Analysis Of One-Class and Binary Classification For COVID-19 Image Data

Huihan Cui *

Department of Computer Science, University College London, London, United Kingdom

* Corresponding Author Email: zcabhcu@ucl.ac.uk

Abstract. With the recent Coronavirus Disease 2019 (COVID-19) outbreak, there has been a growing interest in the utilization of machine learning models to aid in medical diagnostics. However, during the initial phases of the pandemic, the available datasets were severely imbalanced due to a scarcity of positive cases. These datasets predominantly comprised normal cases, with only a limited number of disease instances. This inherent data imbalance restricted the effectiveness of binary classification, prompting an exploration of one-class classification. This study aims to compare the performance of one-class and binary classification models during the early stages of COVID-19 development. The objective is to provide healthcare professionals with the most effective machine learning-assisted solutions for handling future outbreaks of a similar nature. The experiment incorporates three distinct models: One-Class Support Vector Machine (OCSVM) for one-class classification, alongside Categorical Boosting (CatBoost) and Extreme Gradient Boosting (XGBoost) for binary classification. These models undergo a rigorous tuning process employing grid search, using training data composed of features extracted from compressed chest X-ray images. In summary, the research findings unequivocally demonstrate the superior performance of one-class classification over binary classification in all early stages, achieving an impressive F1 score of 0.46, compared to binary classification's modest score of 0.063. Consequently, it is strongly recommended to employ a one-class classification model to enhance real-life diagnostic processes. However, it is anticipated that as more positive cases are incorporated into the dataset, achieving a more balanced distribution, binary classification may exhibit greater potential. Thus, medical practitioners should reconsider the applicability of binary classification in later stages when the dataset attains improved balance.

Keywords: Machine learning; imbalanced data; one-class classification; binary classification.

1. Introduction

In the field of medical science, it is widely recognized that acquiring sufficient data for rare or novel diseases poses a significant challenge. Consequently, classification algorithms used for diagnostic assistance often grapple with imbalanced training data [1]. For instance, during the initial outbreak of the COVID-19 pandemic, the available chest X-ray dataset predominantly consisted of samples from individuals without the disease, labeled as the negative class. Conversely, there was a scarcity or even an absence of images from individuals diagnosed as positive for the disease. Despite over twenty years of dedicated research addressing imbalanced data, the field continues to face numerous unresolved challenges [2]. Conventional binary classification algorithms frequently exhibit subpar performance and exhibit biases toward the dominant class [3]. In such scenarios, the adoption of one-class classification becomes prevalent. This technique exclusively trains on data points from a single class with the aim of constructing a model capable of distinguishing between the two classes. For example, through training solely with legitimate Twitter accounts, one-class classification can identify novel malicious accounts [4]. This research primarily focuses on a meticulous comparison of the performance between one-class and binary classification models in the context of COVID-19 detection. A study by Long Gao et al. highlights the effectiveness of the one-class support vector machine (OCSVM), particularly showcasing promising results in datasets related to kidney lesions and breast cancer [5]. On the other hand, insights from Joffrey L. Leevy et al. assert that binary classification demonstrates enhanced proficiency in detecting instances of fraud, with CatBoost emerging as the most optimal classifier for their study [6]. These contrasting findings emphasize the

variability in classifier performance across diverse datasets. In light of this, our research aims to determine the highest-scoring classifier for COVID-19 chest X-ray image data, thereby facilitating its prospective application in diagnosing future pandemics similar to COVID-19. The study employs three distinct classifiers: OCSVM for one-class classification, CatBoost, and Extreme Gradient Boosting (XGBoost) for binary classification. The former two have been mentioned in the aforementioned studies, while the potent capabilities of XGBoost can be further harnessed to effectively address challenges posed by imbalanced data [7]. Furthermore, the research involves manipulating the training data point ratio within the two classes to simulate various stages of pandemic progression. This approach is subsequently employed for comparative evaluations across different ratios. These analyses provide more detailed and context-specific recommendations for practical applications.

