Journal of Engineering Science and Technology Review 16 (1) (2023) 85 - 91

Research Article

JOURNAL OF
**Engineering Science and Technology Review**

www.jestr.org

# A Triple-negative Breast Cancer Classification Algorithm based on High-dimensional Big Data of Breast Ultrasound Radiomics

**Wen Liu [1,2,3], Hui Yao [1], Xiaoling Leng [4], Yutong Xue [3, 5] and Binlin Ma[4,*]**

[1]*College of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China*
[2]*Artificial Intelligence and Smart Mine Engineering Technology Center, Xinjiang Institute of Engineering, Urumqi 830023, China*
[3]*Xinjiang Changsen Data Technology Co., Ltd, Urumqi 830011, China.*
[4]*Department of Ultrasound, the Affiliated Tumor Hospital of Xinjiang Medical University, Urumqi 830011, China*
[5]*Institute of Noise Control Engineering (INCE-USA), Reston, VA 20191, United States*

Received 17 December 2022; Accepted 20 February 2023

***Abstract***

In computer-aided diagnosis, breast cancer classification accuracy of artificial intelligence (AI) method might be influenced by imbalanced classification samples. Model training efficiency also decreases with the increase in data size. To decrease the influences of sample classification imbalance and increase data size on breast cancer classification efficiency, this study proposed a triple-negative breast cancer classification algorithm (TNBCC) based on similarity query. Similar breast cancer feature data could be stored in a similar region by establishing the TNBCC index. A pruning algorithm was designed based on the TNBCC index, which decreases similarity measurement with nodes of the index tree significantly during data input and query. Users could find data similar to the query object quickly and accurately through the TNBCC index using pruning algorithm, and they could obtain the molecular subtype of the query object according to the searched data. Results demonstrate that, on the real dataset provided by the Affiliated Tumor Hospital of Xinjiang Medical University, the TNBCC algorithm decreases the influences of sample classification imbalance on accuracy effectively compared with the existing AI algorithm, and the classification accuracy reaches 91%. Given the mass high-dimensional data, it takes 49 s to build the TNBCC index tree, with an efficiency higher than that of the existing AI algorithm. This study provides a good reference to improve the performances of breast cancer classification algorithm.

*Keywords:* Molecular subtypes of breast cancer, Radiomics, Big data, Similarity query, Index tree

## 1. Introduction

Breast cancer is a common cancer in woman around the world. According to statistics of International Agency for Research on Cancer of the World Health Organization, the global population diagnosed with breast cancer was about 2.26 million in 2020, which was higher than the population diagnosed with lung cancer (about 60,000) [1]. With the continuous development of medical imaging technologies and computer-aided diagnosis, determining the molecular subtype of breast cancer by artificial intelligence (AI) technology becomes the current focus of study [2]. Since the AlexNet model has won the championship of the ImageNet Competition in 2012, deep learning algorithm has been studied continuously in the field of image processing [3]. In the medical field, deep learning algorithm is often applied to classification tasks of molecular subtype of breast cancer [4]. Ha R et al. predicted the molecular subtype of breast cancer by using deep learning [5]. Data with few classifications were input in multiples to decrease the influences of classification imbalance in the dataset, and the prediction accuracy was 70%. Jiang et al. trained the ResNet50 model on single center dataset, and its accuracy on two external test sets were 88% and 92%, respectively [6-7].

With the image data accumulation of breast cancer, directly using mass data with multi-type features is difficult. For example, scholars usually balanced sample classification by using oversampling and undersampling methods under situations of sample classification imbalance [8]. AI algorithm also costs abundant memory and central processing unit (CPU) (or graphics processing unit: GPU) resources and time in the model training stage. For instance, convolutional neural network takes 43 s to train 300 pieces of 35-dimensional data in an environment with Intel(R) Core i5-5200U 2.20 GHz CPU and 8G memory.

On this basis, sample classification imbalance and improvement of model training efficiency of AI algorithm have been widely reported [9-11]. However, sample classification imbalance is still processed using sampling and synthesis method on real dataset. This approach still has a disadvantage of long training time for mass data classification model.

To address this problem, the molecular subtype classification algorithm of breast cancer based on similarity query [12] was proposed. First, features of breast cancer were extracted from relevant historical images in hospitals by using radiomics technology [13]. Then, they were stored in a database by the triple-negative breast cancer classification (TNBCC) algorithm. Next, similarity of query objects was measured according to radiomics features of breast cancer through the TNBCC algorithm and data in the database. This way aims to search the objects with the shortest feature data distance to the query objects and return the molecular subtype of breast cancer. The TNBCC

*E-mail address: mbldoctor@126.com
ISSN: 1791-2377 
doi:10.25103/jestr.161.11

algorithm classifies molecules of breast cancer according to similarity query of radiomics features of breast cancer for high accuracy of molecular subtyping. Thus, references to improve the performances of the algorithm for breast cancer classification are provided.

