, select $\{x_i, y_i\} \in TR$, where $y_i = y^*$ with probability $1/N_j$. Lastly, sample x^* from $K_{H_j}(\cdot, x_i)$ where K_{H_j} is probability distribution that centred at x_i and covariance matrix H_j [34].

2.6. Performance evaluation

The machine learning models were evaluated using 10-fold cross validation technique. Cross validation was used to assess predictive models by dividing the original data into training and testing dataset for ten times. Typically, the data were divided in a ratio of 70:30. Although a universal guideline does not exist, the ratio of 70:30 are the most frequently for evaluation of predictive models [35]. In this study, seven distinct metrics were considered: balanced $accuracy = (Sensitivity + Specificity)/2$, $accuracy = (TP + TN)/(TP + FP + FN + TN)$, $specificity = TN/(TN + FP)$, $sensitivity = TP/(TP + FN)$, $precision = TP/(TP + FP)$, $f - measure = (2 \times Precision \times Recall)/(Precision + Recall)$ and area under the curve $(AUC) = (1 + TPR - FPR)/2$ where TPR is true positive rate and FPR is false positive rate. These metrics were determined using different values given in the confusion matrix as shown in Table 2.

Table 2. Confusion matrix for binary classification

		Predicted	
		Clean	Slightly Polluted
Actual	Clean	True Negative (TN)	False Positive (FP)
	Slightly Polluted	False Negative (FN)	True Positive (TP)

3. RESULTS AND DISCUSSION

3.1. Imbalance ratio (IR)

Imbalance ratio is the most common measure used to describe the extent of the imbalance of a dataset. It is defined as the number of majority class over the number of minority class [36]. The imbalance ratio in this study is 3.84 which means the data is moderately imbalanced. The imbalanced scenario between clean and slightly polluted classes are shown in Figure 1.

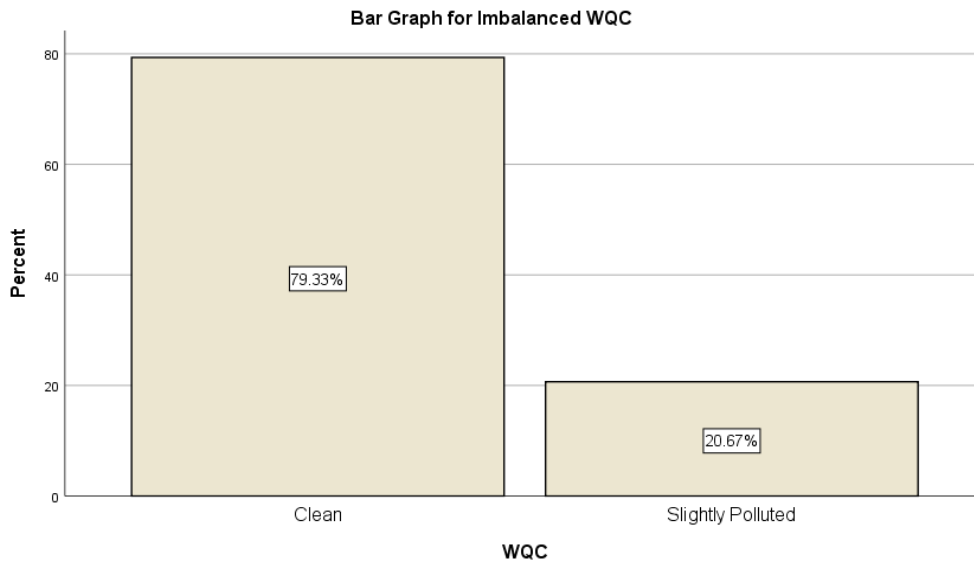


Figure 1. Bar graph for water quality classification

3.2. Comparison of ensemble models and conventional machine learning

This subsection presents the performance results of the five conventional machine learning and the two ensemble models without resampling the original data. Based on the output in Table 3, the performance of the two ensemble models which are RF and GB are better than the other conventional machine learning models in term of accuracy, f-measure and AUC. A clear superiority of GB model which uses ensemble boosting method over the other machine learning models. This is followed by RF which use ensemble bagging method. This means that boosting and bagging models have enhanced the performance of classifiers.

