Computing and Informatics, Vol. 43, 2024, 756–776, doi: 10.31577/cai_2024_3_756

CORRELATION ANALYSIS ALGORITHM FOR MASSIVE ULTRA-HIGH-DIMENSIONAL BREAST ULTRASOUND RADIOMICS FEATURE DATA IN A DISTRIBUTED ENVIRONMENT

Yuehong Tang

Tumor Hospital Affiliated to Xinjiang Medical University
Urumqi, Xinjiang, China
&
School of Public Health, Xinjiang Medical University
Urumqi, Xinjiang, China

Yan Chen

The Medical School of Jiaxing University
Jiaxing Zhejiang, China

Wen Liu*

Artificial Intelligence and Smart Mine Engineering Technology Center
Xinjiang Institute of Engineering, Urumqi, China
&
Xinjiang Changsen Data Technology Co., Ltd, Urumqi 830011, China
e-mail: 627952@qq.com

Zheng Gu

Artificial Intelligence and Smart Mine Engineering Technology Center
Xinjiang Institute of Engineering, Urumqi, China

Hui Yao

Xinjiang Changsen Data Technology Co., Ltd.
Urumqi 830011, China

* Corresponding author

Abstract. Radiomics is a technology that extracts a large number of quantitative features from high-throughput medical images and has become a focus of research. It can help in disease diagnosis, therapy planning, and prognosis evaluation through Big Data analysis algorithms. Radiomics technology can extract hundreds or even tens of thousands of quantifiable data features from medical images, which can no longer fit into the memory of one machine. Therefore, we propose a distributed correlation analysis algorithm (DFCA) based on a MapReduce distributed computing framework for breast ultrasound radiomics feature datasets. Each compute node will produce massive intermediate data while the DFCA calculates the Pearson correlation coefficient of radiomics features. With the increase of feature data and dimensions, the data transmission cost will be in a square growth. To reduce the cost, we propose a distributed correlation estimation algorithm (DFCEA) for radiomics features based on DFCA. The DFCEA algorithm estimates the Pearson correlation coefficient using an iterative method, which can further reduce the I/O cost. The experiment proved that our algorithms are more effective compared to the algorithms in the literature.

Keywords: Radiomics, massive high-dimensional data, correlation analysis, distributed computing

1 INTRODUCTION

In 2012, Lambin et al. [1] first proposed radiomics, which includes the following three stages:

1. segmentation of the image to determine the tumor region;
2. extraction of tumor features to produce quantized data; and
3. construction of a classification and prediction model for analyzing the feature data.

Analyzing high-dimensional data of imaging radiomics features can refine the analysis of small information in breast tumor imaging. Scholars constantly expand and improve radiomics technology, which can extract thousands of high-dimensional data features from medical images and play an important role in improving diagnosis [2, 3], clinical decision [4, 5], and prognostication [6, 7].

We use Matlab's wavelet filter (which has 48 wavelet bases) to process images. Each wavelet base has eight different filtering methods for image filtering in low-pass and high-pass modes. Each filtering method uses 14 first-order statistics. Each image will generate 5 376 $(48 * 8 * 14)$ features. If various filters and classification features are combined, each image can extract more features. For instance, if the above methods are combined with 63 texture-based features, the image will extract $48 * 8 * (14 + 63) = 29\,568$ features [8].

With time, doctors accumulate a large amount of imaging data when diagnosing patients. For instance, the Affiliated Tumor Hospital of Xinjiang Medical University receives more than 2 500 new patients with breast cancer annually. Moreover, computed tomography, magnetic resonance imaging (MRI) and other medical imaging examination data are also increasing yearly. When calculating the Pearson correlation coefficient of the massive ultra-high-dimensional radiomics features data, one machine has a poor processing speed. Therefore, we propose a distributed correlation analysis (DFCA) algorithm. To further improve the efficiency of the DFCA algorithm, we propose the distributed correlation estimation algorithm (DFCEA), which estimates the Pearson correlation coefficient according to an iterative method, to reduce the large input/output (I/O) cost. In summary, the contributions of this paper are as follows:

1. To analyze the association of massive ultra-high-dimensional radiomics feature data, we propose a DFCA algorithm;
2. To reduce the I/O cost of the DFCA algorithm in the Shuffle stage, we propose the DFCEA algorithm, which raises the computing efficiency of the correlation analysis of the massive high-dimensional radiomics feature data;
3. We evaluate the effectiveness of the algorithm on a real dataset and demonstrate that our algorithm is superior to the algorithm in the literature.

The rest of this paper is organized as follows. Section 2 briefly discusses the relevant research of the correlation analysis. Section 3 introduces the definitions and preliminary knowledge of the DFCA and DFCEA algorithms. Section 4 presents the DFCA and DFCEA algorithms, while Section 5 analyzes the experimental results. Section 6 comprises the paper's conclusions.

2 RELATED WORK

2.1 Study on the Correlation of Imaging Features

To analyze the patient's condition more comprehensively, relevant research experts use different processing methods to obtain ultra-high-dimensional image feature information [9]. Redundant features in ultra-high-dimensional image feature data will degrade the performance of the prediction algorithm [10]. By analyzing the correlation between features, we can not only find the correlation among features but also select appropriate and effective features to improve the accuracy of the classification and prediction algorithm.

A study [11] extracted dynamic contrast-enhanced-MRI image features of the breast through radiomics and evaluated the correlation between low- and high-Ki-67 expression image features using the Mann–Whitney U test statistics method, which provided a non-invasive detection method for medical experts to determine the spread of breast cancer in patients. Fu et al. [12] used the minimum redundancy and maximum correlation algorithm for radiomics feature selection, in which the

minimum redundancy is realized between the selected feature and other features, and the maximum correlation selected the feature with the highest correlation with the development of coronavirus disease 2019. The least absolute shrinkage and selection operator regression algorithm was combined to complete the feature selection process. In the literature [8], the data processed by various wavelets are divided into subsets in a single-machine environment. Pearson correlation coefficient is subsequently used to analyze the correlation of radiomics features with the same name between subsets by normalizing the data, and radiomics features with correlations greater than 0.9 are removed to reduce the "dimension disaster" and increase the diagnostic effectiveness of the algorithm. Oubel et al. [13] used the mutual information method to observe the variability between multi-information and global extremum of image features changing with time, to evaluate the repeatability between features and achieve feature selection.

