An Anomaly Detection Model Based on Pearson Correlation Coefficient and Gradient Booster Mechanism

Tuo Ding[1], He Sui[2]*
National Minorities Energy and Technology Co., Ltd, Beijing, China[1]
Civil Aviation University of China, Tianjin, China[2]

***Abstract*—Anomaly detection aims to build a decision model that estimates the class of new data based on historical sample features. However, the distance between samples in the feature space is very close sometimes, resulting in samples being invisible to the detection model that is the class overlap problem. To address this issue, an anomaly detection model based on Pearson correlation coefficient and gradient booster mechanism is proposed in this paper. Different from traditional resampling methods, the proposed method groups and sorts features from different dimensions such as feature correlation, feature importance, and feature exclusivity firstly. Then, it selects features with higher correlation and lower importance for deletion to improve the training accuracy of the detector. Furthermore, through the unilateral gradient sampling mechanism, ineffective or inefficient training samples can be further reduced to improve the training efficiency of the detector. Finally, the proposed method was compared with three feature selection methods and six anomaly detection ensemble models on six datasets. The experimental results showed that the proposed method has significant advantages on feature selection, detection performance, detection stability, and computational cost.**

***Keywords*—Anomaly detection; class overlap; Pearson correlation coefficient; gradient booster mechanism**

I. INTRODUCTION

Anomaly detection aims to build a decision model that estimates the class of new data based on historical sample features. However, samples in a large amount of data may have very close distances in the feature space, with overlapping areas of features, resulting in some samples being invisible to the detection model, that is the problem of feature overlap. Feature overlap can critically affect the definition of decision boundaries [1], but unfortunately, feature overlap always accompanies the occurrence of small sample problems, as small samples have a higher probability of being located to these overlapping areas. Therefore, feature overlap and small sample coupling are a more challenging scenario for data anomaly detection models. For example, when network attacks are hidden in large-scale normal network behavior, abnormal traffic or access data can easily escape detection by the detection system [2].

At present, there are two main approaches to solve the problem of feature overlap: data preprocessing methods and cost sensitive algorithms [3]. The former focuses on preprocessing training dataset in the feature space to alleviate feature overlap [4], while the latter provides a guidance for detection models that lean towards overlapping samples, especially for imbalanced overlapping samples [5]. Generally, the former has a wider range of application, and the most widely used data preprocessing method is resampling. The resampling methods can solve the problem of small data samples by generating minority class samples. However, feature overlap is a more complex and challenging scenario that involves multiple factors [6]. Therefore, feature overlap presents greater challenges for data anomaly detection [7], especially for datasets with complex distribution and higher level of noises. For small sample problems (imbalance problem), the detection model should consider the minority class where the small sample is located as a whole and try to learn its global distribution characteristics to better generate high-quality samples. As for the problem of feature overlap, the detection model should pay more attention to the local distribution characteristics of each small sample. The above contradiction leads to existing methods having better performance on specific domain datasets, while their detection accuracy decreases and generalization ability is insufficient on other domain datasets. The reason is that they have not fundamentally solved the problem of overlap from the perspective of feature distribution.

In response to the above issues, this paper proposes a lightweight anomaly detection model for overlapping data based on Pearson correlation coefficient and gradient booster mechanism, PG-LightGBM. First, PG-LightGBM calculates the correlation between various features and establish a feature overlap matrix based on the Pearson correlation coefficient. Then, it calculates the importance of all features and ranking them with gradient boosting decision tree (GBDT). Furthermore, based on the correlation and importance of all features, it removes some ones with higher correlation and lower importance to alleviate feature overlap. In addition, the unilateral gradient sampling mechanism is used to further reduce invalid or inefficient samples and to improve the training efficiency of the detector. The main contributions of this paper are as follows:

1) To quantify the degree of feature overlap, Pearson Correlation Coefficient (PCC) is introduced to calculate the correlation between two feature variables. The overlap matrix is then obtained based on this correlation calculation, and then the overlapping feature set is obtained through numerical quantification;

2) To select worthier feature to remove, Gradient Boosting Decision Tree (GBDT) is introduced to calculate the importance of overlapping features. At the same time, the feature importance values are accumulated and sorted to obtain important and non-important feature sets.

3) To solve the problem of insufficient detection ability of weak learning machines, a unilateral gradient sampling mechanism is introduced. The detection model is trained by selecting a certain proportion of large and small gradient samples to reduce data size, improving training efficiency, and achieving performance enhancement of weak learning machines through iterative training.

The remainder of this article is organized as follows. In Section II, we review the main methods of anomaly detection for overlapping data. In Section III, we outline the proposed PG-LightGBM in detail. In Section IV, we present the experimental methodology including benchmarked datasets, baseline methods and evaluation metrics. Additionally, in Section V, we report on and analyze the experimental results. Finally, we conclude this paper and look forward to future work in Section VI.

II. Related Work

There are three main approaches to solving the problem of feature overlap: data sampling, feature selection, and model optimization. The following is an overview of related work from these three aspects.

A. Data Sampling Methods

The most classic oversampling method is the Synthetic Minority Oversampling Technique (SMOTE) based on linear interpolation, and its variant algorithm overcomes noise related degradation problems through weighted clustering, such as the NI-MWMOTE (Noise Immunity Majority Weighted Minority Oversampling Technique) algorithm [8]. In addition, Zhu et al. [9] also used positional feature aware interpolation algorithms to segment samples and provided different interpolation strategies for different categories, effectively improving the sampling effect. In recent years, sampling methods based on Generative Adversarial Networks (GANs) have been developed due to their better ability to generate diverse samples [10]. For example, Gayathri et al. [69] further improved the quality of GAN generated samples by using auxiliary information. Engelmann et al. [11] applied Wasserstein distance to GAN models to sample data of specified categories and achieved classifier training optimization on strongly nonlinear datasets. Zheng et al. [12] further introduced penalty coefficients into the GAN model, which significantly improved its stability. Dlamini and Fahim [13] proposed a conditional GAN model with KL divergence. This method not only guides the model to learn the features of minority class samples, but also overcomes the problem of gradient vanishing. Zhu et al. [14] proposed a new GAN based mixed sampling method to handle the classification problem of small sample data. It not only generates samples that match the actual data distribution, but also significantly reduces the impact of feature overlap.

The most classic undersampling method is nearest neighbor search and its variant algorithms, such as Tomelink [15]. Undersampling has shown significant advantages in dealing with feature overlap issues. Kumar et al. [16] proposed an entropy and improved k-nearest neighbor search based undersampling (ENU) method, which overcomes the problems of over elimination and information loss by only removing normal samples with low entropy scores. Dai et al. [17] proposed a multi granularity relabeled undersampling algorithm (MGRU) based on the Tomeklink method for small sample datasets. This algorithm fully considers local information in the granularity subspace and detects potential local overlapping samples in the dataset. Then, eliminate overlapping samples based on the globally re labeled index values. Farshidvard et al. [18] divided large class (normal) samples into multiple clusters in undersampling, so that each cluster did not contain small class (abnormal) samples and controlled the size of each cluster. Zheng et al. [19] proposed a three-stage undersampling framework that integrates functions such as denoising, fuzzy C-means clustering, and representative sample selection to improve the final anomaly detection performance by removing noise and unrepresentative samples. Mayabadi and Saadatfar [20] further reduced the number of large class data, eliminated overlap, and removed noise. Some researchers have also transformed the undersampling problem into other problems to explore new solutions. Dai et al. [21] proposed a method to solve feature overlap through Schur matrix factorization, attempting to obtain global similarity to identify potential overlapping samples, and using matrix factorization to handle feature overlap problems. Soltanzadeh et al. [22] and Le et al. [23] both consider undersampling as an optimization problem and use clustering based surrogate models for optimization processing.