Table 3. Performance metrics of conventional machine learning and ensemble models

Algorithm	Accuracy (%)	Balanced Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)	AUC (%)
KNN	90.82	82.15	67.50	96.79	84.38	75.00	89.86
SVM	92.86	85.29	72.50	98.08	90.62	80.56	92.85
ANN	93.37	89.33	82.50	96.15	84.62	83.54	93.85
DT	86.22	80.19	70.00	90.38	65.12	67.47	87.76
NB	90.31	85.54	77.50	93.59	75.61	76.54	92.52
RF	93.88	88.72	80.00	97.44	88.89	84.21	98.27
GB	94.90	89.36	80.00	98.72	94.12	86.49	98.61

3.3. Comparison of performance metrics for all machine learning after resampling

Next, this study compares the performance of the seven machine learning using ROS, RUS and hybrid sampling ROS-RUS. The method without resampling which means no established method of processing imbalance was also included as a baseline performance reference. Based on the output in Table 4, the best method was GB with Original data for almost all metrics except balanced accuracy and sensitivity, followed by RF with ROS-RUS, in term of accuracy, balanced accuracy, specificity, precision and f-measure. While, the method that showed the worst results was NB with ROS-RUS followed by DT with RUS. Focusing on each performance metric separately, the results of specificity and precision for some classifiers which are KNN, SVM and GB tend to reveal that it is better not to resample the data since resampling approach did not improve the classifiers. However, for classifiers like RF, ANN, DT and NB, the results were improved after resampling.

The results for sensitivity showed improvement for all classifiers except NB when using resampling dataset as shown in Figure 2.

Table 4. Performance metrics for all machine learning after resampling using ROS, RUS and ROS-RUS

Algorithm	Sampling	Accuracy (%)	Balanced Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F-measure (%)	AUC (%)
KNN	Without resampling	90.82	82.15	67.50	96.79	84.38	75.00	89.86
	ROS	89.29	81.19	67.50	94.87	77.14	72.00	81.19
	RUS	94.39	92.76	90.00	95.51	83.72	86.75	98.02
	ROS-RUS	90.82	86.79	80.00	93.59	76.19	78.05	86.79
SVM	Without resampling	92.86	85.29	72.50	98.08	90.62	80.56	92.85
	ROS	92.86	88.08	80.00	96.15	84.21	82.05	93.22
	RUS	86.22	83.91	80.00	87.82	62.75	70.33	90.88
	ROS-RUS	88.78	87.37	85.00	89.74	68.00	75.56	93.43
ANN	Without resampling	93.37	89.33	82.50	96.15	84.62	83.54	93.85
	ROS	95.92	93.72	90.00	97.44	90.00	90.00	97.84
	RUS	93.88	91.51	87.50	95.51	83.33	85.37	95.18
	ROS-RUS	90.31	87.40	82.50	92.31	73.33	77.65	89.04
DT	Without resampling	86.22	80.19	70.00	90.38	65.12	67.47	87.76
	ROS	87.76	83.94	77.50	90.38	67.39	72.09	92.14
	RUS	84.18	79.84	72.50	87.18	59.18	65.17	77.99
	ROS-RUS	86.22	81.12	72.50	89.74	64.44	68.24	78.82
NB	Without resampling	90.31	85.54	77.50	93.59	75.61	76.54	92.52
	ROS	86.73	82.37	75.00	89.74	65.22	69.77	90.11
	RUS	91.33	87.12	80.00	94.23	78.05	79.01	95.21
	ROS-RUS	83.67	79.52	72.50	86.54	58.00	64.44	88.80
RF	Without resampling	93.88	88.72	80.00	97.44	88.89	84.21	98.27
	ROS	91.84	87.44	80.00	94.87	80.00	80.00	97.12
	RUS	89.80	88.94	87.50	90.38	70.00	77.78	97.19
	ROS-RUS	94.39	89.97	82.50	97.44	89.19	85.71	97.58
GB	Without resampling	94.90	89.36	80.00	98.72	94.12	86.49	98.61
	ROS	93.37	89.33	82.50	96.15	84.62	83.54	98.38
	RUS	91.84	91.15	90.00	92.31	75.00	81.82	97.04
	ROS-RUS	92.86	90.87	87.50	94.23	79.55	83.33	97.28

