Proceedings of Machine Learning Research vol 245:1–10, 2024 2024 International Conference on Machine Learning and Intelligent Computing

# Research on Imbalanced Classification Problem Based on Optimal Random Forest Algorithm

**Shan Yue*** 1109793282@QQ.COM
*School of Computer Science and Technology, Kashi University, kashi, China*

**Hui Liu**
*College of Medical Engineering Technology, Xinjiang Medical University, Urumqi, China*

**Zheng He**
*Southern Xinjiang Military Region, Shule,China*

**Qiaoling Yong**
*School of Computer Science and Technology, Kashi University, kashi, China*

**Yali Wang**
*School of Computer Science and Technology, Kashi University, kashi, China*

**Editors:** Nianyin Zeng and Ram Bilas Pachori

## Abstract

In order to solve the binary classification problem of imbalanced data, an optimal random forest algorithm GWORF (Grey Wolf Optimizer Random Forest) is pro-posed. The algorithm first uses BLSMOTE (BorderLine SMOTE) technology to oversample the imbalanced data set to make the positive and negative data equivalent, and then uses the Grey Wolf optimization algorithm to calculate the optimal parameters, and then puts the calculated optimal parameters into the forest for modeling training. Through testing on four imbalanced data sets, the effectiveness of the GWORF algorithm in the study of imbalanced binary classification problems is verified.

**Keywords:** Imbalanced; Binary classification; Oversampling; GWORF

## 1. Introduction

In the study of classification problems, there is such a kind of data that cannot be ignored. Some types of data account for a large proportion, sometimes even more than 90%, while some types of data account for a very small proportion. However, it is these small proportions The data plays a decisive role, and once the classification error is made, it will cause serious losses. For example, when a financial institution makes credit predictions, if it predicts an honest user in the past as a dishonest user, the result is only a loss of one customer. Conversely, if it predicts an untrustworthy user as an honest user and lends to this person, the institution may face relatively large property losses. For another example, the above-mentioned problems exist in the medical field. When a patient's physical function is monitored, most of the time is normal, and abnormal data of only a few seconds may cause major trauma to the patient's body.

This kind of problem is called quasi-non-equilibrium problem. Classes with large amounts of data in them are called negative classes, and classes with small amounts of data are called positive classes (Das, 2024).

In the field of data mining and artificial intelligence, there are many researches on classification problems under the condition of imbalanced distribution of data instances (Das, 2024), such as fault detection, anomaly detection, spam filtering, image recognition, etc. (Hao and Liu, 2020; Lee et al.,

2024; Zhao et al., 2020). Highly imbalanced data is considered difficult to learn in the field of machine learning (Das, 2024).

For the classification problem of the above-mentioned feature data, the traditional machine learning method that only considers the accuracy rate often does not obtain the optimal solution (Guo et al., 2022). Even if all the data are simply classified as negative classes without any processing, the accuracy rate of the classification algorithm can reach at least half. If the positive class data only accounts for 1%, then the accuracy rate of the classification algorithm will be as high as 99%. It can be seen that the performance of classification algorithms that only consider the accuracy rate is not ideal on imbalanced data sets. At present, the classification optimization of imbalanced data mainly focuses on data level, algorithm level and mixed methods (Al-Stouhi and Reddy, 2016).

## 2. Definition and Evaluation Index of "Imbalanced Data"

### 2.1. Problem Definition

When studying binary classification problems, when the number of available data samples is sufficient and the degree of disequilibrium is high, the data set is classified as an imbalanced data set, The ratio of the number of negative class samples to the number of positive class samples is defined as the degree of disequilibrium. Only when the importance of positive class samples is similar to or greater than that of negative class samples, the data set is regarded as having the problem of "imbalanced data".

### 2.2. Evaluation index

As mentioned above, if the accuracy rate is simply used as the only reference value, there is no research value. The reason is that the proportion of certain types of data is small, and sometimes there are very few cases, such as most users in the banking system data can repay on time, and a very small part has overdue repayment. What we need to do is to mark this type of users who will repay overdue in advance. Therefore, for the imbalanced classification problem, not only the accuracy, but also the precision and recall should be considered.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \tag{1}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

Among them, TP is True-positive, which means that it is True, and the number of samples whose prediction result is positive; FP is False-positive, which means that it is False, but the prediction result is positive; TN is True-negative, which represents True itself, and the number of samples predicted as negative, FN is False-negative, which represents the number of samples that are False and whose prediction result is negative.