B. Feature Selection Methods

Liu et al. [24] proposed a hybrid method C-E-MWELM (COFS and Ensemble Modified WELM) based on the Weighted Extreme Learning Machine (WELM) to address the imbalance problem of cancer data at the feature and algorithm levels. The classification results on multiple gene datasets show that this method achieves good classification performance, higher balance, and has advantages in detecting and classifying high-dimensional imbalanced data. Wang et al. [25] designed a novel hybrid ensemble classification strategy SFSHEL (Sample and Feature Selection Hybrid Ensemble Learning), and constructed the SFSHEL-RF (Random Forest) classification model based on a random forest classifier. SFSHEL-RF selects both a sample subset and a feature subset, uses a clustering based hierarchical random undersampling method to undersample the majority class samples, and combines them with minority class samples to obtain a sample subset. Surani et al. [26] proposed the Principal Component Loading Feature Selection (PCLFS) method to extract the feature subset with the highest amount of information from imbalanced data. This method sorts the features using the sum of the absolute values of the first k principal component loadings, and then uses the sequential feature selection method to extract the optimal feature subset. Maldonado et al. [27] proposed two embedded feature selection methods, KP-CSSVM (KP Cost Sensitive SVM) and KP-SVDD (KP Support Vector Data Description), for high-dimensional imbalanced data based on kernel penalty Kernel Penalized and KP-SVM. By using a strategy similar to scaling factors to penalize Cardinality in the feature set, and combining cost sensitive SVM and support vector data to

describe SVDD, the predictive performance of SVM based model methods in handling high-dimensional imbalanced data is achieved. Linear and Gaussian kernels were experimentally validated on 12 high-dimensional imbalanced datasets, and both proposed methods achieved the highest average predictive performance. Moayedikia et al. [28] proposed a new feature selection algorithm SYMON (Symmetrical Uncertainty and Harmony Search) for imbalanced data. This method is divided into two stages. In the first stage, SYMON uses Symmetrical Uncertainty SU (Symmetrical Uncertainty) to balance the dependency between features and class labels, and assigns corresponding importance weights to features based on the dependency; in the second stage, SYMON uses Harmony Search (HS) to transform feature selection into an optimization problem, and selects the potential optimal subset of features through vector optimization algorithms. The results indicate that SYMON exhibits comparable or better performance compared to other benchmark feature selection algorithms on different high-dimensional datasets. Du et al. [29] proposed a risk prediction method JICFS (Joint imbalanced classification and feature selection) by combining imbalanced data classification and feature selection. This method uses the Large Margin framework to construct a loss function, which handles the problem of data imbalance by assigning different penalty weights to the majority and minority class samples. It also optimizes the function and achieves feature selection by adding a ℓ 1-norm regularization term to the loss function. In addition, based on the designed iterative optimization scheme, it converges to the global optimal value, and finally, SVM is used for classification prediction. The results on six real medical datasets indicate that the proposed method has certain advantages compared to some more advanced methods. Sun et al. [30] designed a feature selection method AFNFS (Adaptive fuzzy neighborhood-based feature selection) for imbalanced data adaptive synthesis oversampling based on fuzzy neighborhood. This method constructs a balanced decision system through an improved adaptive synthesis of minority class oversampling method, and introduces tolerance parameters into the feature subset selection algorithm of adaptive fuzzy neighborhood to obtain the optimal feature subset. The classification model is trained on a sub training set based on this feature subset. The results indicate that AFNFS can select feature subsets with stronger classification performance.

C. Model Optimization Methods

Tao et al. [31] proposed a density sensitive SVDD classifier DSMSM-SVDD (Density Sensitive SVDD classifier based on Maximum Soft Margin) based on support vector data to describe SVDD. This method optimizes the objective function through penalty weights based on relative density, so that training samples with high relative density are located as much as possible inside the hypersphere, thereby eliminating the influence of noisy data on traditional SVDD. In addition, by introducing the maximum soft interval regularization term, the optimal description boundary is more biased towards minority class samples. This method combined with the AdaBoost ensemble classifier, improves the generalization performance and stability of handling imbalanced data, and outperforms other methods in multiple performance metrics. Rezvani et al. [32] proposed a class imbalance learning method called CIL-FART-IFTSVM (Class Imbalance Learning using Fuzzy Adaptive Resolution Theory and Intuitionistic Fuzzy Twin SVM) for the classification problem of noisy data, outliers, and large-scale imbalanced data. It uses fuzzy ART as the clustering algorithm for imbalanced data. After data processing, train IFTSVM with the retained data and find the optimal hyperplane. The experimental results show that CIL-FART-IFTSVM outperforms other SVM based methods on large-scale imbalanced datasets with noisy data and outliers. Tao et al. [33] proposed a SVM cost sensitive ensemble framework SCW-SVM-CE (Self adaptive Cost Weights based SVM Cost sensitive Ensemble) based on adaptive cost weights for classification research of imbalanced data. This method is based on SVM as the classifier and can adaptively consider the different contributions of minority class samples to SVM. At each iteration, only misclassified minority class samples and correctly classified boundary minority class samples will be assigned higher cost weights, which will have a significant decision impact on the classifier in subsequent iterations. As a result, the final classification boundary will be slightly offset towards the minority class samples. Maurya et al. [34] proposed a large-scale distributed sparse class imbalanced learning algorithm called CILSD (Class imbalanced Learning problem on large scale sparse data in a distributed setting). This algorithm divides imbalanced datasets into different sub datasets and assigns each sub-dataset to different processing nodes. Each node runs the cost sensitive learning distributed learning algorithm FISTA like, which can accelerate the convergence speed of CILSD. The results indicate that CILSD demonstrates its effectiveness and advantages in using multi-core computing on multiple imbalanced test datasets. Wang et al. [35] proposed two improved methods based on AdaBoost, namely Enhanced AdaBoost and Reinforced AdaBoost. The key to these two improvement methods is to adjust the weighted voting parameters of the weak classifier while considering the imbalanced rate of the dataset. The results indicate that if the data imbalance rate is high, Enhanced AdaBoost can achieve good classification performance. If the data imbalance rate is small, the classification performance of Reinforced AdaBoost is better. Fu et al. [36] proposed an ensemble classifier EREC (ER based Ensemble Classifier) based on Evidence Reasoning (ER). This method first divides the training set into n equally sized sub training sets, and then uses an oversampling method based on AP (Affinity Promotion) to balance n sub training sets and train n ER based sub classifiers. The decision weights of each sub classifier are determined by their performance on OOB (Out of Bag) data, and the final decision classification result is determined by the n sub classifiers together. O'Brien et al. [37] proposed a q* classifier based on data density ratio to address the issue of data imbalance. As the q* classifier is implemented based on a random forest classifier, it is also known as an RFQ (Random Forests Quantity) classifier. RFQ optimizes both true positive rate and true negative rate simultaneously, and is equivalent to a cost weighted Bayesian classifier, thus minimizing weighted risk. Raghuwanshi et al. [38] proposed a kernel based ELM classification method called UBKELM (Underbagging based kerneled ELM) based on Underbagging ensemble. UBKELM obtains multiple balanced sub training sets by randomly undersampling the majority class samples, and then uses multiple kernel-based ELMs as sub

classifiers for each balanced sub training set. The final classification results are obtained by combining Majority Voting and Soft Voting methods for each classifier. This method performs better than other contrastive classifiers in the KEEL dataset library.

D. Summary and Motivation

All of data sampling methods, feature selection methods and model optimization methods have proven to be effective in certain situations. Among them, feature selection methods may have a wider application prospect, since it fundamentally solves the problem of overlap from the perspective of feature distribution.

However, the gap the existing methods with the task target is also big. Although the existing methods have realized that feature selection is the fundamental method to solve the overlapping problem, they are still focused on the operational level of how to do feature processing. Not enough attention has been paid to the more important question of which characteristics should be addressed.

The motivation of this paper is to consider the above issue. Especially, we consider not only in terms of the impact on overlap, but also in terms of the useful information that the feature is rich in. That is, we try to balance the role of features in overlap mitigation and knowledge learning, proposing a more widely used anomaly detection method.

III. PG-LightGBM Method

A. PG-LightGBM Process

The process flow of the proposed PG-LightGBM method is shown in Fig. 1.

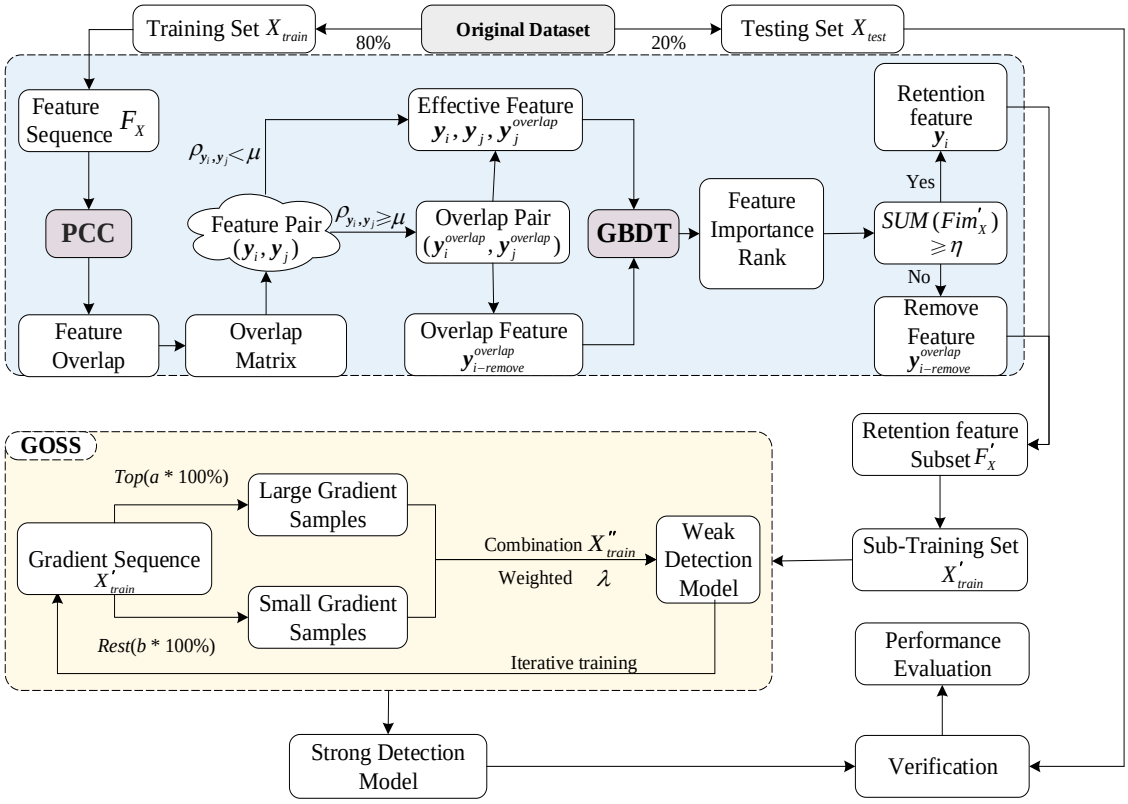

Fig. 1. The process flow of the proposed PG-LightGBM method.

