IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 8, 2020, accepted January 14, 2021, date of publication February 8, 2021, date of current version February 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057693

An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients

MENG WANG[1], XINGHUA YAO[2], AND YIXIANG CHEN[1]
[1]East China Normal University, Shanghai 200241, China
[2]School of Basic Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Corresponding author: Xinghua Yao (xhyao@shutcm.edu.cn)

This work was supported in part by the National Key Research and Development Project of China under Grant 2018YFB2101300.

ABSTRACT Early predicting heart attack out of stroke patients in a view of data analysis is an approach to reduce a high mortality rate. Stroke-patient data in Intensive Care Unit are imbalanced due to that stroke patients with heart attack are in the minority of stroke patients. How to predict heart attack in the stroke-patient data becomes a challenge. For processing the imbalanced data, this paper designs an algorithm by leveraging random undersampling, clustering and oversampling techniques, which is called undersampling-clustering-oversampling algorithm (shortly, UCO algorithm). The UCO algorithm generates nearly balanced data which are utilized to train machine-learning models for predicting heart attack. Over the database of Medical Information Mart for Intensive Care III, extensive experiments are conducted to evaluate the UCO algorithm. A setting of undersampling number of 120 in the algorithm UCO, denoted UCO(120), shows good performance in helping machine-learning classifiers extract features. Five classifiers are separately deployed to predict heart attack based on outputs of the UCO(120). Our results show that random forest classifier achieves the best predicting performance with an *accuracy* of 70.29%, and *precision* of 70.05%. It could be well-predicted using UCO(120) and random forest that whether a stroke patient will have heart attack or not.

INDEX TERMS Undersampling, clustering, oversampling, imbalanced data, stroke, heart attack.

I. INTRODUCTION

Stroke, also known as ''ischemic stroke'', refers to ischemic necrosis or softening of localized brain tissue caused by cerebral blood supply, ischemia and hypoxia. The main clinical manifestations are sudden collapse, mental coma, unclear speech, and hemiplegia [1]. Heart attack is a myocardial necrosis caused by acute and persistent ischemia and hypoxia of coronary artery which manifestations are arrhythmia, shock or heart failure, which can be fatal [2]. Stroke complicated with heart attack is cerebral infarction accompanied by heart attack. As we know, the stroke complicated by heart attack was 30%, and the mortality rate was as high as 54%[3].The main causes of death are ventricular arrhythmia, acute left heart failure and cardiogenic shock.

Troponin is an effective indication to detect heart attack [4]–[6]. In clinic, it is also commonly used. A drawback of troponin is that troponin starts changing just four hours after heart attack. There exists a time delay of four hours for the troponin changes that signify the happened heart attack. On the other side, the onset of heart attack is rapid, and sudden deaths easily happen on the heart attack patients. This paper attempts to predict heart attack for the stroke patients based on analyzing medical indications except the troponin. Such a prediction is to gain more treatment time for the stroke patients with heart attack. Normally the medical indications include 34 items such as Spo2, Resprate, Heartrate, Glucose, Creatinine, Sysbp. The result is bad when we use all of the medical indications to predict, so we select eight medical indications which are Heartrate_Min, Heartrate_Mean, Resprate_Max, Glucose_Min, Glucose_Max, Glucose_Mean, Creatinine, Sysbp. Based on the indication data, a data processing algorithm is developed to make prediction of heart attack for the stroke patients.

In clinical practice, stroke patients with heart attack are much less than stroke patients without heart attack. The two types of stroke patients form an imbalanced dataset. In the database Medical Information Mart for Intensive Care III (shortly, MIMIC III), there are 2,406 stroke samples without heart attack which is nearly 30 times more than 82 stroke samples with heart attack. Simply using machine learning

The associate editor coordinating the review of this manuscript and approving it for publication was Amir Masoud Rahmani.

methods usually makes a training procedure focus on features in the majority data samples, and the minority data are neglected. Experimental results of high accuracy and low precision usually occur [7].

In the existing studies, There are two common imbalanced data processing methods, i.e., oversampling and undersampling [8]. But for the imbalanced data of stroke samples, classic oversampling and undersampling algorithms can not produce well-balanced data for feature extraction. In this paper, we using resampling and clustering methods, use the UCO algorithm to process the data, and train the model to predict the possibility of stroke patients with heart attack.Our experimental results show that best in the performances among the predictions is Random Forest model with the 1-Recall of 75.59%, 0-Recall of 63.95%, Accuracy of 70.29%, and Precision of 70.05%. As a result, a stroke patient in Intensive Care Unit (shortly, ICU) whether have heart attack can be predicted according to the daily monitoring indicators in advance base on UCO(120) and Random Forest model.

To the best of our knowledge, there is no previous research on the prediction heart attack of stroke patients. A new algorithm is proposed to process imbalanced stroke-patient data. Our contributions in this study are as follows.

(1)A new algorithm is designed by leveraging undersampling, clustering, and oversampling techniques to process imbalanced stroke-patient data.

(2)A composition of UCO(120) and random forest classifier has good capability of predicting heart attack in stroke patients.

A combination of the proposed algorithm UCO(120) and machine-learning model of random forest (RF) can predict that whether ICU patients will have heart attack or not. Such a prediction helps physicians make precautions and intervene in advance. So the mortality rate of stroke patients with heart attack could be reduced. Because there are no available ICU data of Asian stroke patients at the present. The proposed approach is evaluated over the stroke-patients data in American, and it has potential in application to American medical system.

The rest of this paper is organized as follows. Section II analyzes related work including heart attack prediction and imbalanced data processing methods. In Section III, a data set of stroke patients is described and analyzed. Based on data analyses in stroke patients, a data-processing algorithm is designed for heart attack prediction in Section IV. Section V executes simulations to analyze the proposed algorithm and to evaluate the performances of the algorithm. Also, heart-attack-prediction capabilities of classical machine-learning models are explored based on outputs of the proposed algorithm in Section V. Section VI discusses our proposed method for heart attack prediction. Finally, Section VII concludes this study and presents future work.

II. RELATED WORK

The study of predicting heart attack in stroke patients is to reduce the mortality of stroke patients with heart attack. For the prediction problems, machine learning methods are extensively utilized in recent years. How to process well imbalanced data is a challenge in the prediction of heart attack in stroke patients. In this section, research work about heart attack prediction in stroke patients and imbalanced data processing are presented separately.

A. HEART ATTACK PREDICTION

There are many studies on the prediction of heart attack, which are mainly divided into two types: one is from the perspective of medicine [9]–[11], the other is using patient medical record information to predict the heart attack by the artificial intelligence technology [12].

With the perspective of medicine,Fenglan Zhang predicted heart attack by observing the early and late stage QTc dispersion (QTcd) of patients with severe heart attack [9]. By analyzing the changes of electrocardiogram QTcd in 44 patients with severe heart attack, they observed the difference of QTcd between the death group and the survival group and identified the high risk group of heart attack by the value of QTcd in the group. Yuejin Yang et al observed the improvement of segmental contractile function during exercise, and compared the viable myocardial segments detected by the experiment and calculated the Dob-2DE to predict the accuracy of spontaneous improvement of segmental systolic function in patients with severe heart attack [10]. Jianping Qi analyzed the high frequency electrocardiogram (HFECG) of 10 patients with heart attack [11]. It was found that the HFECO in the early first week of heart attack was significantly higher than the HFECG in the early first week, and HFECG had a certain predictive value for the occurrence of heart attack.