At present, there are many reference models for the research of imbalanced classification problems, such as: F1 value, G-means value and AUC value, etc. In this paper, F-measure value is used as the evaluation of the performance of the algorithm, and its calculation formula is as follows:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

## 3. Random forest algorithm

The random forest algorithm is a Bagging ensemble learning algorithm (Kim et al., 2021), which integrates multiple decision trees to classify the samples to be tested. The algorithm evenly extracts multiple sets of samples from a given training set, and each set of samples builds a decision tree, and then forms a forest, and votes the results generated by each decision tree in the forest, and the category with the most votes is the final classification result.

The basic structure of the random forest algorithm is shown in Figure 1.

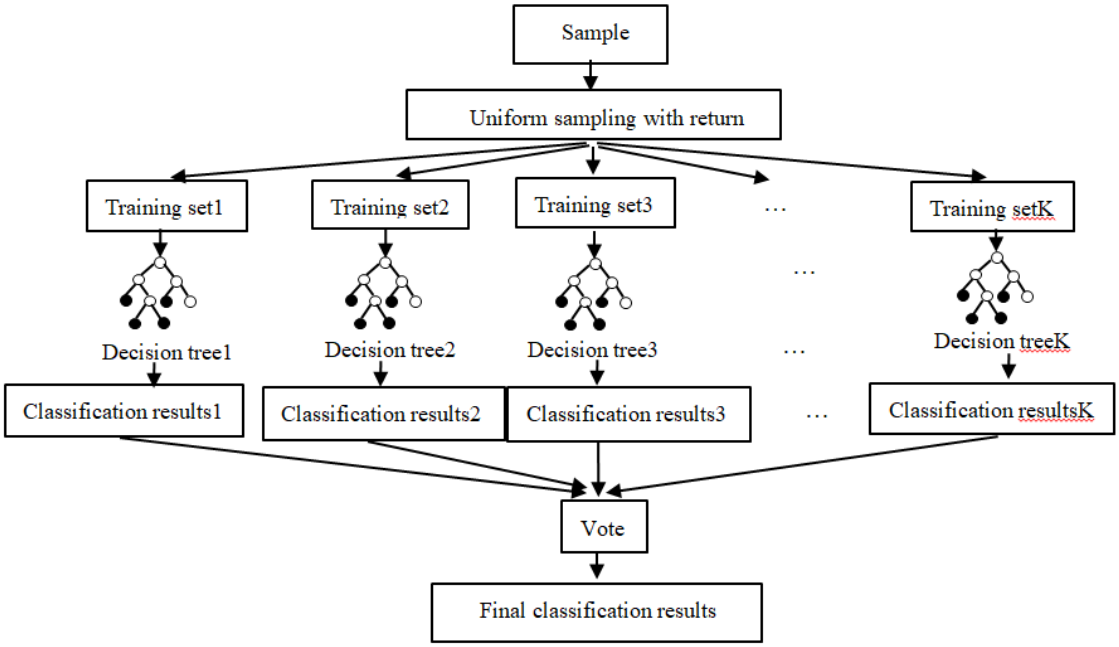


Figure 1: Basic structure of random forest algorithm.

In this paper, four classical non-equilibrium data sets are selected for binary classification research. In order to prevent overfitting, this experiment selects 50% of all datasets as the training set.

When using the random forest algorithm to model and classify on the four datasets eciol, zoo, yeast, and winequality and view the performance, it is found that the accuracy rate of the four datasets is high. The main reason for the analysis is that in the imbalanced data Set, the data allocation is unreasonable, with less positive type data and the majority of negative type data. Random forest algorithm is not good for positive data classification, and there will be misclassification. However, due to the relatively small amount of such data, positive data classification errors will not affect the accuracy of the algorithm to a large extent.

Considering the accuracy rate, precision rate, recall rate and F1 value for comparison, as shown in Figure 2.

