IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 January 2023, accepted 8 February 2023, date of publication 13 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244682

RESEARCH ARTICLE

Semi-Supervised Gaussian Processes Active Learning Model for Imbalanced Small Data Based on Tri-Training With Data Enhancement

CHENXIAO ZHOU AND LIANYING ZOU
School of Electrical and Information, Wuhan Institute of Technology, Wuhan 430205, China
Corresponding author: Chenxiao Zhou (2653464846@qq.com)

ABSTRACT To solve the problem that some imbalanced small sample datasets only contain a few labeled samples, a semi-supervised gaussian processes active learning model based on improved tri-training with enhanced data is proposed. Firstly, the label samples are balanced and enhanced, and we present a quantitative enhanced data evaluation criteria based on the JS distance and the similarity of information entropy between enhanced data and original data to select the best enhanced data. Secondly, an improved semi-supervised learning method based on tri-training is proposed to find the unlabeled samples which have high confidence, so the certainty of the labeled samples group can be increased, in order to ensure that the three classifiers of tri-training have both difference and robustness, random forest is introduced to divide the features of the dataset into three groups with equal contribution, and each classifier trains different combinations of two feature groups. Thirdly, in order to query and classify the most informative unlabeled samples more precisely, active learning based on the Gaussian process and JS distribution range is structured, because of the high uncertainty of the unlabeled samples predicted by active learning, the similarity distribution range of JS distance is introduced to compare the similarity of unlabeled samples and labeled samples in active learning's classifier, so the model can classify more diverse samples. The final experimental results show that compared with several traditional models, the proposed model performs better on artificial datasets and imbalanced small-size UCI datasets.

INDEX TERMS Active learning, imbalaned small dataset, Gaussian procession, semi-supervised learning, tri-training.

I. INTRODUCTION

The traditional machine learning classification tasks are usually divided into two types: one is supervised learning, and the model uses labeled samples for training; another is unsupervised learning, the model clusters unlabeled samples. Nevertheless, in practical application, several datasets only contain a few labeled samples, which require a huge time and labor to mark the unlabeled samples [1]. Not only that, but some datasets also have the defects of small sample size and imbalanced samples [2]. Therefore, how to effectively predict unlabeled samples by training only a few samples has become a key problem in current machine learning, this problem is called semi-supervised learning problem.

At present, there are two solutions to this problem. One is semi-supervised learning(SSL), and another is active learning(AL) [3]. SSL attempts to find the unlabeled samples with the highest confidence in prediction, mark them and put them into labeled dataset, then continue to predict the remaining unlabeled samples by the new labeled dataset until all unlabeled samples are predicted. AL aims to query the most informative unlabeled samples and mark them to expand the labeled dataset and repeat the program until all unlabeled samples are marked.

The difference-based SSL model is the most mainstream SSL model, originating from co-training [4]. Co-training divides the data attributes into two groups that are

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

conditionally independent of each other, and the two groups of data are trained by two classifiers. Dalva et al. summarized the co-training strategy into three types: agreement, disagreement and self-combined [5]. The agreement strategy hypothesizes unlabeled data as a class by both classifiers with a confidence score, sorts the agreed samples according to the sum of confidence scores, then select the samples whose confidence score exceeds a certain threshold. The disagreement strategy aims to classify the hard sample which is decided by one classifier but another classifier is indecisive. In the disagreement strategy, unlabeled samples are sorted according to the absolute difference of absolute confidence scores, then the sample whose confidence score of the corresponding hypothesis exceeds a certain threshold is selected. The self-combined strategy allows two classifiers to select high confidence samples independently, then select the unlabeled samples that are classified into the same class by two classifiers. However, co-training has several shortages, such as neglecting learning model's relevance and dataset's characteristics [6]. Therefore, researchers have been improving co-training for many years, and the most famous variant of co-training is tri-training [7]. Tri-training was proposed by Zhou et al. in 2005, it's a semi-supervised classification model with ensemble thinking. This model can fully use the feature set of data to improve the efficiency of semi-supervised learning. For a long time, tri-training was considered an effective SSL model. However, like other SSL models, tri-training will introduce noise due to false predicting of unlabeled samples in iterative training when lacking enough data, leading to degradation of classification performance [8].

AL actively finds the most informative unlabeled samples to increase the diversity of labeled training samples. In the active learning process, the learning machine actively searches the unlabeled samples with the most information through the query strategy, trains and classifies these unlabeled samples, and then adds these samples into the labeled training samples group. These new labeled samples can significantly reduce the wrong classification information, thereby improving the classification accuracy of the classifier [9]. However, the most informative unlabeled samples are also the most uncertain samples, which contain more noise. If these samples are not effectively classified, the generalization ability of the model will be significantly affected. Therefore, active learning sometimes has high uncertainty.

The full prediction ability of Tri-training on data can effectively avoid the risk of wrong prediction in AL, and AL can select the best samples for the classifier to predict. Therefore, some researchers try to combine SSL with AL [2]. Xu et al. proposed a QBC and tri-training based on the active SVM model [3]; this model uses an improved tri-training algorithm to label the unlabeled samples with the highest confidence, and then uses an AL algorithm based on QBC to select these new labeled samples with the highest inconsistent to increase the generalization performance of the model. However, the threshold setting of this model depends on manual operation, which will undoubtedly affect the model's classification performance. Zhang et al. introduced tri-training algorithm in the CEAL model to select the most confident unlabeled samples, and improved the AL strategy in the CEAL based on voting entropy [10]. Nevertheless, this model can not select the pseudo-label samples with high precision. Although these algorithms that combine SSL and AL have more advantages than SSL and AL alone, these algorithms will select too many redundant samples, and need to search the entire sample space when querying samples, thus increasing the complexity and running time of the algorithms. In addition, most similar studies rarely consider the imbalanced small sample problem.

The imbalanced small sample problem is also a significant difficulty in semi-supervised learning. Due to the imbalance of training data and the scarcity of sample size, the classification results of machine learning models tend to favor the majority of class samples, lacking the learning of minority class samples, thus affecting the model's generalization ability. Zhao et al. ingeniously proposed a semi-supervised learning algorithm based on mixed sampling for imbalanced data classification [2], this algorithm can effectively improve the classification ability of semi-supervised model for small-size imbalance samples. However, it does not pay enough attention to minority samples, and the effect on the binary classification task is not ideal.

In order to solve the above problem, based on the model proposed in literature [2], we propose a semi-supervised model suitable for two-class imbalanced small sample datasets. In this model, we combine the robustness of tri-training and the diversity of AL. We have innovated the feature allocation technology of tri-training and improved the query strategy and classifier in AL. The proposed model is called a semi-supervised gaussian processes active learning model based on improved tri-training.

The main contributions of this study are summarized as follows:

1)A new semi-supervised learning model with imbalanced samples is proposed, which is suitable for binary classification. In this model, tri-training and AL are combined.

2)The classifier feature assignment mechanism of tri-training is improved. The features are divided into three groups with the same contribution value. The three groups of features are combined in pairs, so the three classifiers have sufficient prediction ability and certain difference.

3)The query strategy of AL is improved to Gaussian processes, and the similarity distribution range of JS distance is introduced into the classifier of AL. The Gaussian process can better measure the uncertainty of samples than the traditional voting entropy and KL divergence, and the improved classifier, which introduces the distribution range of JS distance, can effectively help judge the class of unlabeled samples.

4)A quantitative enhanced data evaluation criteria is proposed to measure the quality of enhanced data, the JS distance between the original sample and the enhanced sample is

used to measure the quality of the extended data, and the JS distance of information entropy between the original sample and the enhanced sample is used to measure the diversity of the extended data digitally.

5)In order to solve the problem that all prediction results of the three classifiers of tri-training are not the same, cause the training process can't be continued. A total classifier is introduced to predict all the remaining samples.

Experiments on two artificial datasets and five UCI datasets prove the effectiveness of the proposed model.

II. RELATIVE WORK

A. TRI-TRAINING

Tri-training is an improved co-training algorithm, it uses three classifiers to identify the label of each unlabeled sample. Therefore, tri-training has strong robustness. In the training process, classifiers are used for cooperative training, and unlabeled samples with high confidence are selected for labeling. Although the fall prediction will introduce noise into labeled samples, Zhou et al. proved in his paper that when there are enough new data, the impact of noise can be offset [7].

In recent years, many scholars have been improving and expanding the application of Tri-training. Inspired by the asymmetric tri-training framework for unsupervised domain adaptation, Saito et al. proposed a model-agnostic meta-learning method which is applied to the recommender system [11]. Mo et al. improved tri-training by using ladder network [12], allocating different weights to the new labeled data, and expanding the training set. Liu et al. introduced the theory of teacher-student model in tri-training [13]. Zhang et al. introduced the convex optimization method into tri-training to reduce the noise label [14], replaced the error rate with cross-entropy, proposing a Safe Tri-training Algorithm Based on Cross Entropy. Zhang et al. implemented the Tri-Training algorithm in cost-effective active learning to improve generalization performance on image classification problems [10]. Tseng et al. proposed a tri-training decision module based on the judgment of probability threshold [15].