2. Related Work

Prior research efforts have conducted in-depth comparative analyses of one-class and binary classification methodologies. Notably, the investigation by Xueqing Deng et al. comprehensively compares the performance of five distinct classifiers, including models like presence and background learning and maximum likelihood classifiers, which are not included in this study [8]. However, this study considers other powerful boosting models and holds increased significance in the context of recent medical concerns, thereby enhancing its potential implications. Further exploration encompasses additional studies centered on enhancing and comparing models within the context of imbalanced data. Laura Schlieker and her colleagues present a novel sampling method to rectify imbalances within datasets. Their study evaluates the behavior of powered partial least squares discriminant analysis and random forests [9]. Utilizing datasets related to Systemic Lupus Erythematosus and Rheumatoid Arthritis, the authors propose that cost-sensitive learning is an effective method to improve the performance of random forests. Another perspective, provided by Marwa Chabbouh et al., highlights that the greedy algorithm of oblique decision trees (ODT) fails to address the issue of imbalanced data [10]. To overcome this challenge, they advocate using an evolutionary optimization method to refine the ODT.

The study conducted by Long Gao et al. identifies a bottleneck in the effectiveness of conventional feature mapping techniques due to the complexity of clinical images [11]. Their solution involves recommending deep learning models, which demonstrate superior proficiency in handling the complexity of image data. The validity of their findings is substantiated through experiments on four datasets. Concurrently, the research conducted by Mohammad Sabokrou et al. ventures down a similar path, focusing on the innovative integration of deep models [12]. They introduce two neural networks for adversarial training. One network act as a discriminator, while its counterpart generates adversarial examples during the training phase and cooperates during the testing process. Regarding the establishment of effective evaluation criteria, Jafar Tanha et al. propose that the Mean Matthews Correlation Coefficient is better suited for imbalanced data than the Mean Area Under Curve or G-mean [13]. In this study, the accuracy score serves as the evaluation metric, and the advanced and complex models listed above are not included.

3. Methodology

The methodology of this study is thoughtfully organized into three interrelated phases, each serving a pivotal role in the comprehensive analysis. In the initial phase, the research focuses on data processing, which lays the foundation for subsequent stages. Subsequently, models are meticulously constructed and fine-tuned to optimize the most effective hyperparameters. Lastly, the outcomes are assessed using three different evaluation metrics. These intricately woven phases culminate in a cohesive research endeavor, illustrated in Fig.1.

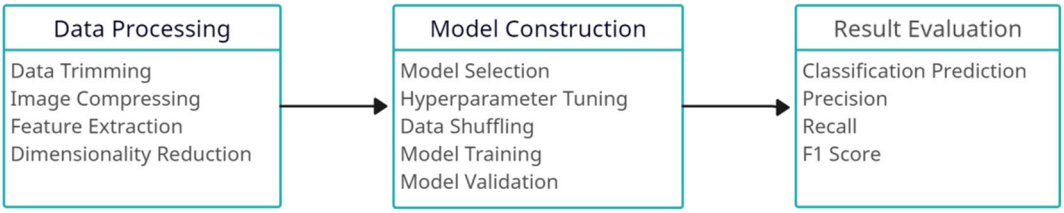

Fig. 1 The flow chart of this study

3.1. Data Processing

This study's dataset includes COVID-19 chest X-ray images from Kaggle, with an example shown in Fig. 2. Each image is labeled with class and source. To simplify, source labels are removed. The dataset contains various formats and sizes (1024x1024 to 439x391 pixels). All are standardized to 128x128 pixels, ensuring uniformity, aiding comparisons, and reducing memory load. This resizing enhances computational efficiency and speeds up machine learning tasks. After resizing, raw pixel data is processed using the Histogram of Oriented Gradients (HOG) technique, generating large 15876-dimensional arrays. To manage this dimensionality and prevent overfitting, Principal Component Analysis (PCA) from Scikit Learn reduces the feature set to 50 key components.

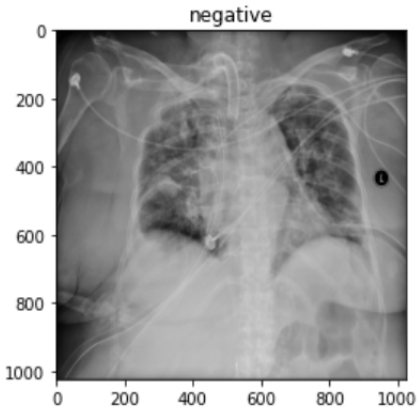

Fig. 2 Sample X-ray image

3.2. Model Principles and Optimization

This study incorporates three distinct models. The OCSVM serves as the model for one-class classification, while CatBoost and XGBoost are utilized for binary classification purposes. Optimizing hyperparameters is facilitated by employing a grid search methodology. This entails an exhaustive exploration of a predetermined set of hyperparameters, aiming to find the optimal combination that yields superior performance. In this study, five-fold cross-validation is used, and this signifies that each hyperparameter configuration undergoes five rounds of training. During each iteration, the training data is divided into five segments, with four of them serving as the training set, and the remaining one is dedicated to cross-validation. The overall performance is assessed by calculating the average score across all five rounds, with the evaluation metric being the F1 score. In the presence of a zero-division error, the F1 score is adjusted to 1.