By using the artificial intelligence technology, Yaowang Lin et al used XGBoost to assess the risk of coronary heart disease complicated with heart attack [12]. The dataset is from the 4049 medical records obtained from Department of Cardiology, Shenzhen People's Hospital, Guangdong, the information was selected manually by the doctors and XGBoost was used to predict whether coronary heart disease could cause heart attack.

However, the current researches of heart attack prediction are not for a group of stroke patients. In such a stroke-patients group, patients with heart attack are much less than patients without heart attack. The available data is very imbalanced. On the other side, the analyzed medical indicators in the existing work are more than such indicators as heart rate, blood sugar, respiratory rate and blood pressure. In clinic, the above four medical indicators are common and easy to obtain. Based on such common medical indicators, developing methods for the heart attack prediction will be an improvement on the aspect of clinic convenience.

B. IMBALANCED DATA PROCESSING

Resampling techniques, including the undersampling method [13]–[15] and the oversampling method [16]–[20], are usually used to process imbalanced data.

Little M A came up with random undersampling algorithm to randomly select a certain amount of samples and

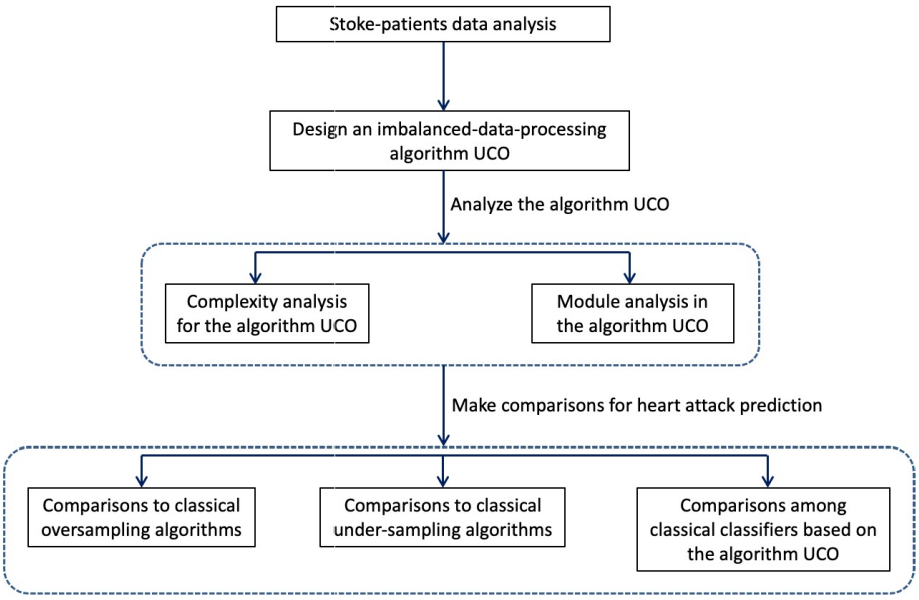


FIGURE 1. Workflow of this study.

eliminate [13], Qi Fan et al adopted all samples in the training process, and dynamically determined whether a majority sample should be used for the classifier learning and call the algorithm One-sided Dynamic Undersampling (ODU) algori thm [14], Shaik.Nagul et al designed an effective K-means undersampling algorithm for imbalance data using precise reduction sampling [15].

Oversampling's mainly methods contain random copying of original minority samples, SMOTE and ADASYN. Moreo A randomly copied some original minority samples and put them into the dataset, which made the two lables' samples number equal to each other [16]. Blagus R proposed an algorithm called SMOTE which improved random oversampling but its behavior on high-dimensional data has not been thoroughly investigated [17]. Qi W et al designed a new synthetic minority oversampling technique to incorporate the borderline information and called the algorithm Bagging of Extrapolation Borderline-SMOTE, it was proposed in dealing with imbalanced data learning problems [18], [19].The SMOTE algorithm can effectively reduce the overfitting and improve the generalization ability of the classifier.He H et al proposed ADASYN algorithm which reducing the bias introduced by the class imbalance and adaptively shifting the classification decision boundary toward the difficult examples [20].

Based on the above analyses of related work, we propose a workflow for the study of heart attack prediction over stroke patients. Our workflow is showed in FIGURE 1. In the workflow, analyzing stroke-patients data is firstly done. It is obtained that heart attack patients in the group of stroke patients are much less than the stroke patients without heart attack. For the imbalanced data, a new data-processing algorithm is designed. Related analyses are done in order to make clear role of each module and to measure complexity. Then, comparing to classical imbalanced-data-processing algorithms, effects of the proposed algorithm are explored. And performances of classical classifiers are investigated based on data-processing results of the proposed algorithm for the heart attack prediction.

III. DATASET ANALYSIS AND DATA PREPROCESSING

For evaluating the proposed algorithm UCO, the database MIMIC-III is taken as data source. MIMIC-III which was developed by MIT laboratory is a database of critical medicine. The database contains 26 CSV files collecting medical data from 46520 patients, in which 38606 patients are adults. For example, the CSV file of Chartevents records medical examination data, the file of Icustays collects time information about patients entering and leaving ICU.

PostSQL was used to get access to the database MIMIC-III. In that database d_icd_diagnoses chart record the disease categories of all patients. In this chart, the icd9_Code value of Cerebral Infarction is taken as filtering condition to select all the stroke samples. Select the samples which age is more than 18 and admission days are more than 12 hours to make further filter, and 2488 stroke samples are achieved. Among these samples, the indicator of troponin is utilized to determinate whether a stroke sample is with heart attack or not [23]–[25], The troponin value of 1 means stroke patients have heart attack, and 0 means stroke sample without heart attack. Then 2488 stroke samples are divided into 82 stroke samples with heart attack and 2406 samples without heart attack.

Statistics for missing data in the 2488 stroke samples are shown in FIGURE 2. Two items, i.e., Triglycerides and X.Hemoglobin.A1e, are deleted, which are in purple. Because they have too many missing datas. For each preserved medical indicator, the related missing data are filled