It is not difficult to find that the accuracy rate of the four data sets is above 0.9, and the accuracy rate and accuracy rate of the eciol data set are almost the same, and the recall rate is relatively high. The zoo dataset has 0 precision and recall, which is the worst performance. The precision of yeast data set is about 0.75, but the recall rate is low, less than 0.3. The precision rate of the winequality

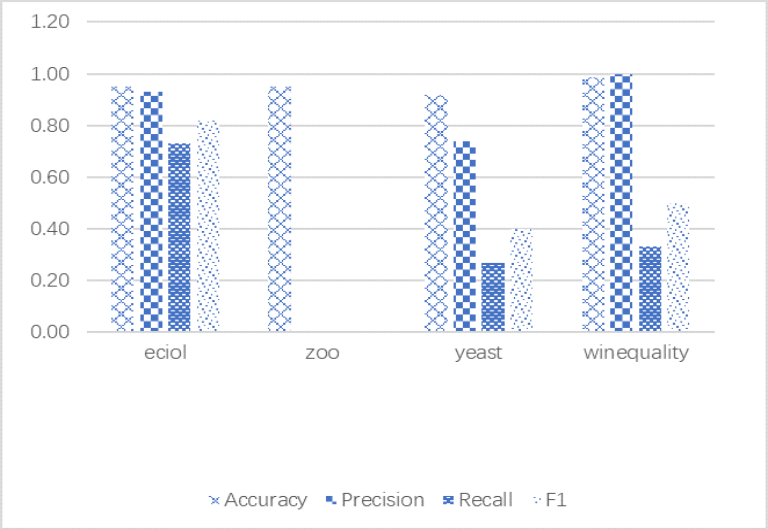


Figure 2: Accuracy, Precision, Recall and F1 Values on Random Forest Models.

data set is higher than the accuracy rate, close to 1, while the recall rate is about 0.3. Among the four datasets, only eciol performed better at around 0.8, the F1 values of the other three datasets did not exceed 0.5, and the F1 values of the zoo dataset were even 0. It is not difficult to find that the simple use of random forest algorithm for classification on imbalanced data sets is very unsatisfactory.

## 4. Optimize the structure of GWORF

### 4.1. Data oversampling

The BorderLineSMOTE algorithm (Han et al., 2005) is a method specially proposed by HAN H et al. in 2005 for oversampling of imbalanced datasets.

Algorithm description:

(1) The Euclidean distance between each minority class sample and the nearest neighbor sample is calculated, and the distance list is obtained by sorting them.

(2) Traversing each minority class sample, for each sample, determine whether it is located at the decision boundary. If most of its k nearest neighbor samples belong to the same class (majority class), the sample is considered to be on the decision boundary.

(3) For the samples located on the decision boundary, select one of the nearest neighbor samples, and calculate the difference vector between the two samples.

(4) Based on the difference vector and a random number between 0 and 1, a new composite sample is generated. The generation method can be linear interpolation or random interpolation.

(5) Repeat steps (3) and (4) , until a sufficient number of synthetic samples are generated.

### 4.2. Optimal parameter calculation

According to the way gray wolves hunt, the wolves were divided into 4 grades: $\alpha\beta\delta\omega$. The whole predation process is led by $\alpha$. The predation process is divided into: stalking and approaching the

prey, harassing, chasing and surrounding the prey until the prey stops moving, and finally attacking the prey (Elewi et al., 2024; Jagadeesh et al., 2023; Tian and Wang, 2023).

That is, the solutions of $\alpha\beta\delta\omega$ are obtained respectively, and these four solutions have a priority order from left to right, the hunting process is guided by $\alpha\beta\delta$, and $\omega$ follows the three wolves. That is, always go and find the three best solutions, then search around the area to find a better solution and then update $\alpha\beta\delta$.

**(1) Surround prey:**

Wolf pack rounding up prey is defined as follows:

The distance formula between the individual and prey:

$$\mathrm{D} = |\mathrm{C} \cdot X_p(\mathrm{t}) - X(\mathrm{t})| \tag{5}$$

Grey Wolf Location Update Formula:

$$X(\mathrm{t}+1) = X_p(t) - A \cdot D \tag{6}$$

Vector of coefficients:

$$A = 2a \cdot r1 \tag{7}$$

$$C = 2 \cdot r2 \tag{8}$$

Where: $t$ is the number of iterations, $D$ is the distance vector between the individual and the hunter, $X_p$ is the prey position vector, $X$ is the gray wolf position vector, $a$ is the convergence factor, $r1$ and $r2$ are random vectors.