## 2. State of the art

Nowadays, scholars are actively studying the molecular subtype of breast cancer based on AI algorithm. Convolutional neural network is a common method of image classification which trains model after processing of the original data [14-15]. This method requires multiple preprocessing of data, which takes a lot of time. Zhu Z et al. proved through an experiment that the GoogLeNet model was superior to other algorithms in terms of luminal A-type breast cancer classification [16-17]. However, the dataset of this study had imbalance in molecular subtypes of breast cancer, which influenced the prediction accuracy of deep learning algorithm significantly. Yang et al. trained the breast cancer classification model based on magnetic resonance imaging (MRI) by using the traditional convolutional neural network and convolutional long short-term memory (CLSTM), and the accuracy of CLSTM reached 91% [18]. This algorithm took a long time and consumed high software and hardware resources for model training on mass datasets, and the experiment was only applicable to MRI images. Moon et al. trained breast cancer models by using different deep learning models and chose the optimal models as the reference model for classification; then, the images were classified through ensemble learning method [19]. Virmani et al. classified ultrasound images of breast cancer by combining algorithms such as VGG19 and GoogleNet and transfer learning [20]. This algorithm combined different models to improve the accuracy of corresponding datasets in this study, which was inapplicable to different dataset. Since the radiomics was proposed, cancer features were transferred into measurable data, which built a bridge between descriptive and forecasting models of molecular subtype of breast cancer [21-22]. However, the forecasting efficiency of models might be influenced if all features were applied to the model training of machine learning algorithm. Thus, features have to be selected before model training. The combination of breast cancer radiomics features and machine learning algorithm becomes the major research topic at present [23]. Li W et al. classified breast cancer molecules based on extracted feature data through logic regression, random forest, gradient-boosted tree, and support vector machines (SVMs) [24]. For high-throughput extraction features of breast cancer in MR, Hui et al. chose radiomics features related with breast cancer molecular subtypes [25]. Fan et al. extracted cancer features through contrast-enhanced MRI (DCE-MRI) of breast cancer and predicted molecular subtypes of breast cancer using multi-class logistic regression classifier [26]. Laajili R et al. classified benign and malignant breast cancers by using feature selection and machine learning model, and the classification accuracy was 85% [27]. Different molecular subtypes of breast cancer become a factor that influences the accuracy of this algorithm. Preprocessing should be conducted before model training to improve accuracy of the model for decreasing the influences of data sample classification imbalance and abnormal values [28-29]. However, this method still has poor performances in radiomics of hybrid imaging. The abovementioned analyses on radiomics features of breast cancers have the problem of classification sample imbalance.

The aforementioned studies focus on breast cancer classification, but few works are available on increasing utilization of source data, decreasing generation of synthesized data, and improving training efficiency of breast cancer classification model. In this study, an ultrasound breast cancer classification method based on similarity query was proposed. It classifies breast cancers by searching feature data which are the closest to the query object through similarity query. Based on improvement in radiomics feature data utilization and accuracy of breast cancer, it further increases similarity query efficiency by building the similarity index and pruning algorithm.

The remainder of this study is organized as follows. Section 3 introduces definitions needed by TNBCC of breast cancer based on similarity query. Meanwhile, production of the TNBCC index structure during inputting of radiomics feature data of breast cancer and TNBCC process based on similarity query are introduced. Section 4 analyzes the experimental results. Section 5 summarizes the conclusions.

## 3. Methodology

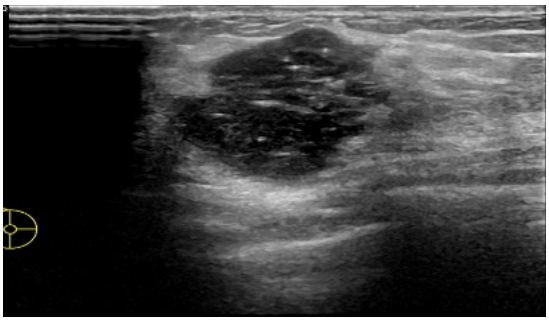
The definitions needed to build the TNBCC index were introduced first in this study, and they are the foundation to build the TNBCC index tree. Next, the construction of the TNBCC index tree was described. Finally, similarity query of radiomics feature data of breast cancer was realized through TNBCC index to conduct molecular subtyping of breast cancers.