All of the methods mentioned above process data in a single-machine environment. The processing speed will be reduced as the data size and dimensions rise because calculations involving massive ultra-high-dimensional data will take longer and require more central processing unit (CPU) and memory resources.

2.2 Research on the Correlation of Distributed Computing

Hadar and Shayevitz [14] used Gaussian correlation to estimate the cross-correlation matrix of a few vectors by calculating the mean of each node and jointly observed the unknown vector correlation coefficient of the Gaussian scalar under random variables. A study [15] used the K-means++ algorithm to obtain the centroid of the partitioned data and subsequently constructed a secondary storage index mechanism to partition the data using the centroid. This method is suitable for cutting data horizontally. If vertical cutting is used, we need to transpose the data, and the number of rows and columns will be fixed. Palma-Mendoza et al. [16] used vector partition, similar to the index table, to take features as indices and each data object as index value and used spark combined with CFS feature selection algorithm to perform a correlation analysis to realize feature selection. When every data object is consistent and there is no missing value, the process of partitioning will also generate several calculations. CCCA-LTS algorithm is proposed in the literature [17], which uses Euclidean distance to calculate the relationship between the data processed by normal distribution and Pearson correlation under the original data to estimate the correlation of the time series. The CCCA-LTS algorithm needs two MapReduce calculations. For the first calculation, it needs to calculate the mean value and standard deviation of the feature columns in each memory block and estimate the mean value and standard deviation of the standardized sequences in each Map according to the mean value and standard deviation. For the second calculation, it uses the relationship between the Euclidean distance between any two features that conform to normal distribution and the estimated mean value and standard deviation of each partial sequence to perform correlation estimation.

To address the problems, such as the slow calculation efficiency of processing massive ultra-high-dimensional radiomics features and the limitation of distributed processing formula correlation, we propose a DFCA algorithm to calculate the correlation between massive ultra-high-dimensional radiomics features in a distributed way. To reduce the I/O cost of a large amount of data generated in the calculation process during data transmission between nodes, we propose a DFCEA algorithm to estimate the relationship between the Pearson correlation coefficient and threshold and compare with the CCCA-LTS algorithm.

3 RELEVANT DEFINITIONS AND FUNDAMENTAL CONCEPTS

This chapter mainly introduces the definitions and fundamental concepts of the DFCA and DFCEA algorithms, which serve as the building blocks for creating distributed correlation analysis algorithms for massive ultra-high-dimensional radiomics feature data.

3.1 Related Definitions

Definition 1. According to the standard statistical division of the Image Biomarker Standardization Initiative [18], radiomics features are frequently divided into shape features, first-order statistics features, texture-based features, high-order features, and features based on model transformation (filters). Taking PyRadiomics [19] as an example, features can be divided into seven categories, as shown in Table 1. Before feature extraction, we can use filtering to preprocess the medical images. Table 2 shows that this paper used various image filters. Through various filters, each image can extract $n * m$ features (n is the sum of the number of features, and m is the sum of the number of filtering methods).

Feature Type	Number of Features
First-order statistics features	19
Shape features (2D)	10
Gray level co-occurrence matrix (GLCM) features	24
Gray level size zone matrix (GLSZM) features	16
Gray level run length matrix (GLRLM) features	16
Neighboring gray-tone difference matrix (NGTDM) features	5
Gray level dependence matrix (GLDM) features	14

Table 1. Feature classification

For instance, we processed the breast ultrasound images using the wavelet filter. The Grayscale of the image has four different filtering methods (LH, HL, HH, and LL), and extracts first-order (standard deviation is not enabled by default, and there are only 18 features when extracting), GLCM, GLSZM, GLRLM, NGTDM, and GLDM features. The original filtered images can extract $(1 + 4) * (18 + 24 + 16 + 16 + 5 + 14) = 465$ features. Table 3 shows the partial feature data.

Feature Type	Number of Filtering Results
Original	1
Wavelet	4
LoG	1
Square	1
Square Root	1
Exponential	1
Gradient	1
LocalBinaryPattern2D	1

Table 2. Filters

Original_firstorder _10Percentile	Original_firstorder _90Percentile	Wavelet-LH_firstorder _10Percentile	Wavelet-LH_firstorder _90Percentile
26	95	−9.550913443	9.650366591
0	127	−7.731168297	7.770054805
27	110	−7.908291615	7.866153714
3	100	−5.427456432	5.456422723
17	135	−8.326565507	8.386594786
3	92	−5.827297771	5.920504416
37	92	−6.201063772	6.291238555
36	95	−6.673283964	6.573962048
29	83	−3.910080637	3.929090028

Table 3. Extracted partial feature data of breast ultrasound image

Definition 2 (Pearson correlation coefficient)**.** Two feature sequences are $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. n denotes the quantity of data items. x_i, y_i $(1 \leq i \leq n)$ denotes the X and Y feature values for the i^{th} entity. The two features' Pearson correlation coefficient is as follows:

$$\rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{\sigma_X}\right)\left(\frac{y_i - \bar{Y}}{\sigma_y}\right). \tag{1}$$

The mean and standard deviation of X and Y respectively are and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The results of the Pearson correlation coefficient are as follows: $\rho(X,Y) = 0$ demonstrates that the two features are linearly independent; $\rho(X,Y) > 0$ indicates that the two features are positively correlated; and $\rho(X,Y) < 0$ indicates that the two features are negatively correlated.

