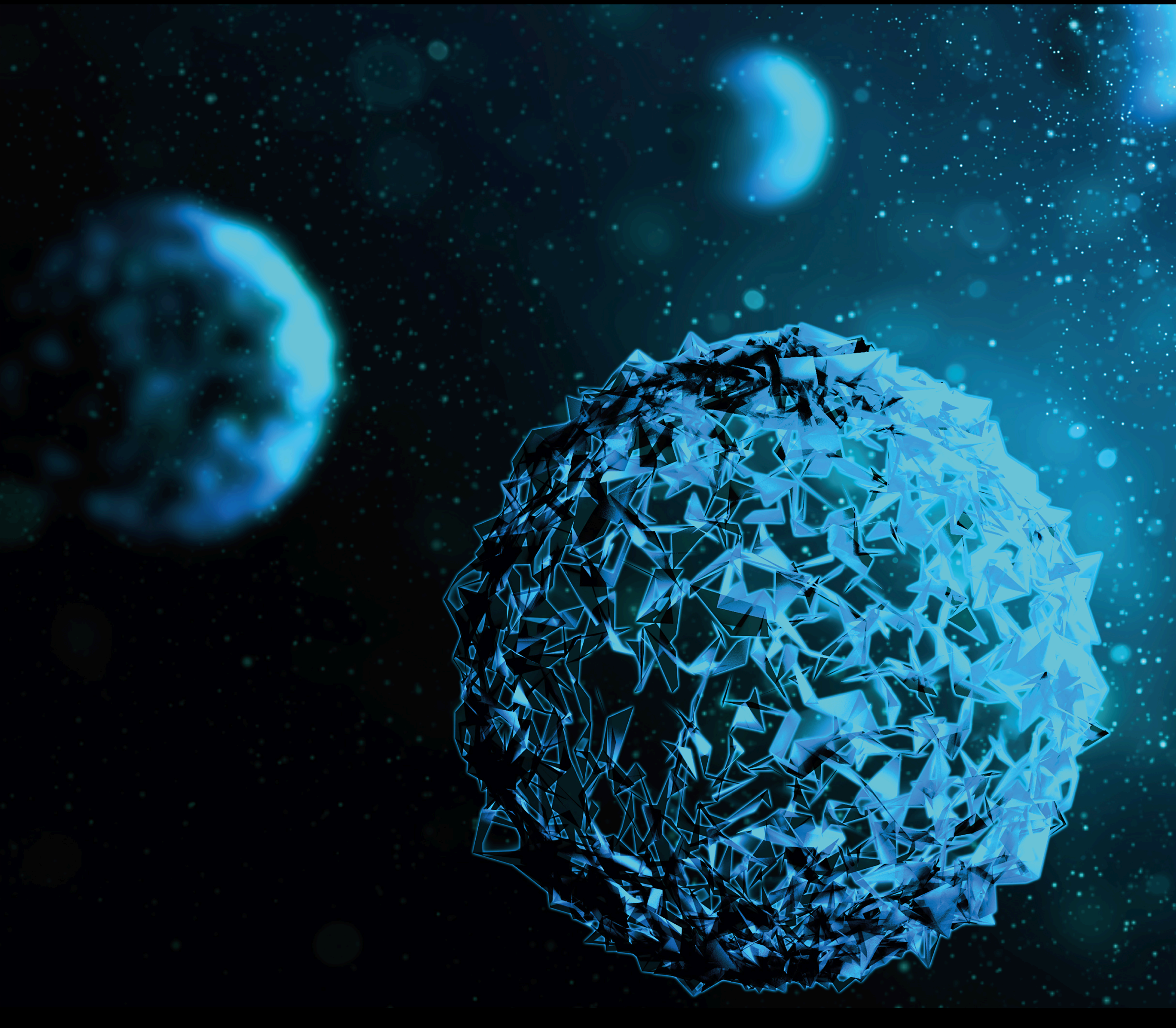
BioMed Research International

Representation Learning in Radiology

Lead Guest Editor: Shaode Yu
Guest Editors: Zhiguo Zhou, Erlei Zhang, and Wenjian Qin