### 3.1 Definitions related with the TNBCC index

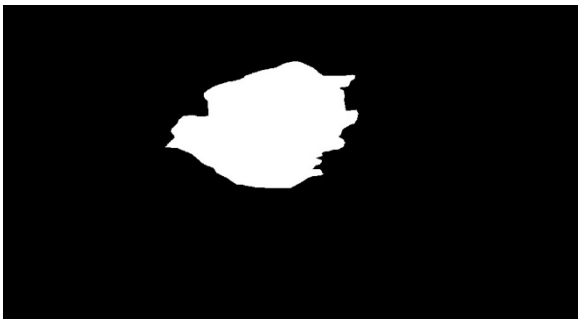
**Definition 1**: Extraction of ultrahigh-dimensional ultrasound radiomics feature data of breast cancer. Radiomics features included shape features, fusion and fractal features, first-order histogram features, and texture features. The ultrasonic image of breast cancer and corresponding mask pattern of breast cancer were given. The given ultrasonic image of breast cancer was filtered using different filters, and the filtered image may produce n images. A total of n*m (m is the number of extracted cancer features) of breast cancer features were extracted through different filtering images.

For example, breast cancer features were extracted using pyradiomics [30]. Ultrasound images of breast cancer and corresponding mask pattern are shown in Figs. 1(a) and 1(b), respectively

Next, Harr wavelet processing was performed to Fig. 1(a), which produced four different filtering images (Fig. 2). Subsequently, images were processed using Original, Wavelet, LoG, Square, SquareRoot, Exponential, Gradient, and LocalBinaryPattern2D filters (all can produce one filtering result except that wavelet can produce four filtering results). Finally, First Order Statistics (18 features), Gray Level Co-occurrence Matrix Features (GLCM, 24 features), Gray Level Run Length Matrix Features (GLRLM, 16 features), and Gray Level Dependence Matrix Features (GLDM, 14 features) were extracted, which produced a total of 3960 (5*11*72) features. Some data are shown in Table 1.



(a) Original ultrasonic image of breast cancer



(b) Corresponding mask pattern

**Fig. 1.** Ultrasonic image of breast cancer

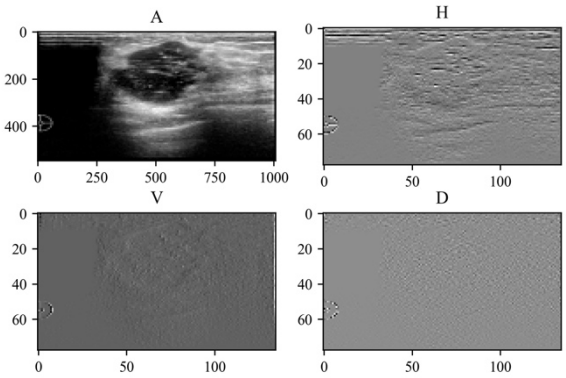



**Fig. 2.** Harr wavelet processing results of original ultrasonic images of breast cancer

**Table 1.** Some radiomics feature data of breast cancer

| o_original_firstorder_Mean | o_squareroot_firstorder_Uniformity | cA_lbp-2D_glrlm_RunLength NonUniformity | cH_square_ firstorder_Variance | cH_wavelet-HL_glrlm_ RunPercentage |
|---|---|---|---|---|
| 49.06202 | 0.171907 | 6.77663 | 1.314174 | 0.559408 |
| 47.69668 | 0.164514 | 6.92821 | 1.580532 | 0.555886 |
| 61.5427 | 0.188621 | 8.2809 | 0.275911 | 0.547224 |
| 58.32409 | 0.290501 | 5.979683 | 0.116707 | 0.577135 |

**Definition 2:** Euclidean distance. The given dataset (S) and two groups of data $D_1=\{x_1,x_2,...,x_n\}$ and $D_2=\{y_1,y_2,...,y_n\}$ (where $D_1,D_2 \in S$, $n(n>1)$ were used as dimensions of data; $x_i$ and $y_i$ $(1 \le i \le n)$ numerical values corresponding to feature $i^{th}$ in data objects $D_1$ and $D_2$). The measuring formula of Euclidean distance is $ED(D_1,D_2)=\sqrt{\sum_{i=1}^{|D_1|}(x_i-y_i)^2}$.

**Definition 3:** Molecular subtypes of radiomics feature data of breast cancer based on similarity query. Two groups of ultrasound radiomics feature data of breast cancer $D_1=\{x_1,x_2,...,x_n\}$, $D_2=\{y_1,y_2,...,y_n\}$ and the similarity threshold (Ɛ) were given. If $ED(D_1,D_1) \le \varepsilon$, then $D_1$ and $D_2$ are the same molecular subtypes of breast cancer. Otherwise, $D_1$ and $D_2$ are different molecular subtypes of breast cancer.

**Definition 4:** Triangle inequality. Three data objects $D_1$, $D_2$, and $D_3$ were considered. The three data objects were used as three vertexes, and they meet the triangle inequality principles $D_1D_2+D_2D_3>D_1D_3$ and $D_1D_2-D_2D_3|<D_1D_3$.

