Model optimisation of class imbalanced learning using ensemble classifier on over-sampling data

Yulia Ery Kurniawati, Yulius Denny Prabowo
Department of Informatics, Faculty of Computers Science and Design, Institut Teknologi dan Bisnis Kalbis, Jakarta, Indonesia

Article Info

Article history:

Received May 28, 2021
Revised Dec 23, 2021
Accepted Jan 4, 2022

Keywords:

Adaptive synthetic-nominal
class imbalance learning
Ensemble classifier
Over-sampling
Synthetic minority oversampling technique-nominal

ABSTRACT

Data imbalance is one of the problems in the application of machine learning and data mining. Often this data imbalance occurs in the most essential and needed case entities. Two approaches to overcome this problem are the data level approach and the algorithm approach. This study aims to get the best model using the pap smear dataset that combined data levels with an algorithmic approach to solve data imbalanced. The laboratory data mostly have few data and imbalance. Almost in every case, the minor entities are the most important and needed. Over-sampling as a data level approach used in this study is the synthetic minority oversampling technique-nominal (SMOTE-N) and adaptive synthetic-nominal (ADASYN-N) algorithms. The algorithm approach used in this study is the ensemble classifier using AdaBoost and bagging with the classification and regression tree (CART) as learner-based. The best model obtained from the experimental results in accuracy, precision, recall, and f-measure using ADASYN-N and AdaBoost-CART.

Corresponding Author:

Yulia Ery Kurniawati
Department of Informatics, Faculty of Computer Science and Design, Institut Teknologi dan Bisnis Kalbis
Jalan Pulomas Selatan Kav 22, Kayu Putih, Pulogadung, Jakarta Timur 13210, Indonesia
Email: yulia.kurniawati@kalbis.ac.id

1. INTRODUCTION

One of the problems of machine learning and data mining is imbalanced data. Imbalanced occurs when there is disproportion among the number of examples of each class in the dataset [1] and usually in the most essential and needed entities. It will be a complicated issue when dealing with the multiclass problem. It will be hard to acknowledge a priori of the multi-majority and multi-minority classes that should be stressed during the learning stage. For example, machine learning in data mining has difficulty classifying minority classes or classes with the smallest number of instances because the algorithm assumes that the class distribution is balanced. So that in some cases, there are errors in classifying the results for each class. The result is errors in the classification of minority classes due to the class imbalance that tends to focus on the majority class and ignore the minority class at the time of classification. The imbalanced data can be found in many areas such as medical [2], [3], abnormal electricity consumption [4], price forecasting [5], credit evaluation [6], and cyanobacteria bloom [7].

There are two approaches to solving this problem in dealing with class imbalance: the data level approach, the algorithmic approach, and hybrid-based approaches [8], [9]. The data-level approach can use the sampling method. This data sampling method is divided into two: the sampling method in the minority class (over-sampling) [10], [11], and the majority class sampling method (under-sampling) [12], [13]. Meanwhile, the algorithm approach is an approach by designing new algorithms or refining existing algorithms, and it uses the ensemble method. Ensemble methods use one set of classifiers to make a prediction. The generalisation ability of the ensemble is generally much stronger than the individual ensemble members [8]. There is two ensemble categorisation, parallel and sequential ensemble. The parallel ensemble obtains base learners in parallel, for example, bagging [6], [14]. In comparison, the sequential ensemble produces base learners sequentially, where the previous base learner influences the next generation of learners, for example, by using adaptive boosting (AdaBoost).

Kurniawati *et al.* on adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test dat proposed ADASYN-N in their study in 2018 [2]. It can handle nominal data types that ADASYN proposed by He *et al.* [11] cannot. This study used an over-sampling method to solve cases of class imbalance in the pap-smear result dataset [15]. The over-sampling methods used are synthetic minority oversampling technique-nominal (SMOTE-N), ADASYN-N, and ADASYN-KNN. The result is that ADASYN-N performed better than SMOTE-N on all performance matrices for NBC.