B. ACTIVE LEARNING

Different from semi-supervised learning, active learning actively search and labels the most informative unlabeled samples, that is, the most uncertain samples [16]. There are several query strategies frameworks for active learning to find the most uncertain samples like uncertain sampling, query-by-committee, expected model change and density-weighted methods, etc. In fact, these strategies are querying the unlabeled sample which is most different to discriminate. In recent years, researchers have found that simple query strategies have been difficult to measure the uncertainty of samples, and many researchers have tried to propose new query strategies or combine different query strategies [17]. Xu et al. selected the most inconsistent unlabeled samples while the vote entropy are higher than the threshold and the most consistent unlabeled samples while the vote entropy is lower than the threshold [3]. Gu et al. provided an active learning risk bound based on the informativeness and representativeness of unlabeled samples [17], then propose a novel batch mode active learning combined with semisupervised SVM based on risk bound, improving the generalization ability of the model. Zhao et al. introduced the mixtures of Gaussian processes into active learning [18], and designed three query strategies based on mixtures of Gaussian processes. Compared with other deterministic models or probabilistic models, this model uses the Gaussian processes to select the most uncertain samples from the probability, thus providing a flexible framework for probabilistic regression and classification. This model is especially suitable for binary classification problems.

Dwarikanath et al. improved active learning in the medical image classification task [19], aiming at the problem that active learning cannot be applied to multi-label samples, a new sample selection method based on graph analysis is proposed to identify information samples in multi-label environment. Lee et al. proposed a data acquisition framework based on active learning for the highly unbalanced distribution of property in data-driven metamaterials design [20], aiming to guide the generation of diversity and task-aware data. Hossein et al. proposed Probabilistic Minimax Active Learning (PMAL) [21], which uses the variational method in the likelihood function of logistic regression to approximate the PMAL target, thus minimizing the upper risk limit of the classifier. Luciano et al. uses active learning based on uncertainty in the application of diagnosing unknown industrial faults [22], which helps experts by intelligent fault diagnosis and searching for potential samples of new types of faults.

III. SEMI-SUPERVISED GAUSSIAN PROCESSES ACTIVE LEARNING MODEL BASED ON IMPROVED TRI-TRAINING WITH ENHANCED DATA

Our model is comprises of labeled data enhancement module, high confidence sample classification module, and low confidence sample classification module. A quantitative enhanced data evaluation criteria based on sample similarity and diversity similarity to evaluate the enhanced samples is proposed, and the best enhanced method is selected to improve the model's prediction ability for imbalanced and small-size samples. After data enhancement, all labeled samples are input into the improved tri-training as the train samples, and the improved tri-training predicts all unlabeled samples to find the highest confidence samples. In order to ensure the difference and robustness of the three classifiers of tri-training, each classifier is input features group with similar contribution value. When the prediction results of the three classifiers for an unlabeled sample are same, the sample is classified into the labeled sample set as a new train sample.

When the unlabeled samples whose prediction results of three classifiers are inconsistent, these samples are input into the active learning. First, the most uncertain unlabeled samples are selected by the Gaussian processes. Then decides the

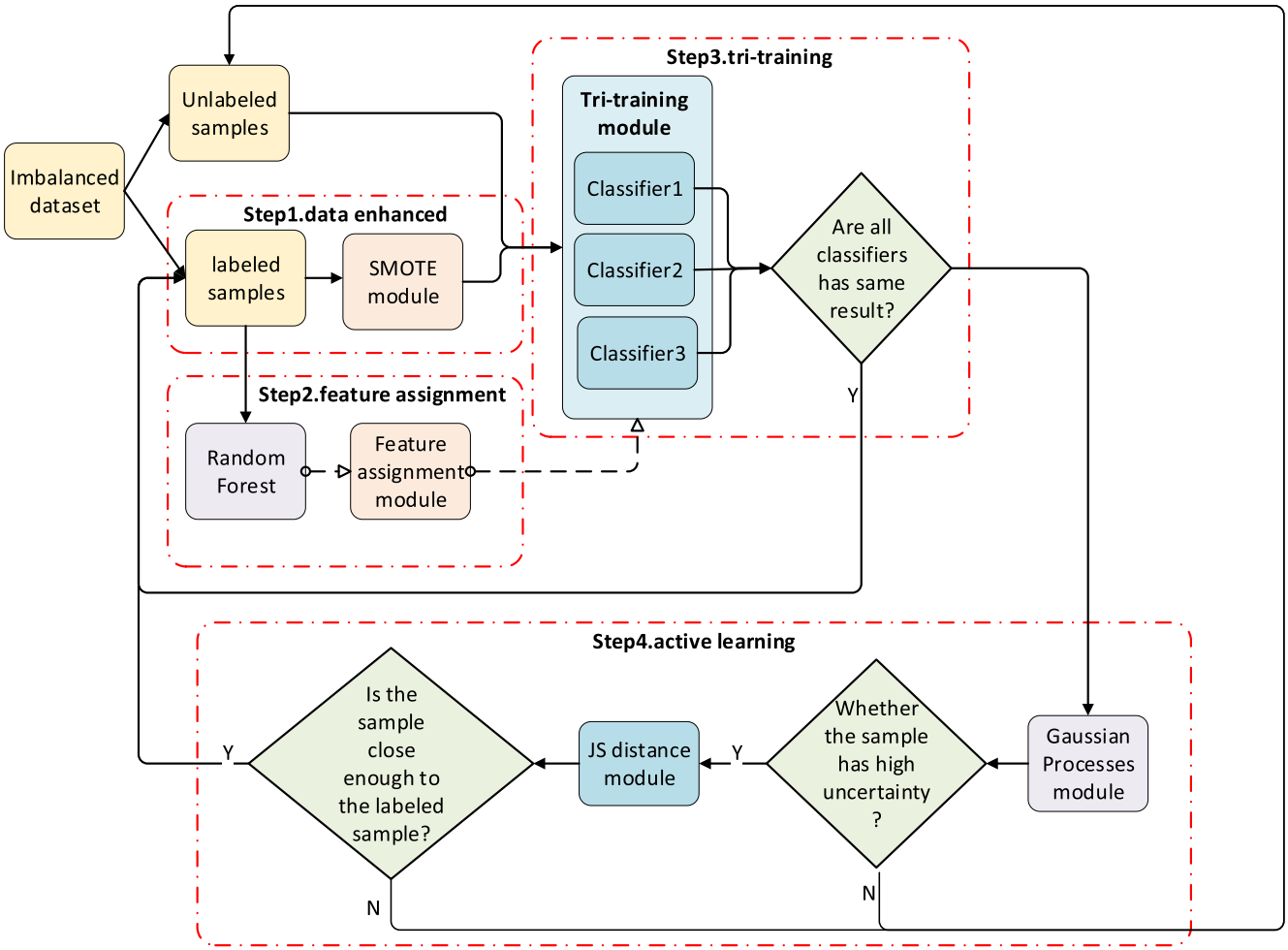

FIGURE 1. Model's flow.

inclined classes of the three classifiers according to the voting entropy of the prediction results of each sample. After that, calculates the JS distance between each unlabeled sample and their inclined labeled samples. Suppose the JS distance between an unlabeled sample and their inclined labeled samples is within a certain threshold range, in that case, the sample can be considered as a true positive sample or a true negative sample. The distribution range of JS distance between labeled samples determines the threshold range. The remaining unlabeled samples are input into the model again until all unlabeled samples are labeled. The model's flow is shown in FIGURE 1.

The description of our model is as follow in Algorithm 1.

A. QUANTITATIVE ENHANCED DATA EVALUATION CRITERIA

In traditional literature, KL divergence is usually used to measure the similarity between the enhanced data and the original data. KL divergence is usually used to calculate the difference between two distributions, and its formula is:

$$KL(P||Q) = \sum P(x) \log \frac{p(x)}{q(x)} \tag{1}$$

However, KL divergence is also asymmetric, which makes KL divergence not flexible in practical application. Therefore, the JS distance is introduced to improve from KL divergence as the measurement standard of sample similarity. Compared with KL divergence, JS distance can distinguish the similarity more accurately and has symmetry, which makes it more flexible than KL divergence. The formula is:

$$JS(P||Q) = \frac{1}{2}KL(P||\frac{P+Q}{2}) + \frac{1}{2}KL(Q||\frac{P+Q}{2}) \tag{2}$$

Meanwhile, the traditional standard for measuring the enhanced data is only to compare the similarity between the enhanced data and the original data. Zang et al. introduced the diversity of enhancement data into the measurement standard [23]. However, literature [23] only relies on the distribution map of samples to measure the diversity, which undoubtedly makes the measurement standard of diversity in literature [23] highly subjective. At the same time, literature [23] believes that better enhancement samples should have better diversity, but the enhancement samples that are too diverse will also deviate from the original samples. Consequently, the diversity of enhancement samples should be close to the original samples.

Algorithm 1 Semi-Supervised Gaussian Processes Active Learning Model Based on Tri-Training With Data Enhancement

Input: Labeled dataset LD and unlabeled dataset UD
Function data enhancement(LD) **begin**:
  synth ← Oversampling(LD)
  If synth has close similarity and diversity with original samples:
    LD ← LD ∪ synth
End
Function improved tri-training(LD, UD) **begin**:
  Features of LD are assigned to tri-training by random forest
  tri-training is trained by LD
  While UD is no empty:
    For each data ud in UD do:
If three classifier of tri-training have same result for ud:
  LD ← ud
Else if three classifier of tri-training have different result for ud:
  If ud has high uncertainty according to Gaussian processes:
    active learning ← ud
  Else if ud has low uncertainty according to Gaussian processes:
    UD ← ud
End
Function active learning(ud) **begin**:
  C0 ← class of tri-training result tendency of ud
  J0 ← Calculating the JS distance between C0 samples
  Ju ← Calculating the JS distance between C0 samples and ud
  If Ju is in normal distribution of J0:
    LD ← ud
  Else if Ju is in normal distribution of J0:
    UD ← ud
End

Therefore, information entropy is introduced as a digital measurement standard to enhance the diversity of data, and measures the proximity of the diversity of the enhanced data and the original data by comparing the JS distance between the enhanced data information entropy and the original data information entropy. Information entropy was first proposed by Shannon to measure the occurrence frequency of each probability. Its formula is:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{3}$$

According to the above evaluation criteria, our model will select the most similar enhanced data to the original data to balance and expand the labeled samples.