Definition 3 (Basic inequality)**.** a and b satisfy $a * b \leq \frac{a^2+b^2}{2}$ $(a \in R$ and $b \in R)$. R is the set of real numbers.

Problem definition. An ultra-high-dimensional radiomics data D is stored in HDFS.

For any two feature columns $X, Y \in D$ $(X \neq Y)$, the correlation coefficient threshold of X and Y is ε. We used parallel computing to analyze the correlation between X and Y features.

3.2 Preliminary Knowledge

MapReduce is a parallel computing framework of the Hadoop processing platform, which puts the calculation on each storage node with data and finally summarizes it. MapReduce's task includes three steps.

1. **The Map Task:** the total amount of data slices in each storage node determines how many Map Task tasks are needed, which are mostly used to read the file data line by line for associated processing.
2. **The Reduce Task:** the Reduce Task is responsible for summarizing and computing each Map Task that generates the intermediate data and outputting the results.
3. **MRApp Task:** this task is responsible for scheduling and resource coordination during MR execution.

The Shuffle stage exists between the Map Task and the Reduce Task and is used to integrate the results of the Map stage and output the data to be pulled by the Reduce task, as shown in Figure 1.

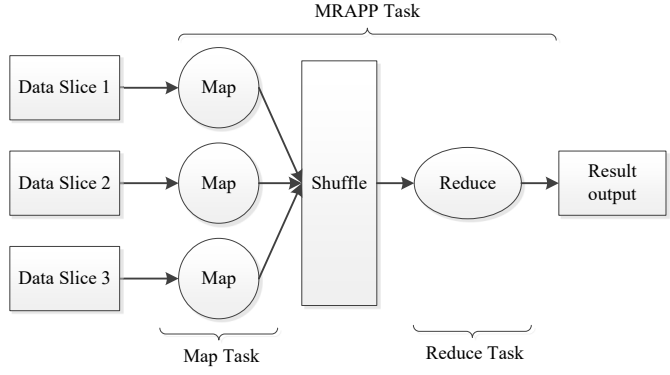


Figure 1. Overview of the MapReduce execution process

4 ALGORITHM PRINCIPLE

In this section, we introduce the DFCA and DFCEA algorithms' principles and implementation procedures, at the same time, we have also demonstrated the correctness of our approach.

4.1 DFCA Algorithm Principle

The DFCA algorithm is a distributed correlation analysis algorithm for massive ultra-high-dimensional breast ultrasound radiomics feature data. According to Definition 2, the Pearson correlation coefficient formula can be modified as follows:

$$\rho(X,Y) = \frac{1}{n}\frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{Y} - \sum_{i=1}^{n} y_i \bar{X} + n\bar{X}\bar{Y}}{\sigma_X \sigma_y}.$$

We replace $\bar{X}$, $\bar{Y}$, σ_X, and σ_Y in $\rho(X,Y) = \frac{1}{n}\frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{Y} - \sum_{i=1}^{n} y_i \bar{X} + n\bar{X}\bar{Y}}{\sigma_X \sigma_y}$. The Pearson correlation coefficient formula is as shown in formula (2):

$$\rho(X,Y) = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \tag{2}$$

According to formula (2), we only need to determine the value of $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n}(x_i)^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n}(y_i)^2$, $\sum_{i=1}^{n} x_i y_i$ to calculate the size of $\rho(X,Y)$, which can be compared with the given correlation threshold.

4.1.1 Correlation Coefficient of Distributed Computing

Distributed calculation of correlation coefficient requires each Map to calculate the value of partial data $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n}(x_i)^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n}(y_i)^2$, $\sum_{i=1}^{n} x_i y_i$. Our approach can reduce the cost of calculating the Pearson correlation coefficient according to the results of the calculation for each node.

Assume that the dataset $D = \{d_1, d_2, d_3\}$ is split into three portions (d_1, d_2, d_3) and stored on each of the three computer nodes N_1,N_2, N_3. Two columns in the dataset D are X and Y. The values of the k^{th} data of the X and Y features at the n^{th} node are X_k^n and Y_k^n. We assume that the dataset is divided into three equal parts $m = \frac{|D|}{3}$. In the experiment, the dataset also cannot be divided equally. $|D|$ is the total number of datasets D. As shown in Figure 2, each node calculates the five values of $\sum_{i=1}^{m} x_i$, $\sum_{i=1}^{m}(x_i)^2$, $\sum_{i=1}^{m} y_i$, $\sum_{i=1}^{m}(y_i)^2$, $\sum_{i=1}^{m} x_i y_i$. We sum up the calculation results of each node and use formula (2) to calculate $\rho(X,Y)$.

Suppose a node has $n\,(n \geq 2)$ feature columns, each node needs to output $2n+C_n^2$ data. Each computer node needs to output 125 750 data if there are 500 features. The number of data transmitted between the nodes in the DFCA algorithm accounts for $(n-1)/(n+3)$ of the total number when calculating the sum of any two-feature data multiplication. To further optimize the DFCA algorithm, we propose the DFCEA algorithm.

4.2 DFCEA Algorithm Principle

The DFCEA algorithm uses the inequality of the arithmetic and geometric means principle to estimate the Pearson correlation coefficient, which only needs to trans-