**(2) Hunting:**

When the gray wolf identifies the prey, it encircles the prey under the leadership of the three leaders. Where an individual gray wolf tracks prey is defined as follows:

$$\begin{aligned} D_\alpha &= |C_1 \cdot X_\alpha - X| \\ D_\beta &= |C_2 \cdot X_\beta - X| \\ D_\delta &= |C_3 \cdot X_\delta - X| \end{aligned} \tag{9}$$

$D_\alpha$, $D_\beta$ and $D_\delta$ represent the distance between the three leading wolves and the other individuals, $X_\alpha$, $X_\beta$ and $X_\delta$ represent the current positions of the three leading wolves, $C_1$, $C_2$ and $C_3$ are random vectors, and $X$ is the current position of the wolf.

$$\begin{aligned} X_1 &= X_\alpha - A_1 \cdot (D_\alpha) \\ X_2 &= X_\beta - A_2 \cdot (D_\beta) \\ X_3 &= X_\delta - A_3 \cdot (D_\delta) \end{aligned} \tag{10}$$

$$X_{t+1} = \frac{X_1 + X_2 + X_3}{3} \tag{11}$$

$X_1$, $X_2$ and $X_3$ denote the step size and direction that $\omega$ advances towards $\alpha$, $\beta$, $\delta$, respectively, and $X_{t+1}$ denotes the final position of $\omega$.

**(3) Attack prey:**

When the movement of the prey stops, the gray wolf can carry out an attack.

### 4.3. Building an Optimal Forest

The specific algorithm process is as follows:

(1) Input N groups of sample data, each group of data has $M$ attributes, $(X_{11}, X_{12}, X_{13}, \ldots, X_{1m})$, $(X_{21}, X_{22}, X_{23}, \ldots, X_{2m})$,... $(X_{n1}, X_{n2}, X_{n3}, \ldots, X_{nm})$, the classification labels of the samples are $Y_1, Y_2$.

(2) Determine the training set and the test set.

(3) A random forest model was constructed with optimal parameters and trained. (L decision trees, the classification result of each tree for the t-th sample is: $C_l = f_l(X_{t1}, X_{t2}, \ldots, X_{tm})$, the final classification result is: $C = \sum_{l=1}^{L} f_l(X_{t1}, X_{t2}, \ldots, X_{tm})$, The model needs to solve L $f_l[X_{t1}, X_{t2}, \ldots, X_{tm}]$).

(4) Test the model with a test set.

(5) Output the classification results corresponding to the samples to be classified.

### 4.4. Algorithm description

Input: Imbalanced dataset
Output: Accuracy, Precision, Recall and F1 values
Oversampling datasets with BLSMOT
Initialize $\alpha\beta\delta$ position
Initialize the value of the $\alpha\beta\delta$ objective function
Iterative solution:
    Iterate each wolf position
    Build a random forest model and train it
    Take the cross-validation mean to minimize the error rate
    The loop updates the location of grey wolf individuals:
    $\alpha$ Position Update
    $\beta$ Position Update
    $\delta$ Position Update
    Other Grey Wolf Location Updates
Constructing Random Forest Classification Model with Optimal Parameter Values
Perform data classification

## 5. Experimental evaluation and analysis

### 5.1. Experimental dataset

In order to verify the performance of the optimized random forest algorithm, this paper investigates KEEL (Alcala-Fdez et al., 2011), a commonly used public dataset in the field of data mining and machine learning. The KEEL dataset has a special module for storing imbalanced datasets, and the imbalanced degree of these datasets varies from 1.5 to 129.44. In this paper, four data sets are selected from eciol, yeast, zoo and winequality. The disequilibrium degrees of these four data sets range from Guinea Bissau to Dozens to One, and the disequilibrium degrees are quite different. The number of negative and positive data in each dataset is shown in Table 1.

Table 1: Experimental dataset

| Dataset | Positive | Negative |
|---|---|---|
| eciol | 52 | 284 |
| yeast | 51 | 477 |
| zoo | 5 | 96 |
| winequality | 18 | 837 |

### 5.2. Result analysis

Experiment 1 Data Disequilibrium Contrast

The BLSMOTE algorithm is used to oversample the small class of samples to make the data reach an equilibrium state.

Figure 3 shows the original imbalanced degree and the current imbalanced degree of the four imbalanced data sets. Before oversampling techniques, the disequilibrium of the eciol dataset was 5.46, the disequilibrium of the yeast dataset was 9.35, the disequilibrium of the zoo dataset was 19.2, and the disequilibrium of the winequality dataset was 46.5. After BLSMOTE oversampling technology, the non-equilibrium degree of the four data sets reaches 1, that is, the number of large-class samples and small-class samples is the same.