B. TRI-TRAINING BASED ON RANDOM FOREST'S FEATURE ASSIGNMENT

Tri-training is a prediction method based on the difference between classifiers. However, if the classification ability of one classifier is too weak, the totality classification effect will decrease and noise will be introduced. Therefore, all three classifiers should have similar classification ability.

Inspired by this idea, the random forest is introduced to calculate the contribution value of each feature of the dataset. Random forest divides these features into three groups of features with equal total contribution value, so that the three classifiers have similar classification performance while ensuring that the three classifiers have differences. At the same time, in order to improve the classification ability of the three classifiers, the three groups of features are combined in different pairs, so there are three different combinations of two feature groups. Then each classifier is trained with a combination so that the tri-training has both the difference and better classification ability. The structure of tri-training based on random forest feature assignment is shown in FIGURE 2.

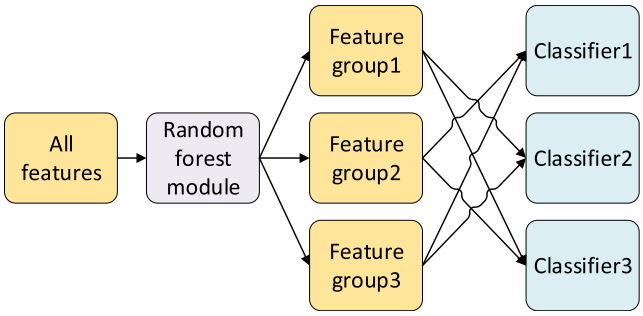

FIGURE 2. Feature assignment mechanism of tri-training.

C. GAUSSIAN PROCESS ACTIVE LEARNING WITH DISTRIBUTION RANGE OF JS DISTANCE

The active learning module is composed of Gaussian process and classifier. Gaussian processes is a random process in which the observed values appear in a continuous domain [24]. In the Gaussian process, every point in the continuous input space is associated with a normally distributed random variable. Each finite set of these random variables has a multidimensional normal distribution. That is to say, the distribution of Gaussian process is the joint distribution of all random variables.

Suppose there are N training points. For all, if obey multivariate Gaussian distribution, can be said to be a Gaussian process, the formula is:

$$f(\vec{x}) \sim N(\mu(x), K(x, x)) \tag{4}$$

Gaussian process is usually used as regression method, but its principle can also be applied to classification problems. Gaussian process regression method can be used for binary classification problems by taking positive and negative labels as output. The classification is performed by determining the sign of the average value of the prediction distribution. If the average value exceeds a certain threshold, the test points are classified as positive, otherwise, the opposite is true.

The distribution formula of the Gaussian process is as follows:

$$\mathrm{m}(x*) = \sum_{k=1}^{T} p_k m_k(x*) \tag{5}$$

$$\sigma(x*) = \sum_{k=1}^{T} p_k(\sigma_k(x*) + m_k^2(x*)) - (\sum_{k=1}^{T} p_k m_k(x*))^2 \tag{6}$$

In (5) and (6), m(x) is the predicted mean, and σ(x) is the predicted variance. It can be seen from (1) that the mean value

of the labels of the two classes of samples is the decision boundary, and the sample nearest to the decision boundary can be found by calculating the difference between the predicted mean value of the samples and the decision boundary. It can be seen from (2) that the sample with the highest degree of deviation, that is, the sample with the lowest confidence, can be found by comparing the size of the sample prediction variance.

According to the above derivation, three query strategies based on Gaussian processes is used to select the most informative samples [18].

(1)Select the sample closest to the classification boundary according to the mean value of the prediction probability. The formula is:

$$\widehat{X} = \arg\min_{x* \in X_U} |\sum_{k=1}^{T} p_k m_k(x*)| \tag{7}$$

(2) Select the sample with the lowest confidence according to the predicted probability variance. The formula is:

$$\widehat{X} = \arg\max_{x* \in X_U} \sum_{k=1}^{T} p_k(\sigma_k(x*) + m_k^2(x*)) - (\sum_{k=1}^{T} p_k m_k(x*))^2 \tag{8}$$

(3)Select the sample whose category cannot be determined most according to the variance and mean value of the prediction probability using the cumulative distribution function of a standard Gaussian distribution N(0, 1). The formula is:

$$\widehat{X} = \arg\min_{x* \in X_U} |\sum_{k=1}^{T} p_k \psi(\frac{m_k(x*)}{\sqrt{\sigma_k(x*)}})| \tag{9}$$

In formulas (7-9), it means all unlabeled data, m(x) is the predicted mean, and σ(x) is the predicted variance of all unlabeled data, and is the prediction probability of each unlabeled sample.

As for classification tasks, traditional active learning uses QBC(query by committee). The principle of QBC is similar to tri-training, which uses two classifiers for prediction. If the classification results are consistent, the samples can be considered true. However, in this model, the function of QBC will coincide with tri-training, thus increasing model's redundancy. Therefore, we attempt to combine the original classification results of tri-training, determining the class of sample by combining the voting entropy of the classification results of unlabeled samples in tri-training with the similarity of each class of labeled samples. For example, suppose the tri-training classification result of an unlabeled sample is biased toward the positive sample, and the similarity with the positive labeled sample is higher than a certain threshold. In that case, the unlabeled sample can be considered as the positive sample.

Therefore, how to set the similarity threshold becomes a key point. Because there are differences among all samples, even among the same class samples have different similarities, hence, the similarity between all samples with the same label should be in a certain range.

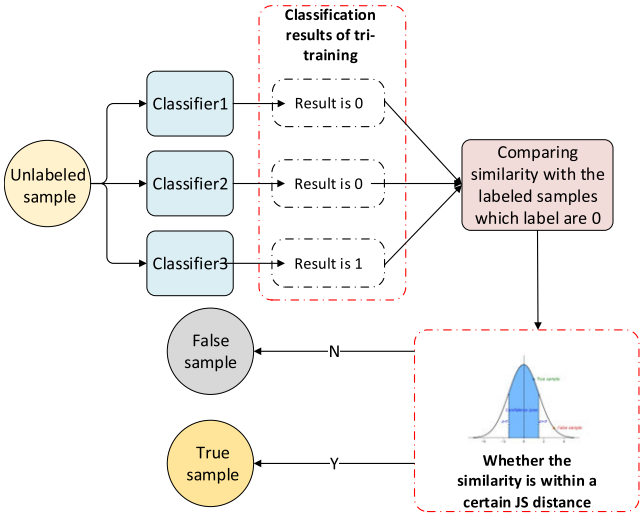

FIGURE 3. Similarity distance between samples.

Calculating the JS distance between each sample in the same label to obtain the similarity distribution range of each class of samples by calculating the maximum, minimum, and average of these JS distances. The average of the population maximum and population average is defined as the upper bound, and the average of the population minimum and population average is defined as the lower bound. The range consisting of the upper bound and the lower bound is the similarity range. Suppose the average JS distance between an unlabeled sample and a labeled sample is within the similarity distribution range. In that case, the unlabeled sample can be considered to belong to the labeled sample.

FIGURE 3 shows the classification process of active learning.

D. THE TERMINATION STRATEGIES OF MODEL

One problem of tri-training is that when three classifiers predict different result for all samples, the model won't continue to train. Although the introduction of active learning will help tri-training complete the training, there will still be samples that cannot be queried by active learning. Therefore, a classifier trained by data with all features is introduced to predict the remaining unlabeled samples. Since most of the unlabeled samples have been predicted to be labeled in the previous tri-training and active learning processes, the prediction at this time has been transformed into a traditional supervised learning classification problem.

IV. EXPERIMENTS AND RESULTS ANALYSIS

In this section, two experimental groups are conducted to demonstrate the validity of the proposed quantitative enhanced data evaluation criteria and proposed model. First, the best enhancement data is selected according to the proposed quantitative enhanced data evaluation criteria, then compare the prediction performance of tri-training after training by different enhancement data to verify the validity of selected enhanced data. After that, the proposed model is

TABLE 1. Information summary of two artificial datasets used in experiments.

Dataset	Instance number	Feature number	Redundant feature number	Proportion of positive and negative samples
Artificial Disease 1	250	5	1	20%
Artificial Disease 2	500	8	2	20%

TABLE 2. Information summary of five UCI datasets used in experiments.

Dataset	Instance number	Feature number	Proportion of positive and negative samples
CMC	627	9	25.24%
Vehicle	317	18	31.54%
WDBC	569	11	37.25%
Diabetes	768	7	34.89%
Heart Disease	297	9	46.12%

compared with other semi-supervised models on different datasets to verify whether the proposed model has better prediction performance on imbalanced small datasets.

A. DATASETS

So far, there is no publicly available and generally agreed benchmark dataset for semi-supervised classifier, researchers often used other common datasets for semi-supervised classification experiments. However, for some datasets whose labeled samples are not representative, their unlabeled samples are also unavailable [25]. In order to obtain a comprehensive statistical analysis and fairly compare the performance of the proposed models and other listed model, we constructed two artificial datasets by using the make_classification in sklearn(V.0.0). These two artificial datasets contain problems in actual prediction, The sample number of each dataset is not more than 1000, and the imbalanced rate is 20%. Considering the common noise in datasets, redundant features are set in artificial datasets.