For OCSVM, the kernel used is the radial basis function (rbf). Data points are mapped into a high-dimensional space by the OCSVM learner. Within this expanded dimensionality, a hyperplane is found. It aims to separate the origin from all the data points, so a definitive decision boundary can be established for the learner. The two hyperparameters focused on this study are nu and gamma. Nu governs the balance between maximizing the margin while allowing certain training points to either fall within the margin or become misclassified. Gamma has a substantial influence on the flexibility of the model as it shapes the decision boundary.

CatBoost is an extension of the gradient boosting algorithm. It integrates a variety of inventive methodologies aimed at enhancing both performance and robustness, and its specialty is to handle categorical features. XGBoost is another highly effective gradient boosting classifier. It possesses the ability to harness the multi-threading capabilities inherent in CPUs for concurrent processing. Additionally, it has been enhanced to improve precision and accuracy. Iterations, learning rate, and depth are the three hyperparameters tuned for both models. Iteration, or number of trees, represents the number of boosting rounds or trees constructed. Within each round, a novel tree focuses on rectifying errors committed by its predecessor. The learning rate defines the stride length within gradient descent. Depth plays a crucial role in balancing complexity and generalization by quantifying the maximum edges extending from the root to the leaf within a decision tree.

3.3. Evaluation Metrics

The final performances of the models undergo assessment through three metrics: precision, recall, and the F1 score. These metrics are fundamental to the field of machine learning evaluation. Precision serves as a gauge of the proficiency of the model in positive predictions which is defined in formula 1, and it delineates the proportion of predicted positive instances that align with true positives. This metric emerges as particularly advantageous when the impact of erroneous positive predictions bears a substantial weight, while the implications of false negatives remain comparatively minor.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{1}$$

In contrast, recall, as elucidated in formula 2, evaluates the model's capacity to accurately identify positive instances among all the actual positive cases. It computes the ratio of true positives to the entirety of positive cases. In contrast to precision, recall becomes important when it's crucial to catch as many real positive cases as possible, even if it means allowing some false positives along the way.

$$\text{Recall} = \frac{\text{True Positive}}{\text{Positive}} \tag{2}$$

Finally, the F1 score is calculated in formula 3 which converges precision and recall into a singular measure, offering a comprehensive view of the capabilities of the model. This approach ensures a balanced assessment of performance, especially in situations where both false positives and false negatives hold importance in the overall evaluation context. To avoid zero-division errors, all scores are adjusted to 1 in such scenarios.

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision + Recall}} \tag{3}$$

4. Experiment and Results

4.1. Setup

The setup section of this paper describes the essential hardware and software used to create a strong and repeatable experimental setup, and these important details are presented in Table 1 and Table 2. By explaining the tools and technologies used, this part helps grasp the basis for the following analysis and results. The setup details not only guide replication but also make the research results more trustworthy, ensuring they rely on a steady and clear experimental foundation.

Table 1. Hardware setup of the experiment

Hardware component	Version
Central Processing Unit (CPU)	12th Gen Intel(R) Core (TM) i7-12700H
Random Access Memory	16.0 GB
Graphics Processing Unit 0 (GPU)	Intel(R) Iris(R) Xe Graphics
GPU 1	NVIDIA GeForce RTX 3060 Laptop GPU
Storage drive	TOSHIBA External USB 3.0 USB Device
Operating System	Windows 11 22H2

Table 2. Software setup of the experiment

Python	3.6.5
Numpy	1.19.5
Pandas	1.1.5
Matplotlib	3.3.4
Python Imaging Library	8.4.0
Scikit-Image	0.17.2
Scikit-Learn	0.24.2
CatBoost	1.1.1
Extreme Gradient Boosting	1.5.2

4.2. Dataset

The chest X-ray image dataset used for this experiment includes two parts: the training data contains 15994 examples for the positive class and 13992 for the negative class, while the test data contains 200 images for both classes. To simplify the experiment, 13000 images from the negative class are utilized for training, and the number of images from the positive classes changes according to the specific ratio. All images are sequentially utilized according to the default order in the dataset.