Fithrasari *et al.* on handling imbalance data in classification model with nominal predictors in 2020, studied handling imbalanced data in classification models with nominal predictors [16]. They used Survei Kinerja dan Akuntabilitas Kependudukan Keluarga Berencana dan Pembangunan Keluarga (SKAP KKBPK) data Jawa Timur Province in 2018. ADASYN-N, SMOTE-N, and SMOTE-N-ENN were used for imbalanced dataset handling then tested using classification and regression trees (CART). ADASYN-N gave the best average area under the curve (AUC) compared with SMOTE or synthetic minority oversampling technique-nominal edited nearest neighbor (SMOTE-N-ENN). It could increase accuracy from 0.737 to 0.963.

Rachburee and Punlumjeak on oversampling technique in student performance classification from engineering course, conducted a study to combine oversampling methods with several classifier models [17]. The oversampling methods that were used were SMOTE, Borderline-SMOTE, SVMSOTE, and ADASYN. The classifier models were applied using MLP, gradient boosting, AdaBoost, and random forest. The result was Borderline-SMOTE gave the best result among other models.

The absence of further research to find the best model based on the pap-smear result dataset [15], so this study will combine the over-sampling methods, which is a data-level approach with an algorithm approach. This algorithm approach used an ensemble classifier, AdaBoost.M1 and bagging, and based learner used CART. This study chose CART because CART and decision tree are unstable learning algorithms, and the ensemble method can improve the generalisation performance and accuracy of unstable learning algorithms [18].

2. METHOD

This study was how to optimise the model of imbalanced class learning using ensemble learning on over-sample data. Figure 1 shows the research flow in this research. The imbalance dataset was the pap smear results dataset [15]. Data level and algorithm approach used to optimise class imbalance handling that was. The first approach was data level one using oversampling. It uses SMOTE-N [10] and ADASYN-N [2] to handle nominal features. SMOTE-N can solve the overfitting problem in random oversampling [10]. While ADASYN, the algorithm was proposed by He *et al.* improve SMOTE to generate the synthesis instances based on the idea of adaptively generating minority data samples based on their distribution [11]. But, ADASYN can only compute the numerical data, so for this study, ADASYN-N [2] were used because the dataset is nominal. The second one is the algorithm approach. It used an ensemble classifier using CART combined with AdaBoost and Bagging. Thirty stages of 10-fold cross-validation were used to validate the model. It divided the dataset into tenfold, then one-fold will be data test, and the rest will be data training. It will repeat until all folds become testing data. The evaluation matrices were accuracy, precision, recall, f-score.

2.1. Dataset

The dataset used is a dataset of pap smear results conducted by Kurniawati *et al.* [15]. It was used because it has a huge difference between the minority and majority classes. There is no further research to improve the classifier's performance on this dataset using over-sampling and ensemble classifiers. The dataset contains 38 features: microscopic features of the anatomical pathology results of the Pap smear test and 75 instances divided into seven classes. Table 1 is the list of the seven classes in the dataset.

Figure 2 shows the number of instances from each class in the dataset. It showed that Chronic Inflammation had the most amount with 31 instances, and the Ca Cervix Suspect had the lowest amount with two ones. Thus, the ratio of the most amount and the lowest one was 31:2. The impact of the imbalance ratio for each class is poor performance when classifying the minority class.

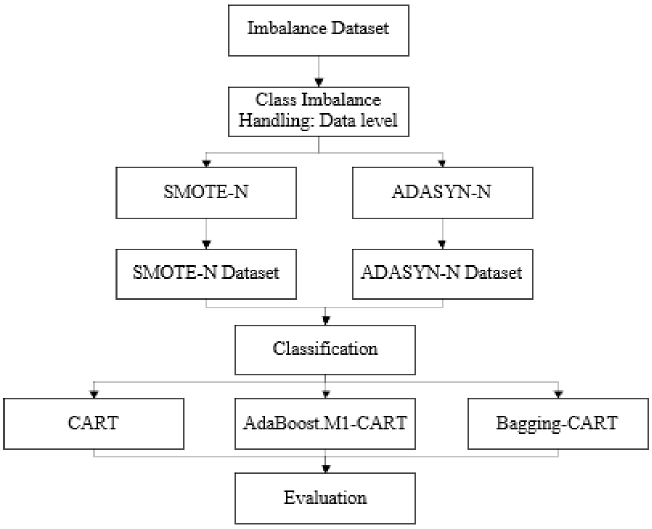

Figure 1. Research flow

Table 1. Classes

No	Class
1	Chronic Imflamation
2	Cervical Polyp
3	Epidermoid Carcinoma
4	Normal
5	Papillary Adenocarcinoma
6	Squamous Carcinoma
7	Ca Cervic Suspect