To further explore the performance of the proposed model and its ability to solve practical problems, five commonly used UCI (University of California, Irvine) datasets were used in experiments [3]. These five datasets are CMC, Vehicle, WDBC, Diabetes and Heart Disease datasets. Note that, in view of the fact that there will be a large number of features in real problems, the Vehicle dataset with 18 features was used in the experiment.

The information of artificial datasets is shown in TABLE 1.

The information of UCI datasets is shown in the TABLE 2.

TABLE 3. JS distance between the generated data of the four models and the original data.

	CMC	Vehicle	WDBC	Diabetes	Heart Disease
ADASYN	0.0272	0.0045	0.0082	0.1652	0.0093
BorderlineSMOTE	0.0269	**0.0044**	0.0083	**0.1571**	0.0096
SMOTE	0.0259	0.0047	0.0054	0.1701	**0.0092**
SMOTETomek	**0.0257**	0.0047	**0.0051**	0.1665	0.0099

B. EVALUATING INDICATOR

In traditional machine learning classification experiments, accuracy is usually used to evaluate the classification effect of the classifier. However, for imbalanced datasets, because of the small proportion of minority samples in the overall sample, accuracy is difficult to evaluate the classification effect of the classifier for minority samples.

For imbalanced data, four descriptions, *TP*, *TN*, *FP* and *FN* are usually used for evaluation [2]. The meaning of these four descriptions is shown below.

•*TP*: True Positive. A positive sample is classified as a positive sample.

•*TN*: True Negative. A negative sample is classified as a negative sample.

•*FP*: False Positive. A negative sample is classified as a positive sample.

•*FN*: False Negative. A positive sample is classified as a negative sample.

Based on these four descriptions, the true positive rate(Abbreviated to *TPR*) and the false positive rate(Abbreviated to *FPR*) can be calculated as shown below.

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{TN + FP} \tag{11}$$

where, the *TPR* means the proportion of real positive samples to all the samples predicted as positive samples, and the *NPR* means the proportion of real negative samples to all the samples predicted as negative samples.

Sorting the samples according to the prediction results of the model, and the samples are predicted as positive samples in order. Then the *FPR* of each sample is taken as the abscissa and the *TPR* of each sample as the ordinate for plotting. So the ROC curve is got, and AUC(Area Under Curve) can be obtained by calculating the area under the ROC curve. AUC can both effectively measure the classification results of positive and negative samples.

In addition, the F-measure is also introduced to evaluate the model's generalization ability. The F-measure is calculated as shown below.

$$F - measure = \frac{(\beta^2 + 1) \times (TP/(TP + NP)) \times (TP/(TP + FN))}{\beta^2 \times (TP/(TP + NP)) + (TP/(TP + FN))} \tag{12}$$

TABLE 4. JS distance between the diversity of the data generated by the four methods and the diversity of the original data.

	CMC	Vehicle	WDBC	Diabetes	Heart Disease
ADASYN	**0.0013**	0.0004	7.5665	0.0027	0.1093
BorderlineSMOTE	0.0019	0.0003	0.0005	0.0036	0.1112
SMOTE	0.0022	0.0003	7.1201	0.0042	0.1315
SMOTETomek	0.0022	**0.0002**	**0.0001**	**0.0024**	**0.0985**

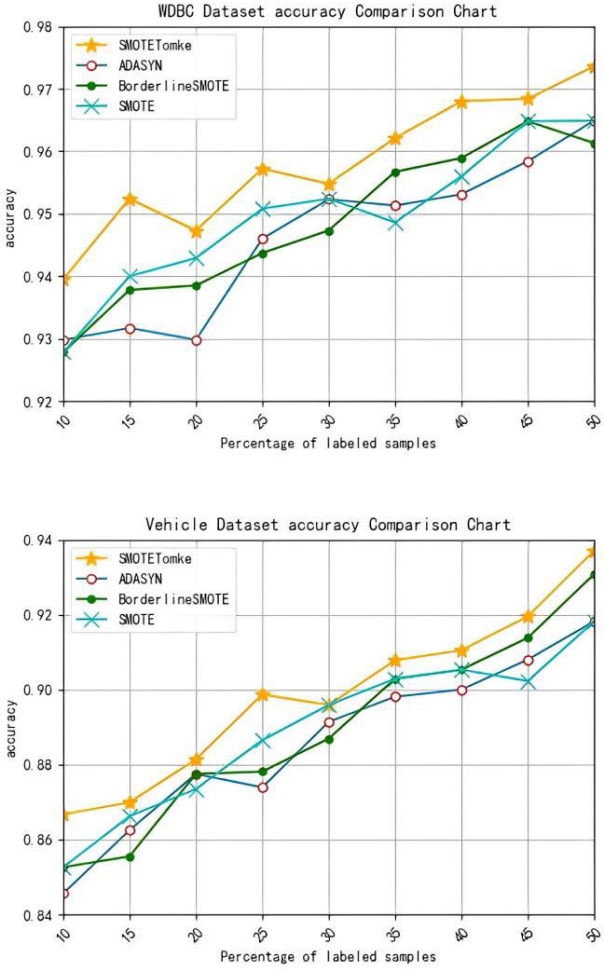

FIGURE 4. Comparison of prediction accuracy of tri-training after training by five kinds of enhanced data.

where $(TP/(TP + NP))$ means the calculation of precision rate, $(TP/(TP + FN))$ means the calculation of recall rate, β is a parameter in F-measure, in this paper we set β as 1. It can be seen from (12) that the F-measure is defined as the harmonic average of the precision rate and recall rate.

C. EFFECTIVENESS OF THE PROPOSED QUANTITATIVE ENHANCED DATA EVALUATION CRITERIA

The four most commonly used oversampling methods are selected for comparison according to the enhanced data evaluation criteria proposed in Section IV-C. These five methods

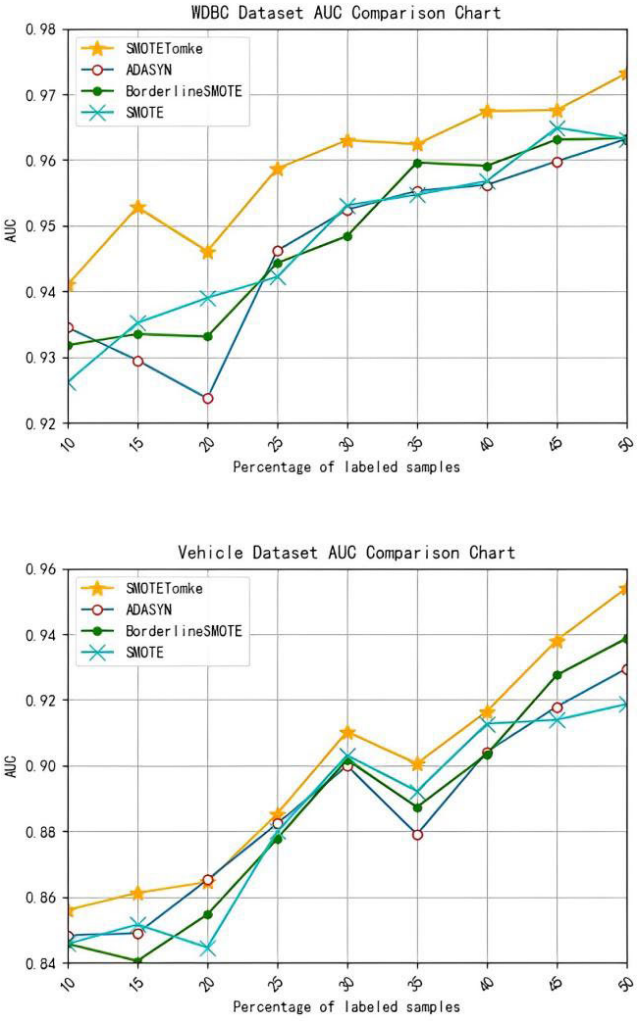

FIGURE 5. Comparison of prediction AUC of tri-training after training by five kinds of enhanced data.

are SMOTE [26], Borderline-SMOTE [27], ADASYN [28] and SMOTETomek [29].

According to (2), the JS distance between the generated data of the four methods and the original data is shown in the TABLE 3.

According to (2) and (3), the JS distance between the diversity of the data generated by the four methods and the diversity of the original data is shown in the TABLE 4.

It can be seen from TABLE 3 and TABLE 4 that compared with other enhanced data, the enhanced data of SMOTETomek and Borderline-SMOTE have the same similarity to the original data. However, when comparing similarity with the diversity of the original sample, the information entropy value of enhanced data of SMOTETomek is the most similar to that of the original data. Based on the above results, the enhanced data of SMOTETomek can be considered the best enhanced data.