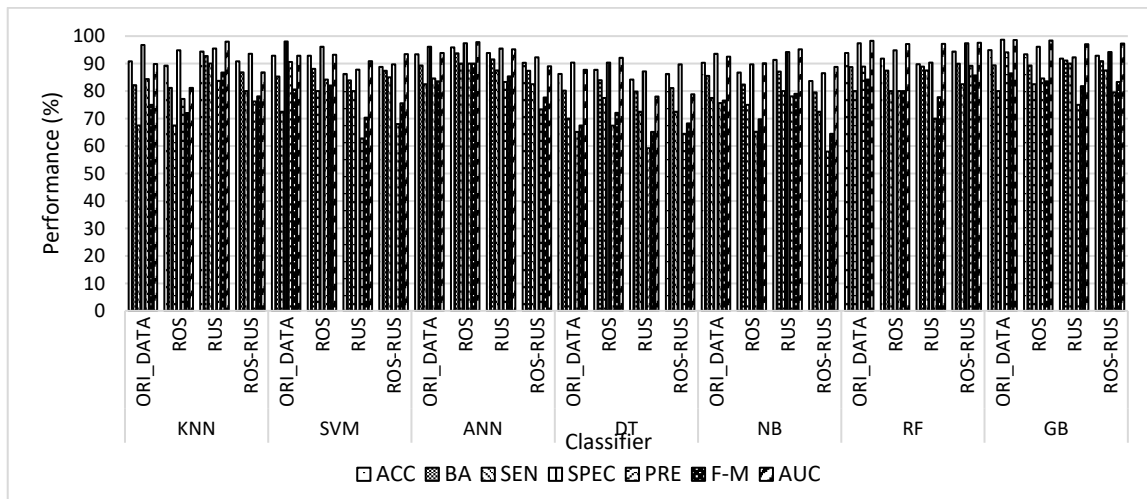


Figure 2. Comparison of classifier performance by resampling method

Moreover, it is worth noting the results of some conventional classifiers highlighted a better performance when resampling methods were used. Sensitivity improves for KNN (ROS-RUS and RUS), SVM (ROS-RUS-highest), ANN (ROS-highest), DT (ROS-highest) and NB (RUS), as shown in Figure 3. On the other hand, f-measure metric revealed that some resampling contributed to overcome the imbalance compared to without resampling. The improvement was also observed for ensemble classifier of RF. Sensitivity improves for both RF and GB, especially under RUS sampling method, as shown in Figure 4.