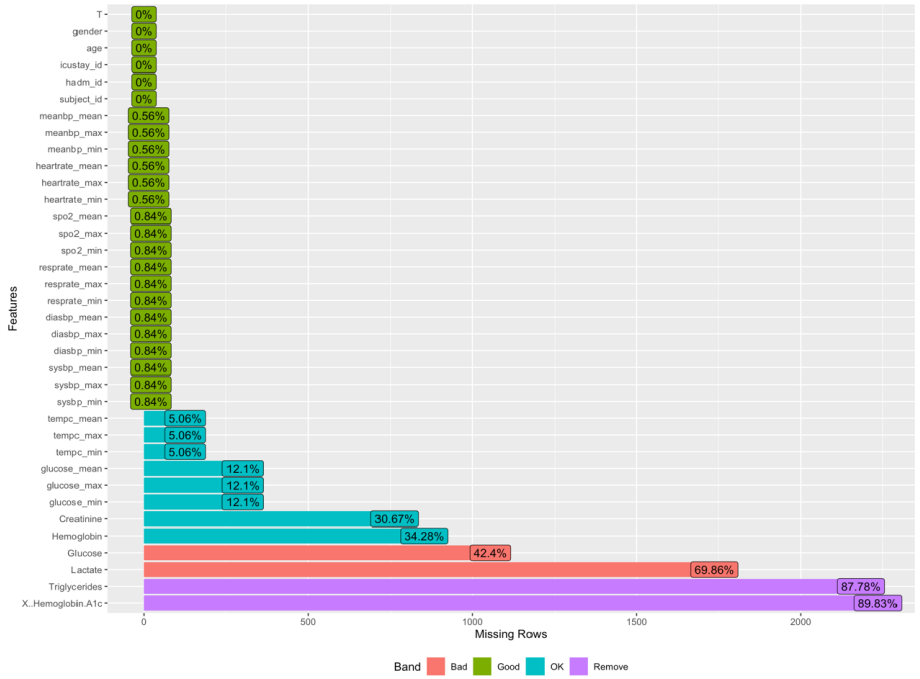


FIGURE 2. Proportion of the missing data.

in by using the average of the existing values of the indicator. So we can get a dataset which contains 2488 stroke patients and 34 medical indexes.

More medical indexes are not necessarily good for data analysis and making prediction. One disadvantage is that more computing cost is needed. Perturbations among indexes maybe exist. So T-test is conducted to select the features.

The influence of each medical index on Troponin is measured in T-test. First, two hypotheses was established: H0 : $\mu = \mu_0$, which assume that the samples are same. H1 : $\mu \neq \mu_0$, which assume that the samples are different. Then we use double sample T-test to test whether there is difference between the average and overall of two samples. The T value was calculated by formula (8) and the range of p was determined by the table of p-t [29]. If the p value was less than the threshold, H1 is right, indicating that the medical index is closely related to Troponin. It can predict the possibility of heart attack.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{1}$$

Among them, $\bar{X}_1$ and $\bar{X}_2$ are the average values of the two sets of samples, n_1 and n_2 are the sizes of the two sets of samples, and S_1 and S_2 are the sample standard deviations of the two sets of samples.

We use 0.05 as threshold and compare the p values of the 34 medical indexes with the threshold. Select the medical indexes that p value is less than the threshold and remove the medical indexes that p value is bigger than the threshold value.Then eight medical indexes which are Heartrate_Min, Heartrate_Mean, Resprate_Max, Glucose_Min, Glucose_Max, Glucose_Mean, Creatinine, Sysbp are selected.

IV. DESIGN OF UCO ALGORITHM

In this paper, UCO algorithm is designed to process imbalanced data of stroke patients.UCO algorithm can process the dataset of section 3 and output a nearly balanced dataset for training.

A. OVERVIEW OF UCO ALGORITHM

An algorithm is developed to process imbalanced stroke data by using resampling and clustering techniques. The resampling includes undersampling and oversampling. The undersampling is to filter out data from the majority samples. The oversampling is to expand the minority samples. The clustering is to group the data and make samples in each group have similar features.

FIGURE 3 presents main operations in the proposed algorithm. In the algorithm, undersampling operation is executed over the set of samples with majority labels. Clustering is conducted over the sample set of minority labels. The obtained sample set by the undersampling operation and each obtained cluster are separately divided into training set, validation set and testing set according to the ratio 8:1:1. The obtained training sets are merged. And so are done the achieved validation sets and testing sets. Then, the algorithm SMOTE is deployed to over-sample the samples with minority labels in the merged training set, and the final training set is obtained. In the final training set, the number of positive samples is nearly equal to

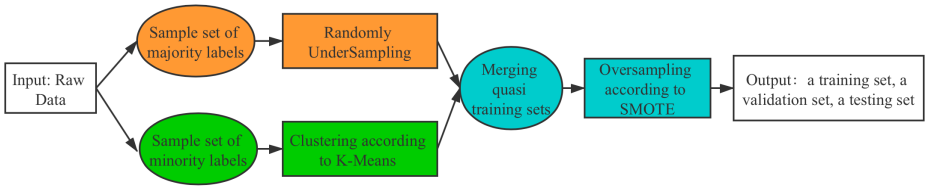


FIGURE 3. Overview of the UCO algorithm.

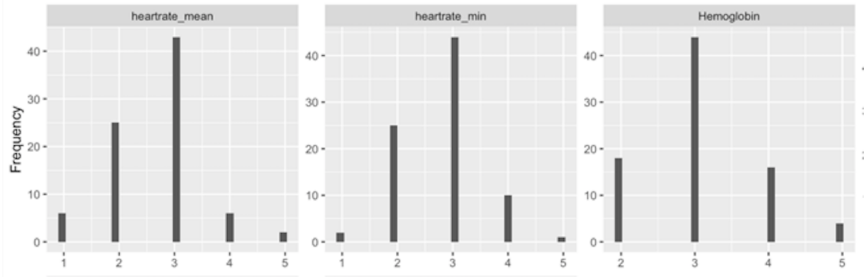


FIGURE 4. Data distribution of stroke patients with heart attack.

the number of negative samples, and each type of positive sample is included. Such a training set provides sufficient samples for training machine learning models to extract features. Finally, the algorithm outputs a nearly balanced training set, validation set and testing set. The proposed algorithm is called Undersampling-Clustering-Oversampling algorithm and it is also called UCO algorithm for being short.

UCO algorithm consists of three modules. Since the number of people with heart attack is much smaller than that without heart attack, the samples without heart attack are undersampled. The sample number selected by the random undersampling is 120. Because the prediction result obtained under the setting of the sampler number of 120 is the best as shown in FIGURE 9. On the other side, the data of patients with heart attack are relatively scattered, as shown in FIGURE 4. In order to achieve features of each group of close-distance samples when training, a clustering operation is executed over the samples with heart attack. Finally, in order to make the number of samples with heart attack nearly equal the number of samples without heart attack in the training procedure, the samples with heart attack are oversampled.

(1)Undersampling on the majority negative samples (stroke without heart attack samples)

The majority samples (i.e., stroke samples without heart attack) are randomly selected to generate a training subset, validation subset, and testing subset. The number of random pick samples is a parameter of the UCO algorithm in this paper, which is related to the dataset and the sample set of minority lable. The undersampling proportion of samples with heart attack to samples without heart attack in the MIMIC-III database is 2:3. For the MIMIC-III database's stroke samples dataset, that means random pick 2/3 samples from stroke without heart attack samples.

(2)Clustering the sample set of minority lable (Stroke with heart attack samples).

We cluster minority samples (i.e., stroke samples with heart attack) into several groups. Then partition each cluster into training set, validation set, and testing set according to the ratio 8:1:1. FIGURE 4 shows the data distribution of stroke patients with heart attack, clearly the samples are quite different. For example, for the indicator of 12-hour heart rate, the highest value is 126 and the lowest value is 44. In order to make the training set contain each type of stroke paitient samples, we use K-means algorithm to cluster the stroke patients data [26]. We calculate the mean square error (SSE) to get the number of K [27]. Equation (2) is the calculation of the SSE. C_i is the i-*th* cluster, P is the sample in C_i, and m_i is the center of C_i.

$$SSE = \sum_{p \in C_i} |p - m_i|^2 \tag{2}$$

For the sample of stroke patients, to select cluster number k, we calculate SSE and draw the "elbow map" in FIGURE 5. For K being less than 3, SSE will decreases significantly with that K increases. When k is over 3, SSE will decrease slightly with the increase of K. According to elbow rule, the clustering effect is the best at the elbow of cluster number of 3. So, the cluster number is selected to be 3.