4.3. Results

To simulate different stages in the pandemic, the experiment is divided into three parts, each with a different ratio of data points between the two classes. For the first part which simulates the earliest stage, no image from the positive class is included. As the outbreak progresses, the ratio between positive and negative classes becomes 1:1000. For the final part, the ratio extends to 1:50. In each experiment, the hyperparameters are tuned for the three models, and the resulting scores for the three-evaluation metrics are compared across models.

In the segments of the experiment that contain positive class images, the even distribution of positive examples is tried to be ensured within the training data. This is achieved by randomly shuffling the training data, thereby eliminating biases and promoting a more balanced representation of positive instances.

Throughout the entire experiment, the tunable range of model hyperparameters remains fixed. For OCSVM, the selectable range for nu spans from 0.1 to 0.5 with intervals of 0.1, and the range for gamma includes 0.01, 0.1, 0.5, 1, 2, 5, and 10. For binary classifiers, the iterations hyperparameter is limited to 500, 1000, and 2000. The learning rate can be chosen from 0.01, 0.05, 0.1, 0.2, and 0.3, while the depth has a range from 4 to 8 with intervals of 1. In each experiment, the hyperparameters are tuned using grid search to identify the optimal set that performs best within the specific dataset.

4.3.1 Lack of Positive Instances

At the earliest stage of the pandemic, no positive cases can be obtained, so the training set only contains 13000 images from the negative class. In this circumstance, the optimal hyperparameter set for OCSVM includes nu equal to 0.3 and gamma equal to 0.5. The tuned hyperparameters for XGBoost are 500 for iterations, 0.01 for learning rate, and 4 for depth. CatBoost training fails because the training set contains only the negative class, which does not meet its requirement for at least two classes.

According to Fig.3, the validation scores for both OCSVM and XGBoost are 1.0, indicating the models correctly predict every training image as negative. However, the test scores for the three-evaluation metrics are not as great. The precision scores for both models are 1 due to zero division error, while the recall and F1 scores are 0. Without enough information, the models fail to distinguish images from the positive class and predict all images as negative. This is surprising for OCSVM since it is designed to be trained with data from only one-class, but it is possible that positive images are too analogous to be discriminated.

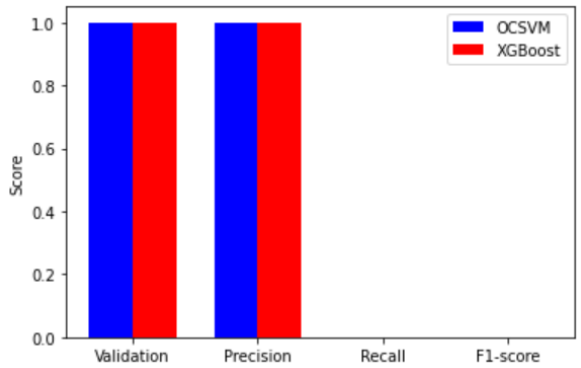

Fig. 3 The results of OCSVM and XGBoost

The models are not meaningful in this setting, and the high precision score is a result of zero division errors. Neither OCSVM nor XGBoost offer diagnostic capabilities for COVID-19 based on chest X-ray images.

4.3.2 Adjusting the Positive-Negative Ratio to 1:1000.

As the pandemic progresses, some people start to be diagnosed as positive for the disease. To simulate this situation, the ratio between positive and negative classes becomes 1:1000, which means 13 positive images are added to the training set. The new hyperparameters selected for OCSVM include nu equals 0.5, and gamma equals 0.01. For XGBoost and CatBoost, the iteration found is 500, the learning rate is 0.01, and the depth is 4.

The scores for the three models are shown below in Fig.4. The validation scores are very low for all models because zero division error for recall no longer exists during validation. XGBoost and CatBoost still do not have enough information to classify any test images as positive, while OCSVM performs a lot better. Although the precision score decreases to 0.71, recall and F1 score surge to 0.35 and 0.46. The OCSVM model starts to distinguish positive cases. Though not precise enough, it is much more useful and meaningful in diagnosis than the binary classifiers.