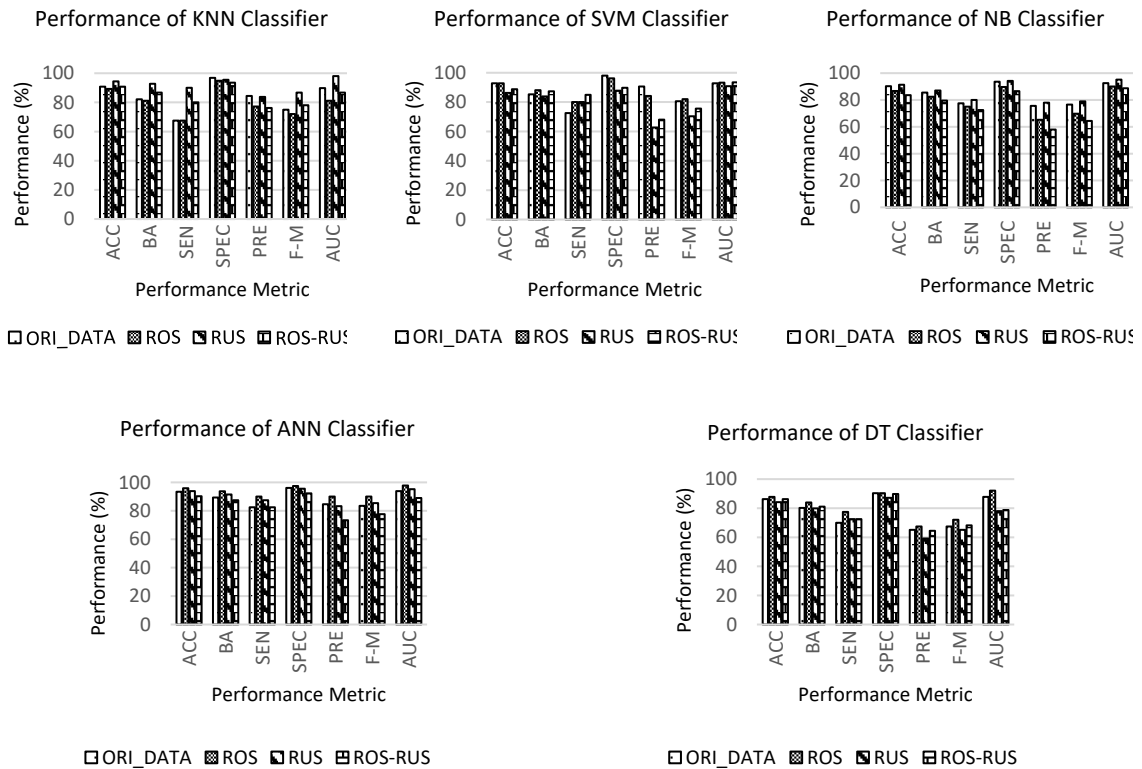


Figure 3. Comparison of conventional classifier performance by resampling method

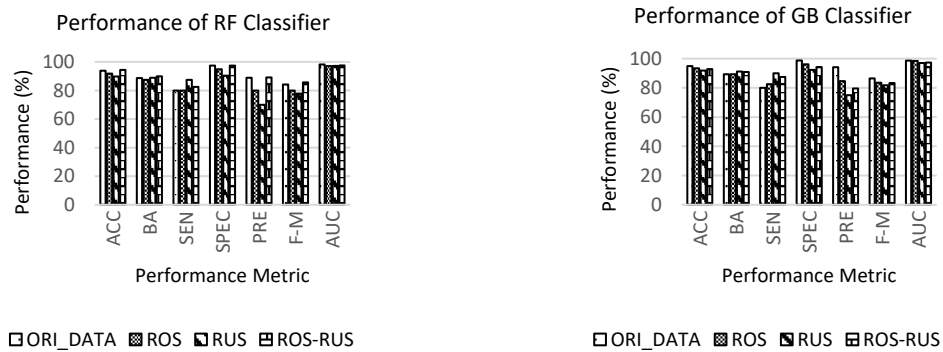


Figure 4. Comparison of ensemble classifier performance by resampling method

4. CONCLUSION

This paper illustrated the impact of using data-sampling approaches for developing predictive model for imbalanced water quality data. These approaches involve primarily the use of preprocessing techniques such as RUS, ROS and ROS-RUS (hybrid sampling) to transform an imbalanced dataset into a balanced dataset. The analysis was conducted to emphasize the effect of resampling techniques on the performance of two ensemble families: bagging (random forest) and boosting (gradient boosting). The ensemble boosting method, while it requires more computing power, has clearly outperformed the bagging method. Surprisingly, the training of the ensembles on the original dataset without any change offered quite good results overall, especially for gradient boosting. For resampling techniques, ROS generally performed better, but with minimal advantage, closely followed by RUS. A very interesting conclusion of the study is the importance of using different assessment metrics when addressing imbalance issues. This is preferable because every metric uses the values of the confusion matrix in a specific way and thus has its own strengths and weaknesses. Therefore,

the use of more than one measure provides a more informed view of the results and an improved assessment of a single classifier's performance.