There are three modules for the proposed PG-LightGBM method: overlap degree computing module (PCC), feature importance calculation module (GBDT), and lightweight detection module (GOSS).

As for overlap degree computing module, The Pearson Correlation Coefficient (PCC) is used. The specific design idea is as follows: assuming there is an imbalanced dataset X = {x1, x2,…, xn}, divide the dataset X into a training set Xtrain and a testing set Xtest in a certain proportion (8: 2). From the training set Xtrain, the feature sequence FX = {y1, y2,…, ym} can be obtained, and the Pearson correlation coefficient PCC is used to calculate the feature overlap ρ_{y_i,y_j} between each pair of features in FX. According to the ρ_{y_i,y_j}, the feature overlap matrix can be further obtained, and the upper triangular region is extracted to determine whether the overlap value of all feature pairs is higher than the predetermined threshold μ. If it is higher than μ, it is considered that there is overlap between the pairs of features, and one feature is marked as overlapping feature $y_{i-remove}^{overlap}$ and the other as effective feature; otherwise, it is considered that there is no overlap between the pairs of features and they are all considered valid features.

As for feature importance calculation module, by analogy until all feature pairs are determined, then the gradient boosting decision tree GBDT is used to calculate the importance values

of all features. Furthermore, the feature importance values are sorted and accumulated to obtain the cumulative feature importance value $SUM(Fim_X^{'})$. It is determined whether $SUM(Fim_X^{'})$ has reached the predetermined cumulative threshold η. If the threshold η is reached, the non-cumulative features will continue to be marked as overlapping features $\boldsymbol{y}_{i-remove}^{overlap}$, and the accumulated features will be marked as retention features.

As for lightweight detection module, the single-sided gradient sampling (GOSS) is introduced. All overlapping features marked as $\boldsymbol{y}_{i-remove}^{overlap}$ are discarded, and the detection model is trained with only the sub training set $X_{train}^{'}$ composed of the remaining effective feature subset $F_X^{'}$. The GOSS mechanism utilizes the sample gradient of $X_{train}^{'}$ for training, and by selecting a certain proportion of large and small gradient samples to train weak learners, it can reduce the data size during the training process.

As a result, the PG-LightGBM method, which combines overlapping feature selection with gradient boosting ensemble learning, achieves high performance in anomaly detection, while maximizing the preservation of effective features and information. The following will provide a detailed explanation of the implementation of the PG-LightGBM model.

B. Overlap Quantization Based on Pearson Correlation Coefficient

The Pearson Correlation Coefficient (PCC) [39] is widely used to measure the degree of correlation between two variables or vectors, with correlation values ranging from [-1,1]. For the convenience of calculation, the Square Pearson Correlation Coefficient (SPCC) is generally used to participate in the subsequent calculation process, with SPCC correlation values between [0, 1].

Assuming $\boldsymbol{a}=[a_1,a_2,...,a_n]^T$ and $\boldsymbol{b}=[b_1,b_2,...,b_n]^T$ are two random vectors, whose mean real is 0 and length is *n*. Then the SPCC between $\boldsymbol{a}$ and $\boldsymbol{b}$ is:

$$\rho^2(\boldsymbol{a},\boldsymbol{b})=\frac{E^2(\boldsymbol{a}^T\boldsymbol{b})}{E(\boldsymbol{a}^T\boldsymbol{a})E(\boldsymbol{a}^T\boldsymbol{b})} \quad (1)$$

Let $\prod_{\boldsymbol{a}}$ and $\prod_{\boldsymbol{b}}$ to be two permutation matrices. If $\prod_{\boldsymbol{a}}=\prod_{\boldsymbol{b}}$, then there is $\rho^2(\prod_{\boldsymbol{a}}\boldsymbol{a},\prod_{\boldsymbol{b}}\boldsymbol{b})=\rho^2(\boldsymbol{a},\boldsymbol{b})$; Otherwise, If $\prod_{\boldsymbol{a}}\neq\prod_{\boldsymbol{b}}$, it is obviously $\rho^2(\prod_{\boldsymbol{a}}\boldsymbol{a},\prod_{\boldsymbol{b}}\boldsymbol{b})\neq\rho^2(\boldsymbol{a},\boldsymbol{b})$. According to Eq. (1), it can be seen that $\rho^2(\boldsymbol{a},\boldsymbol{b})\geq 0$. For the case of $\rho^2(\boldsymbol{a},\boldsymbol{b})\leq 1$, it can be defined as:

$$E[(\boldsymbol{a}-c\boldsymbol{b})^T(\boldsymbol{a}-c\boldsymbol{b})]\geq 0 \quad (2)$$

where, *c* is a real number, and the expansion Eq. (2) is:

$$E[(\boldsymbol{a}-c\boldsymbol{b})^T(\boldsymbol{a}-c\boldsymbol{b})]=E(\boldsymbol{a}^T\boldsymbol{a})-2cE(\boldsymbol{a}^T\boldsymbol{b})+c^2E(\boldsymbol{b}^T\boldsymbol{b}) \quad (3)$$

Specifically, for $c=\frac{E(\boldsymbol{a}^T\boldsymbol{a})}{E(\boldsymbol{a}^T\boldsymbol{b})}$, it can be inferred:

$$E(\boldsymbol{a}^T\boldsymbol{a})-2E(\boldsymbol{a}^T\boldsymbol{a})+\frac{E^2(\boldsymbol{a}^T\boldsymbol{a})E(\boldsymbol{b}^T\boldsymbol{b})}{E^2(\boldsymbol{a}^T\boldsymbol{b})}\geq 0 \quad (4)$$

that is:

$$\frac{E(\boldsymbol{a}^T\boldsymbol{a})E(\boldsymbol{b}^T\boldsymbol{b})}{E^2(\boldsymbol{a}^T\boldsymbol{b})}\geq 1 \quad (5)$$

Therefore, it can be concluded that $\rho^2(\boldsymbol{a},\boldsymbol{b})\leq 1$, therefore $0\leq\rho^2(\boldsymbol{a},\boldsymbol{b})\leq 1$. If $\rho^2(\boldsymbol{a},\boldsymbol{b})=0$, then vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are uncorrelated; If $\rho^2(\boldsymbol{a},\boldsymbol{b})$ is closer to 1, then the correlation between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is stronger.

Based on the above analysis, for the training set X_{train}, assuming its feature sequence is $F_X=\{\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_m\}$, for each pair of features in the feature sequence F_X, the feature overlap $\rho_{\boldsymbol{y}_i,\boldsymbol{y}_j}$ is first calculated through the Pearson correlation coefficient SPCC, where $\rho_{\boldsymbol{y}_i,\boldsymbol{y}_j}\in[0,1]$. The calculation method is shown as follows:

$$\rho_{\boldsymbol{y}_i,\boldsymbol{y}_j}=\left|\frac{\mathrm{cov}(\boldsymbol{y}_i,\boldsymbol{y}_j)}{\sigma_{\boldsymbol{y}_i}\sigma_{\boldsymbol{y}_j}}\right|=\left|\frac{E(\boldsymbol{y}_i\boldsymbol{y}_j)-E(\boldsymbol{y}_i)E(\boldsymbol{y}_j)}{\sqrt{E(\boldsymbol{y}_i^2)-E^2(\boldsymbol{y}_i)}\sqrt{E(\boldsymbol{y}_j^2)-E^2(\boldsymbol{y}_j)}}\right| \quad (6)$$