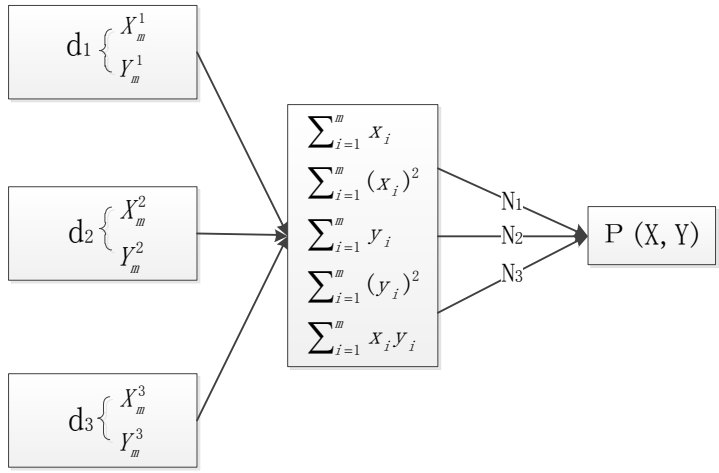


Figure 2. Process of the distributed calculation of the Pearson correlation coefficient

mit four data $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} (x_i)^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} (y_i)^2$ between nodes (the dataset contains only two feature columns, X and Y. $n > 0$ is the number of data pieces). Each node only needs to output $2m$ data if the data is in the $m\,(m \geq 2)$ dimension, which reduces the number of C_m^2 data produced by the multiplication of any two dimensions. The following is the estimation formula:

The threshold is ε. ρ_1 is the Pearson correlation coefficient value under formula (1). ρ_2 is the value of the estimated Pearson correlation coefficient and is always $\geq \rho_1$ (equal sign is taken when the two-feature data coincide consistently). The relationship between ρ_2 and ε will produce the following two results:

Result 1: When $\rho_2 < \varepsilon$, then $\rho_1 < \varepsilon$, the two features do not have a high correlation.

Result 2: When $\rho_2 \geq \varepsilon$, we cannot judge the relationship between ρ_1 and ε.

The DFCEA algorithm can obtain the relationship between ρ_1 and ε in the form of estimation without calculating the real Pearson correlation coefficient. The principle of the DFCEA algorithm is proven as follows:

Proof. First, three prerequisites are listed.

Condition 1: The two feature sequences are $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. $n(n > 0)$ denotes the number of the data size. x_i, y_i correspond to the two feature values of X and Y of the i^{th} $(1 \leq i \leq n)$ data object.

Condition 2: According to Definition 3, $ab \leq \frac{a^2+b^2}{2}$.

Condition 3: The standard deviation of the data must be > 0.

When $i = 1$: $a = x_1$, $b = y_1$. Then $x_1 y_1 \leq \frac{(x_1)^2+(y_1)^2}{2}$.

When $i = 2$: $a = x_2$, $b = y_2$. Then $x_2 y_2 \leq \frac{(x_2)^2+(y_2)^2}{2}$.

$$\vdots$$

When $i = n$: $a = x_n$, $b = y_n$. Then $x_n y_n \leq \frac{(x_n)^2+(y_n)^2}{2}$.
We sum $x_i y_i \leq \frac{(x_i)^2+(y_i)^2}{2}$ from $i = 1$ to $i = n$ to generate formula (3):

$$\sum_{i=1}^{n} x_i y_i \leq \sum_{j=1}^{n} \frac{(x_j)^2 + (y_j)^2}{2}. \tag{3}$$

According to conditions 2 and 3, we convert the left part of formula (3) into the form of formula (2):

$$\begin{aligned}
&\frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \\
&\qquad \leq \frac{n\sum_{i=1}^{n}\left((x_i)^2 + (y_i)^2\right) - 2\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{2\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}, \\
&\frac{n\sum_{i=1}^{n} x_i y_i}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \\
&\qquad \leq \frac{n\sum_{i=1}^{n}\left((x_i)^2 + (y_i)^2\right)}{2\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.
\end{aligned} \tag{4}$$

According to formula (4), the formula of ρ_1 is as follows:

$$\rho_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.$$

The estimated value of ρ_2 is as follows:

$$\rho_2 = \frac{n\sum_{j=1}^{n}(x_j)^2 + n\sum_{j=1}^{n}(y_j)^2 - 2\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{2\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}.$$

According to formula (4), we can get $\rho_1 \leq \rho_2$. The relationship between ρ_1 and threshold ε is shown in Figure (3) ($\varepsilon = 0.8$). When $\rho_2 < \varepsilon$, as Result 1 in Figure (3) shows, ρ_1 must be $< \varepsilon$. When $\rho_2 \geq \varepsilon$, as shown in Result 2 in Figure (3), the relationship with X and Y cannot be determined at this time. If the mean value – $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – and standard deviation – $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i)^2}{n} - \left(\bar{X}\right)^2}$,

$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i)^2}{n} - \left(\bar{Y}\right)^2}$ – of these two features are calculated according to $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n}(x_i)^2$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n}(y_i)^2$ after the completion of the Reduce task, Figure (4) describes the execution process of the second MapReduce task. Each computing node receives the mean – $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – and standard deviation – $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i)^2}{n} - \left(\bar{X}\right)^2}$, $\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i)^2}{n} - \left(\bar{Y}\right)^2}$ – of these two feature columns. In the Shuffle stage, each node calculates the Pearson correlation coefficient of some data $r = \sum_{i=1}^{t}\left(\frac{x_i-\bar{X}}{\sigma_X}\right)\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)$, $t = \frac{|D|}{3}$, $1 \le i \le t$, $|D|$ is the total number of data. In the Reduce stage, the Reduce Task summarizes the data of each node and obtains the correlation $\rho(X,Y) = \frac{\sum_{j=1}^{num} r_j}{|D|}$ (num is the number of nodes, representing the features of the j^{th} node X and Y $(1 \le j \le \text{num})$) between X and Y features according to formula (1). The computing time will grow as the second MapReduce task run. □

The DFCEA algorithm uses the method of expanding ρ_1 to get ρ_2 and expands the multiplication result of any two eigenvalues in each data object and sums the expanded values. ρ_2 must be $> \rho_1$. As a result, a threshold that is fair will reduce the number of MapReduce tasks and increase calculating efficiency. To determine whether any two features are correlated, this study expands the threshold and compares ρ_2 with the expanded threshold ε_1.