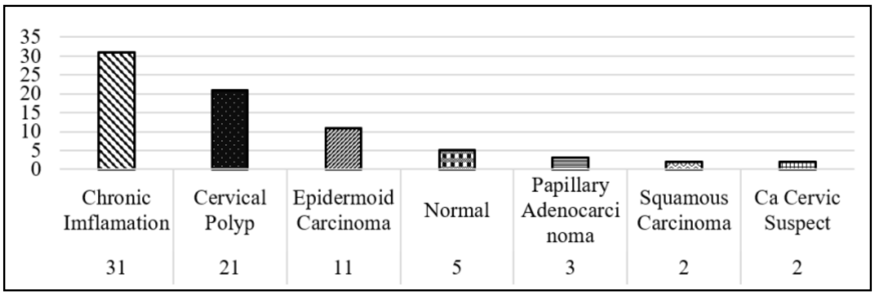

Figure 2. The instance number of each class

2.2. Class imbalanced handling

Class imbalance learning, also known as CIL, is learning with a class imbalance. The dataset has class imbalanced if it disproportionates the number of instances from each class in the dataset [19] or one class instance is higher than the other [20]. Datasets, where the most common class is less than twice the most minor class, will only be slightly unbalanced. In contrast, the dataset with an imbalance ratio of 10: 1 will be imbalanced, and the dataset with an imbalance ratio of 1000:1 will be very unbalanced [8].

The impact of imbalance is the learning and ability in rare classes. There are two aspects of the approach in dealing with imbalanced datasets, namely data level and algorithms [8], [21], [22]. The first approach to overcoming class imbalance is sampling the minority class (over-sampling). Over-sampling is a method of balancing class distribution by randomly replicating instances of minority classes. However, over-sampling increases the likelihood of overfitting occurring because it duplicates the instances exactly. In 2002, Chawla *et al*. [10] proposed a solution to deal with overfitting in the over-sampling method, namely SMOTE. SMOTE makes use of the nearest neighbours and the desired amount of over-sampling. Meanwhile, under-sampling is a method to balance the class by reducing instances in the majority class randomly. However, the under-sampling method has a disadvantage, namely the loss of information and data that is considered necessary for the decision-making process by machine learning. The second approach is an algorithm. One of the algorithm approaches is the ensemble method. The ensemble method uses a set of classifiers to make predictions. An ensemble's generalizability is generally more robust than individual ensemble members [8]. There are two categories of ensembles, namely parallel ensembles and sequential ensembles. Parallel ensembles produce parallel base learners, for example, bagging. Consecutive ensembles make base learners sequentially, whereas previous base learners have influenced subsequent learner generations, for instance, AdaBoost. In this study, data level and algorithm will use to handle the class imbalance problem: the data level and algorithm approaches. The data level approach will use an over-sampling method with SMOTE-N and ADASYN-N. The algorithm approach will use ensemble methods.

2.2.1. SMOTE-N

SMOTE-N is a development of SMOTE used for nominal features with nominal features proposed by Chawla as the development of SMOTE [10]. At SMOTE-N, the modified version value difference metric (VDM) was proposed by Cost and Salzberg. It was used to calculate the nearest neighbour. New set feature values can be created by taking the majority vote of the feature vector considering its k nearest neighbour to generate new minority class feature vectors.

2.2.2. ADASYN-N

ADASYN is a method for oversampling approach to learning with an unbalanced dataset proposed by He *et al.* [11]. The main idea of ADASYN is to use distribution weights for data on minority classes based on the level of learning difficulty. Synthesised data are generated from minority classes that are difficult to learn compared to minority data that are easier to learn. ADASYN enhances learning in two ways. First, it reduces the bias caused by class imbalances, and the second adaptively shifts the boundaries of classification decisions towards data difficulty. ADASYN-N is a development of ADASYN with a nominal type data approach called ADASYN-N developed by Kurniawati *et al.* [2]. The nearest neighbour in ADASYN-N was calculated using a modified version of VDM as in SMOTE.