where, cov($\boldsymbol{y}_i$, $\boldsymbol{y}_j$) represents the covariance between features $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, while $\sigma_{\boldsymbol{y}_i}$ and $\sigma_{\boldsymbol{y}_i}$ are the standard deviations of features $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$, respectively. After calculating the feature overlap degree $\rho_{\boldsymbol{y}_i,\boldsymbol{y}_j}$ of all feature pairs ($\boldsymbol{y}_i$, $\boldsymbol{y}_j$), the feature overlap matrix of feature sequence F_X can be obtained. Based on the upper triangular region of the overlap matrix, all overlapping feature pairs ($\boldsymbol{y}_i^{overlap}$, $\boldsymbol{y}_j^{overlap}$) with feature overlap degree higher than the predetermined threshold μ can be statistically calculated, and the subsequent overlapping features to be discarded can be marked by $\boldsymbol{y}_i^{overlap}$. From this aspect, it can be seen that for a single pair of overlapping features, that is, $\boldsymbol{y}_i$ only forms an overlapping feature pair ($\boldsymbol{y}_i^{overlap}$, $\boldsymbol{y}_j^{overlap}$) with $\boldsymbol{y}_j$, then $\boldsymbol{y}_i$ is considered as the overlapping feature $\boldsymbol{y}_{i-remove}^{overlap}$, and $\boldsymbol{y}_j$ is considered as the effective feature; For the case of multiple pairs of overlapping features, that is, there are multiple overlapping feature pairs ($\boldsymbol{y}_i^{overlap}$, $\boldsymbol{y}_j^{overlap}$), ($\boldsymbol{y}_i^{overlap}$, $\boldsymbol{y}_k^{overlap}$) composed of $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$,

y_k, and other features, the processing method is similar to the former, still marking y_i as the overlapping feature $y_{i-remove}^{overlap}$ to be discarded, and treating y_j and y_k as valid features.

C. Feature Importance Calculation based on Gradient Booster Mechanism

Gradient Boosting Machine (GBM) is a type of Boosting mechanism. The main idea of GBM is to construct multiple base learners. During the gradient boosting iteration, the goal of each base learner is to fit the negative gradient of the cumulative model loss function before, and then add the base learner to the cumulative model and gradually reduce the function loss of the cumulative model. In addition, GBM will also use different weights to linearly combine base learners, so that better performing base learners occupy a larger proportion of decision-making.

The tree-based gradient boosting mechanism mainly uses the Gini index (Gi) to calculate feature importance, which is represented by the Feature Importance Measure (Fim). For a certain feature y_i of feature sequence $F_X = \{y_1, y_2, ..., y_m\}$, its Gi index at a node p in the kth tree (k∈K) is:

$$Gi_p^k = 1 - \sum_{C=1}^{|C|} \left(q_{p,C}^k\right)^2 \tag{7}$$

where, C represents the number of categories, and $q_{p,C}^k$ represents the proportion of C in node p of the kth tree. According to Eq. (7), it can be seen that the importance of feature y_i at node p in the kth tree is represented as:

$$Fim_{i,p}^k = Gi_p^k - Gi_l^k - Gi_r^k \tag{8}$$

where, Gi_l^k and Gi_r^k are the Gi indices of the two new nodes after splitting. Based on $Fim_{i,p}^k$, it can be seen that the importance of feature y_i in the kth tree is:

$$Fim_i^k = \sum_{p \in P} Fim_{i,p}^k \tag{9}$$

where, P represents the set of nodes where feature y_i appears in the kth tree. Then, it can be inferred that the final feature importance of feature y_i is:

$$Fim_i = \frac{\sum_{k=1}^{K} Fim_i^k}{\sum_{i=1}^{m}\sum_{k=1}^{K} Fim_i^k} \tag{10}$$

According to Eq. (10), the feature importance sequence of feature sequence $F_X = \{y_1, y_2, ..., y_m\}$ can be obtained, that is, $Fim_X = \{Fim_1, Fim_2, ..., Fim_m\}$. Then, all features in F_X are sorted in descending order based on the value of feature importance. The sorted feature sequence is $F_X' = \{y_1', y_2', ..., y_m'\}$, and the feature importance sequence is $Fim_X' = \{Fim_1', Fim_2', ..., Fim_m'\}$. Then, the feature importance of Fim_X' is accumulated, that is:

$$SUM(Fim_X') = \sum_{i=1}^{t} Fim_i', \quad t \le m \tag{11}$$

In the process of feature importance accumulation, if the feature accumulation value $SUM(Fim_X')$ reaches the predetermined accumulation threshold β, the accumulated features are marked as retained features, and the non-accumulated features are marked as further overlapping features $y_{i-remove}^{overlap}$ to be discarded. Afterwards, all overlapping features marked as $y_{i-remove}^{overlap}$ are discarded to obtain a training set X_{train}' with a reserved feature subset as the feature. Then, the training set X_{train}' is combined with LightGBM for the next training operation.

D. Detection Model Lightweight Based on Unilateral Gradient Sampling

Gradient Booster (GBM) is a general algorithm that can select different base learners $h(x, \theta)$ and loss functions $L(y, F)$ according to actual situations, in order to adapt to different scenarios and evolve into different algorithms. Due to the important role of samples with larger gradients in calculating information gain, single-sided gradient sampling (GOSS) eliminates a larger proportion of small gradient samples, allowing for very accurate information gain estimates with smaller data sizes and accelerating the learning process. With the support of GOSS, the algorithm has significant advantages in terms of computational speed and memory consumption model accuracy.

During the training process, the unilateral gradient sampling mechanism GOSS of LightGBM will utilize the sample gradient of training set X_{train}' to accelerate the training process. Based on the sample gradient sequence of training set X_{train}', GOSS combines the sampled large gradient samples and the remaining small gradient samples to obtain the sub training set X_{train}'. This is used to train a weak classifier and iterate through a loop:

$$X_{train}'' = Top(a*100\%) + Rest(b*100\%)*\lambda \tag{12}$$

where, $Top(a*100\%)$ is the large gradient sample size for sampling, and a represents the sampling ratio; $Rest(b*100\%)$ is the number of small gradient samples

sampled except for large gradient samples, and b represents the sampling ratio; λ is the weight coefficient of small gradient samples, with a value of $(1-a)/b$. λ can increase the learning ability of weak learners on small gradient samples. After the training is completed, the test set X_{test} is subjected to overlapping feature selection and model classification detection on the trained detection model.

IV. Experimental Design

The experiment used six publicly available datasets from the fields of industrial control systems and network security anomaly detection to validate the method proposed in this chapter. Among them, the Power dataset is the power transmission system dataset, which records sensor data and measurement data related to network attack behavior in the power transmission system. The BATDAL dataset is a dataset used to detect network attacks in water supply systems. The ISCX-URL dataset is a dataset used for analyzing and detecting malicious website links. The NSLKDD dataset is an optimized dataset of the famous KDDCUP99 network security anomaly detection dataset, which solves the serious problem of excessive redundant data in the original KDDCUP99 dataset. The WST (Water Storage Tank) dataset is a network attack traffic dataset for water storage tank systems. The UNSW-NB15 dataset is a comprehensive dataset on network intrusion detection systems collected and created by the University of New South Wales. Table I shows the basic information of six datasets: dataset name, data size, feature dimension, number of majority classes, number of minority classes, and imbalance rate (IR) and overlap degree (OR).

TABLE I. Basic Information of Datasets

Dataset	Scale	Features	Maj#	Min#	*IR*	*OR*
Power	5570	128	3921	1648	2.379	0.532
BATADAL	12938	43	12719	219	58.078	0.43
ISCX-URL	18982	79	13796	5186	2.660	0.259
NSLKDD	148517	42	77054	71463	1.078	0.158
WST	236179	23	172415	63763	2.704	0.127
UNSW-NB15	257673	42	164673	93000	1.771	0.484

The evaluation indicators used in the experiment include Accuracy, Precision, Recall, F1 score, ROC AUC value, and PR-AUC. Among them, the horizontal axis of the ROC curve represents specificity (FPR), and the vertical axis represents sensitivity (TPR); The horizontal axis of the PR curve represents Recall, and the vertical axis represents Precision. The relevant solution formula is shown as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

The comparative classification methods used in the experiment are all ensemble learning classification models, which can combine several single weak learning models to obtain a strong learning model with high accuracy, robustness, and stability. The six contrastive ensemble learning classification models used in this chapter are AdaBoost, CatBoost, GBDT (Gradient Boosting Decision Tree), RandomForest, XGBoost, and LightGBM.

The comparative feature selection methods used in the experiment cover three categories of feature selection engineering, namely Filter, Wrapper, and Embedded. Representative feature selection methods from each category are selected, namely ANOVA (Analysis of Variance), RFE (Recursive Feature Elimination), and L1-BFS (L1 Based Feature Selection). Three feature selection methods and the integrated classification model LightGBM constitute the ANOVA-LightGBM, RFE-LightGBM, and L1-BFS-LightGBM classification models.

V. Results and Discussion

A. Feature Processing Capability

1) Results: In order to compare the feature selection capability of the proposed PG-LightGBM method in each dataset more clearly, the experimental results of feature dimension changes for each dataset were recorded, as shown in Table II. The dimension changes, as well as the amount and rate of change, were recorded for six datasets before and after PG-LightGBM processing.