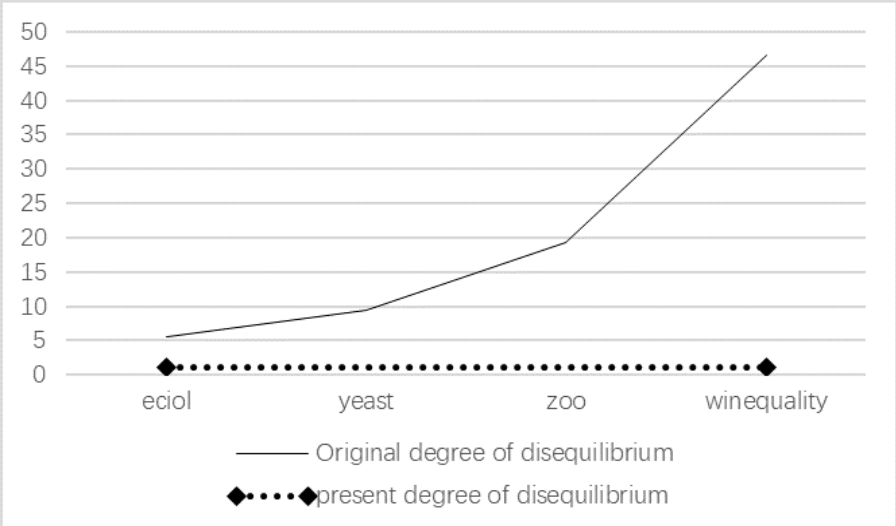


Figure 3: Comparison of data disequilibrium.

Experiment 2 Performance of GWORF algorithm

Use the balanced data set in Experiment 1 to select the optimal parameters, and obtain the maximum depth of each tree in the GWORF algorithm and the minimum number of samples contained by leaf nodes, as shown in Figure 4.

The comparison of algorithm performance is performed, as shown in Figure 5.

It can be found that the performance of GWORF algorithm on the four data sets is better. Whether it is accuracy, precision, recall or F1 value, most of them can reach more than 0.95.

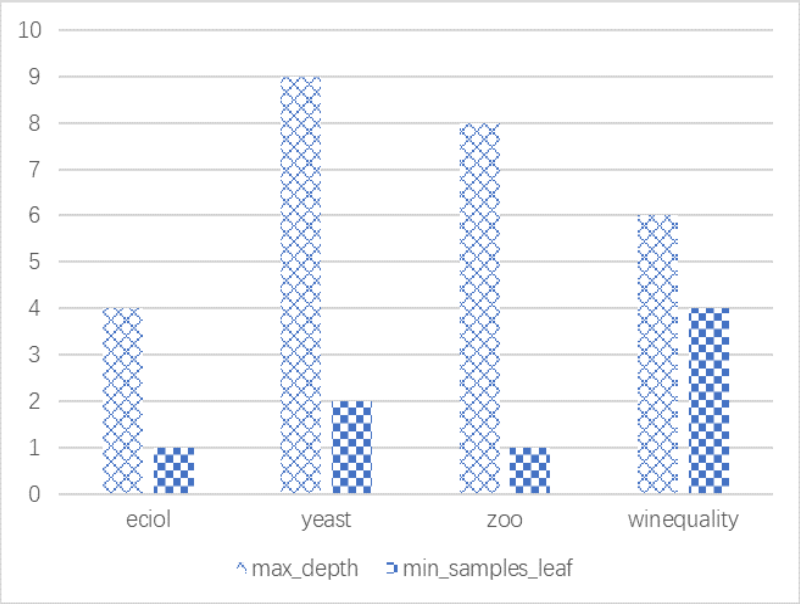


Figure 4: Maximum depth and minimum number of samples contained by leaf nodes.