ACKNOWLEDGEMENTS

The authors are gratefully acknowledged Department of Environment Malaysia for providing the water quality data. This research was funded by Geran Insentif Penyelidikan, Universiti Teknologi MARA, grant number 600-RMC/GIP 5/3 (065/2021) and supported by The Journal Support Fund, UiTM.




REFERENCES

- [1] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.
- [2] B. Siswoyo, Z. A. Abas, A. N. C. Pee, R. Komalasari, and N. Suyatna, "Ensemble machine learning algorithm optimization of bankruptcy prediction of bank," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 679-686, 2022, doi: 10.11591/ijai.v11.i2.pp679-686.
- [3] W. Yotsawat, P. Wattuya, and A. Srivihok, "Improved credit scoring model using XGBoost with Bayesian hyper-parameter optimization," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5477-5487, 2021, doi: 10.11591/ijece.v11i6.pp5477-5487.
- [4] A. S. Desuky, A. H. Omar, and N. M. Mostafa, "Boosting with crossover for improving imbalanced medical datasets classification," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 5, pp. 2733-2741, 2021, doi: 10.11591/eei.v10i5.3121.
- [5] N. Rachburee and W. Punlumjeak, "Oversampling technique in student performance classification from engineering course," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3567-3574, 2021, doi: 10.11591/ijece.v11i4.pp3567-3574.
- [6] H. I. Lin and M. C. Nguyen, "Boosting minority class prediction on imbalanced point cloud data," *Applied Sciences*, vol. 10, no. 3, 2020, doi: 10.3390/app10030973.
- [7] N. P. Shetty, J. Shetty, R. Narula, and K. Tandona, "Comparison study of machine learning classifiers to detect anomalies," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 5445-5452, 2020, doi: 10.11591/ijece.v10i5.pp5445-5452.
- [8] N. H. A. Malek, W. F. W. Yaacob, S. A. M. Nasir, and N. Shaadan, "Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques," *Water*, vol. 14, no. 7, 2022, doi: 10.3390/w14071067.
- [9] L. L. Joffrey, M. K. Taghi, A. B. Richard, and S. Naeem, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, p. 42, 2018, doi: 10.1186/s40537-018-0151-6.
- [10] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, 2014, pp. 13-22, doi: 10.1007/978-981-4585-18-7_2.
- [11] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid methods for class imbalance learning employing bagging with sampling techniques," in *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, 2017, pp. 1-5, doi: 10.1109/CSITSS.2017.8447799.
- [12] S. M. J. Moghaddam and A. Noroozi, "A novel imbalanced data classification approach using both under and over sampling," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 5, pp. 2789-2795, 2021, doi: 10.11591/eei.v10i5.2785.
- [13] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484, 2011, doi: 10.1109/TSMCC.2011.2161285.
- [14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119-139, 1997, doi: 10.1006/jcss.1997.1504.
- [15] F. Aziz and A. Lawi, "Increasing electrical grid stability classification performance using ensemble bagging of C4. 5 and classification and regression trees," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, p. 2955-2962, 2022, doi: 10.11591/ijece.v12i3.pp2955-2962.
- [16] E. De Leon Evangelista and B. D. Sy, "An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 5226-5235, 2022, doi: 10.11591/ijece.v12i5.pp5226-5235.
- [17] T. O. Priasni and T. Oswari, "Comparative study of standalone classifier and ensemble classifier," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 5, pp. 1747-1754, 2021, doi: 10.12928/telkomnika.v19i5.19508.
- [18] Y. Lu, Y.-m. Cheung, and Y. Y. Tang, "Hybrid sampling with bagging for class imbalance learning," *PAKDD 2016: Advances in Knowledge Discovery and Data Mining*, 2016, pp. 14-26, doi: 10.1007/978-3-319-31753-3_2.
- [19] M. H. I. Bijoy, S. A. Akhi, M. A. A. Nayeem, M. M. Rahman, and M. J. Mia, "Prediction of internet user satisfaction levels in Bangladesh using data mining and analysis of influential factors," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 2, pp. 926-935, 2022, doi: 10.11591/eei.v11i2.3617.
- [20] T. Mohd, S. Jamil, and S. Masrom, "Machine learning building price prediction with green building determinant," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 3, pp. 379-386, 2020, doi: 10.11591/ijai.v9.i3.pp379-386.
- [21] N. A. Mashudi, N. Ahmad, and N. M. Noor, "Classification of adult autistic spectrum disorder using machine learning approach," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 3, pp. 743-751, 2021, doi: 10.11591/ijai.v10.i3.pp743-751.
- [22] A. C. Alhadi, A. Deraman, M. M. A. Jalil, W. N. J. W. Yussof, and R. Mohamad, "A computational analysis of short sentences based on ensemble similarity model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5386-5394, 2019, doi: 10.11591/ijece.v9i6.pp5386-5394.
- [23] H. Shamsudin, M. Sabudin, and U. K. Yusof, "Hybridisation of RF (Xgb) to improve the tree-based algorithms in learning style prediction," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 422-428, 2019, doi: 10.11591/ijai.v8.i4.pp422-428.