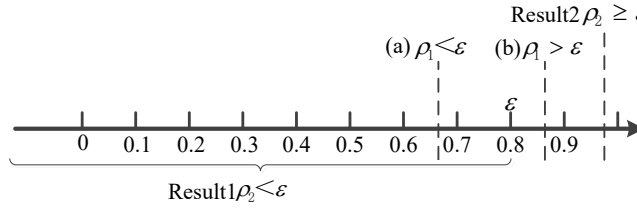


Figure 3. Relationship between ρ_2 and threshold ε

5 EXPERIMENTAL RESULTS

First, the configuration needed for the experiment is initially described in this chapter, including the number of clusters, the settings for each computing node, the threshold for the Pearson correlation coefficient, and the software version. Second, we analyze the correlation of massive ultra-high-dimensional radiomics feature data. Finally, the effectiveness, computational efficiency, and ability to solve practical problems of the DFCA and DFCEA algorithms are analyzed in detail.

5.1 Experimental Configuration

Cluster settings: This experiment runs on five machines with CentOS7.5 operating system. Each machine has a 4-core CPU and 8 gigabyte (GB) main memory.

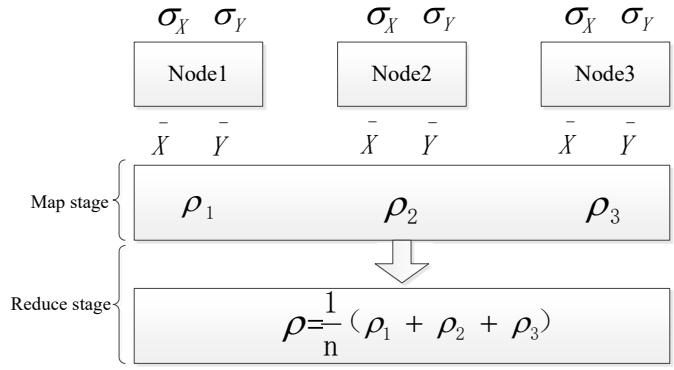


Figure 4. The processing flow of the second MapReduce task

The version of Hadoop in the distributed cluster environment is 2.7.3. One master node (the master node runs NameNode and ResourceManager processes) and four computing nodes (the computing nodes run DataNode and NodeManager processes) are set in the cluster. The physical memory of Map and Reduce containers is configured as 4 GB. Yarn's virtual memory to physical memory ratio is set to 6.

Dataset: Ultrasound images of breast tumors from the Affiliated Tumor Hospital of Xinjiang Medical University were used. First, we used Harr wavelet to process the ultrasound image, which generated four filtered images.

Subsequently, we used the filter in Table 2 to process the four filter images and the original image, which produced 11 sub-filter images. We extracted all the features in Table 1 except GLSZM, GLDM, and shape features from 11 sub-filter images (a total of 63 features). In the end, we obtained 20 000 data pieces (each data has 3 465 ($5 * 11 * 63$ features).

Parameter setting: The Pearson correlation coefficient's threshold is 0.9.

5.2 Experimental Comparison Algorithm

In this experiment, the following algorithms were compared to the DFCA and DFCEA algorithms:

1. Correlation analysis algorithm in single-machine environment is a common small sample analysis method in the medical analysis. We checked the execution efficiency of the algorithm on a large-sample dataset in one machine, which can be seen as the DFCA algorithm in one machine (SFCA algorithm).
2. When the DFCEA algorithm is run in a single-machine environment, the DFCEA algorithms can be thought of as a single node computing partial data in a distributed environment (SFCEA algorithm).

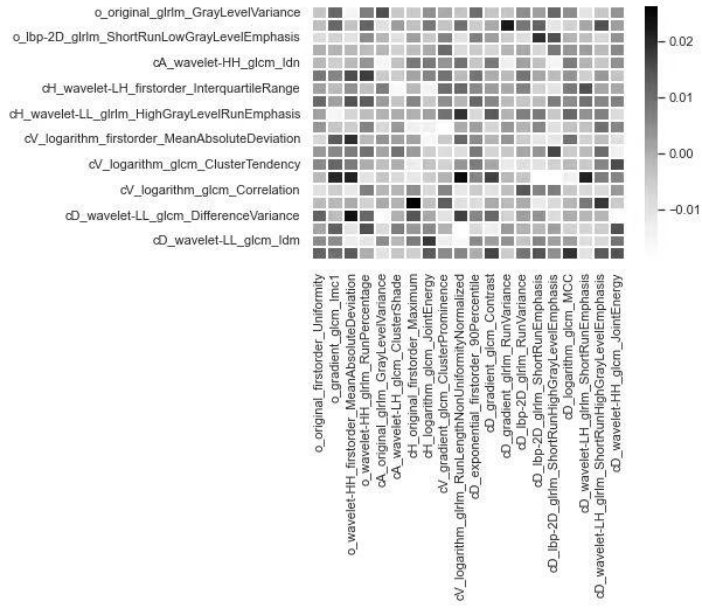
3. The experiment was contrasted with the DFCA and CCCA-LTS algorithms because there is no other research on the Pearson correlation coefficient of massive ultra-high-dimensional data in distributed computing outside the CCCA-LTS algorithm [17].