### 3.2 Building TNBCC index based on stored radiomics feature data of breast cancer

In this section, a three-layer TNBCC tree was built using the dataset $\{(D_1,L_1),(D_2,L_2),(D_3,L_3),...,(D_n,L_n)\} \in S(n>0)$. Nodes in each layer were also placed in an ascending order according to distance to the father node. Specifically, $D_i$ and $L_i(1 \le i \le n)$ are the $i^{th}$ image of breast cancer and the corresponding molecular subtype. The structure of the TNBCC tree is shown in Fig. 3.

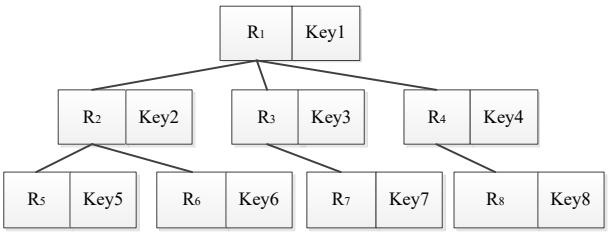



**Fig. 3.** Structure of TNBCC tree

The key structure is Level@Disparent@Rid, where Level is the layer number of the reference node in the TNBCC, @ is the separator, Disparent is the threshold of distance from the reference node to the father node, and Rid

is the id of the reference node data. When $D_i$ is similar to a reference point, its key uses the value of key attribute in the reference node, splices it with the id of $D_i$, and stores it.

Theorem 1: Pruning principle. Three pieces of data (R, D, and Q) were given. R is the reference node, D is the subnode of R, and Q is the query node. The similarity threshold (Ɛ) was given. According to Definitions 2 and 3, R, D, and Q form a triangle (Fig. 4). If $ED(R,Q) \le \varepsilon$, then Q and R are similar. If $ED(R,Q) > \varepsilon$, then the subnode of R was pruned. According to the triangle inequality, let $|ED(R,Q) - ED(R,D)| < \varepsilon$, which yields $|\varepsilon - ED(R,Q)| < ED(R,D) < \varepsilon + ED(R,Q)$. Thus, the subnodes with a distance to the father node larger than $|\varepsilon - ED(R,Q)|$ were pruned.

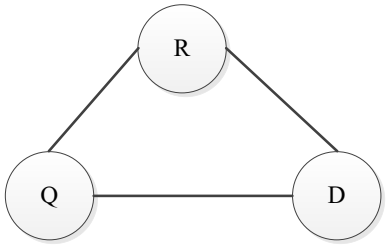



**Fig. 4.** Pruning principle

Each node in the index tree is a reference node of one region of similarity. When the database is empty, the TNBCC index tree is empty. The first piece of radiomics feature data of breast cancer $D_1$ was input. The TNBCC index tree viewed $D_1$ as the root node (or known as the reference point $R_1$). Meanwhile, the index of $D_1$ was built and stored. During the input of $D_i$, the similarity between $D_i$ and $D_1$ was measured first according to Definitions 2 and 3. If $ED(D_1,D_i) \le \varepsilon$, then the query index of $D_i$ was built and stored into the database. Otherwise, pruning calculation of the reference node of the root node was conducted using triangle inequality. If $2\varepsilon > ED(D_1,D_i) > \varepsilon$, then similarity between $D_i$ and $R_2$ was calculated. $D_i$ was used as the reference point $R_2$ when $R_2$ is empty. If $3\varepsilon > ED(D_1,D_i) > 2\varepsilon$, then similarity between $D_i$ and $R_3$ was calculated. $D_i$ was used as the reference point $R_3$. when $R_3$ is empty. The rest was performed in the same way. During the input of $D_j$, the distance between $D_j$ and $R_1$ is supposed to be $2\varepsilon > ED(D_1,D_i) > \varepsilon$. Then, similarity between $D_j$ and $R_2$ was measured. Under this circumstance, if $ED(D_j,R_1) > \varepsilon$, then $R_2$ was used as the root node and the subnode of $R_2$ was obtained. The construction algorithm of TNBCC is shown in Algorithm 1. The batch storage of historical ultrasonic data of breast cancer in hospitals is shown in Algorithm 2.

Algorithm 1 Building the index tree for radiomics features similarity query of breast cancer TNBCCA (Node root, Node newNode).