In order to verify the effectiveness of the proposed quantitative enhanced data evaluation criteria, the above oversampling methods are respectively used to enhance training

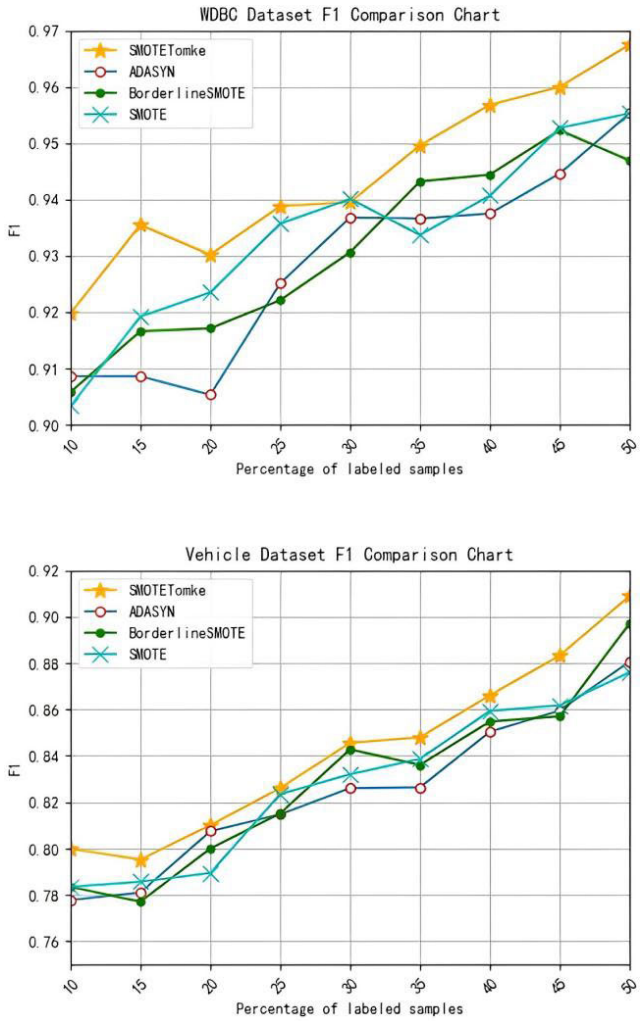

FIGURE 6. Comparison of prediction F-measure of tri-training after training by five kinds of enhanced data.

data, and the training data after enhancement is used to train the tri-training. The experimental datasets are WDBC and Vehicle datasets. In order to evaluate the effect of each kind of enhancement data on the prediction performance of tri-training when dealing with different proportions of labeled samples, the labeled samples of the experimental datasets are set with different proportions. The experimental results are shown in FIGURE 4 to FIGURE 6.

From FIGURE 4 to FIGURE 6, it can be seen that compared with other listed enhanced data, SMOTETomek's enhanced data can better improve the prediction performance of tri-training. This result proves the effectiveness of the proposed quantitative enhanced data evaluation criteria.

Combining the results of the above experiments, we believe that SMOTETomek is more suitable for data enhancement.

D. PERFORMANCE COMPARISONS BETWEEN PROPOSED MODEL AND OTHER SEMI-SUPERVISED METHODS

In order to evaluate the prediction performance of the proposed model, it is compared with the original XGBoost model [30], tri-training model [7], tri-training model with local convex optimization(Abbreviated to TRLOC) [13], tri-training model with Gaussian process(Abbreviated to TRGP) [18], and semi-supervised gaussian processes active learning model based on tri-training and improved QBC model without data enhancement(Abbreviated to TRGPQ) [3]. In order to obtain a comprehensive comparison of the performance of the listed semi-supervised models, different proportions of labeled samples in the whole experimental samples are set. We try to explore the predictive ability of the listed models when the ratios of labeled and unlabeled samples are different.

Aiming to compare the prediction ability and the ability to deal with practical problems of listed models more clearly,

TABLE 5. Accuracy, AUC and f-measure of proposed models on heart disease datasets in cross validation experiment.

Percentage of labeled samples						
Accuracy	1-Fold Cross	2-Fold Cross	3-Fold Cross	4-Fold Cross	5-Fold Cross	Average
10	0.8127	0.7753	0.8015	0.7491	0.7154	0.7708
15	0.746	0.8333	0.8214	0.7143	0.7738	0.7778
20	0.8186	0.827	0.8143	0.7257	0.7511	0.7873
25	0.8063	0.7973	0.7973	0.7703	0.7973	0.7937
30	0.7778	0.7633	0.8164	0.8164	0.7681	0.7884
35	0.8342	0.8031	0.8135	0.8187	0.8135	0.8166
40	0.7809	0.8427	0.8202	0.7753	0.8202	0.8079
45	0.8344	0.7914	0.816	0.8405	0.8405	0.8246
50	0.8041	0.7568	0.8041	0.7635	0.7905	0.7838
AUC	1-Fold Cross	2-Fold Cross	3-Fold Cross	4-Fold Cross	5-Fold Cross	Average
10	0.8086	0.772	0.7923	0.743	0.7103	0.7652
15	0.7441	0.8239	0.8176	0.7033	0.766	0.7710
20	0.8143	0.8233	0.8138	0.7447	0.7344	0.7861
25	0.8002	0.795	0.797	0.7639	0.7934	0.7899
30	0.774	0.7557	0.8098	0.812	0.7677	0.7838
35	0.8289	0.7997	0.8126	0.8168	0.7995	0.8115
40	0.7745	0.8374	0.8196	0.776	0.8138	0.8043
45	0.8331	0.794	0.812	0.84	0.8501	0.8258
50	0.8026	0.7613	0.792	0.7647	0.7799	0.7801
F-measure	1-Fold Cross	2-Fold Cross	3-Fold Cross	4-Fold Cross	5-Fold Cross	Average
10	0.7881	0.75	0.758	0.7074	0.6752	0.7357
15	0.7193	0.7941	0.7982	0.6505	0.7324	0.7389
20	0.7943	0.8019	0.8	0.751	0.6667	0.7628
25	0.7749	0.7783	0.7907	0.733	0.7716	0.7697
30	0.7356	0.72	0.7841	0.7865	0.7474	0.7547
35	0.8072	0.7738	0.8	0.8	0.7632	0.7888
40	0.7417	0.8082	0.8118	0.7531	0.7895	0.7809
45	0.8188	0.7848	0.7887	0.8243	0.8289	0.8091
50	0.7883	0.7143	0.7563	0.7445	0.7438	0.7494

TABLE 6. Comparison of accuracy of listed models on two artificial datasets.

Percentage of labeled samples						
Artificial Disease 1	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.8381	0.8291	0.814	0.8336	0.8353	**0.8434**
15	0.8491	0.8689	0.8415	**0.8774**	0.8717	0.8415
20	0.853	0.868	0.858	0.864	0.869	**0.87**
25	0.8909	0.9101	0.8941	0.9102	0.9091	**0.9166**
30	0.904	0.9143	0.8986	0.9154	0.9143	**0.9177**
35	0.9	0.9049	0.8963	0.9025	0.9062	**0.9099**
40	0.896	0.9075	0.8888	0.9075	0.9061	**0.9115**
45	0.8890	0.8905	0.8949	0.8861	0.8905	**0.9051**
50	0.9328	0.9312	0.9184	0.9312	0.9312	**0.9344**
Average	0.8837	0.8916	0.8783	0.892	0.8926	**0.8945**
Artificial Disease 2	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.8756	0.8644	0.8667	0.8733	0.8644	**0.8978**
15	0.8824	0.8871	0.8965	0.8965	0.8871	**0.8992**
20	0.9075	0.915	0.9075	0.92	0.915	**0.9205**
25	0.9093	0.904	0.896	0.912	0.904	**0.912**
30	0.8971	**0.9171**	0.8943	**0.9171**	**0.9171**	0.9057
35	0.9077	0.9108	0.9046	**0.9169**	0.9108	**0.9169**
40	0.8833	0.9067	0.9067	0.9133	0.9067	**0.92**
45	0.9091	0.9091	0.9055	0.92	0.9091	**0.9236**
50	0.9	0.912	**0.916**	**0.916**	0.9120	0.9
Average	0.8969	0.9029	0.8993	0.9095	0.9029	**0.9106**

TABLE 7. Comparison of AUC of listed models on two artificial datasets.

Percentage of labeled samples						
Artificial Disease 1	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.732	0.7403	0.7078	0.7486	0.7442	**0.7964**
15	0.7167	0.7432	0.7228	0.7431	0.7364	**0.7962**
20	0.7748	0.7823	0.7687	0.7708	0.783	**0.8224**
25	0.8039	0.8251	0.8094	0.8199	0.8244	**0.8474**
30	0.8365	0.8566	0.8378	0.8573	0.8566	**0.8864**
35	0.823	0.8359	0.8143	0.8269	0.8389	**0.8663**
40	0.8188	0.8354	0.8121	0.8354	0.8345	**0.8687**
45	0.8043	0.796	0.8084	0.7856	0.796	**0.852**
50	0.8806	0.8823	0.8694	0.8793	0.8823	**0.9059**
Average	0.7989	0.8108	0.7945	0.8074	0.8107	**0.8491**
Artificial Disease 2	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7401	0.7492	0.7264	0.7387	0.7492	**0.8146**
15	0.7937	0.8052	0.8111	0.8026	0.8052	**0.8182**
20	0.8269	0.8592	0.8269	0.8669	0.8592	**0.8804**
25	0.8365	0.823	0.8231	0.8535	0.823	**0.8637**
30	0.8232	0.8518	0.8107	0.8518	0.8518	**0.8661**
35	0.8234	0.8253	0.8089	0.8354	0.8253	**0.8606**
40	0.79	0.8242	0.8242	0.8349	0.8242	**0.8849**
45	0.8647	0.8503	0.8409	0.8571	0.8503	**0.9024**
50	0.8485	0.8395	0.842	0.842	0.8395	**0.8648**
Average	0.8163	0.8253	0.8127	0.8314	0.8253	**0.8617**

we conducted experiments on the artificial datasets and UCI datasets respectively. In order to obtain an overall evaluation effect and analyze model performance statistically, we calculated the average of the prediction results of each model after dealing with each dataset.