(3)The partitioning and merging of clusters.

We separate The undersampled samples and each cluster into training sets, validation sets, and testing sets according to a ratio of 8:1:1 and merged them into a new training set, validation set, and testing set, respectively. The three new sets are quasi training set, quasi validation set, and quasi testing set.

(4)Oversampling the quasi training set by SMOTE

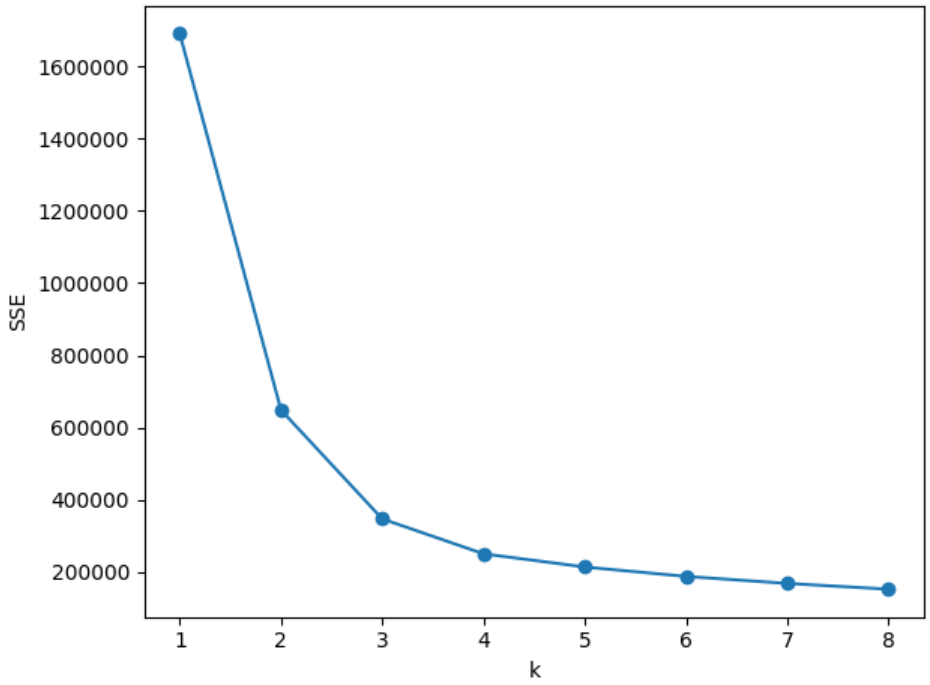


FIGURE 5. k-SSE Curve.

The SMOTE algorithm is a sampling algorithm for the sample set of minority label. It is improved from random sampling algorithm [28]. Input the quasi training set to the SMOTE algorithm. Using the algorithm, we randomly select samples with minority label, and chose the nearest neighbors. Then we generate new samples with minority label among the centering samples and the neighbors. The algorithm runs until the number of samples with minority label equals to the number of samples with majority label and output the new training set.

FIGURE 6 shows the flow chart of the data processing UCO algorithm, UCO algorithm's process as Algorithm 1.

Among them $C_i^{\text{train},1}$ is the training set of C_i. $C_i^{\text{valid},1}$ is the validation set of C_i. $C_i^{\text{test},1}$ is the testing set of C_i.

B. ALGORITHMIC COMPLEXITY

In the UCO algorithm, we conbine the random undersampling algorithm, K-means clustering algorithm and SMOTE algorithm. N is the total number of samples, n is the number of undersampling. In the UCO algorithm, the time complexity for the module of partitioning data set is $O(N)$. The time complexity of random undersampling is $O(n)$ and $O(N)$. The time comlexity of the K-Means clustering algorithm is $O(N)$. The time complexity of the SMOTE algorithm is $O(N)$. The time complexity of the oversampling algorithm is $O(\text{N}^2)$. So the time complexity of the UCO algorithm is $O(\text{N}^2)$.

Algorithm 1 UCO Algorithm

Require: Stroke patient data set; Number of undersampling samples n; Cluster number K;

Ensure: Train set X^{train}; validation set X^{valid}; Test set X^{test};

1: Divide all the samples according to stroke without heart attack dataset S_0 and stroke with heart attack dataset S_1;
2: Using random undersampling algorithm to select n samples from S_0 and put them into S_{00}.Then divide S_{00} into training set $X^{\text{train},0}$, validation set $X^{\text{valid},0}$, test set $X^{\text{test},0}$ according to ratio 8:1:1;
3: Using $\text{K} - \textit{Means}$ algorithm to clustering dataset S_1 into cluster $\text{C}_1, \text{C}_2, \cdots\cdots, \text{C}_\text{k}$. Then divide each C_i into train set $C_i^{\text{train},1}$, validation set $C_i^{\text{valid},1}$, test set $C_i^{\text{test},1}$ according to ratio 8:1:1;
4: Put dataset $X^{\text{train},0}$, $C_1^{\text{train},1}$, $C_2^{\text{train},1}$, $\cdots$, $C_k^{\text{train},1}$ into the quasi train set X^{pretrain}; Put dataset $X^{\text{valid},0}$, $C_1^{\text{valid},1}$, $C_2^{\text{valid},1}$, $\cdots$, $C_k^{\text{valid},1}$ into the validation set X^{valid}; Put dataset $X^{\text{test},0}$, $C_1^{\text{test},1}$, $C_2^{\text{test},1}$, $\cdots$, $C_k^{\text{test},1}$ into the test set X^{test};
5: Using SMOTE algorithm to oversampling the quasi train set X^{pretrain} to get train set X^{train};
6: **return** X^{train}. X^{valid}. X^{test}.

V. SIMULATION ANALYSIS

A. METRICS

We use *Accuracy*, *Precision*, 1_*Recall*, 0_*Recall*, $F1$_*Score*, *AUC* as classification metrics, and calculate the mean and mean square error of the 300 experiments' results as the measurement criteria [21].