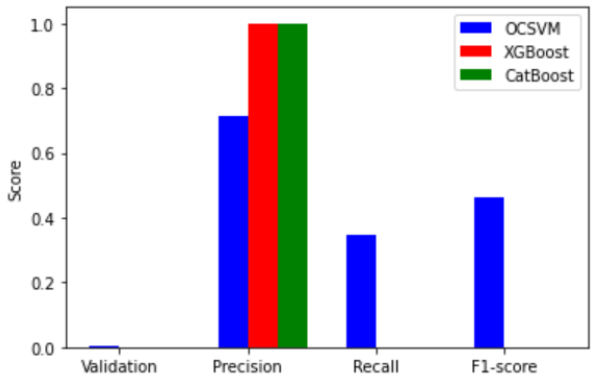

Fig. 4 scores of three models at a 1:1000 ratio

4.3.3 Adjusting the Positive-Negative Ratio to 1:50

In the next stage, the ratio between positive and negative classes is 1:50, and 260 positive images are included in the training data. The chosen hyperparameters for OCSVM remain consistent with Experiment 2, including a new value of 0.5 and a gamma value of 0.01. The best iteration and learning rate for both XGBoost and CatBoost are 500 and 0.3 respectively, while the best depth is 8 for XGBoost, and 4 for CatBoost.

It can be seen from Fig.5 that while the validation scores are generally low, CatBoost slightly outperforms the other two models. However, for test scores, the OCSVM has demonstrated significantly better performance across all three criteria compared to XGBoost and CatBoost, reaching an F1 score of 0.46.

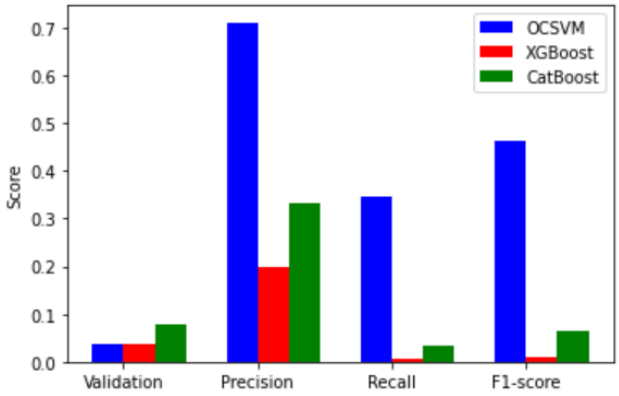

Fig. 5 scores of three models at a 1:50 ratio

4.4. Analysis

Since the F1 score provides a comprehensive view of precision and recall, the analysis is done based on the F1 score. The scores are summarized in Fig.6, with test scores on the left and validation scores on the right. Besides the three experiments demonstrated above, a ratio of 1:500 is added to further enrich the contents and aid analysis.

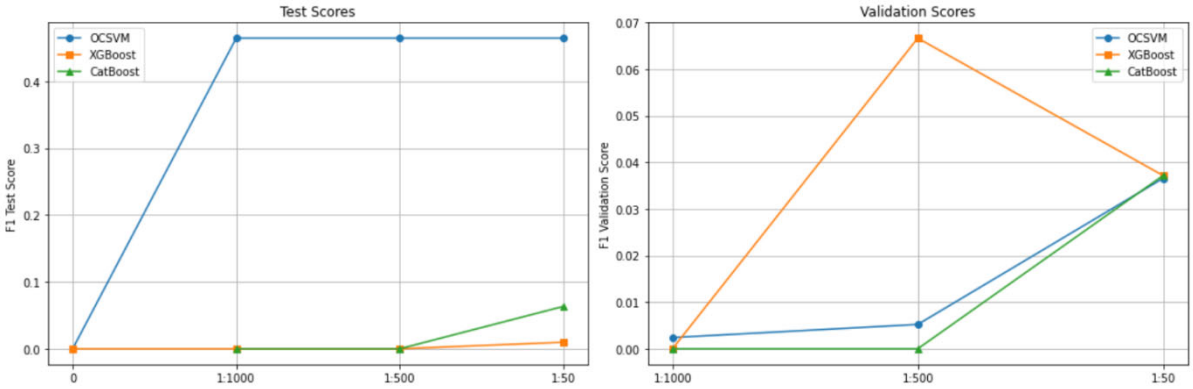

Fig. 6 scores for all models across different scenarios

Except for the first experiment where both OCSVM and XGBoost classifies every image as negative, the performance of OCSVM is significantly better than others. The F1 score for OCSVM is held at 0.46. On the other hand, the F1 score for CatBoost is only 0.063 when the ratio is 1:50. These results confirm that in the presence of imbalanced data, one-class classification outperforms binary classification by a significant margin.