As to further verify the effectiveness of feature selection of PC mechanism in PG-LightGBM, ANOVA -LightGBM, RFE-LightGBM, and L1-BFS-LightGBM were selected for experimental comparison. To ensure the fairness of the experiment, the feature selection dimensions of the comparative method on each dataset were consistent with those of PG-LightGBM on each dataset. The experimental results are shown in Table III.

TABLE II. DATA DIMENSIONAL CHANGES AFTER PROCESSING BY PG-LIGHTGBM

Dimensional change	Power	BATADAL	ISCX-URL	NSLKDD	WST	UNSW-NB15
Before	128	43	79	42	23	42
After	68	23	55	26	8	31
Variation	60	20	24	16	15	11
Change Rate	-46.8%	-46.5%	-30.4%	-38.1%	-65.2%	-26.2%

TABLE III. PERFORMANCE COMPARISON OF DIFFERENT FEATURE SELECTION METHODS COMBINED WITH LIGHTGBM ON 6 DATASETS WITH THE SAME DIMENSIONALITY

	Dataset	ANOVA-LightGBM	RFE-LightGBM	L1-BFS-LightGBM	PG-LightGBM
Accuracy	Power	0.9455	0.9575	0.9449	**0.9623**
	BATADAL	0.9915	0.9905	0.9902	**0.9920**
	ISCX-URL	0.9828	0.9824	0.9822	**0.9837**
	NSLKDD	0.7650	0.7469	0.7707	**0.8022**
	WST	0.8831	0.9419	0.8229	**0.9684**
	UNSW-NB15	0.9597	0.9609	0.9603	**0.9615**
Precision	Power	0.9482	**0.9579**	0.9367	0.9567
	BATADAL	0.9932	0.9917	0.9917	**0.9935**
	ISCX-URL	0.9697	0.9678	0.9672	**0.9717**
	NSLKDD	0.6526	0.6417	0.6582	**0.6927**
	WST	0.8625	0.9790	0.8612	**0.9962**
	UNSW-NB15	0.9640	0.9650	0.9643	**0.9659**
Recall	Power	0.8676	0.8992	0.8775	**0.9170**
	BATADAL	0.9982	**0.9987**	0.9984	0.9984
	ISCX-URL	0.9659	0.9666	0.9666	0.9672
	NSLKDD	0.9719	0.9341	**0.9730**	0.9719
	WST	**1.0000**	0.9409	0.9040	0.9605
	UNSW-NB15	0.9732	**0.9740**	0.9738	**0.9740**
F1 score	Power	0.9060	0.9276	0.9061	**0.9364**
	BATADAL	0.9957	0.9952	0.9950	**0.9960**
	ISCX-URL	0.9678	0.9672	0.9669	**0.9694**
	NSLKDD	0.7809	0.7607	0.7852	**0.8089**
	WST	0.9262	0.9596	0.8821	**0.9781**
	UNSW-NB15	0.9686	0.9695	0.9690	**0.9699**

2) Discussion: As shown in Table II, after the feature correlation and feature importance selection by PG-LightGBM, the feature dimensions of each dataset were significantly reduced. Among them, the feature dimension of the WST dataset decreased the most, reaching 65.2%, which means that more than 60% of the features were removed. However, significantly removing features does not mean sacrificing the classification performance of the detection model; The UNSW-NB15 dataset has the smallest dimensional change, but there is also a 26.2% decrease; the feature dimension reduction of the remaining datasets remains between 30% and 50%.

According to Table III, compared to other methods, PG-LightGBM achieved the highest accuracy on all six datasets, with an accuracy of over 96% on all datasets except for the NSLKDD dataset. In terms of accuracy, RFE-LightGBM narrowly outperformed PG-LightGBM with a slight advantage of 0.0012, but PG-LightGBM achieved the highest accuracy on other datasets. In terms of recall rate, ANOVA- LightGBM achieved a recall rate of 1.0000 on the WST dataset. However, a high recall rate does not necessarily mean high performance, and other performance indicators need to be considered simultaneously. It can be observed that other indicators of ANOVA-LightGBM are relatively low, indicating that ANOVA-LightGBM classifies a large number of negative class samples as positive class samples during classification, resulting in poor overall performance and model instability; On the UNSW-NB15 dataset, PG-LightGBM has the same performance as RFE-LightGBM, but on the BATADAL and NSLKDD datasets, PG-LightGBM is slightly inferior to RFE-LightGBM and L1-BFS-LightGBM with a slight disadvantage of 0.0003 and 0.0011, respectively. However, PG-LightGBM performs the best on the remaining datasets. The F1 score is a comprehensive indicator for evaluating the overall performance of a classification model, taking both the accuracy and recall of the classification model into account. It can be seen from the table that PG-LightGBM has the highest F1 score on all six datasets. This means that the PG-LightGBM proposed in this chapter has the best comprehensive performance in feature selection compared to the other three feature selection methods. It can effectively detect overlapping and low importance features in imbalanced data and maintain strong robustness on complex and diverse datasets.

B. Anomaly Detection Performance

1) Results: In order to further analyze the classification detection ability of PG-LightGBM on imbalanced data, this section selected six ensemble learning classification models, AdaBoost, CatBoost, GBDT, RandomForest, XGBoost, and LightGBM, as comparative classification models. In addition,

to further validate the ability of PG-LightGBM, another two similar methods of PG-GBDT and PG-XGBoost are also used, that are all tree-based ensemble detection method with PC mechanism. This comparison can also verify the advantage of LightGBM. The performance evaluation indicators still use accuracy, precision, recall, and F1 score. The relevant experimental results are shown in Table IV.

In order to more intuitively demonstrate the comprehensive detection ability of PCC-GBDT-COSS on imbalanced data, the ROC curves of AdaBoost, CatBoost, GBDT, RandomForest, XGBoost, LightGBM, and PG LightGBM on six datasets were plotted in the experiment, and the area under the ROC curve (ROC-AUC value) was used as the comprehensive detection and evaluation indicator for seven integrated classification models. The ROC curve can intuitively reflect the impact of different classification thresholds on the generalization performance of the classification model, which helps to select the optimal classification threshold. Moreover, the fuller and closer the ROC curve is to the upper left corner, the larger the ROC-AUC value, indicating that the comprehensive detection ability of the classification model is stronger. The ROC curves of seven ensemble classification models on 6 datasets are shown in Fig. 2.

2) Discussion: Observing Table IV, it is worth mentioning that for the WST dataset, after PG LightGBM selection of data features, the dimensionality decreased from 23 dimensions to 8 dimensions, with a decrease of up to 65.2%. As mentioned in the effectiveness analysis of feature selection in Section V (A), significantly removing features does not mean sacrificing the detection performance of the classification model. By observing the experimental results in Table IV, this can be confirmed: based on the experimental data in the table, although the four performance indicators of PG-LightGBM on the WST dataset are not the best, it can be observed that the experimental results of LightGBM and PG-LightGBM on the WST dataset are surprisingly consistent. Significantly reducing the dimensionality of data features, but achieving the same performance, further validates the effectiveness of PG-LightGBM in the feature selection process. In addition, the performance of PG-LightGBM on the WST dataset has certain practical significance in the storage and detection classification of massive data, which can save a lot of space and computational resources.