Accuracy means the proportion of the number of correctly predicted samples to the total number of all the samples; *Precision* means the proportion of correctly predicted samples with heart attack to predicted samples with heart attack;

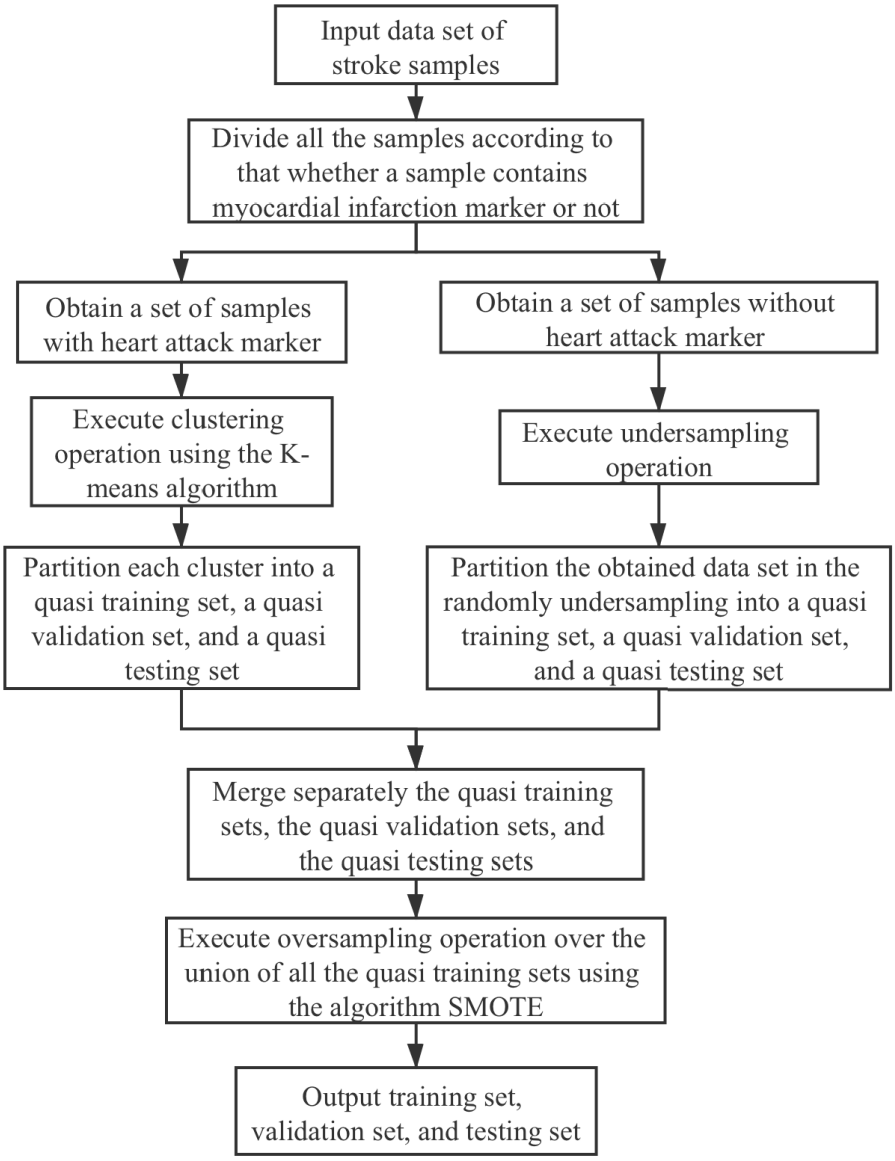


FIGURE 6. Flowchart of the UCO algorithm.

1_Recall describes that how many true heart attack samples are correctly predicted; 0_Recall measures that how many true non-heart-attack samples are correctly predicted. $F1\_Score$ is the harmonic mean of *Precision* and 1_Recall; *AUC* refers to the area under the receiver operating characteristic (shortly, ROC) curve. The ROC curve is a curve drawn by the 1_Recall as the ordinate and the false positive rate (*FPR*) as the abscissa. The metrics are separately calculated according to the following equations (3)–(8) [22]:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$1\_Recall = \frac{TP}{TP + FN} \tag{4}$$

$$0\_Recall = \frac{TN}{FP + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$F1\_Score = \frac{2 * Precison * 1\_Recall}{Precision + 1\_Recall} \tag{8}$$

TP (True Positive) is the number of correctly predicted heart attack in the result; FP (False Positive) is the number of incorrectly predicted as heart attack; TN (True Negative) is correctly predicted as being without heart attack; FN (False Negative) is the number that is incorrectly predicted as having no heart attack.

B. UCO MODULES SELECTION

UCO algorithm contains three modules, i.e., Undersampling module, Clustering module, and Oversampling module.

TABLE 1. Results using method UCO(120)+RF.

Metrics	Mean value	Mean square error
$Accuracy$	70.29%	0.0845
$Precision$	70.05%	0.1199
1_Recall	75.59%	0.1282
0_Recall	63.95%	0.1116
$F1\_Score$	0.6613	0.0919
AUC	0.6977	0.0831

[1] The metric of $Accuracy$ is calculated according to Equation (3), $Precision$ by using Equation (6), 1_Recall according to Equation (4), 0_Recall by using Equation (5), and $F1\_Score$ according to Equation (8). The metric of AUC is an area under the receiver operating characteristic (ROC) curve. The ROC curve is in a Cartesian coordinate taking 1_Recall as an ordinate and the false positive rate (FPR) as abscissa. Calculation of metrics in Tables 2-5 are the same as in Table 1.

In this section, we remove one of the modules and use models to predict in order to analyze the effect of data processing. It shows that using UCO(120) to process the data and RF to predict, the result is best.

1) RESULTS

Based on the selected medical indicators,UCO algorithm was used to obtain the training set, validation set and testing set. Then Random Forest algorithm is used to predict the risk of heart attack in stroke patients. For using the UCO algorithm, the number of samples in the undersampling need be fed. We use a symbol UCO(n) to denote that the undersampling module selects n samples. This paper choose 120 to be the UCO(n), UCO(120) outputs a training set, validation set, and testing set and put them into the model RF to do prediction. The whole procedure is called Undersampling-Clutering-Oversampling-RF (120), it can also be called UCO(120)+RF algorithm for short. Then we do RF for 300 times experiment. TABLE 1 shows the mean and mean square error of *Accuracy*, *Precision*, 0_*Recall*, 1_*Recall*, *F*1_*Score* and *AUC* in the 300 times experiment results. FIGURE 7 shows one of the ROC curve in an experiment.The X-axis is FPR and Y-axis is 1_*Recall*(TPR),The diagonal line from (0, 0) to (1, 1) divides the ROC space into upper left and lower right regions. The points on the line means the FPR is equals to TPR which represent random classification. The points above this line represent a good classification result which means better than random classification, And the points below this line represent poor classification results which means worse than random classification.

2) MODULES ANALYSES IN THE UCO ALGORITHM

The UCO algorithm consists of a randomly-undersampling module, a clustering module and an oversampling module. This section analyzes impacts of the three modules on predicting heart attack events for stroke patients. The predictions are conducted using the model RF.

In FIGURE 9 the X-axis is the number of undersampling, and the Y-axis is the Metrics, It shows the prediction of different random undersampling numbers, we can see that

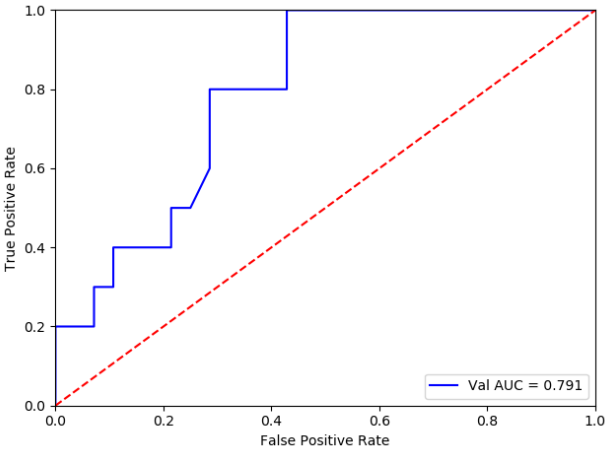


FIGURE 7. ROC for UCO(120)+RF.