Validation scores increase as the number of positive images increases for OCSVM and CatBoost but reach the peak at 0.067 for XGBoost when the ratio is 1:500. One possible explanation for this behavior is overfitting. Compared with 1:50, the experiment with a ratio 1:500 includes fewer positive images, so the model might find it easier to memorize the positive class data. At the same time, noise might exist. This could lead to overfitting, and, consequently, better performance on the training set. However, the overfitted model fails to generalize in the test set and thus performs poorly during testing.

As the number of positive instances increases, the test scores for binary classifiers slightly improve, while the test scores for OCSVM remain consistent. As the pandemic develops, it is possible that binary classification may outperform one-class classification. However, it is important to note that binary classification tends to perform poorly when dealing with imbalanced data.

5. Conclusion

This study highlights the efficacy of one-class classification in early COVID-19 detection with imbalanced data. OCSVM achieves an impressive F1 score of 0.46, outperforming CatBoost and

XGBoost, which score much lower at 0.009 and 0.063, respectively, under a 1:50 class ratio. Notably, OCSVM's F1 score remains constant at 0.46 as the number of positive instances increases, unlike binary classifiers that gradually improve with balanced data. The results suggest that during the initial COVID-19 stages when positive instances are scarce, one-class classification is beneficial for diagnosis. Adding a few positive cases to a predominantly healthy dataset enhances the one-class classification's effectiveness. However, as the pandemic evolves and data balance shifts, binary classification models become more viable. While offering valuable insights, this study has limitations. It examined only three models with narrow hyperparameter tuning and used a relatively small dataset. Future research should explore a wider range of models, conduct extensive hyperparameter tuning, and expand the dataset for improved performance. Additionally, the impact of image compression and feature extraction on data preprocessing should be further investigated, considering the possibility of retaining essential information by utilizing images directly.

References

[1] Mazurowski M A, Habas P A, Zurada J M, et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance [J]. Neural networks, 2008, 21(2-3): 427-436.

[2] Krawczyk B. Learning from imbalanced data: open challenges and future directions [J]. Progress in Artificial Intelligence, 2016, 5(4): 221-232.

[3] Seliya N, Abdollah Zadeh A, Khoshgoftaar T M. A literature review on one-class classification and its potential applications in big data [J]. Journal of Big Data, 2021, 8(1): 1-31.

[4] Rodríguez-Ruiz J, Mata-Sánchez J I, Monroy R, et al. A one-class classification approach for both detection on Twitter[J]. Computers & Security, 2020, 91: 101715.

[5] Gao L, Yang L, Arefan D, et al. One-class classification for highly imbalanced medical image data[C]//Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. SPIE, 2020, 11318: 342-347.

[6] Leevy J L, Hancock J, Khoshgoftaar T M. Comparative analysis of binary and one-class classification techniques for credit card fraud data [J]. Journal of Big Data, 2023, 10(1): 118.

[7] Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost [J]. Pattern Recognition Letters, 2020, 136: 190-197.

[8] Deng X, Li W, Liu X, et al. One-class remote sensing classification: one-class vs. binary classifiers [J]. International Journal of Remote Sensing, 2018, 39(6): 1890-1910.

[9] Schlieker L, Telaar A, Lueking A, et al. Multivariate binary classification of imbalanced datasets—A case study based on high-dimensional multiplex autoimmune assay data [J]. Biometrical Journal, 2017, 59(5): 948-966.

[10] Chabbouh M, Bechikh S, Hung C C, et al. multi-objective evolution of oblique decision trees for imbalanced data binary classification [J]. Swarm and Evolutionary Computation, 2019, 49: 1-22.

[11] Gao L, Zhang L, Liu C, et al. Handling imbalanced medical image data: A deep-learning-based one-class classification approach [J]. Artificial intelligence in medicine, 2020, 108: 101935.

[12] Sabokrou M, Fathy M, Zhao G, et al. Deep end-to-end one-class classifier [J]. IEEE transactions on neural networks and learning systems, 2020, 32(2): 675-684.

[13] Tanha J, Abdi Y, Samadi N, et al. Boosting methods for multi-class imbalanced data classification: an experimental review [J]. Journal of Big Data, 2020, 7: 1-47.