5.3 Experimental Results and Analysis

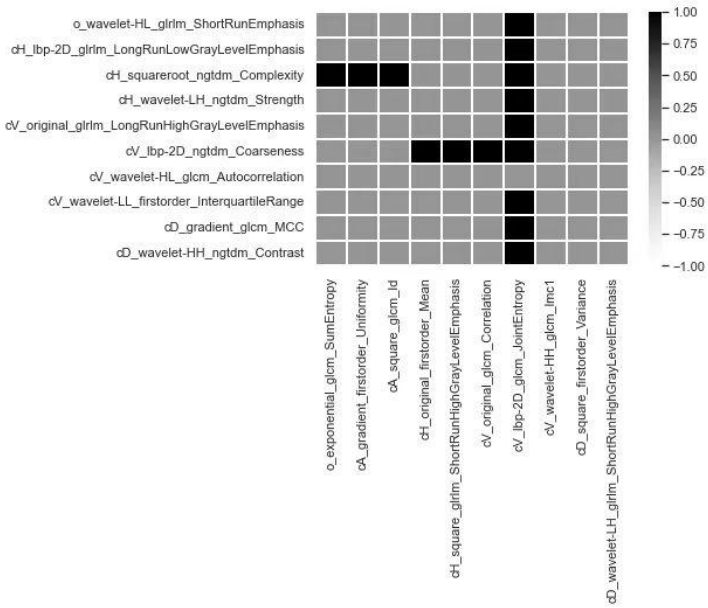
1.Effectiveness of the algorithm. With the use of the aforementioned dataset, we confirmed the efficacy of the DFCA and DFCEA algorithms. The DFCA algorithm was run to obtain the correlation coefficient of any two columns in the dataset, as shown in Figure 5 a). (Each feature name is separated by "_". The first part represents the preprocessing operation of the image before feature extraction. o represents the unprocessed operation. cA, cH, cV, and cD represent the sub-images generated from the Haar wavelet-processed images; the second part is the selection of the filter before feature extraction, as shown in Table 2. The third and fourth parts represent specific features under a certain feature class). A darker cell indicates a higher correlation. The correlation between o_wavelet-HH_fir-storder_MeanAbsoluteDeviation and cD_wavelet-LL_glcm_Difference Variance is significantly higher than o_original_glrlm_Gray Level Variance. The DFCEA algorithm is an estimation algorithm, which can judge the relationship between correlation coefficient and threshold of any two features without accurately calculating Pearson correlation coefficient, as shown in Figure 5 b). The DFCEA algorithm adopts the form of estimation. The results of any two features either be correlated or uncorrelated. In Figure 5 b), black indicates correlation and gray indicates noncorrelation.

2. Runtime evaluation. We contrasted the DFCA and DFCEA algorithms with the SFCA, SFCEA, and CCCA-LTS algorithms.

Through the analysis of the aforementioned dataset, Figure 6 shows the results of the computing time. When dealing with high-dimensional data, DFCA requires to randomly choose two feature data to multiply, which will generate data that will be stored in memory and cause a memory overflow. Therefore, one machine is not suitable to analyze the massive ultra-high-dimensional data. The DFCEA algorithm uses distributed parallel computing, which is superior to other algorithms since the DFCEA reduces the CPU and memory usage of a single node and increases computing efficiency. In distributed computing, the number of data transmitted between the nodes by the DFCEA algorithm is less than that of the DFCA algorithm, which greatly reduces the I/O cost caused by the data transmission between nodes. The DFCEA algorithm offers the largest advantage of employing estimation to reduce the execution of MapReduce tasks and enhance processing efficiency when compared to the CCCA-LTS algorithm, which necessitates two MapReduce processing stages.



a) DFCA algorithm



b) DFCEA algorithm

Figure 5. Correlation between some features

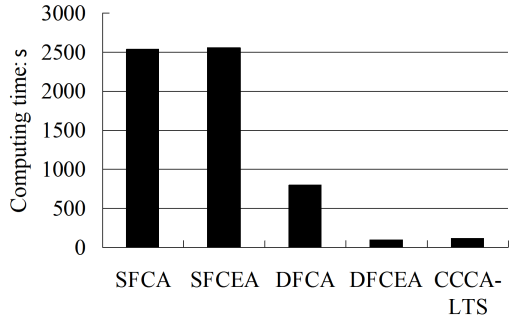


Figure 6. Comparison of each algorithm's computing times

3. Threshold expansion's effect on error rate. By repeatedly altering the multiples of the threshold expansion, as seen in Figure 7, the probability of estimate inaccuracy of the DFCEA algorithm while predicting Pearson correlation coefficient is confirmed. Since the threshold of the Pearson correlation coefficient is adopted by the CCCA-LTS algorithm and the threshold does not change during estimation, the error rate of CCCA-LTS algorithm is 12.8 % at the aforementioned dataset. To reduce the significant, I/O cost in the Shuffle stage, the DFCEA algorithm converts the multiplication of any two feature values in each data item into the sum of the squares of the two feature values. With the expansion of the threshold, the error rate gradually decreases and becomes flat. The threshold for the lowest error rate is required in this algorithm.

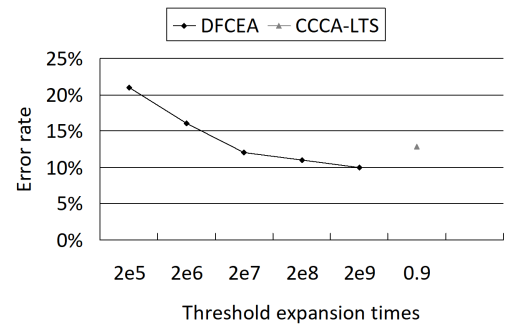
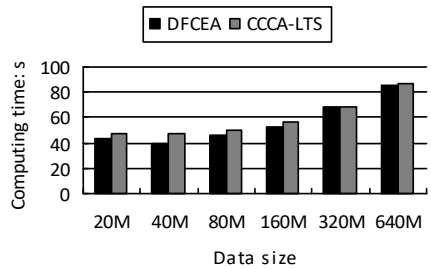


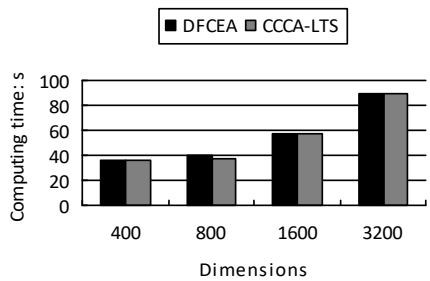
Figure 7. Expanding threshold's effect on the estimation results

4. Scalability. This section verifies the scalability of the DFCEA algorithm. With 3 465 dimensions unchanged, we verified the impact of the data size changes on the execution efficiency of the algorithm. The execution results are shown in Figure 8 a). When the dimensions constantly change and the data size remains at 20 000, the impact of data dimensions on the DFCEA algorithm is shown in