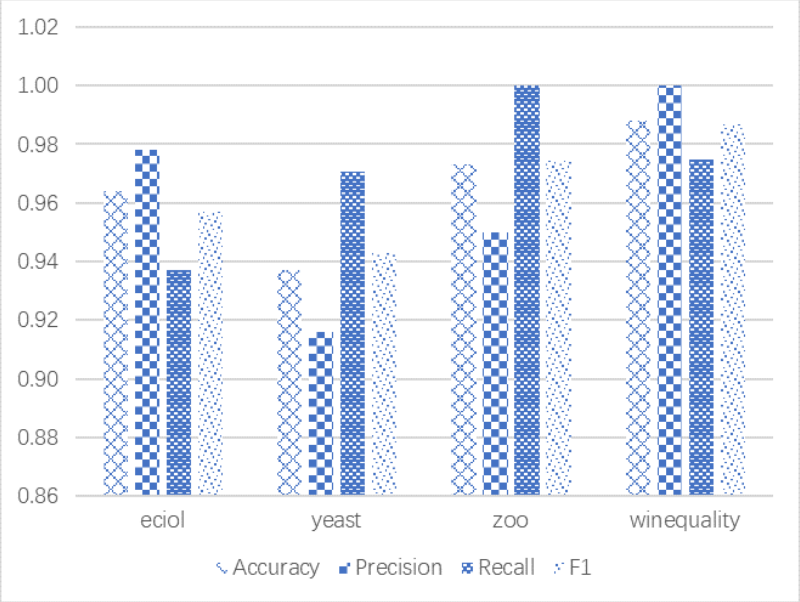


Figure 5: Performance comparison of GWORF algorithm in different datasets.

## 6. Conclusion

In this paper, the optimal stochastic forest algorithm is used to study the binary classification problem of divided and balanced data. The oversampling technology is used to preprocess the data first, and then the optimal forest is constructed by calculating the optimal parameters. Finally, through performance comparison on four classical imbalanced data sets, it is found that the algorithm performs well on imbalanced data classification problems.

The next work will focus on the performance optimization of each tree and the optimization of undersampling technology for a large number of data objects. The optimization algorithm proposed in this paper can still improve the performance of each tree.

## Acknowledgments

Fund Project: Kashi University Intramural Project: Research on Imbalanced Classification Problem Based on Optimized Random Forest Algorithm ((2021) 2742) Kashi University Teaching Research and Teaching Reform Project: Research on Innovation of Applied Talent Training Model from the Perspective of Undergraduate Education-Taking "Python Language Programming" Course as an example (KJBY2201).

## References

Samir Al-Stouhi and Chandan K. Reddy. Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems*, 48(1):201–228, 2016. doi: 10.1007/s10115-015-0870-3.

J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17(2-3):255–287, 2011.

Sufal Das. A new technique for classification method with imbalanced training data. *International Journal of Information Technology*, 16(4):2177–2185, 2024. doi: 10.1007/s41870-024-01740-1.

Abdullah Elewi, Semih Kahveci, and Erdinç Avaroğlu. Image contrast enhancement using a low-discrepancy population initialized gray wolf optimization algorithm. *Multimedia Tools and Applications*, 83(17):50307–50328, 2024. doi: 10.1007/s11042-023-17366-7.

Yinan Guo, Jiawei Feng, Botao Jiao, Ning Cui, Shengxiang Yang, and Zekuan Yu. A dual evolutionary bagging for class imbalance learning. *Expert Syst. Appl.*, 206(C), nov 2022. doi: 10.1016/j.eswa.2022.117843.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Wei Hao and Feng Liu. Imbalanced data fault diagnosis based on an evolutionary online sequential extreme learning machine. *Symmetry*, 12(8), 2020. doi: 10.3390/sym12081204.

G. Jagadeesh, J. Gitanjali, J. Vellingiri, M. Pounambal, E. Sathiyamoorthy, and Celestine Iwendi. Mdrogwl: modified deep reinforcement oppositional wolf learning for group key management in iot environment. *J. Supercomput.*, 80(8):10223–10254, dec 2023. doi: 10.1007/s11227-023-05809-9.

Byung-Chul Kim, Jingyu Kim, Ilhan Lim, Dong Ho Kim, Sang Moo Lim, and Sang-Keun Woo. Machine learning model for lymph node metastasis prediction in breast cancer using random forest algorithm and mitochondrial metabolism hub genes. *Applied Sciences*, 11(7), 2021. doi: 10.3390/app11072897.

Sung-Jae Lee, Hyun Jun Oh, Young-Don Son, Jong-Hoon Kim, Ik-Jae Kwon, Bongju Kim, Jong-Ho Lee, and Hang-Keun Kim. Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning. *BMC Oral Health*, 24 (1):161, 2024. doi: 10.1186/s12903-024-03898-3.

Hailin Tian and Fang Wang. Application of gray wolf optimization algorithm in urban electricity load forecasting model. *Journal of Physics: Conference Series*, 2592, 2023.

Chensu Zhao, Yang Xin, Xuefeng Li, Yixian Yang, and Yuling Chen. A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. *Applied Sciences*, 10(3), 2020. doi: 10.3390/app10030936.