2.2.3. Ensemble methods

The ensemble method trains base learners from the training data to make predictions and then combine them to make the final decision. In contrast, the standard machine learning method only produces one learner [8]. An ensemble can increase the learner with better performance than random guess into the learner with strong generalizability and very successful in many machine learning challenges for real-life applications [23].

The base learner is often referred to as the weak learner. It indicates that the base learner can have weak generalizability in the ensemble method. However, most learning algorithms, such as decision trees, neural networks, or other machine learning methods, can be called to train the base learner, and ensemble methods can improve performance [8].

The ensemble method can be categorised into parallel and sequentially based on how the base learner is generated. The parallel one produces the base learner in parallel, for example, Bagging. The sequentially one produces the base learner sequentially, where the base learner influences the next generation, for instance, AdaBoost.

Bagging is a method for generating multiple versions of a predictor and using predictors to get a combined predictor [24]. Bagging is a representation of a parallel ensemble method. Bagging should be used with an unstable learner, for example, a decision tree, because the more unstable the base learner is, the better its performance will be. However, it turned out that bagging resulted in a combined model that performed better than a single model built from the original training data and never substantially worse off [25]. Here is the Bagging pseudocode [26]:

Algorithm bagging for classification

```
**Input**: $S$: Training set; $T$: number of iterations; $n$: number of bootstrap; $I$: weak learner
**Output**: Bagged classifier: $H(x) = sign\left(\sum_{t=1}^{T} h_t(x)\right)$ where $h_t \in [-1,1]$ is induced classifier
for $t = 1$ to $T$ do
$S_t$ ←RandomSampleReplacement $(n, S)$
$h_t \leftarrow I(S_t)$
end for
```

AdaBoost is a machine learning technique to improve the performance of the weak learner. The method called the weak learner iteratively, the training data used is taken from several subsets of the entire database. A single robust classifier is then constructed by combining the resulting weak learner with the resampling training set [27]. There are many boosting variations, one of which is AdaBoost.M1, specially designed for classification. Adaboost.M1 is a simple generalisation of AdaBoost for more than two classes or multiclass [1], [27], which has the same algorithm as AdaBoost for the multiclass base instead of the binary learner. Here is the AdaBoost.M1 pseudocode [26]:

Algorithm Adaboost.M1

```
Input: Training set $S = \{x_i, y_i\}$, $i = 1, \dots, N$; dan $y_i \in \mathbb{C}, \mathbb{C} = \{c_1, \dots, c_m\}$; $T$: number of iterations; $I$: weak learner
Output: Boosted classifier:
$H(x) = \arg\max_{y \in \mathbb{C}} \sum_{t=1}^{T} ln\left(\frac{1}{\beta_t}\right) [h_t(x) = y]$ where $h_t, \beta_t$ is the induced classifier (with $h_t(x) \in \mathbb{C}$) and give weight to each
$D_1(i) \leftarrow \frac{1}{N}$ for $i = 1, \dots, N$
for $t = 1$ to $T$ do
    $h_t \leftarrow I(S, D_t)$
    $\varepsilon_t \leftarrow \sum_{i=1}^{N} D_t(i)[h_t(x_i) \neq y_i]$
    if $\varepsilon_t > 0{,}5$ then
        $T \leftarrow t - 1$
        return
    end if
    $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$
    $D_{t+1}(i) = D_t(i).\beta^{1-[h_t(x_i) \neq y_i]}$ for $i = 1, \dots, N$
    Normalise $D_{t+1}$ for the proper distribution
end for
```

2.2.4. CART

Classification and regression tree algorithm or CART is a regression tree and classification tree method that will produce a classification tree if it consists of categorical attributes and create a regression tree if it consists of continuous attributes [28]. CART will select several attributes and interactions between the most dominant attributes in determining the attributes' results depending on the binary sorting procedure. In choosing the best splitter, CART strives to maximise the average purity of the two child nodes. The way to measure purity can be selected freely, and it can be by the criteria of splitting or the splitting function. The most common splitting function is the Gini index. Gini index calculation as shown in (1):

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \tag{1}$$

where $P(i|t)$ is the relative frequency of class i at node t, and c is the number of classes. Thus, the calculation will get the highest if the distribution is from a uniform class and the smallest if it contains identical class records.