1) CROSS VALIDATION

Considering the small number of samples, aiming to avoid experimental errors, cross validation technology is applied in our semi-supervised model comparison experiment [32]. When setting different proportions of labeled samples and unlabeled samples for the experiment, each dataset is randomly assigned to labeled samples and unlabeled samples five times. We take the average of the five prediction results of the model as the overall prediction result of a proportion.

As an example, TABLE 5 shows the cross validation experimental prediction results and their average result of the proposed model on the heart disease dataset.

2) EXPERIMENTAL RESULTS OF ARTIFICIAL DATASETS

We set the percentage of labeled samples to unlabeled samples from 10% to 50%, with each percentage increasing by 5%. The accuracy comparison of listed models on two artificial datasets is shown in TABLE 6.

The AUC comparison of listed models on the two artificial datasets is shown in TABLE 7.

The F-measure comparison of listed models on the two artificial datasets is shown in TABLE 8.

As shown in TABLE 6 to TABLE 8, it can be observed from the experimental results of two listed artificial datasets that when the number of labeled samples in datasets is too small, the original model will not obtain better prediction effect. Whereas the semi-supervised model can significantly improve the prediction effect when dealing with such datasets, which proves the effectiveness of semi-supervised learning. Among them, the proposed model has the best average prediction effect on listed datasets, which shows that the

TABLE 8. Comparison of F-measure of listed models on two artificial datasets.

Percentage of labeled samples						
Artificial Disease 1	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.5614	0.5764	0.5252	0.5853	0.5852	**0.6319**
15	0.5345	0.58	0.5298	0.5893	0.5745	**0.6423**
20	0.6253	0.6432	0.6233	0.618	0.645	**0.6842**
25	0.6896	0.7376	0.7015	0.7343	0.7351	**0.7708**
30	0.7431	0.7755	0.7373	0.7777	0.7755	**0.7976**
35	0.7382	0.7542	0.7225	0.7432	0.7578	**0.7821**
40	0.724	0.7559	0.71	0.7559	0.7529	**0.7846**
45	0.6932	0.6877	0.7051	0.6725	0.6877	**0.7519**
50	0.8253	0.8238	0.7945	0.8227	0.8238	**0.8405**
Average	0.6816	0.7038	0.6721	0.6999	0.7042	**0.7429**
Artificial Disease 2	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.6267	0.6258	0.6	0.6225	0.6258	**0.7294**
15	0.6914	0.7073	0.725	0.7179	0.7073	**0.7273**
20	0.7517	0.7848	0.7517	**0.7975**	0.7848	0.7609
25	0.7571	0.7391	0.7273	0.7724	0.7391	**0.7785**
30	0.7313	**0.782**	0.7176	**0.782**	**0.782**	0.7724
35	0.7368	0.7434	0.7207	0.7611	0.7434	**0.7769**
40	0.6789	0.7407	0.7407	0.7593	0.7407	**0.8**
45	0.7706	0.7619	0.75	0.7843	0.7619	**0.8142**
50	0.7423	0.7556	**0.764**	**0.764**	0.7556	0.7525
Average	0.7208	0.7378	0.7219	0.7512	0.7378	**0.768**

prediction ability of the proposed model is generally better than that of other listed models. Especially on AUC, the proposed model has more obvious advantages than other models, showing that the semi-supervised classifier after training by enhanced data has the same strong prediction ability for both majority samples and minority samples.

3) EXPERIMENTAL RESULTS OF UCI DATASETS

We set the percentage of labeled samples to unlabeled samples from 10% to 50%, with each percentage increasing by 5%. The accuracy comparison of listed models on the five UCI datasets is shown in TABLE 9.

The AUC comparison of listed models on the five UCI datasets is shown in TABLE 10.

The F-measure comparison of listed models on the five UCI datasets is shown in TABLE 11.

TABLE 9. Comparison of accuracy of listed models on five uci datasets.

Percentage of labeled samples						
CMC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7525	0.7083	0.6820	0.7217	0.7544	**0.7641**
15	0.7395	0.6841	0.6554	0.7105	0.7591	**0.7812**
20	0.7551	0.6989	0.6902	0.7158	0.7605	**0.7662**
25	0.7497	0.7008	0.6828	0.7076	**0.7636**	0.7456
30	0.7541	0.6869	0.6628	0.7138	**0.7779**	0.7569
35	0.7491	0.6810	0.6569	0.6989	0.7721	**0.7732**
40	0.7405	0.6922	0.6833	0.7147	**0.7618**	0.7384
45	0.7578	0.6844	0.6681	0.7059	**0.7754**	0.7538
50	0.7473	0.6701	0.6498	0.6937	**0.7657**	0.7488
Vehicle	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	**0.8274**	0.8126	0.8224	0.8147	0.8147	0.8204
15	0.8558	0.8677	0.864	**0.8721**	0.8714	0.8691
20	**0.8814**	0.8712	0.8538	0.8751	0.8719	0.868
25	0.9089	0.8987	0.892	0.8971	0.8979	**0.9097**
30	0.8959	0.8805	0.8833	0.8914	0.8887	**0.9086**
35	0.9146	0.9	0.8932	0.9126	0.901	**0.9184**
40	0.9074	0.9091	0.9042	0.9126	**0.9126**	0.9079
45	0.9126	0.9138	0.9126	0.9172	0.9161	**0.9167**
50	0.9215	0.9139	0.8949	0.9165	0.9152	**0.9241**
WDBC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.9363	0.941	0.9391	**0.9437**	0.9422	0.9391
15	**0.9453**	0.9412	0.9416	0.9428	0.9437	0.942
20	0.9468	0.9516	0.9477	0.9543	0.9543	**0.9578**
25	0.9376	0.9432	0.9408	0.9469	**0.9484**	**0.9484**
30	0.9512	0.9472	0.9472	0.9533	0.9518	**0.9528**
35	0.9588	0.9566	0.9566	0.9615	**0.9621**	0.9593
40	**0.9601**	0.9548	0.9566	0.956	0.956	0.9589
45	0.9583	0.9609	0.9615	**0.9667**	0.9602	0.9654
50	0.9585	0.9584	0.9556	0.9577	0.957	**0.9606**
Diabetes	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7132	0.7308	0.6831	**0.7331**	0.7314	0.7308
15	0.7184	0.7307	0.6853	0.7294	0.7402	**0.7451**
20	0.741	**0.7583**	0.7052	0.7498	0.7531	0.7476
25	0.717	0.7281	0.6774	0.7212	0.7285	**0.7403**
30	0.7386	0.7438	0.6998	**0.7475**	0.7464	0.7374

TABLE 9. *(Continued.)* **Comparison of accuracy of listed models on five uci datasets.**

35	0.7387	0.7507	0.7074	0.7547	**0.7551**	**0.7551**
40	0.7295	0.7456	0.6917	0.7413	0.7465	**0.7516**
45	0.736	0.7407	0.7019	0.7398	**0.7460**	0.7417
50	0.7396	**0.7484**	0.7016	0.7412	0.7448	0.7453
Heart Disease	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7191	0.7334	0.6974	0.7423	0.7438	**0.7708**
15	0.7659	0.7746	0.7452	**0.7825**	0.7809	0.7778
20	0.7679	0.7713	0.7367	0.7747	0.7738	**0.7873**
25	0.7676	0.7865	0.7423	0.7919	0.7892	**0.7937**
30	0.7826	**0.7894**	0.7478	0.7884	0.7894	0.7884
35	0.7938	0.7969	0.7606	0.8031	0.8072	**0.8166**
40	0.7967	0.8034	0.7584	0.8056	0.7989	**0.8079**
45	0.7914	0.811	0.7656	0.8061	0.8086	**0.8246**
50	**0.7973**	0.7892	0.7514	0.7797	0.7879	0.7838

Summarizing the experimental results in TABLE 9 to TABLE 11, the average accuracy, AUC and F-measure of the five UCI datasets are shown in TABLE 12.

In order to verify the significance of the experimental results in TABLE 12, we introduce the Friedmanchisquare test for statistical analysis. Friedmanchisquare test is often used to examine whether the performance of different models is the same in machine learning. In Friedmanchisquare test, models with the same performance are considered to have the same rank value [32]. We assume that k models are compared on N datasets, and r represents the rank value of the i-th model. Assuming that the rank value of each model follows a normal distribution, the corresponding chi-square statistic is:

$$\tau_{x^2} = \frac{k-1}{k} \times \frac{12N}{k^2} \sum_{i=1}^{k} \left(r_i - \frac{k+1}{2} \right)^2 \tag{13}$$

where, τ_{x^2} obey the chi-square distribution with k-1 degree of freedom. Models that statistic exceed the statistical threshold of chi-square distribution and have a P-value is less than 0.05 can be considered to have significant differences. According to chi-square distribution table, in our experiment, the statistical threshold is 2.711.

TABLE 13 shows the statistics and P-values of the accuracy, AUC and F-measure in TABLE 12.

As shown in TABLE 13, the statistical values of accuracy, AUC and F-measure in Table 12 are greater than the statistical threshold, while the P-values is less than 0.05. The result in TABLE 13 indicates that the listed models have significant differences. The performance of these listed models is significantly different.