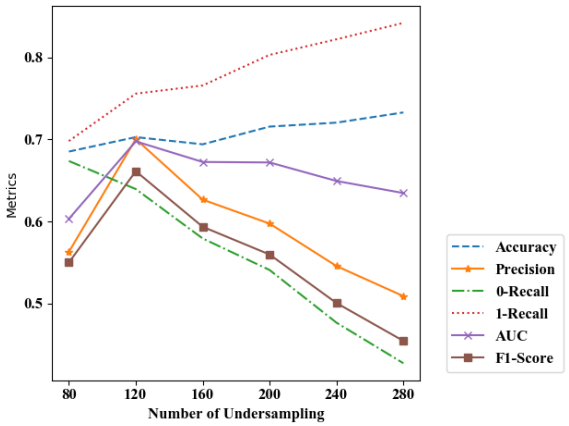
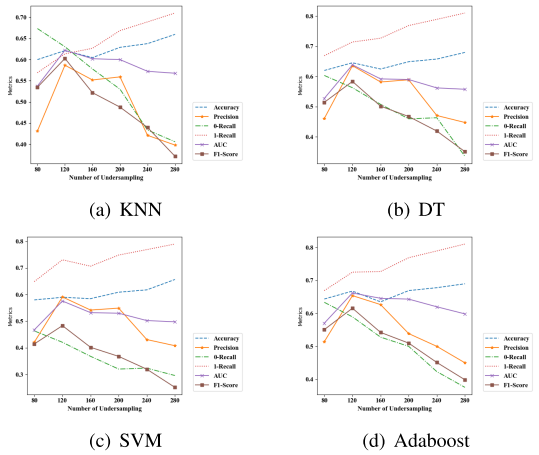


FIGURE 8. Results using UCO(m)+RF, m = 80, 120, 160, 200, 240, 280.



(a) KNN (b) DT

(c) SVM (d) Adaboost

FIGURE 9. Results using UCO(m)+KNN/DT/SVM/Adaboost, m = 80, 120, 160, 200, 240, 280.

only when the number of undersampling number is 120, the metrics is maximum. When the number of undersampling is less or bigger than 120, the Accuracy and AUC are relatively low, indicating that the prediction is over fitting. In FIGURE 8, the X-axis is the number of undersampling,

TABLE 2. Results of modules analyses.

	1_Recall	*0_Recall*	*Precision*	*Accuracy*	*F1_Score*	*AUC*
UCO(120)+RF	$75.59\% \pm 0.1282$	$\mathbf{63.95\% \pm 0.1116}$	$\mathbf{70.05\% \pm 0.1199}$	$\mathbf{70.29\% \pm 0.0845}$	$\mathbf{0.6613 \pm 0.0919}$	$\mathbf{0.6977 \pm 0.0831}$
Undersampling(120)+Oversampling+RF	$75.42\% \pm 0.1301$	$58.24\% \pm 0.1290$	$65.66\% \pm 0.1512$	$68.06\% \pm 0.0984$	0.6094 ± 0.1162	0.6684 ± 0.0976
Undersampling(120)+Clustering+RF	$\mathbf{80.05\% \pm 0.1236}$	$52.63\% \pm 0.1026$	$67.58\% \pm 0.0789$	0.5948 ± 0.0949	$70.76\% \pm 0.1382$	0.6633 ± 0.0772

TABLE 3. Results using Oversampling+RF.

	1_Recall	*0_Recall*	*Precision*	*Accuracy*	*F1_Score*	*AUC*
UCO(120)+RF	$\mathbf{75.59\% \pm 0.1282}$	$\mathbf{63.95\% \pm 0.1116}$	$\mathbf{70.05\% \pm 0.1199}$	$70.29\% \pm 0.0845$	$\mathbf{0.5948 \pm 0.0919}$	$\mathbf{0.6977 \pm 0.0831}$
SMOTE[18]+RF	$96.77\% \pm 0.0053$	$0.752\% \pm 0.0298$	$0.737\% \pm 0.0294$	$\mathbf{93.68\% \pm 0.0053}$	0.0074 ± 0.0295	0.4876 ± 0.0153
Random_Oversampling[16]+RF	$98.79\% \pm 0.0078$	$0.688\% \pm 0.0436$	$0.654\% \pm 0.0576$	$95.87\% \pm 0.0087$	0.0068 ± 0.0277	0.4329 ± 0.0254
ADASYN[20]+RF	$98.43\% \pm 0.0065$	$0.733\% \pm 0.0354$	$0.711\% \pm 0.0476$	$94.66\% \pm 0.0058$	0.0070 ± 0.0342	0.4832 ± 0.0276

and the Y-axis is the Metrics, It shows when separately using KNN, DT, SVM, Adaboost models on predict different random undersampling numbers data.It can be seen that the prediction results of these models are the best when use UCO(120).So the best undersampling number is 120.

TABLE 2 shows the result of RF model predict on the dataset which data processing don't use SMOTE Algorithm or clustering. When remove SMOTE module from the data processing, the *Accuracy* is reduced by 2%, the *Precision* is reduced by 5%, and the 0_*Recall* is reduced by 5%. When remove clustering module from the data processing, 1_*Recall* is 80.05%, 0_*Recall* is 52.63%, and there is a big difference between 1_*Recall* and 0_*Recall*. This indicates that the prediction result is obviously biased towards the one with heart attack, that is, the majority of samples are predicted to have heart attack. So the data is still imbalanced, using SMOTE algorithm and clustering in data processing can greatly help balance the data.

C. PREDICTION ONLY USING OVERSAMPLING METHODS BASED ON PROCESSING RESULTS

SMOTE algorithm proposed by Qi W and ADASYN proposed by He H are the best oversampling methods at the present, The two algorithms are applied widely.Randomly oversampling method, the SMOTE algorithm, and the ADASYN algorithm used process the data separately. Then put the processed results into the model RF to do prediction, respectively. Calculate the mean and mean square error of *Accuracy*, *Precision*, 1_*Recall*, 0_*Recall*, $F1$_*Score*, and *AUC* in 300 experiments. The best result is using SMOTE algorithm to process the data, and the specific experimental results are shown in TABLE 3.

In TABLE 3, it is shown that the obtained results of *Accuracy* by using SMOTE, RandomOversampling and ADASYN are all above 90%, and their results of *Precision* are all less than 1%. Such small precisions indicate that many negative samples are falsely predicted. For the UCO algorithm, the achieved results, including *Accuracy*, *Precision*, 1-*Recall*, 0-*Recall*, and *AUC*, are all above 60%. In the UCO algorithm, the clustering operation makes sure that each type of positive sample could be oversampled, and the undersampling operation reduces the negative samples to help make feature extraction. So, the UCO algorithm generates nearly balanced data set, which contains each type of positive samples. Using only oversampling algorithm does not necessarily sample each type of positive samples. The obtained samples by only using oversampling method do not help machine model to extract good features for differentiating positive samples and negative samples.