2.2.5. Evaluation

The confusion matrix is used to calculate the evaluation matrices such as accuracy, recall, precision, f-score, and receiver operating character (ROC) area as evaluation matrices. Table 2 shows the confusion matrix. TP is the condition where the classifier correctly classifies the positive result. Otherwise, TN is a condition where the classifier correctly classifies a negative result. Meanwhile, FP is a condition in which the classifier identifies a positive result as negative, and FN is a condition where the classifier identifies a negative result as positive.

Table 2. Confusion matrix

Actual class	Predicted class	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

The calculation of accuracy, recall, precision, and f-score using (2)-(5):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F - score = (1+\beta^2) \times \frac{presisi \times recall}{(\beta^2 \times presisi)+recall} \quad (5)$$

ROC area, commonly known as the AUC, is a technical standard for classifier evaluation. The wider the AUC area, the better the model and the excellent interpretation of the probability that the classifier ranks randomly selected positive instances over randomly selected negative instances [26]. Every curve on the ROC curve represents the performance of different classifiers in the dataset. The X-axis represents the false positive rates (FPR), and Y-axis represents the true positive rates (TPR). FPR and TPR calculate using (6) and (7):

$$FPR = \frac{FP}{(TN+FP)} \quad (6)$$

$$TPR = \frac{TP}{(TP+FN)} \quad (7)$$

3. RESULTS AND DISCUSSION

New datasets were generated using SMOTE-N and ADASYN-N then tested using CART, AdaBoost-CART, and Bagging-CART. Implementation with the classifier using 30-Stages 10-Cross Fold Validation. The evaluation matrices used were accuracy, recall, precision, f-score, and ROC area. Table 3 shows the algorithm configuration.

Table 3. Algorithm configuration

Algorithm	Configuration	
	Variable	Value
SMOTE-N	KNN	5
	%N	Adjusted to the number of instances
ADASYN-N	KNN	5
	d_{th}	0.75
	β	1

Both SMOTE-N and ADASYN-N used the five nearest neighbours. ADASYN-N used 0.75 for d_{th} or maximum tolerance level of imbalance class ratio and 1 for β or level balance. As for SMOTE-N, the value of the %N adjusted depends on the number of instances generated by ADASYN-N. Figure 3 shows each optimisation's performance from both datasets and all the classifications used in this study. The optimisation using data level or data level and algorithm could improve the accuracy from 89.34% to 96.39%. The imbalance dataset using CART had the lowest mean accuracy, equal to 77.87%. The best accuracy used ADASYN-N with AdaBoost-CART as the classifier obtained 96.39%. The mean accuracy from SMOTE-N and ADASYN-N datasets are better than the imbalanced dataset.

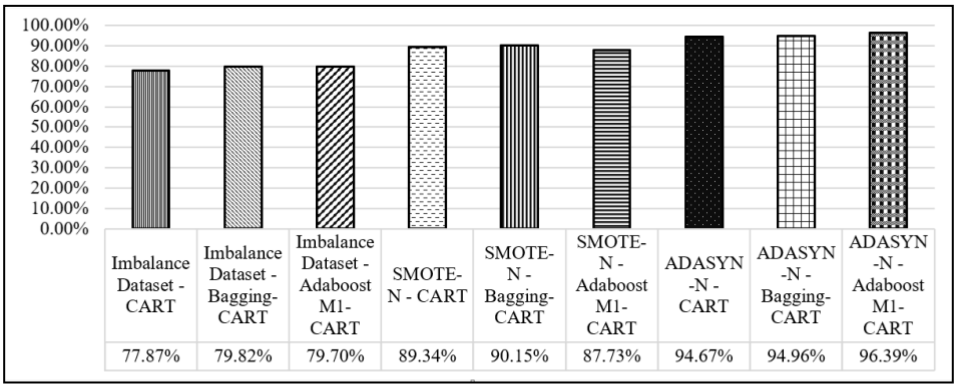

Figure 3. Mean accuracy

Table 4 shows all the performance matrices used in this study, such as precision, recall, f-measure, and ROC Area. ADASYN-N with AdaBoost-CART had the best performance matrices in precision, recall, and f-measure. However, the best ROC area was obtained by ADASYN-N with Bagging-CART. It was reasonable because the model obtained using Boosting is iterative. Therefore, the new model was affected by the previous model's performance. It encourages new models to become experts for instances that the previous model correctly handles by assigning a greater weight to their instances.