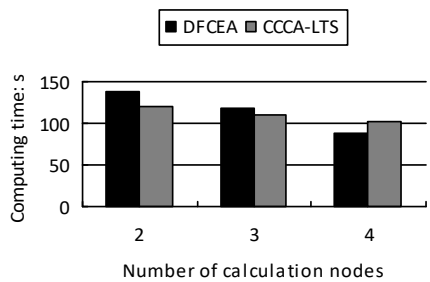
Figure 8 b). With the increase of dimension or data size, the running time of the DFCEA algorithm shows a linear growth trend. The calculation quantity of each node is decreased when the number of computing nodes is increased. When the 20 000 pieces of data (each data contains 3 465 dimensions) were calculated, the number of nodes increased while the running time is reduced. Figure 8 c) shows the experimental results.



a) The influence of the data volume on algorithm



b) The influence of the dimensions on algorithm



c) The influence of the number of nodes on the algorithm

Figure 8. Correlation between some features

5. Algorithm running time in different stages. In this section, the Map, Shuffle, and Reduce stage running times of the DFCA and DFCEA algorithms are compared. The results are shown in Figure 9. The task of the DFCA algorithm in the Map stage is to output $x_i, (x_i)^2, y_i, (y_i)^2, x_i y_i$ (i is the number of data, x and y are the two features) of each data object (assuming that the object has two features). The DFCEA algorithm only needs to compute four values $x_i, (x_i)^2, y_i, (y_i)^2$ in the Map stage and this reduces the I/O cost to improve the processing efficiency in the Shuffle stage. In the Reduce stage, the running time of the two algorithms is relatively close.

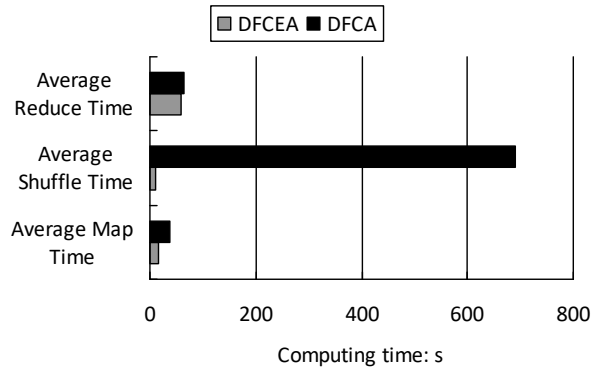


Figure 9. Expanding threshold's effect on the estimation results

6 CONCLUSION

This paper proposes the DFCA for ultra-high-dimensional radiomics feature data. We studied how to use the Pearson correlation coefficient to make the correlation calculation more efficient in a distributed environment. Also, we proposed an estimation approach, named DFCEA. DFCEA uses a MapReduce task to calculate the relationship between the Pearson correlation coefficient and the correlation threshold without accurately calculating it. Different from the CCCA-LTS algorithm, DFCEA only needs the form of estimation to reduce the amount of data processed in the Shuffle stage of MapReduce. The experimental results showed that the efficiency of the DFCEA algorithm is better than that of the traditional methods.

Data Sharing Agreement. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Conflicting Interests. The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding. This work is supported by Tianshan Talent of Xinjiang Uygur Autonomous Region – Young Top Talents in Science and Technology (No. 2022TSYCCY0008) and NSFC under Grant No. 61962058.

REFERENCES

[1] Lambin, P.—Rios-Velazquez, E.—Leijenaar, R. T. H.—Carvalho, S.—van Stiphout, R. G. P. M.—Granton, P. V.—Zegers, C. M. L.—Gillies, R. J.—Boellard, R.—Dekker, A.—Aerts, H. J. W. L.: Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. European Journal of Cancer, Vol. 48, 2012, No. 4, pp. 441–446, doi: 10.1016/j.ejca.2011.11.036.

[2] de Jong, E. E. C.—Sanders, K. J. C.—Deist, T. M.—van Elmpt, W.—Jochems, A.—van Timmeren, J. E.—Leijenaar, R. T. H.—Degens, J.—Schols, A. M. W. J.—Dingemans, A. M. C.—Lambin, P.: Can Radiomics Help to Predict Skeletal Muscle Response to Chemotherapy in Stage IV Non-Small Cell Lung Cancer? European Journal of Cancer, Vol. 120, 2019, pp. 107–113, doi: 10.1016/j.ejca.2019.07.023.

[3] Le, E.: Radiomics and Machine Learning in the Prediction of Cardiovascular Disease. Ph.D. Thesis. University of Cambridge, 2021, doi: 10.17863/CAM.66688.

[4] Granata, V.—Grassi, R.—Fusco, R.—Galdiero, R.—Setola, S.—Palaia, R.—Belli, A.—Silvestro, L.—Cozzi, D.—Brunese, L.—Petrillo, A.—Izzo, F.: Pancreatic Cancer Detection and Characterization: State of the Art and Radiomics. European Review for Medical and Pharmacological Sciences, Vol. 25, 2021, No. 10, pp. 3684–3699, doi: 10.26355/eurrev_202105_25935.

[5] Wang, J.—Tang, S.—Mao, Y.—Wu, J.—Xu, S.—Yue, Q.—Chen, J.—He, J.—Yin, Y.: Radiomics Analysis of Contrast-Enhanced CT for Staging Liver Fibrosis: An Update for Image Biomarker. Hepatology International, Vol. 16, 2022, No. 3, pp. 627–639, doi: 10.1007/s12072-022-10326-7.

[6] Fave, X. J.—Zhang, L.—Yang, J.—Mackin, D. S.—Stingo, F. C.—Followill, D. S.—Balter, P. A.—Jones, A. K.—Gomez, D. R.—Court, L. E.: TU-D-207B-02: Delta-Radiomics: The Prognostic Value of Therapy-Induced Changes in Radiomics Features for Stage III Non-Small Cell Lung Cancer Patients. Medical Physics, Vol. 43, 2016, No. 6, pp. 3750–3750, doi: 10.1118/1.4957510.