D. PREDICTION ONLY USING UNDERSAMPLING METHODS BASED ON PROCESSING RESULTS

We use Random undersampling and K-means cluster undersampling to do data processing respectively and input the results of data processing into the RF prediction model to conduct experiments. Table 4 shows the mean and mean square error of *Accuracy*, *Precision*, 1_*Recall*, 0_*Recall*, $F1$_*Score*, and *AUC* in 300 experiment. Using 80,120,160,200,240 as Random undersampling number to get the best results of experiments (the number of samples is 80) to record respectively.

In TABLE 4, it is shown that the result using random undersampling algorithm is better than that using Kmeans undersampling algorithm, and its obtained *precision* is 4.356%. Such a bad precision result indicates that only using undersampling method does not provide sufficient data to classifiers to extract features. The prediction results of UCO(120)+RF model is the best. In the algorithm UCO, clustering procedure and oversampling procedure increase positive samples. Such an increment could not be achieved by only executing undersampling operation.

E. PREDICTING-HEART-ATTACK CAPABILITY OF CLASSIC MACHINE LEARNING MODELS

Based on the same settings in the UCO algorithm(120), this section analyzes predicting-heart-attack performances of four models, including Adaboost, kNN, DT, and SVM. TABLE 5 shows the mean and mean square error of *Accuracy*, *Precision*, 1_*Recall*, 0_*Recall*, $F1$_*Score*, and *AUC* of the four models in 300 experiment.

Compared with the predictions of the other four models (Adaboost, kNN, DT, SVM), RF has better performances. For the model RF, the obtained 1_*Recall* is 2% bigger than the other models, *Precision* being 4% bigger, *AUC* being 3% bigger, *Accuracy* being 3% bigger, and $F1$_*Score* being 4% bigger.

TABLE 4. Results using Undersampling+RF.

	1_Recall	*0_Recall*	*Precision*	*Accuracy*	*F1_Score*	*AUC*
UCO(120)+RF	$\mathbf{75.59\% \pm 0.1282}$	$\mathbf{63.95\% \pm 0.1116}$	$\mathbf{70.05\% \pm 0.1199}$	$\mathbf{70.29\% \pm 0.0845}$	$\mathbf{0.5948 \pm 0.0919}$	$\mathbf{0.6977 \pm 0.0831}$
Random_Undersampling[13]+RF	$70.04\% \pm 0.0104$	$41.10\% \pm 0.0576$	$4.356\% \pm 0.0059$	$69.11\% \pm 0.0101$	0.0788 ± 0.0107	0.5557 ± 0.0291
Kmeans_Undersampling[15]+RF	$40.19\% \pm 0.0232$	$65.43\% \pm 0.0540$	$3.508\% \pm 0.0031$	$41.00\% \pm 0.0225$	0.0666 ± 0.0058	0.5281 ± 0.0294

TABLE 5. Results using five different classifiers.

	1_Recall	*0_Recall*	*Precision*	*Accuracy*	*F1_Score*	*AUC*
UCO(120)+RF	$\mathbf{75.59\% \pm 0.1282}$	$\mathbf{63.95\% \pm 0.1116}$	$\mathbf{70.05\% \pm 0.1199}$	$\mathbf{70.29\% \pm 0.0845}$	$\mathbf{0.6613 \pm 0.0919}$	$\mathbf{0.6977 \pm 0.0831}$
UCO(120)+Adaboost	$72.49\% \pm 0.2062$	$59.87\% \pm 0.1983$	$65.37\% \pm 0.2072$	$66.75\% \pm 0.0932$	0.6156 ± 0.0946	0.6618 ± 0.0988
UCO(120)+kNN	$61.37\% \pm 0.2324$	$63.04\% \pm 0.2134$	$58.67\% \pm 0.2481$	$62.13\% \pm 0.0898$	0.6028 ± 0.0995	0.6221 ± 0.0932
UCO(120)+DT	$71.42\% \pm 0.1756$	$56.29\% \pm 0.1780$	$63.59\% \pm 0.1874$	$64.55\% \pm 0.0789$	0.5837 ± 0.0856	0.6386 ± 0.0957
UCO(120)+SVM	$73.08\% \pm 0.1532$	$42.07\% \pm 0.1479$	$59.11\% \pm 0.1652$	$58.98\% \pm 0.0987$	0.4832 ± 0.0879	0.5758 ± 0.0963

VI. DISCUSSION

In the UCO algorithm, the undersampling module is to reduce the negative samples, and the oversampling module is to add positive samples. Because of the undersampling and the oversampling, the gap between the number of negative samples and the number of positive samples is narrowed.The number of samples in the undersampling is 120, and the ratio between majority samples and minority samples is 3:2. The clustering module is to make different types of positive samples be well-distributed in the training set, so that the samples are relatively balanced. When performing K-means algorithm on patients with heart attack, it is also tried to cluster patients without heart attack. Since patients without heart attack are randomly selected each time, the number of related clusters is different. In such a way, the cluster number need be recalculated before each clustering. Much time increases, and prediction results are not improved. Therefore, UCO algorithm performs K-means clustering operation only over positive samples.

When using the model to make predictions, the RF model has the best predictive ability. The RF model evolved from the DT model. The DT model uses a decision tree for prediction. The RF model is a combination of multiple decision trees, so it could reduce overfitting and has good generalizing ability. The Adaboost model uses the construction of multiple classifiers, and gives a larger weight to the classifier with a smaller classification error rate, and gives a smaller weight value to the classifier with a larger classification error rate, so it is more sensitive to abnormal samples. Adaboost is possible to obtain a higher weight value in the iterative process, which will ultimately affect the model effect. Because the data of ICU stroke patients differs greatly, so there are some abnormal samples, the prediction of the Adaboost model is not good. The core idea of the kNN algorithm is that if most of the k nearest samples in the feature space of a sample belong to a certain category, the sample will belongs to this category and has the characteristics of the samples in this category. This algorithm is suitable for classification problems with a large sample size. The sample size of the data set used in this article is not large, so kNN model is not suitable for this kind of data. SVM is a two-category model. Its purpose is to find a hyperplane to divide the sample. The principle of segmentation is to maximize the interval, which is finally transformed into a convex quadratic programming problem to solve, but the SVM model is sensitive to missing values, The data set used in this article has missing values, and the mean value filling method makes the prediction results of the SVM model not accurate.

In clinic practice, physicians could utilize the proposed UCO algorithm to process their obtained stroke-patients data, and train a random forest classifier based on the processing results. For a stoke patient whose heart-attack risk is not known, physicians feed medical indexes to the trained-well classifier, including minimal Heartrate, mean of Heartrate, maximal Resprate, minimal Glucose, maximal Glucose, mean of Glucose, Creatinine, and Sysbp. The classifier will give a prediction result about heart attack.

Our proposed algorithm UCO is evaluated only on the database MIMIC-III. By a lack of stroke-patients data, it has not been investigated that what the performance of UCO over other stroke-patients data set is.