Further analyzing the experimental results, as shown in TABLE 9 to TABLE 12, the proposed model outperforms

TABLE 10. **Comparison of AUC of listed models on the five uci datasets.**

Percentage of labeled samples						
CMC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.6238	0.5064	0.5056	0.5051	0.5397	**0.6501**
15	**0.626**	0.4896	0.4868	0.4977	0.5691	0.6252
20	0.6182	0.5014	0.5096	0.5078	0.5611	**0.6193**
25	0.6234	0.5091	0.4981	0.4996	0.5716	**0.6473**
30	0.6303	0.494	0.4921	0.4957	0.5829	**0.682**
35	0.6422	0.494	0.4809	0.4944	0.5909	**0.6869**
40	0.6121	0.5095	0.5068	0.5057	0.5696	**0.6352**
45	0.6482	0.4991	0.4882	0.5008	0.5925	**0.652**
50	0.6372	0.5064	0.4841	0.5105	0.611	**0.6583**
Vehicle	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7963	0.7849	0.8008	0.7892	0.7915	**0.8048**
15	0.8413	0.8619	0.8543	0.8619	0.8679	**0.8728**
20	0.865	0.854	0.841	0.8587	0.855	**0.8698**
25	0.9005	0.8876	0.8828	0.8804	0.8863	**0.9138**
30	0.8704	0.8513	0.8582	0.8635	0.8615	**0.9005**
35	0.8979	0.8759	0.8712	0.8911	0.8758	**0.9139**
40	0.8892	0.8911	0.8868	0.8975	0.8974	**0.9042**
45	0.8984	0.9044	0.9064	0.9088	0.9061	**0.915**
50	0.9143	0.9024	0.8838	0.9046	0.9027	**0.9258**
WDBC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.9291	0.9368	0.935	0.9386	0.9353	**0.9396**
15	**0.9401**	0.9383	0.9382	0.9399	0.941	0.9397
20	0.9439	0.9464	0.9454	0.9483	0.9489	**0.9535**
25	0.9321	0.937	0.9361	0.9406	0.9414	**0.9454**
30	0.9484	0.9426	0.942	0.9488	0.9467	**0.9493**
35	0.9541	0.9527	0.9524	0.9564	0.9563	**0.9566**
40	0.9525	0.9478	0.9491	0.9485	0.9483	**0.9531**
45	0.9568	0.9585	0.9592	0.9627	0.9568	**0.966**
50	0.9564	0.9541	0.9521	0.9536	0.9517	**0.9576**
Diabetes	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.6687	0.6846	0.6382	0.6814	0.6793	**0.6943**
15	0.6898	0.6897	0.6531	0.6834	0.6945	**0.7228**
20	0.707	0.7128	0.6664	0.697	0.7008	**0.7222**
25	0.6828	0.6706	0.6309	0.6596	0.6685	**0.7071**

TABLE 10. *(Continued.)* **Comparison of AUC of listed models on the five uci datasets.**

30	0.7101	0.7043	0.6654	0.7074	0.7077	**0.7206**
35	0.6965	0.6979	0.6575	0.6946	0.6998	**0.7214**
40	0.6987	0.7104	0.6632	0.7038	0.7091	**0.7323**
45	0.7015	0.6906	0.6598	0.6851	0.6958	**0.7152**
50	0.7107	0.7121	0.6710	0.6992	0.7065	**0.7236**
Heart Disease	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7206	0.731	0.6961	0.7399	0.7412	**0.7652**
15	0.7657	0.774	0.7432	**0.7800**	0.7791	0.771
20	0.7715	0.7736	0.739	0.7761	0.7752	**0.7861**
25	0.7649	0.782	0.7384	0.7872	0.7837	**0.7899**
30	0.7813	0.7879	0.7452	0.7844	**0.788**	0.7838
35	0.7899	0.7926	0.7575	0.7982	0.803	**0.8115**
40	0.7955	0.8011	0.7552	0.8017	0.7956	**0.8043**
45	0.7922	0.8124	0.7668	0.8035	0.8056	**0.8258**
50	**0.7945**	0.7861	0.7486	0.777	0.7868	0.7801

other listed models in most cases, which shows that the proposed model can be effectively applied to practical problems. Precisely, by comparing the average predicted result of each model, it can be seen from TABLE 12 that the proposed model obtains higher average accuracy and average AUC than other listed models. This excellent prediction result is attributed to the strong classification prediction ability of the proposed model. And the proposed model performs much better especially on AUC, which means the proposed model can effectively predict both the majority samples and the minority samples. In addition, TABLE 12 exhibits that on F-measure, the proposed model also performs much better than other listed models. The highest F-measure of proposed model can be attributed to data enhancement by SMOTETomek. The F-measure result shows that the proposed model has better generalization ability than other listed models when dealing with practical problems.

We think the reason why the proposed model performs better in the experiment than other listed models is that the proposed model combines the advantages of two main semi-supervised methods, and strengthens the prediction ability of each module. When the proposed model is in the training process, the enhancement of training data increases the generalization ability of the model, so that the model has similar prediction ability for positive and negative samples, therefore the proposed model exhibits the best AUC and F-measure in the experiment. In the iterative process of the proposed model, semi-supervised learning and active learning are combined to make the model both robust and diverse. At the same time, the feature assignment mechanism is not only

TABLE 11. **Comparison of F-measure of listed models on the five uci datasets.**

Percentage of labeled samples						
CMC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.8424	0.8236	0.8015	0.8345	**0.8552**	0.8483
15	0.8306	0.8053	0.7806	0.8259	0.8548	**0.8654**
20	0.8454	0.8168	0.808	0.8295	0.8542	**0.8573**
25	0.8401	0.8168	0.8036	0.8239	**0.8585**	0.8333
30	0.8432	0.8075	0.7869	0.8292	**0.8676**	0.8381
35	0.8369	0.8027	0.7841	0.8175	**0.8624**	0.8501
40	0.8342	0.8096	0.8025	0.8288	**0.8571**	0.8285
45	0.8437	0.8044	0.7922	0.8217	**0.8646**	0.8405
50	0.8361	0.7911	0.7771	0.8110	**0.8551**	0.8349
Vehicle	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7121	0.688	0.7181	0.6943	0.6942	**0.7183**
15	0.7741	0.7984	0.791	0.7956	0.8038	**0.8076**
20	**0.8079**	0.7918	0.7695	0.7977	0.7927	0.8019
25	0.8591	0.8416	0.8332	0.8359	0.8403	**0.8667**
30	0.8323	0.8061	0.8129	0.8245	0.8209	**0.8617**
35	0.864	0.8366	0.8275	0.8581	0.8378	**0.8764**
40	0.8485	0.8475	0.8447	**0.8591**	0.8589	0.8559
45	0.8558	0.8596	0.8598	0.8654	0.8627	**0.8705**
50	0.875	0.8625	0.8336	0.8659	0.8642	**0.8836**
WDBC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.9142	0.9212	0.9189	**0.9245**	0.9217	0.9206
15	**0.9267**	0.9225	0.9228	0.9245	0.9258	0.9236
20	0.929	0.9344	0.9307	0.9377	0.9377	**0.9428**
25	0.9178	0.9247	0.9224	0.9296	0.9311	**0.9328**
30	0.9346	0.9288	0.9284	**0.9369**	0.9347	0.9366
35	0.9436	0.9407	0.9406	0.9471	0.9477	**0.9445**
40	**0.945**	0.938	0.9406	0.9396	0.9394	0.9443
45	0.9434	0.9466	0.9474	**0.954**	0.9455	0.9531
50	0.9442	0.9436	0.9402	0.9427	0.9415	**0.9468**
Diabetes	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.5546	0.5745	0.5092	0.5659	0.564	**0.5926**
15	0.5953	0.5866	0.5467	0.5758	0.5904	**0.6385**
20	0.6114	0.6155	0.5565	0.5896	0.5955	**0.6335**
25	0.5828	0.5449	0.5038	0.5242	0.5397	**0.6194**

TABLE 11. *(Continued.)* **Comparison of F-measure of listed models on the five uci datasets.**

30	0.6194	0.6043	0.5599	0.6099	0.609	**0.6348**
35	0.5968	0.5881	0.5382	0.584	0.5935	**0.6332**
40	0.6015	0.614	0.5565	0.6028	0.6106	**0.6464**
45	0.6057	0.5826	0.5464	0.5725	0.5895	**0.6254**
50	0.6258	0.6234	0.5716	0.6028	0.6149	**0.644**
Heart Disease	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
10	0.7086	0.7068	0.674	0.7179	0.7185	**0.7357**
15	0.7479	0.7566	0.7201	0.7584	**0.7594**	0.7389
20	0.7581	0.7571	0.724	0.7592	0.7573	**0.7628**
25	0.745	0.7583	0.7106	0.7625	0.7565	**0.7697**
30	0.7572	**0.7642**	0.714	0.7568	0.7631	0.7547
35	0.7668	0.7683	0.7301	0.7733	0.7794	**0.7888**
40	0.773	0.7779	0.7246	0.777	0.7698	**0.7809**
45	0.7753	0.7944	0.7492	0.7802	0.7811	**0.8091**
50	**0.7734**	0.7599	0.727	0.7488	0.7618	0.7494

TABLE 12. **Comparison of average accuracy, AUC and F-measure of listed models on the five uci datasets.**

Average accuracy	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
CMC	0.7495	0.6896	0.6701	0.7092	**0.7656**	0.7587
Vehicle	0.8917	0.8853	0.88	0.8899	0.8877	**0.8937**
WDBC	0.9503	0.9505	0.9496	0.9537	0.9529	**0.9538**
Diabetes	0.7302	0.7419	0.6948	0.7398	0.7436	**0.7439**
Heart Disease	0.7758	0.784	0.745	0.786	0.7866	**0.7945**
Average AUC	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
CMC	0.629	0.5011	0.4947	0.5019	0.5765	**0.6507**
Vehicle	0.8748	0.8682	0.865	0.8729	0.8716	**0.8912**
WDBC	0.9459	0.946	0.9455	0.9486	0.9474	**0.9512**
Diabetes	0.6962	0.697	0.6562	0.6902	0.6958	**0.7177**
Heart Disease	0.7751	0.7823	0.7433	0.7831	0.7842	**0.7909**
Average F-measure	original mode[30]	tri-training model[7]	TRLOC [13]	TRGP [18]	TRGPQ [3]	Proposed model
CMC	0.8392	0.8086	0.7929	0.8247	**0.8592**	0.8437
Vehicle	0.8254	0.8147	0.81	0.8218	0.8195	**0.8381**
WDBC	0.9332	0.9334	0.9324	0.9374	0.9361	**0.9383**
Diabetes	0.5993	0.5927	0.5432	0.5808	0.5897	**0.6294**
Heart Disease	0.7561	0.7604	0.7193	0.7593	0.7608	**0.7656**

used to strengthen the prediction ability of semi-supervised learning, the JS range is also used to avoid the impact of the uncertainty of active learning on prediction. These methods that enhance the prediction ability in each module make the proposed model perform better in the experiments than other listed models.