- [24] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," *ICDT 1999: Database Theory — ICDT'99*, 1999, pp. 217-235, doi: 10.1007/3-540-49257-7_15.
- [25] M. Behzad, K. Asghari, M. Eazi, and M. Palhang, "Generalization performance of support vector machines and neural networks in runoff modeling," *Expert Systems with applications*, vol. 36, no. 4, pp. 7624-7629, 2009, doi: 10.1016/j.eswa.2008.09.053.
- [26] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857-900, 2019, doi: 10.1007/s10462-017-9611-1.
- [27] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," *Water Quality Research Journal*, vol. 53, no. 1, pp. 3-13, 2018, doi: 10.2166/wqrj.2018.025.
- [28] J. R. Quinlan, "Decision trees and decision-making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 2, pp. 339-346, 1990, doi: 10.1109/21.52545.
- [29] H. Zhang, "The optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 2004, pp. 562-567.
- [30] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water (Switzerland)*, vol. 11, no. 11, pp. 1-14, 2019, doi: 10.3390/w11112210.
- [31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
- [32] R. Prakash, V. P. Tharun, and S. R. Devi, "A comparative study of various classification techniques to determine water quality," 2018: IEEE, pp. 1501-1506, doi: 10.1109/ICICCT.2018.8473168.
- [33] J. F. Díez-Pastor, J. J. Rodríguez, C. Garcia-Osorio, and L. I. Kuncheva, "Random balance: ensembles of variable priors classifiers for imbalanced data," *Knowledge-Based Systems*, vol. 85, pp. 96-111, 2015, doi: 10.1016/j.knosys.2015.04.022.
- [34] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018, doi: 10.1007/978-3-319-98074-4.
- [35] K. Khosravi, L. Mao, O. Kisi, Z. M. Yaseen, and S. Shahid, "Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile," *Journal of Hydrology*, vol. 567, pp. 165-179, 2018, doi: 10.1016/j.jhydrol.2018.10.015.
- [36] R. Zhu, Y. Guo, and J.-H. Xue, "Adjusting the imbalance ratio by the dimensionality of imbalanced data," *Pattern Recognition Letters*, vol. 133, pp. 217-223, 2020, doi: 10.1016/j.patrec.2020.03.004.