VII. CONCLUSION

In this paper, a data processing UCO algorithm was proposed by combining three methods, including undersampling method, clustering method, and oversampling method. The algorithm can deal with imbalanced stroke patients data.From the original data, eight medical indicators that affect heart attack were selected. Then this paper compare the performance of several machine learning models in predicting heart attack. Experimental results show that RF is the best model to predict the possibility of heart attack on the MIMIC-III database of stroke patients datasets. The *Accuracy* is 70.29%, and the *Precision* is 70.05%.

Through the UCO(120)+RF method of this paper, We can use clinical monitoring data from the ICU to predict whether stroke patients will have heart attack, and we don't require doctors filter the data.

Predicting heart attack from daily detection indicators will be of great help to clinical diagnosis and treatment, and will greatly reduce the mortality of stroke patients complicated by heart attack. It can warn doctors and patients in advance and help doctors take treatment measures to prevent the occurrence of heart attack and reduce the possibility of serious illness.

In the future, application to other imbalanced data for the proposed UCO algorithm will be investigated. Also, its combination with deep neural network is our next research direction. Besides, a software system for heart attack prediction in stroke patients will be developed based on the algorithm UCO and random forest. The software system is to provide clinical support for taking precautions against heart attack in stroke patients.

ACKNOWLEDGMENT

The authors would like to thank the editors and reviewers for their constructive comments and thank to everyone on the project team for their advice and help.

REFERENCES

[1] L. Junfan, *Observation of Clinical Effect of 'Stroke Integration' in the Treatment of Ischemic Stroke*. Beijing, China: Beijing Univ. Chinese Medicine, 2019.
[2] S. Diekmann, L. Hörster, S. Evers, M. Hiligsmann, G. Gelbrich, K. Gröschel, R. Wachter, G. F. Hamann, P. Kermer, J. Liman, M. Weber-Krüger, J. Wasem, and A. Neumann, "Economic evaluation of prolonged and enhanced ECG holter monitoring in acute ischemic stroke patients," *Current Med. Res. Opinion*, vol. 35, no. 11, pp. 1859–1866, Nov. 2019.
[3] Y. Rongfeng and X. Minhui, "A report of 9 cases of acute stroke complicated with acute myocardial infarction," *Hunan Med.*, to be published.
[4] G. V. Dous, A. C. Grigos, and R. Grodman, "Elevated troponin in patients with acute stroke—Is it a true heart attack," *Egyptian Heart J.*, vol. 69, no. 3, pp. 165–170, 2017.
[5] D. Bhatnagar, I. Kaur, and A. Kumar, "Ultrasensitive cardiac troponin i antibody based nanohybrid sensor for rapid detection of human heart attack," *Int. J. Biol. Macromolecules*, vol. 95, pp. 505–510, Feb. 2017.
[6] B. Week, "Heart disorders and diseases; Data on heart attack described by researchers at capital medical University (over expression of protein kinase C epsilon improves retention and survival of transplanted mesenchymal stem cells in rat acute myocardial infarction)," Tech. Rep., 2016.
[7] Q. Wang, Y. Zhou, W. Zhang, Z. Tang, and X. Chen, "Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis," *Expert Syst. Appl.*, vol. 152, Aug. 2020, Art. no. 113334.
[8] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," in *Proc. IEEE 30th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2018, pp. 137–142.
[9] Z. Fenglan, *Electrocardiogram QTcd Changes and Prognosis in Different Periods of Acute Myocardial Infarction*. Shanxi, China: Shanxi Clinical Medicine, 2001.
[10] Y. Yuejin *et al.*, "Spontaneous improvement of exercise abnormality in acute myocardial infarction and the predictive value of low-dose dobutamine echocardiographic test," *Chin. J. Circulat.*, to be published.
[11] Q. Jianping and H. Hong, "Analysis of the predictive value of high-frequency electrocardiogram on acute myocardial infarction," *Biomed. Eng. Res.*, to be published.
[12] L. Yaowang *et al.*, "Application of machine learning algorithm in prediction of coronary heart disease and myocardial infarction," *Int. Med. Health Guidance News*, to be published.
[13] M. A. Little, "Random sampling," in *Machine Learning for Signal Processing*, 2019.
[14] Q. Fan, Z. Wang, and D. Gao, *One-Sided Dynamic Undersampling No-Propagation Neural Networks for Imbalance Problem*, Oxford, U.K.: Pergamon Press, 2016.
[15] S. Nagul and R. K. Kumar, "An effective K-means approach for imbalance data clustering using precise reduction sampling," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 3, pp. 65–70, Mar. 2018.
[16] A. Moreo, A. Esuli, and F. Sebastiani, "Distributional random oversampling for imbalanced text classification," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 805–808.
[17] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 14, no. 1, Dec. 2013.
[18] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–11, Jan. 2017.
[19] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, Berlin, Germany: Springer, 2005, pp. 878–887.
[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE World Congr. Comput. Intell.*, Jun. 2012, pp. 1322–1328.
[21] T. C. W. Landgrebe and R. P. W. Duin, "Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 810–822, May 2008.
[22] Z. Elouedi, E. Lefevre, and D. Mercier, "Discountings of a belief function using a confusion matrix," in *Proc. 22nd IEEE Int. Conf. Tools with Artif. Intell.*, Oct. 2010, pp. 287–294.
[23] Z. Yuqiong and W. Huanjun, "Diagnostic significance of determination of troponin I and T for acute myocardial infarction," *Exp. Lab. Med.*, 2008.
[24] Z. Honglian, L. Luming, and B. Zhenghe, "Diagnostic value of cardiac troponin T for acute myocardial infarction," *Contemp. Med.*, to be published.
[25] Y. Xiaoli, *Research on the Application Value of High-Sensitivity Troponin T Detection in Myocardial Infarction*. Dalian, China: Dalian Medical Univ., 2011.
[26] P. R. Letters, "An empirical comparison of four initialization methods for the K-means algorithm," Tech. Rep., 1999.
[27] P. Hansen, E. Ngai, B. K. Cheung, and N. Mladenovic, "Analysis of global k-means, an incremental heuristic for minimum sum-of-squares clustering," *J. Classification*, vol. 22, no. 2, pp. 287–310, Sep. 2005.
[28] J. C. De Winter, "Using the student's t-test with extremely small sample sizes," *Practical Assessment Res. Eval.*, vol. 18, no. 1, p. 10, 2013.

MENG WANG was born in Tangshan, Hebei, China, in 1996. She received the B.S. degree from Xiangtan University. She is currently pursuing the M.S. degree with the School of Computer Science and Software Engineering, East China Normal University, China.

Her main research interests include machine learning, deep learning, biomedical engineering, and participation in the prediction of medical problems.

XINGHUA YAO is currently an Assistant Professor with the School of Basic Medicine, Shanghai University of Traditional Chinese Medicine, China. His research interests include data mining in healthcare, deep learning, and formal methods.

YIXIANG CHEN is currently a Full Professor with the School of Software Engineering, East China Normal University, China, where he is also a coordinating trust- worthy software, the Internet of Things and human- cyber-physical systems related research activities. He used to be the Director of the MoE Engineering Research Center for Software/Hardware Co- design Technology and Application. He is also the Vice Chairman of the Fuzzy Systems and Fuzzy Mathematics Committee of the Chinese Society for Systems Engineering and the Vice Chairman of the Shanghai Zhangjiang Pudong Internet of Things Association for now.