Input: TNBCC and query node (Q)
Output: TNBCC
1. **if** TNBCC is null **then**
2. TNBCC.root=new Node(Q)
3. **else then**
4. **if** TNBCC.root.getSons() is not null **then** // getSons() acquire the subnode
5. rNode=TNBCC.root.getSons().getRf(ED(TNBCC.root,new Node(d)) // getRf() acquire the reference node
6. **if** ED(rNode,new Node(d))<=Ɛ **then**
7. insertDatabase(d) // insertDatabase() store data into the database
8. **else**
9. **for** n ∈ rNode **do**
10: TNBCCA(n,newNode) //iterate
11: **endfor**
12. **endif**
13. **endif**
14. **endif**

Algorithm 2 Building TNBCC index tree.

Input: Dataset $\{D_1,D_2,D_3,...,D_n\} \in S$, similarity query threshold (Ɛ).
Output: TNBCC index
1. **for** i in S **do**
2. d=SoftMax(i) //Standard normalization by using softMax function
3. TNBCCA (Node root, new Node(d))
4. **endfor**

### 3.3 Molecular subtype of radiomics features of breast cancer

According to the TNBCC index tree produced in Section 3.2, this section describes the molecular subtyping of breast cancer by using the TNBCC index tree and pruning algorithm.

The HBase key value database was taken as an example [31]. The stored data order of HBase is defaulted as the dictionary order of RowKey. Thus, similar ultrasound radiomics feature data of breast cancer were stored together according to RowKey, which was reconstructed according to the TNBCC index tree when building the TNBCC index tree. As a result, the region of similarity was formed.

The query process was similar to the abovementioned process of index building. Fig. 3 was taken as an example. Step 1: Measure similarity between Q and $R_1$. If $ED(Q,R_1) \le \varepsilon$, then search within the key of $R_1$. Meanwhile, similarity between data in the region of similarity which uses $R_1$ as the reference node and Q was measured, and the closest node to Q was returned, as shown in 1 in Fig. 5. If $ED(Q,R_1) > \varepsilon$, then pruning of $R_1$ was performed according to the pruning principle. If $ED(R_1,R_3) > |\varepsilon - ED(Q,R_1)|$, then $R_3$ and $R_4$ were pruned, as shown in 2 in Fig. 5. The TNBCC index tree was iterated using $R_2$ as the root node to search the region of similarity of Q. The molecular subtyping of radiomics feature data of breast cancer based on similarity query is shown in Fig. 5.

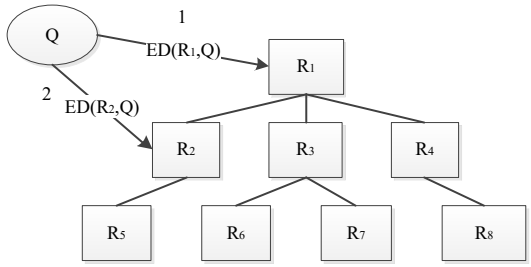



**Fig. 5.** Molecular subtyping of radiomics feature data of breast cancer based on similarity query

## 4. Result Analysis and Discussion

### 4.1 Experimental settings

Cluster setting: This experiment was operated on 5 servers with an operating system of CentOS 7.5. This cluster of servers has 8 GB memory, 6-core CPU, and 64G nave protocol m.2 disc (reading speed: 2400 MB/S; writing speed: 1750MB/S), including 1 master node and 5 calculation nodes. Hadoop 2.7.3, Zookeeper 3.4.6, and HBase 1.3.1 were applied.

Dataset: The image dataset of breast cancer with 9500 images (privacy information has been processed) provided by the Affiliated Tumor Hospital of Xinjiang Medical University was used. According to Definition 1, a total of 3960 features were extracted from each image.

Parameter setting: Radiomics feature data of breast cancer in this experiment were all normalized. Similarity threshold after data normalization was set to 19

### 4.2 Experimental results

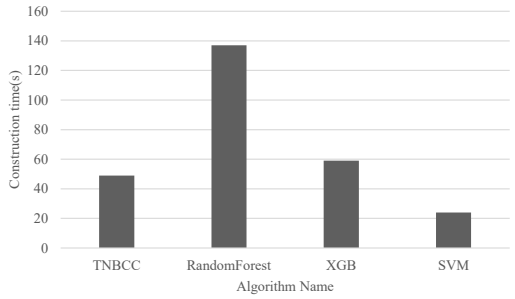
Based on the TNBCC algorithm described in Section 3, the experiment compared the proposed TNBCC algorithm with machine learning algorithm in terms of accuracy. No study is available yet on molecular subtyping of ultrasound radiomics feature data of breast cancer based on similarity query. Thus, the TNBCC was compared with Brute force Query (BQ) and Not Prune-Three negative breast cancer classification (NP-TNBCC) in this experiment in terms of data size and dimension.