Table 4. Evaluation

Model	Precision	Recall	F-Measure	ROC Area
Imbalance Dataset - CART	67.16%	77.87%	71.88%	84.25%
Imbalance Dataset - Bagging-CART	72.29%	79.82%	75.45%	90.50%
Imbalance Dataset - AdaboostM1-CART	80.29%	79.70%	79.47%	91.67%
SMOTE-N - CART	90.05%	89.34%	89.19%	95.11%
SMOTE-N - Bagging-CART	90.71%	90.15%	90.01%	95.74%
SMOTE-N - AdaboostM1-CART	87.82%	87.73%	87.67%	95.83%
ADASYN-N - CART	95.13%	94.67%	94.42%	98.29%
ADASYN-N - Bagging-CART	95.45%	94.96%	94.72%	99.22%
ADASYN-N - AdaboostM1-CART	96.59%	96.39%	96.27%	98.56%

4. CONCLUSION

Laboratory test data, such as a dataset of pap smear results, most have little data and imbalance. Almost in every case, the least entities are the most important and needed. Based on this study's results, the optimisation of Class Imbalanced Learning using both data level and algorithm using ensemble classifier on over-sampling data could increase all the evaluation matrices performance on laboratory test data. ADASYN-N is better than SMOTE-N for over-sampling the dataset used in this study. The best model was obtained using ADASYN-N with AdaBoost-CART. Moreover, this study will use another based-learner besides CART to get the best model for imbalance laboratory test data.

REFERENCES

[1] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Cham: Springer International Publishing, 2018.

[2] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test data," in *2018 4th International Conference on Science and Technology (ICST),* Aug. 2018, pp. 1–6, doi: 10.1109/ICSTC.2018.8528679.

[3] L. Wang *et al.*, "Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods," *Computer Methods and Programs in Biomedicine*, vol. 196, Nov. 2020, doi: 10.1016/j.cmpb.2020.105567.

[4] H. Qin, H. Zhou, and J. Cao, "Imbalanced learning algorithm based intelligent abnormal electricity consumption detection," *Neurocomputing*, vol. 402, pp. 112–123, Aug. 2020, doi: 10.1016/j.neucom.2020.03.085.

[5] S. Hasmita, F. Nhita, D. Saepudin, and A. Aditsania, "Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling (ADASYN) and k-nearest neighbor (KNN) algorithms," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, Jul. 2019, pp. 434–438, doi: 10.1109/ICOIACT46704.2019.8938525.

[6] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Information Sciences*, vol. 425, pp. 76–91, Jan. 2018, doi: 10.1016/j.ins.2017.10.017.

[7] J. Shin, S. Yoon, Y. W. Kim, T. Kim, B. G. Go, and Y. K. Cha, "Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms," *Ecological Informatics*, vol. 61, Mar. 2021, doi: 10.1016/j.ecoinf.2020.101202.

[8] H. Ye and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*, 1st ed. Wiley-IEEE Press, 2013.

[9] B. S. Raghuwanshi and S. Shukla, "Class imbalance learning using UnderBagging based kernelised extreme learning machine," *Neurocomputing*, vol. 329, pp. 172–187, Feb. 2019, doi: 10.1016/j.neucom.2018.10.056.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[11] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[12] W. A. Luqyana, B. L. Ahmadie, and A. A. Supianto, "K-nearest neighbors undersampling as balancing data for cyber troll detection," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET),* Sep. 2019, pp. 322–325, doi: 10.1109/SIET48054.2019.8986079.