TABLE IV. PERFORMANCE COMPARISON OF EACH CLASSIFICATION METHOD ON 6 DATASETS

	Dataset	Ada Boost	Cat Boost	GBDT	RandomForest	XGBoost	Light GBM	PG-GBDT	PG-XGBoost	PG-GBDT
Accuracy	Power	0.7798	0.9509	0.8630	0.9078	0.8594	0.9617	0.9620	0.9431	**0.9623**
	BATADAL	0.9889	0.9918	0.9884	0.9915	0.9915	0.9915	0.9896	0.9919	**0.9920**
	ISCX-URL	0.9345	0.8758	0.9073	0.9614	0.9730	0.9828	0.9371	0.9740	**0.9837**
	NSLKDD	0.7796	0.7999	0.7703	0.7720	0.7794	0.7935	0.7928	0.7862	**0.8022**
	WST	0.9691	0.9667	0.9692	0.9446	0.9691	0.9684	**0.9846**	0.9690	0.9684
	UNSW-NB15	0.9366	0.9459	0.9447	0.9579	0.9447	0.9613	0.9503	0.9523	**0.9615**
Precision	Power	0.6778	**0.9649**	0.9369	0.9422	0.8928	0.9585	0.9567	0.9376	0.9567
	BATADAL	0.9914	0.9925	0.9901	0.9927	0.9930	0.9930	0.9921	0.9926	**0.9935**
	ISCX-URL	0.8951	0.9584	0.9103	0.9618	0.9549	0.9709	0.9208	0.9623	**0.9717**
	NSLKDD	0.6682	0.6922	0.6589	0.6599	0.6677	0.6837	0.6733	0.6788	**0.6927**
	WST	0.9935	0.9940	0.9935	0.9905	0.9947	**0.9962**	0.9941	0.9863	**0.9962**
	UNSW-NB15	0.9484	0.9407	0.9461	0.9648	0.9434	0.9646	0.9569	0.9466	**0.9659**
Recal	Power	0.5198	0.8695	0.5870	0.7411	0.6087	0.9130	0.8895	0.7987	**0.9170**
	BATADAL	0.9974	**0.9992**	0.9982	0.9987	0.9984	0.9984	0.9884	0.9884	0.9984
	ISCX-URL	0.8557	0.9508	0.7252	0.8911	0.9436	0.9645	0.9382	0.9579	**0.9672**
	NSLKDD	0.9699	0.9646	0.9677	0.9715	0.9715	0.9690	0.9707	0.9715	**0.9719**
	WST	**0.9642**	0.9604	0.9642	0.9334	0.9630	0.9605	**0.9717**	0.9603	0.9605
	UNSW-NB15	0.9523	**0.9766**	0.9685	0.9694	0.9716	0.9750	0.9448	0.9745	0.9740
F1 score	Power	0.5884	0.9148	0.7218	0.8296	0.7239	0.9352	0.8083	0.8177	**0.9364**
	BATADAL	0.9944	0.9958	0.9941	0.9957	0.9957	**0.9960**	0.9561	0.9861	**0.9960**
	ISCX-URL	0.8750	0.9546	0.8073	0.9251	0.9492	0.9677	0.8403	0.9500	**0.9694**
	NSLKDD	0.7913	0.8060	0.7843	0.7859	0.7914	0.8017	0.7876	0.7927	**0.8089**
	WST	0.9786	0.9769	**0.9787**	0.9611	0.9785	0.9781	0.9688	0.9785	0.9781
	UNSW-NB15	0.9504	0.9583	0.9572	0.9671	0.9346	0.9561	0.9608	0.9505	**0.9699**

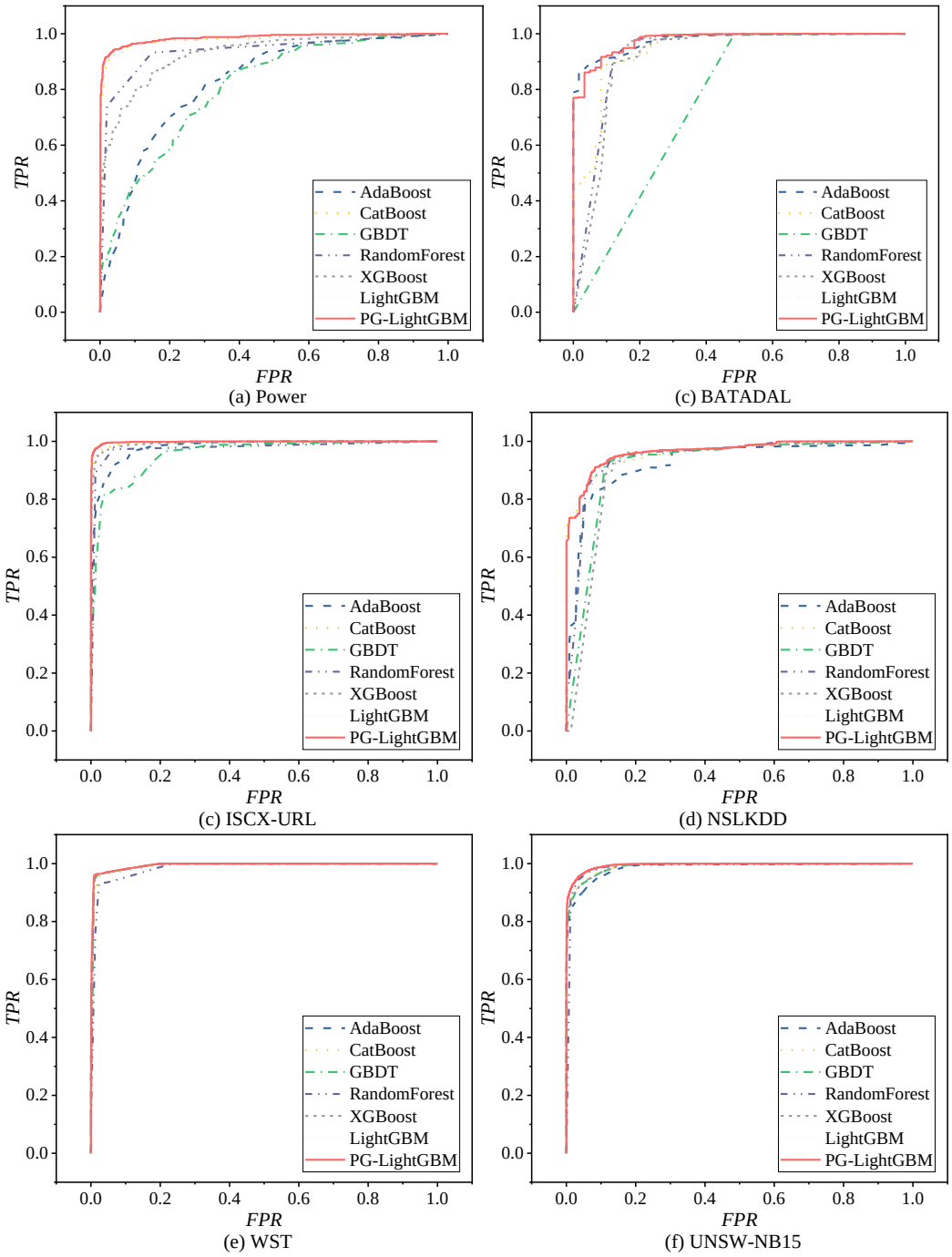

Fig. 2. ROC curves of each classifier on 6 datasets.

Except for the WST dataset, PG-LightGBM achieved the best classification performance in terms of accuracy and F1 score on the other five datasets. In terms of accuracy, PG-LightGBM performs slightly lower than the CatBoost ensemble classification model on the Power dataset, but performs the best on other datasets. In terms of recall, although PG-LightGBM is not as good as the CatBoost integrated classification model on the BATADAL and UNSW-NB15 datasets, the overall difference is small, and its performance is better than the Power, ISCX-URL, and NSLKDD datasets. PG-LightGBM seems to exhibit suboptimal performance in WST dataset. This is because that WST is a low dimensional dataset with only 23 features, and its OR value is also the minimum with only 0.127. The above information indicates that overlap maybe not the major limitation for its anomaly detection. Some conventional detectors can also complete the classification task and identify anomalies. As for the proposed PG-LightGBM method, its selection of features may lead to certain important features being lost. For a high dimensional dataset, it may be acceptable, compared with feature overlap. As for low-dimensional dataset, each feature may contain a lot of useful information, to the feature selection must be more careful. In conclusion, the

proposed PG-LightGBM method is more suitable for high dimensional dataset, especially with higher overlap degree.

As shown in Fig. 2, the solid red line represents the ROC curve of PG-LightGBM. From it, it can be seen more intuitively that the ROC curve of PG-LightGBM maintains a very full curve trend on all six datasets, that is, it maintains a high ROC-AUC value. According to experimental data, PG-LightGBM achieved ROC-AUC values of 0.9854, 0.9777, 0.9982, 0.9659, 0.9943, and 0.9949 on the Power, BATADAL, ISCX-URL, NSLKDD, WST, and UNSW-NB15 datasets, respectively. It only maintained the same highest ROC-AUC value as LightGBM on the WST dataset, but the ROC-AUC values on the other datasets were higher than those of other integrated classification models. This indicates that in anomaly detection of imbalanced data, PG-LightGBM has a relatively high comprehensive detection ability compared to other integrated classification models. The classification performance is better, and the model generalization performance is more impressive.

Compared with the two similar methods of tree-based ensemble detection models with PC mechanism, PG-GBDT and PG-XGBoost, the proposed PG-LightGBM also show clear advantages. Although in some on some datasets, the difference between them is not obvious. Considering all the experimental results, the effectiveness and advance of this method are sufficient to be proven.

C. Stability of Testing Performance

1) Results: In order to further verify the detection stability of PG-LightGBM on imbalanced data, this section of the experiment plotted the PR curves of AdaBoost, CatBoost, GBDT, RandomForest, XGBoost, LightGBM, and PG-LightGBM on six datasets, and used the area under the PR curve (PR-AUC value) as the detection stability evaluation index for seven integrated classification models. The PR curves of seven ensemble learning classification methods on 6 datasets are shown in Fig. 3.

The previous text used ROC curves, which can reflect the comprehensive detection ability and generalization performance of the model and can be applied to most classification and detection scenarios. However, due to the characteristics of ROC curves, they are not very sensitive to the degree of data imbalance. That is, when the degree of data category imbalance is high, the ROC curve cannot well reflect the impact of data imbalance on the classification model. Therefore, this section of the experiment introduces a PR curve, with the horizontal axis representing recall and the vertical axis representing precision. The PR curve is sensitive to the degree of data imbalance and can accurately reflect the stability of the classification model when detecting imbalanced data. Similar to the ROC curve, the fuller and closer the PR curve is to the upper right corner, the larger the PR-AUC value, indicating that the classification model has higher detection stability and better detection performance for imbalanced data.