[7] Bourbonne, V.—Vallières, M.—Lucia, F.—Doucet, L.—Visvikis, D.—Tissot, V.—Pradier, O.—Hatt, M.—Schick, U.: MRI-Derived Radiomics to Guide Post-Operative Management for High-Risk Prostate Cancer. Frontiers in Oncology, Vol. 9, 2019, Art. No. 807, doi: 10.3389/fonc.2019.00807.

[8] Cheng, Z.—Huang, Y.—Huang, X.—Wu, X.—Liang, C.—Liu, Z.: Effects of Different Wavelet Filters on Correlation and Diagnostic Performance of Radiomics Features. Journal of Central South University: Medical Sciences, Vol. 44, 2019, No. 3, pp. 244–250, doi: 10.11817/j.issn.1672-7347.2019.03.003 (in Chinese).

[9] Zhang, S.—Song, G.—Zang, Y.—Jia, J.—Wang, C.—Li, C.—Tian, J.—Dong, D.—Zhang, Y.: Non-Invasive Radiomics Approach Potentially Predicts Non-Functioning Pituitary Adenomas Subtypes Before Surgery. European Radiology, Vol. 28, 2018, No. 9, pp. 3692–3701, doi: 10.1007/s00330-017-5180-6.

[10] Cong, Y.—Liu, J.—Fan, B.—Zeng, P.—Yu, H.—Luo, J.: Online Similarity Learning for Big Data with Overfitting. IEEE Transactions on Big Data, Vol. 4, 2018, No. 1, pp. 78–89, doi: 10.1109/TBDATA.2017.2688360.

[11] Juan, M.—Yu, J.—Peng, G.—Jun, L. J.—Feng, S. P.—Fang, L. P.: Correlation Between DCE-MRI Radiomics Features and Ki-67 Expression in Invasive Breast Cancer. Oncology Letters, Vol. 16, 2018, No. 4, pp. 5084–5090, doi: 10.3892/ol.2018.9271.

[12] Fu, L.—Li, Y.—Cheng, A.—Pang, P.—Shu, Z.: A Novel Machine Learning-Derived Radiomic Signature of the Whole Lung Differentiates Stable from Progressive COVID-19 Infection. Journal of Thoracic Imaging, Vol. 35, 2020, No. 6, pp. 361–368, doi: 10.1097/RTI.0000000000000544.

[13] Oubel, E.—Beaumont, H.—Iannessi, A.: Mutual Information-Based Feature Selection for Radiomics. In: Zhang, J., Cook, T. S. (Eds.): Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations. Proceedings of SPIE, Vol. 9789, 2016, doi: 10.1117/12.2216746.

[14] Hadar, U.—Shayevitz, O.: Distributed Estimation of Gaussian Correlations. IEEE Transactions on Information Theory, Vol. 65, 2019, No. 9, pp. 5323–5338, doi: 10.1109/TIT.2019.2913384.

[15] Lee, K. S.—Tsai, M. F.—Huang, C. S.: Distributed Correlation-Based Clustering Mechanism for Large-Scale Datasets. Preprints, 2020, Art. No. 2020110010, doi: 10.20944/preprints202011.0010.v1.

[16] Palma-Mendoza, R. J.—de Marcos, L.—Rodriguez, D.—Alonso-Betanzos, A.: Distributed Correlation-Based Feature Selection in Spark. Information Sciences, Vol. 496, 2019, pp. 287–299, doi: 10.1016/j.ins.2018.10.052.

[17] Liu, W.—Li, S.—Pan, J.—Xu, L.—Gu, Z.—Hai, L.—Deng, Y.: Design of Management Platform Architecture and Key Algorithm for Massive Monitoring Big Data. Wireless Communications and Mobile Computing, Vol. 2021, 2021, Art. No. 3111844, doi: 10.1155/2021/3111844.

[18] Zwanenburg, A.—Stefan, L.—Vallières, M.—Löc, S.: Image Biomarker Standardisation Initiative - Feature Definitions. CoRR, 2016, doi: 10.48550/arXiv.1612.07003.

[19] PyRadiomic – Welcome to PyRadiomic Documentation. 2022, `https://pyradiomics.readthedocs.io/en/latest/`.

Yuehong Tang received her B.Sc. degree in clinical nutrition from the Xinjiang Medical University, in 2005, and her M.Sc. degree in epidemic and health statistics from the Xinjiang Medical University, Xinjiang, China, in 2008. She is currently a Ph.D. student in the Xinjiang Medical University. Her research interests include epidemiology and health statistics.

Yan Chen received her Ph.D. degree in health toxicology from the Chinese Center for Disease Control and Prevention, Beijing, China, in 2003. She is currently Professor in the Jiaxing University. Her research interests include environment and health, chemical carcinogenic mechanisms and interventions.

Wen Liu received his B.Sc. degree in computer science from the Xinjiang Normal University, Xinjiang, in 2004, and his Ph.D. degree in computer science from the Dalian University of Technology, Dalian, China, in 2009. He is currently Professor in the College of Control Engineering, Xinjiang Institute of Engineering. His research interests include database, stream data processing, and cloud computing.

Zheng Gu received her B.Sc. degree in measurement and control technology and instrumentation from the Tianjin University of Technology and Education, Tianjin, in 2012, and her M.Sc. degree in instrumentation science and technology from the Southwest Petroleum University, Chengdu, in 2015. She is currently Associate Professor in the College of Control Engineering, Xinjiang Institute of Engineering. Her research interests include artificial intelligence and image processing.

Hui Yao received his M.Sc. degree in computer science from the Xinjiang Normal University, Xinjiang, China, in 2023. He is currently a Ph.D. student in the Xinjiang Medical University. His research interests include artificial intelligence and image processing.