[13] Y.-S. Jeon and D.-J. Lim, "PSU: particle stacking undersampling method for highly imbalanced big data," *IEEE Access*, vol. 8, pp. 131920–131927, 2020, doi: 10.1109/ACCESS.2020.3009753.

[14] I. Fakhruzi, "An artificial neural network with bagging to address imbalance datasets on clinical prediction," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, Mar. 2018, pp. 895–898, doi: 10.1109/ICOIACT.2018.8350824.

[15] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results," in *2016 1st International Conference on Biomedical Engineering (IBIOMED)*, Oct.

2016, pp. 1–5, doi: 10.1109/IBIOMED.2016.7869827.

[16] K. Fithriasari, I. Hariastuti, and K. S. Wening, "Handling imbalance data in classification model with nominal predictors," *International Journal of Computing Science and Applied Mathematics*, vol. 6, no. 1, Feb. 2020, doi: 10.12962/j24775401.v6i1.6643.

[17] 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, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3567-3574.

[18] T. Bi, P. Li, J. Huang, and K. Zhang, "Imbalance data classification method based on cluster boundary sampling RF-bagging," in *International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014*, 2014, vol. 2014, no. CP660, pp. 305–311, doi: 10.1049/cp.2014.1580.

[19] V. H. Barella, L. P. F. Garcia, M. C. P. de Souto, A. C. Lorena, and A. C. P. L. F. de Carvalho, "Assessing the data complexity of imbalanced datasets," *Information Sciences*, vol. 553, pp. 83–109, Apr. 2021, doi: 10.1016/j.ins.2020.12.006.

[20] A. Mahmoud, A. El-Kilany, F. Ali, and S. Mazen, "TGT: a novel adversarial guided oversampling technique for handling imbalanced datasets," *Egyptian Informatics Journal*, vol. 22, no. 4, pp. 433–438, Dec. 2021, doi: 10.1016/j.eij.2021.01.002.

[21] K. Li, W. Zhang, Q. Lu, and X. Fang, "An improved SMOTE imbalanced data classification method based on support degree," in *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, Oct. 2014, pp. 34–38, doi: 10.1109/IIKI.2014.14.

[22] F. Koto, "SMOTE-out, SMOTE-cosine, and selected-SMOTE: an enhancement strategy to handle imbalance in data level," in *2014 International Conference on Advanced Computer Science and Information System*, Oct. 2014, pp. 280–284, doi: 10.1109/ICACSIS.2014.7065849.

[23] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion*, vol. 64, pp. 205–237, Dec. 2020, doi: 10.1016/j.inffus.2020.07.007.

[24] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[25] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques*. Elsevier, 2011.

[26] 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 approachess," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: 10.1109/TSMCC.2011.2161285.

[27] O. Herouane, L. Moumoun, and T. Gadi, "Using bagging and boosting algorithms for 3D object labeling," in *2016 7th International Conference on Information and Communication Systems (ICICS)*, Apr. 2016, pp. 310–315, doi: 10.1109/IACS.2016.7476070.

[28] M. T. M. K. Sabariah, S. T. A. Hanifa, and M. T. S. Sa'adah, "Early detection of type II diabetes mellitus with random forest and classification and regression tree (CART)," in *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Aug. 2014, pp. 238–242, doi: 10.1109/ICAICTA.2014.7005947.

BIOGRAPHIES OF AUTHORS

Yulia Ery Kurniawati is currently a lecturer at Informatics Department at Institut Teknologi dan Bisnis Kalbis (Kalbis Institute) Jakarta. She obtained her bachelor degree in informatics at Universitas Sebelas Maret and master degree in information technology from Universitas Gadjah Mada. Her research interest in artificial intelligence focuses on machine learning and class imbalance learning. She can be contacted at email: yulia.kurniawati@kalbis.ac.id.

Yulius Denny Prabowo is a lecturer at the Informatics study program at Institut Teknologi dan Bisnis Kalbis. He obtained a bachelor's degree from Atmajaya University Yogyakarta, a master's degree from the University of Indonesia and is currently pursuing doctoral education. His research focuses on a cross-section of artificial intelligence, machine learning, cognitive intelligence, and the Indonesian language. He can be contacted at email: yulius.prabowo@kalbis.ac.id