2) Discussion: As shown in Fig. 3, the solid red line in the figure represents the PR curve of PG-LightGBM. It can be seen that in the six datasets, the PR curve of PCC-GBDT-GLOSS has a more prominent trend and is quite full. The PR-AUC values of PG-LightGBM on the Power, BATADAL, ISCX-URL, NSLKDD, WST, and UNSW-NB15 datasets were 0.9766, 0.9996, 0.9959, 0.9590, 0.9977, and 0.9971, respectively. PG-LightGBM maintained the best PR-AUC value compared to LightGBM on the WST dataset, while the PR-AUC value on the NSLKDD dataset was slightly lower than LightGBM. However, in other cases, PG-LightGBM had higher PR-AUC values than other ensemble classification models. From this, it can be concluded that PG-LightGBM has strong adaptability to imbalanced data, effectively overcoming data imbalance and overlapping data features. Compared with other integrated classification models, it exhibits strong model stability, further demonstrating the effectiveness of feature selection and its excellent classification detection performance.

D. Calculation Cost of Algorithm

1) Results: Calculation cost is one of the factors that detection methods need to focus on. Assume that a detection method has high detection performance, but consumes a lot of computational costs, this is not friendly for some application scenarios with limited computing resources. Therefore, achieving high detection performance while minimizing computational resources is an ideal state for detection methods. In order to verify the computational cost of PG-LightGBM, this section analyzes the training time cost of PG-LightGBM with AdaBoost, CatBoost, GBDT, RandomForest, XGBoost, and LightGBM ensemble learning detection methods on the Power, BATADAL, ISCX-URL, NSLKDD, WST, and UNSW-NB15 datasets from the perspective of method training time cost. The operating platform used in the experiment is uniformly Apple M1 processor with 16GB of memory. The training time cost results of each ensemble learning detection method obtained are shown in Table V, measured in seconds (s).

TABLE V. TRAINING TIME COST OF EACH ENSEMBLE LEARNING DETECTION METHOD (S)

Dataset	AdaBoost	CatBoost	GBDT	Random Forest	XGBoost	Light GBM	PG-LightGBM
Power	1.106	6.135	5.800	1.249	0.737	0.395	**0.324**
BATADAL	0.960	0.399	5.220	2.078	0.584	0.357	**0.224**
ISCX-URL	1.186	2.287	6.123	1.516	0.944	0.506	**0.465**
NSLKDD	3.498	2.140	15.519	6.044	4.217	0.685	**0.547**
WST	2.267	1.986	7.712	6.983	0.447	0.985	**0.438**
UNSW-NB15	12.549	0.842	55.595	1.064	0.848	0.963	**0.821**

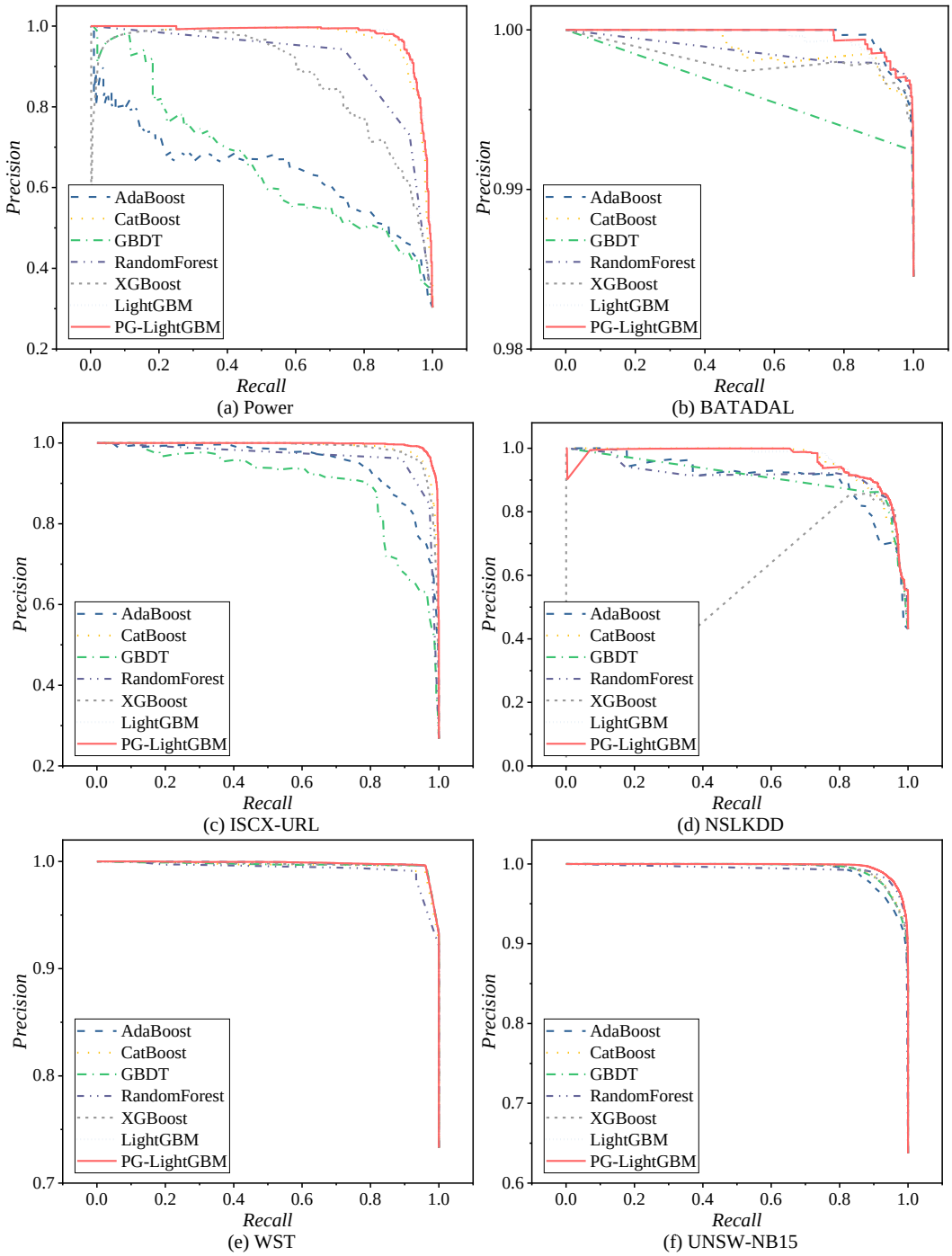

Fig. 3. PR curves of each classification method on 6 datasets.

2) Discussion: From Table V, it can be seen that the training time cost of PG-LightGBM is lower than the other six contrastive ensemble learning detection methods on all six datasets. The main reason for this is that PG-LightGBM can eliminate overlapping features in the data and train models on the basis of effective feature subsets, avoiding unnecessary redundant feature calculations. It also organically combines LightGBM's GOSS mechanisms, further reducing the data and feature size during the training process, accelerating the training process, effectively reducing the consumption of computing resources during the training process, and reducing computational costs.

This chapter focuses on anomaly detection of imbalanced data, and proposes a lightweight gradient boosting ensemble learning detection and classification method PG-LightGBM based on Pearson correlation coefficient (PCC) and gradient boosting machine (GBM) for overlapping feature selection from the perspective of class overlap.

Experimental results have shown that the method proposed in this chapter can effectively detect overlapping features in imbalanced data and select effective features, reducing the interference of redundant features on the performance of classification models, enhancing the learning ability and stability of classification models. At the same time, combining the GOSS mechanisms of LightGBM, PG LightGBM is generally superior to other comparison methods in terms of feature selection effectiveness and comprehensive detection performance on the Power, BATADAL, ISCX URL, NSLKDD, WST, and UNSW-NB15 datasets. In addition, PG-LightGBMalso has strong model stability and robustness, and is suitable for large-scale datasets and highly imbalanced datasets. In the real world where the data scale is increasingly large and rare data is increasingly hidden, PG-LightGBM has good real-world usability.