(1) Time for model building and accuracy comparison

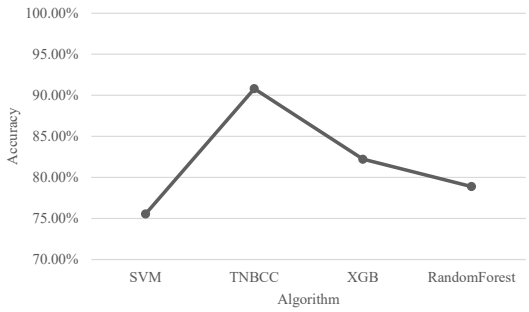
The TNBCC algorithm implements molecular subtyping of breast cancer based on similarity query. It does not involve training of model parameters, but it requires storage of historical data to build the TNBCC index. The time for building TNBCC index was compared with the time for model building (single-machine processing) of machine learning algorithm (Fig. 6(a)). Breast cancer data were transformed into measurable clinical index data through radiomics technology, and the breast cancer feature data with the shortest distance to the query node were measured by Euclidean distance. Comparison of TNBCC, SVM, XGB, and RandomForest algorithms in terms of accuracy is shown in Fig. 6(b). The TNBCC algorithm is superior to XGB and RandomForest algorithms. Although time for building the TNBCC algorithm is longer than that for SVM, the accuracy is improved compared with that of SVM.

(2) Effects of threshold on accuracy of TNBCC algorithm

The TNBCC algorithm measured similarity of ultrasound radiomics feature data of breast cancer by using Euclidean distance, which realized molecular subtyping of breast cancer. The effects of different similarity thresholds on the classification results are shown in Fig. 7. The best accuracy is achieved when the similarity threshold is set to 19.




(a) Time for building TNBCC index and time for building machine learning model




(b) Accuracies of different algorithms

**Fig. 6.** Comparison of evaluation indexes of different algorithms

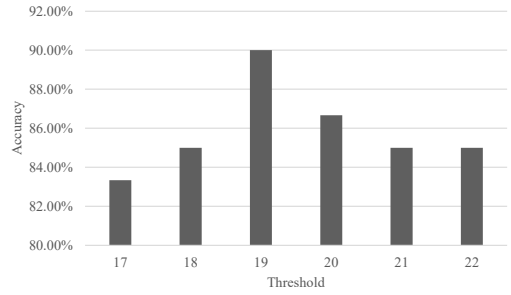
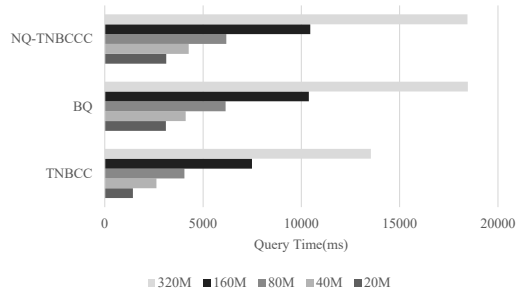


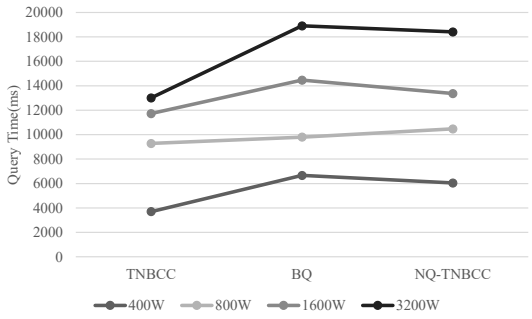

**Fig. 7.** Effects of threshold on the accuracy of the TNBCC algorithm

(3) Performances of the TNBCC algorithm

During similarity query, the BQ algorithm has to transverse all data in the database to measure similarity of ultrasound radiomics feature data of breast cancer. The NQ-TNBCC algorithm takes longer time than the TNBCC algorithm in searching the region of similarity at query nodes given that it has no pruning. The results are shown in Fig. 8(a). As the data dimension increases continuously, the TNBCC algorithm is influenced less by dimensions compared with other algorithms. The results are shown in Fig. 8(b).




(a) Effects of data size on the TNBCC algorithm




(b) Effects of dimension on the TNBCC algorithm

**Fig. 8.** Performances of the TNBCC algorithm

(4) Time complexity of the TNBCC algorithm

Adding pruning algorithm into TNBCC can decrease the times of reference node calculations when searching the region of similarity. The TNBCC algorithm calculates at least one reference node and at most three reference nodes in each layer of the index tree. Therefore, time complexity when searching the region of similarity was $O(m)$, where m refers to the number of layers in the TNBCC index tree, and it is far lower than the data size in the database. After the region of similarity was positioned at the query node, TNBCC has to transverse data in each region of similarity and measure similarity with query points. The time complexity was $O(p)$, where p is the data size in the region of similarity, and it is significantly lower than the data size in the database.