From the above experiments and analysis, compared with the original model and other listed semi-supervised models, the proposed model is the most effective semi-supervised learning method.

TABLE 13. **Statistics and P-values of the accuracy, AUC and F-measure in TABLE 12.**

	Accuracy	AUC	F-measure
statistic/ P-value	22.8/ 1.39E-04	24/ 7.99E-05	22.93/ 1.31E-04

V. CONCLUSION

Aiming at the semi-supervision binary classification problem when dealing with an imbalanced small dataset, a semi-supervised Gaussian processes active learning model based on improved tri-training with data enhancement is proposed. This model selects the best enhancement samples according to a originally quantitative enhanced data evaluation criteria, and enhances the training data by these enhancement samples. Then proposes an improved tri-training based on the random forest's feature assignment to increase the robustness of this model. After that Gaussian processes is introduced in active learning to select the most informative unlabeled samples to increase the diversity of this model, and the distribution range of JS distance is proposed in active learning to help predict the most informative unlabeled samples. This model combines the advantages of tri-training and active learning so that it has stronger prediction ability than tri-training and its variants. Compared with several traditional semi-supervised models, the experimental results show that the proposed model is the most effective. However, since the model is composed of several classification modules and each sample is calculated in detail, the computational complexity is slightly higher. In future work, we will try to reduce the computational complexity of the model.

REFERENCES

[1] W. Xu, J. Tang, and H. Xia, "A review of semi-supervised learning for industrial process regression modeling," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 185–190.

[2] J. Zhao and N. Liu, "Semi-supervised classification based mixed sampling for imbalanced data," *Open Phys.*, vol. 17, no. 1, pp. 975–983, Dec. 2019.

[3] H. L. Xu, L. Y. Li, and P. S. Guo, "Semi-supervised active learning algorithm for SVMs based on QBC and tri-training," *J. Ambient Intell. Humanized Comput.*, vol. 12, pp. 8809–8822, Nov. 2020.

[4] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Jan. 1998, pp. 92–100.

[5] D. Dalva, U. Guz, and H. Gurkan, "Extension of conventional co-training learning strategies to three-view and committee-based learning strategies for effective automatic sentence segmentation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Aug. 2018, pp. 750–755.

[6] T. Zhu, Z. Weng, G. Chen, and L. Fu, "A hybrid deep learning system for real-world mobile user authentication using motion sensors," *Sensors*, vol. 20, no. 14, p. 3876, Jul. 2020.

[7] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.
[8] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: A brief introduction," *Frontiers Comput. Sci.*, vol. 13, no. 4, pp. 669–676, Aug. 2019.
[9] R. Meka, A. Alaeddini, S. Oyama, and K. Langer, "An active learning methodology for efficient estimation of expensive noisy black-box functions using Gaussian process regression," *IEEE Access*, vol. 8, pp. 111460–111474, 2020.
[10] Y. Zhang and S. Yan, "Semi-supervised active learning image classification method based on tri-training algorithm," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Syst. (ICAIIS)*, Mar. 2020, pp. 206–210.
[11] Y. Saito, "Asymmetric tri-training for debiasing missing-not-at-random explicit feedback," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 309–318.
[12] J. W. Mo and P. Jia, "Semi-supervised classification model based on ladder network and improved tri-training," *Acta Automatica Sinica*, vol. 48, no. 8, pp. 2088–2096, Nov. 2022.
[13] Y. Bhalgat, Z. Liu, P. Gundecha, J. Mahmud, and A. Misra, "Teacher-student learning paradigm for tri-training: An efficient method for unlabeled data exploitation," in *Proc. 15th Conf. Natural Lang. Process. (KONVENS)*, Sep. 2019, pp. 262–266.
[14] Y. Zhang, R. Cheng, and J. Zhang, "Safe tri-training algorithm based on cross entropy," *J. Comput. Res. Develop.*, vol. 58, no. 1, pp. 60–69, Apr. 2021.
[15] C.-M. Tseng, T.-W. Huang, and T.-J. Liu, "Data labeling with novel decision module of tri-training," in *Proc. 2nd Int. Conf. Comput. Commun. Internet (ICCCI)*, Jun. 2020, pp. 82–87.
[16] D. Gissin and S. Shalev-Shwartz, "Discriminative active learning," 2019, *arXiv:1907.06347*.
[17] B. Gu, Z. Zhai, C. Deng, and H. Huang, "Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4111–4122, Sep. 2020.
[18] J. Zhao, H. Wang, and Z. Cao, "Promoting active learning with mixtures of Gaussian processes," *Knowl.-Based Syst.*, vol. 188, Jan. 2020, Art. no. 105044.
[19] D. Mahapatra, A. Poellinger, and M. Reyes, "Graph node based interpretability guided sample selection for active learning," *IEEE Trans. Med. Imag.*, early access, Oct. 14, 2022, doi: 10.1109/TMI.2022.3215017.
[20] D. Lee, Y.-C. Chan, W. Chen, L. Wang, A. van Beek, and W. Chen, "T-METASET: Task-aware acquisition of metamaterial datasets through diversity-based active learning," *J. Mech. Design*, vol. 145, no. 3, Mar. 2023, Art. no. 031704.
[21] S. H. Ghafarian, "Local variational probabilistic minimax active learning," *Exp. Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118538.
[22] L. H. P. D. Silva, L. H. S. Mello, A. Rodrigues, F. M. Varejão, M. P. Ribeiro, and T. Oliveira-Santos, "Active learning for new-fault class sample recovery in electrical submersible pump fault diagnosis," *Exp. Syst. Appl.*, vol. 212, Feb. 2023, Art. no. 118508.
[23] D. Zang, J. H. Liu, and F. M. Qu, "Pipeline small leak detection based on virtual sample generation and unified feature extraction," *Measurement*, vol. 184, Aug. 2021, Art. no. 109960.
[24] H. Yanagaimoto and K. Hashimoto, "Review rating prediction with Gaussian process classification," in *Proc. 14th Int. Joint Symp. Artif. Intell. Natural Lang. Process. (iSAI-NLP)*, Oct. 2019, pp. 1–6.
[25] A. Klose and R. Kruse, "Semi-supervised learning in knowledge discovery," *Fuzzy Sets Syst.*, vol. 149, no. 1, pp. 209–233, Jan. 2005.
[26] N. V. Chawla and K. W. Bowyer, "SMOTE: Synthetic minority over sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[27] Y. Chen, R. Chang, and J. Guo, "Effects of data augmentation method borderline-SMOTE on emotion recognition of EEG signals based on convolutional neural network," *IEEE Access*, vol. 9, pp. 47491–47502, 2021.
[28] S. K. Satapathy, S. Mishra, P. K. Mallick, and G.-S. Chae, "ADASYN and ABC-optimized RBF convergence network for classification of electroencephalograph signal," *Pers. Ubiquitous Comput.*, vol. 9, pp. 47491–47502, Mar. 2021.
[29] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining SMOTE with tomek links technique for imbalanced medical data," in *Proc. IEEE Int. Conf. Online Anal. Comput. Sci. (ICOACS)*, May 2016, pp. 225–228.
[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[31] M. Aamir and S. Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 4, pp. 436–444, May 2019.
[32] N. Azman, A. Syarif, M.-E.-A. Brahmia, J.-F. Dollinger, S. Ouchani, and L. Idoumghar, "Performance analysis of RPL protocols in LLN network using Friedman's test," in *Proc. 7th Int. Conf. Internet Things, Syst., Manage. Secur. (IOTSMS)*, Dec. 2020, pp. 1–6.

CHENXIAO ZHOU received the bachelor's degree in automation from Wuchang Shouyi University, China, in 2020. He is currently pursuing the master's degree with the School of Electrical and Information Engineering, Wuhan Institute of Technology. His research interests include artificial intelligence, machine learning, few-shot learning, semi-supervised learning, and active learning.

LIANYING ZOU received the bachelor's degree in communication engineering and the master's and Ph.D. degrees in microelectronics and solid state electronics from the Huazhong University of Science and Technology, China, in 1998, 2003, and 2006, respectively. She is currently an Associate Professor with the School of Electrical and Information Engineering, Wuhan Institute of Technology. Her research interests include embedded system design, FPGA system design, and VLSI integrated circuit design. She has contributed to over 20 peer-reviewed publications in journals, such as *Journal of Huazhong University of Science and Technology* (Natural Science Edition) and the *Journal of China Universities of Posts and Telecommunications*.