VI. CONCLUSION

This paper proposes a method for anomaly detection of overlapping data, PG-LightGBM, based on Pearson correlation coefficient and gradient boosting machine, from the perspective of feature processing. Introducing Pearson correlation coefficient (PCC), calculating the correlation between two feature variables, and obtaining an overlap matrix based on the correlation between different feature pairs to quantify the degree of feature overlap. Introducing gradient boosting decision trees to calculate the importance of overlapping features, while accumulating and sorting feature importance values to obtain important and non-important feature sets, and then removing the intersection features of overlapping and non-important feature sets to solve the problem of feature overlap selection. Introducing a unilateral gradient sampling mechanism, using sample gradients for training, selecting large and small gradient samples in a certain proportion to train the detection model, reducing data size, improving training efficiency, and achieving performance enhancement of weak learning machines through iterative training. The experimental results show that PG-LightGBM can effectively detect overlapping features in the data and select effective features, reducing the interference of redundant features on the performance of classification models, enhancing the learning ability and stability of classification models. At the same time, combined with the GOSS mechanism, PG-LightGBM is generally superior to other comparison methods in terms of feature selection effectiveness and comprehensive detection performance on the Power, BATADAL, ISCX-URL, NSLKDD, WST, and UNSW-NB15 datasets. In addition, PCC-GBDT-COSS also has strong model stability and robustness, and is suitable for large-scale datasets and highly imbalanced datasets. In the real world where the data scale is increasingly large and rare data is increasingly hidden, PG-LightGBM has good real-world usability.

In the PG-LightGBM detection method, its sensitivity to feature dimension is its potential drawback. Because the feature selection mechanism is essentially dimensionality reduction, so certain important information may be lost for lower dimensional datasets. So, it is more suitable for high dimensional dataset, especially with higher overlap degree. In addition, the overlapping feature threshold and feature importance accumulation threshold of the data require human intervention to be set, which may lead to potential feature over elimination and loss of effective feature information. Future work will study on feature stitching instead of feature selection, and design adaptive threshold mechanisms to prevent potential risks caused by human intervention.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Funds for the Central Universities of Civil Aviation University of China (Grant no. 3122023033).

REFERENCES

[1] H. K. Lee, S. B. Kim, “An overlap-sensitive margin classifier for imbalanced and overlapping data”, Expert Syst. Appl., vol. 98, pp. 72–83, 2018.

[2] M. Wan, W. L. Shang, P. Zeng, “Double behavior characteristics for one-class classification anomaly detection in networked control systems”, IEEE T. Inf. Foren. Sec., vol. 12, pp. 3011-3023, 2017.

[3] S. Das, S. Datta, B. Chaudhuri, et al., “Handling data irregularities in classification: foundations, trends, and future challenges,” Pattern Recogn., vol. 81, pp. 674-693, 2018.

[4] P. Vuttipittayamongkol, E. Elyan, “Neighbourhood-based undersampling approach for handling imbalanced and overlapped data,” Inform. Sciences, vol. 509, pp. 47-70, 2020.

[5] Y. D. Zhao, K. R. Hao, X. S. Tang, et al., “A conditional variational autoencoder based self-transferred algorithm for imbalanced classification,” Knowl-Based Syst., vol. 218, p.106756, 2021.

[6] M. S. Santos, P. H. Abreu, N. Japkowicz, et al., “A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research,” Inform. Fusion, vol. 89, pp. 228-253, 2023.

[7] P. Peng, W. J. Zhang, Y. Zhang, et al., “Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis,” Neurocomputing, vol. 407, pp. 232-245, 2020.

[8] J. Wei, H. Huang, L. Yao, et al., “NI-MWMOTE: an improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems,” Expert Syst. Appl., vol. 158, p. 113504, 2020.

[9] T. Zhu, Y. Lin, Y. Liu, “Improving interpolation-based oversampling for imbalanced data learning,” Knowl-Based Syst., vol. 187, p. 104826, 2020.

[10] F. N. Zhou, S. Yang, H. Fujita, et al., “Deep learning fault diagnosis method based on global optimization GAN for unbalanced data,” Knowl-Based Syst., vol. 187, p. 104837, 2020.

[11] R. G. Gayathri, A. Sajjanhar, Y. Xiang, et al., “Multi-class classification based anomaly detection of insider activities,” arXiv:2102.07277, 2021.

[12] J. Engelmann, S. Lessmann, “Conditional wasserstein GAN-based oversampling of tabular data for imbalanced learning,” Expert Syst. Appl., vol. 174, pp. 1-13, 2021.

[13] M. Zheng, T. Li, R. Zhu, Y. H. Tang, et al., “Conditional wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification,” Inform. Sciences, vol. 512, pp. 1009-1023, 2020.

[14] G. Dlamini, M. Fahim, “Dgm: a data generative model to improve minority class presence in anomaly detection domain,” Neural Comput. Appl., vol. 33 (20), pp. 13635-13646, 2021.

[15] B. Zhu, X. Pan, S. V. Broucke, et al., “A GAN-based hybrid sampling method for imbalanced customer classification,” Inform. Sciences, vol. 609, pp. 1009-1023, 2022.

[16] I. Tomek, “Two modifications of CNN,” IEEE T. Syst., Man Cy. B., vol. 6, pp. 769-772, 1976.

[17] A. Kumar, D. Singh, R. S. Yadav, “Entropy and improved k-nearest neighbor search-based under-sampling (ENU) method to handle class overlap in imbalanced datasets,” Concurr. Comp-Pract. E., Online, https://doi.org/10.1002/cpe.7894, 2023.

[18] Q. Dai, J. W. Liu, Y. Liu, “Multi-granularity relabeled under-sampling algorithm for imbalanced data,” Appl. Soft Comput., vol. 124, p. 109083, 2022.

[19] A. Farshidvard, F. Hooshmand, S. A. MirHassani, "A novel two-phase clustering-based under-sampling method for imbalanced classification problems," Expert Syst. Appl., vol. 213(B), p. 119003, 2023.

[20] M. Zheng, T. Li, X. Y. Zheng, et al., "UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification," Inform. Sciences, vol. 576, pp. 658–680, 2021.

[21] S. Mayabadi, H. Saadatfar, "Two density-based sampling approaches for imbalanced and overlapping data," Knowl-Based Syst., vol. 241, p. 108217, 2022.

[22] Q. Dai, J. W. Liu, Y. H. Shi, "Class-overlap undersampling based on Schur decomposition for class-imbalance problems," Expert Syst. Appl., vol. 221, p. 119735, 2023.

[23] P. Soltanzadeh, M. R. Feizi-Derakhshi, M. Hashemzadeh, "Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach," Pattern Recogn., vol. 1, p. 109721, 2023.

[24] H. L. Le, D. Landa-Silva, M. Galar, et al., "UEUSC: A clustering-based surrogate model to accelerate evolutionary undersampling in imbalanced classification," Appl. Soft Comput., vol. 101, p. 107033, 2021.

[25] Z. Liu, D. Tang, Y. Cai, et al., "A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data," Neurocomputing, vol. 266, pp. 641-650, 2017.

[26] Z. Wang, P. Jia, X. Xu, et al., "Sample and feature selecting based ensemble learning for imbalanced problems," Appl. Soft Comput., vol.113(A), p. 107884, 2021.

[27] M. Surani, D. Mike, M. Saman, "Assessing feature selection method performance with class imbalance data." Mach. Learn. Appl., vol. 6, p. 100170, 2021.

[28] S. Maldonado, J. López, "Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification," Appl. Soft Comput., vol. 67, pp. 94-105, 2018.

[29] A. Moayedikia, K. L. Ong, Y. L. Boo, et al., "Feature selection for high dimensional imbalanced class data using harmony search," Eng. Appl. Artif. Intel., vol. 57, pp. 38-49, 2017.

[30] G. Du, J. Zhang, Z. Luo, et al., "Joint imbalanced classification and feature selection for hospital readmissions," Knowl-Based. Syst., vol. 200, pp. 106020, 2020.

[31] L. Sun, M. Li, W. Ding, et al., "AFNFS: Adaptive fuzzy neighborhood-based feature selection with adaptive synthetic over-sampling for imbalanced data," Inform. Sciences, vol. 612, pp. 724-744, 2022.

[32] X. Tao, W. Chen, X. Li, et al., "The ensemble of density-sensitive SVDD classifier based on maximum soft margin for imbalanced datasets," Knowl-Based Syst., vol. 219, p. 106897, 2021.

[33] S. Rezvani, X. Wang, "Class imbalance learning using fuzzy ART and intuitionistic fuzzy twin support vector machines," Inform. Sciences, vol. 578, pp. 659-682, 2021.

[34] X. Tao, Q. Li, W. Guo, et al., "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification," Inform. Sciences, vol. 487, pp. 31-56, 2019.

[35] C. K. Maurya, D. Toshniwal, "Large-scale distributed sparse class-imbalance learning," Inform. Sciences, vol. 456, pp. 1-12, 2018.

[36] W. Wang, D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," Inform. Sciences, vol. 563, pp. 358-374, 2021.

[37] C. Fu, Q. Zhan, W. Liu, "Evidential reasoning based ensemble classifier for uncertain imbalanced data," Inform. Sciences, vol. 578, pp. 378-400, 2021.

[38] R. O'Brien, H. Ishwaran, "A random forests quantile classifier for class imbalanced data," Pattern Recogn., vol. 90, pp. 232-249, 2019.

[39] I. Cohen, Y. Huang, J. Chen, et al., "Pearson correlation coefficient," Noise Re. Speech Process., vol. 1, pp. 1-4, 2009.