## 5. Conclusions

A similarity index tree was built in this study to analyze subtypes of ultrasound radiomics feature data of breast cancer for improving molecular subtyping efficiency of breast cancer. The TNBCC index tree was built by storing historical data of the hospital into the key value database. TNBCC index tree was used for query of historical data similar to breast cancer features. A pruning algorithm was also designed. Calculation of unnecessary reference nodes was decreased by pruning, which increased inputting and query efficiency of radiomics feature data. The following conclusions could be drawn:

(1) Molecular subtyping of breast cancer based on similarity query can classify cancers quickly in mass data, and the classification accuracy is superior to that of the machine learning algorithm.

(2) Given that similarity query is applied, data skewing in AI algorithm may not influence the breast cancer classification accuracy of the TNBCC algorithm. During similarity query, all radiomics features of breast cancer participate in the calculation of similarity measurement.

In this study, a data classification method based on similarity query is proposed by combining theoretical and practice studies. It can provide some references to follow-up studies on clinical therapy of breast cancer. Breast cancer can also be separated to establish an ultrasound radiomics feature data analysis and subtyping system of breast cancers based on similarity query.

**Acknowledgements**

This work was supported by the National Natural Science Foundation of China (61962058), the Tianshan Talent of Xinjiang Uygur Autonomous Region - young Top Talents in Science and Technology (2022198841), the Scientific and Technological Assistance to Xinjiang (2020E0269), the Integration of Industry and Education-Joint Laboratory of Data Engineering and Digital Mine (2019QX0035), and the Bayingolin Mongolian Autonomous Prefecture Science and Technology Research Program (202117).



## References

1. Sung, H., Ferlay, J., R, L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". *CA: A Cancer Journal for Clinicians*, 71(3), 2021,pp.209-249.
2. Nissan, N., Sorin, V., Bauer, E., Anaby, D., Samoocha, D., Yagil, Y., Faermann, R., Halshtok-Neman, S., Shalmon, A., Gotlieb, M., Sklair-Levy, M., "MRI of the Lactating Breast: Computer-Aided Diagnosis False Positive Rates and Background Parenchymal Enhancement Kinetic Features". *Academic radiology*, 29(9), 2022,pp.1332-1341.
3. Krizhevsky, A., Sutskever, I., Hinton, G., "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems*, 60(6),2017,pp.84-90.
4. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Laak, J. A., Ginneken, B., Sánchez, C. I., "A survey on deep learning in medical image analysis". *Medical Image Analysis*, 42, 2017,pp. 60-88.
5. Ha, R., Mutasa, S., Karcich, J., Gupta, N., Sant, E. P. V., Nemer, J., Sun, M., Chang, P., Liu, M. Z., Janbawalikar, S., "Predicting Breast Cancer Molecular Subtype with MRI Dataset Utilizing Convolutional Neural Network Algorithm". *Journal of Digital Imaging*, 32(2), 2019,pp.276-282.
6. Jiang, M., Zhang, D., Tang, S. C., Luo, X. M., Chuan, Z. R., Lv, W. Z., Jiang, F., Ni, X. J., Cui, X. W., Dietrich,C.F., "Deep learning with convolutional neural network in the assessment of breast cancer molecular subtypes based on US images: a multicenter retrospective study". *European Radiology*, 31(6), 2021,pp.3673-3682.
7. He, K. M., Zhang, X. Y., Ren, S.Q., Sun, J., "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA:IEEE, 2016, pp. 770-778.
8. Rodríguez, N., López, D., Fernández, A., García, S., Herrera, F., "SOUL: Scala Oversampling and Undersampling Library for imbalance classification". *SoftwareX*, 15, 2021,pp.100767.
9. Bulut, O., "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining". *Information*, 14(1), 2023,pp.54.
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., "*ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*" [EB/OL] Retrieved from https://arxiv.53yu.com/abs/1909.11942, 2019-09/2022-11.
11. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q., "TinyBERT: Distilling BERT for Natural Language Understanding". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Pontacana, Dominican Republic: Computation and Language 2020, pp.4163-4174.

12. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W., "Efficient and effective Querying by Image Content". *Journal of Intelligent Information Systems*, 3(3), 1994,pp.231-262.
13. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Stiphout, R. G. P. M.V., Granton, P., Zegers, C. M.L., Gilies, R., Boellard, R., Dekker, A. Aerts. H., "Radiomics: Extracting more information from medical images using advanced feature analysis". *European Journal of Cancer*, 43(4), 2012,pp.441-446.
14. Tsochatzidis, L., Koutla, P., Costaridou, L., Pratikakis, L., "Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses". *Computer Methods and Programs in Biomedicine*, 200, 2021,pp.105913
15. Siddeeq, S., Li, J., Bhatti, H., Manzoor, A., Malhi, U., "Deep Learning RN-BCNN Model for Breast Cancer BI-RADS Classification". In: *ICIGP 2021: 2021 The 4th International Conference on Image and Graphics Processing. 2021*, Sanya, China: Association for Computing Machinery, 2021,pp.219-225.
16. Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M. R., Mazurowski, M. A., "Deep leaming for identifying radiogenomic associations in breast cancer". *Computers in Biology and Medicine*, 109, 2019,pp.85-90.
17. Szegedy, C., Liu, W., Jia, Y. Q., Sermannet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., "Going Deeper with Convolutions". *IEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA:IEEE, 2015,pp.1-9.
18. Zhang, Y., Chen, J.H., Lin, Y. Z., Chan, S.W., Zhou, J. J., Chow, D., Chang, P., Kwong, F., Yeh, D. C., Wang, X. X., Rarajuli, R., Mehta, R. S., Wang, M. H., Su, M. Y., "Prediction of breast cancer molecular subtypes on DCE-MRI using convolutional neural network with transfer learning between two centers". *European Radiology*, 31(4), 2021,pp.2559-2567.
19. Moon, W. K., Lee, Y. W., Ke, H. H., Lee, S. H., Huang, C. S., Chang, R. F., "Computer - aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks". *Computer Methods and Programs in Biomedicine*, 190, 2020,pp.105361.
20. Virmani J, Agarwal R., "Deep feature extraction and classification of breast ultrasound images". *Multimedia Tools and Applications*, 79(37), 2020,pp.27257-27292.
21. Limkin, E. J., Sun, R., Dercle, L., Zacharaki, E.I., Robert, C., Reuzé, S., Schernberg, A., Paragios, N., Deutsch, E., Ferté, C., "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology". *Annals of Oncology*, 28(6), 2017,pp.1191-1206.
22. Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., Jong, E. E. C., Timmeren, J. V., Sanduleanu, S., Larue, R. T.H.M., Even, A. J. G., Jochems, A., Wijk, Y. V., Woodruff, H., Soest, J.V., Lustberg, T., Roelofs, E., Elmpt, W. V., Dekker, A., Mottaghy, F. M., Wildberger, J. E., Walsh, S., "Radiomics: the bridge between medical imaging and personalized medicine". *Nature Reviews Clinical Oncology*, 14(12), 2017,pp.749-762.
23. Li, C. L., Song, L. R., Yin, J. D., "Intratumoral and Peritumoral Radiomics Based on Functional Parametric Maps from Breast DCE-MRI for Prediction of HER-2 and Ki-67 Status". *Journal of Magnetic Resonance Imaging: JMRI*, 54(3), 2021,pp.703-714.
24. Li, W., Yu, K., Feng, C.L., Zhao, D. Z., "Molecular Subtypes Recognition of Breast Cancer in Dynamic Contrast-Enhanced Breast Magnetic Resonance Imaging Phenotypes from Radiomics Data". *Computational and Mathematical Methods in Medicine*, 2019,2019,pp.6978650.
25. Li, H., Zhu, Y., Burnside, E. S., Huang, E., Drukker, K., Hoadley, K. A., Fan, C., Conzen, S. D., Zuley, M., Net, J. M., Sutton, E., Whitman, G. J., Morris, E., Perou, C. M., Ji, Y., Giger, M. L., "Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set". *NPJ Breast Cancer*, 2(1), 2016,pp.1-10.
26. Fan, M., Li, H., Wang, S., Zheng, B., Zhang, J., Li, LH., Alessandro, W., "Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer". *PLoS One*, 12, 2017,pp.e0171683.
27. RL, A., MS, B., MT, A., "Application of radiomics features selection and classification algorithms for medical imaging decision: MRI radiomics breast cancer cases study". *Informatics in Medicine Unlocked*, 2021,27: 100801.
28. Xie, C., Du, R., Ho, JWK., Pang, HH., Chiu, KWH., Lee, EYP., Vardhanabhuti, V., "Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients". *European Journal of Nuclear Medicine and Molecular Imaging*, 47, 2020,pp.2826–2835
29. Cysouw, M.C.F., Jansen, B.H.E., van de Brug, T., Oprea-Lager, D. E., Pfaehler, E., Vries, B. M., Moorselaar, R. J. A., Hoekstra, Otto S. Vis, A. N., Boellard, R., "Machine learning-based analysis of [18F]DCFPyL PET radiomics for risk stratification in primary prostate cancer". *European Journal of Nuclear Medicine and Molecular Imaging*, 48, 2021,pp.340-349.
30. Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L., "Computational Radiomics System to Decode the Radiographic Phenotype". *Cancer Research*, 77(21), 2017,pp.e104–e107.
31. George, L., "HBase: The Definitive Guide". *Andre*, 12(1), 2011,pp.1 - 4.