BIOGRAPHIES OF AUTHORS






Nur Hanisah Abdul Malek    is a PhD student in Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Malaysia. She obtained her bachelor degree in Science (Statistics) in year 2011 and master in Applied Statistics from Universiti Teknologi MARA in year 2014. Her research interest is on the imbalanced data issues using various machine learning models. Currently, she is doing a research project on the application of ensemble hybrid sampling of bagging and boosting machine learning in predicting imbalance water quality data. She has published a few papers in water quality and machine learning studies and also presented papers in conferences. She can be contacted at email: hanisahmalek@gmail.com.






Assoc. Prof Dr. Wan Fairos Wan Yaacob    is a statistician by profession. She is a senior lecturer of Department of Statistics, Universiti Teknologi MARA Cawangan Kelantan, Malaysia, Head of Business Datalytics Research Group and currently an associate fellow of Institute for Big Data Analytics and Artificial Intelligence (IBDAAI). She obtained her Master of Science in Statistics from Universiti Kebangsaan Malaysia and PhD in Statistics from Universiti Teknologi MARA. Her area of research interest focuses on statistical modeling of panel count data, data mining and machine learning in dengue disease, road accident, water quality and many other areas. She is a certified trainer in data analyst and machine learning master from RapidMiner. She has published more than 50 articles and papers in various well-known journals and presented papers at conferences. She can be contacted at email: wnfairos@uitm.edu.my.






Prof. Dr. Yap Bee Wah    graduated with a BSc (Hons) (Mathematics Education) from University Sains Malaysia. She then obtained her Masters of Statistics degree from University of California, Riverside, USA and her PhD (Statistics) from University of Malaya. She recently joined UNITAR International University in September 2021, and besides teaching and supervision, spearheads the university research and consultancy unit. Her research interests are in big data analytics and data science, computational statistics and multivariate data analysis. Her research works are in the healthcare, education, environment and business domains. She has published more than 100 papers in Scopus indexed journals and proceedings. She can be contacted at bee.wah@unitar.my.






Dr. Syerina Azlin Md Nasir    received her Ph.D. in Information Technology from Universiti Teknologi MARA (UiTM), Malaysia and her undergraduate studies at University of Salford, United Kingdom. She is a senior lecturer in Faculty of Computer and Mathematical Sciences at UiTM Cawangan Kelantan, Malaysia where she has been a faculty member since 2004. She has been engaged to research works such as conferences, workshops and become a member of Business Datalytics Research Group. She is a certified trainer in data analyst from RapidMiner and data integration from Talend. Her earlier publications are on database technology, ontology construction and mapping and actively involves in research, consultations and publications. The author's primary interest is on data mining, data analytics, text and web mining. She can be contacted at email: syerina@uitm.edu.my.



Dr. Norshahida Shaadan    is a senior lecturer at the Center of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences UiTM Shah Alam. She has been working as a lecturer at UiTM for almost 23 years teaching various subjects such as Business Statistics, Statistical Methods, Probability and Statistics, Research Methodology, Quality Management and Analysis, Statistical Process Control and Statistics for Science and Engineering. She obtained her first degree in Mathematics and Statistics from University of Bradford, United Kingdom and a master degree and PhD in Statistics from the National University of Malaysia. Her areas of research interest are in air pollution modelling, climate change investigation, missing value imputation and outliers' detection whereby her field of expertise are functional data analysis and statistical methods. She can be contacted at email: shahida@fskm.uitm.edu.my.



Assoc. Prof Dr. Sapto Wahyu Indratno    is a lecturer in Faculty of Mathematics and Natural Sciences at Institut Teknologi Bandung, Indonesia. He is currently a research coordinator of University Center of Excellence on Artificial Intelligence for Vision, Natural Language Processing and Big Data Analytics (U-CoE AI-VLB) ITB. He had a basic mathematics degree from Bandung Institute of Technology, Master of Applied Mathematics from Twente University, The Netherlands and PhD in Applied Mathematics from Kansas State University, The USA. His area of research interest focuses on the modeling of risks, the development of statistical learning on symbolic data and machine learning on function spaces. His recent publications are on cyber insurance using graph mining approach and on classification method using family of distribution functions. He can be contacted at email: sapto@math.itb.ac.id.