Representation Learning in Radiology

BioMed Research International

Representation Learning in Radiology

Lead Guest Editor: Shaode Yu
Guest Editors: Zhiguo Zhou, Erlei Zhang, and
Wenjian Qin

Section Editors

Academic Editors

Radiology

Contents

Contents

Hindawi
BioMed Research International
Volume 2021, Article ID 6079163, 9 pages
https://doi.org/10.1155/2021/6079163

Research Article

Evaluation of Feature Selection Methods for Mammographic Breast Cancer Diagnosis in a Unified Framework

Chun-jiang Tian,[1] Jian Lv,[1] and Xiang-feng Xu[2]

[1]Department of Radiology, Tianjin Hospital of ITCWM Nankai Hospital, Tianjin 300100, China
[2]Department of Radiology, Tianjin Central Hospital of Obstetrics and Gynecology, Tianjin 300100, China

Correspondence should be addressed to Jian Lv; glavefall@vip.sina.com

Received 17 June 2020; Revised 10 July 2020; Accepted 18 July 2020; Published 4 October 2021

Guest Editor: Erlei Zhang

Over recent years, feature selection (FS) has gained more attention in intelligent diagnosis. This study is aimed at evaluating FS methods in a unified framework for mammographic breast cancer diagnosis. After FS methods generated rank lists according to feature importance, the framework added features incrementally as the input of random forest which performed as the classifier for breast lesion classification. In this study, 10 FS methods were evaluated and the digital database for screening mammography (1104 benign and 980 malignant lesions) was analyzed. The classification performance was quantified with the area under the curve (AUC), and accuracy, sensitivity, and specificity were also considered. Experimental results suggested that both infinite latent FS method (AUC, 0.866 ± 0.028) and RELIEFF (AUC, 0.855 ± 0.020) achieved good prediction ($\mathrm{AUC} \geq 0.85$) when 6 features were used, followed by correlation-based FS method (AUC, 0.867 ± 0.023) using 7 features and WILCOXON (AUC, 0.887 ± 0.019) using 8 features. The reliability of the diagnosis models was also verified, indicating that correlation-based FS method was generally superior over other methods. Identification of discriminative features among high-throughput ones remains an unavoidable challenge in intelligent diagnosis, and extra efforts should be made toward accurate and efficient feature selection.

1. Background

Feature selection (FS) or variable selection plays an important role in intelligent diagnosis. It is used to identify a subset of features or to weight the relative importance of features in target representation that makes a computer-aided diagnosis model cost-effective, easy to interpret, and generalizable. So far, FS methods have been explored in target recognition [1], logistic regression [2], disease detection and diagnosis [3–6], bioinformatics [7–9], and many industrial applications [10–12].

According to the interaction with machine learning classifiers (MLCs), FS methods can be broadly categorized into three groups [13–16]: (1) filter method that selects features regardless of MLCs. It estimates the correlation between quantitative features and target labels, and the features with strong correlations to data labels are further considered. This kind of approach is efficient and robust to overfitting; however, redundant features might be selected. (2) Wrapper method that uses learning algorithms to select one among the generated subsets of features. It allows for possible interactions between features, while it considerably increases computation time, in particular with a large number of features. (3) Embedded method that is similar to the wrapper method, while it performs FS and target classification simultaneously.

Few studies have addressed the efficiency comparison of FS methods. Wang et al. [17] have compared six filter methods, such as *chi*-square [18] and RELIEFF [19], and ranked features were further analyzed by using different MLCs and performance metrics. Experimental results indicated that the selection of performance metrics is crucial for model building. Furthermore, Ma et al. [20] have examined eight FS methods and found that support vector machine- (SVM-) based recursive feature elimination [6] is a suitable approach for feature ranking. In addition, they strongly suggested performing FS before object classification.

Moreover, Cehovin and Bosnic [21] have evaluated five methods and discovered that RELIEFF [19] in combination to random forest (RF) [21] achieves highest accuracy and reduces the number of unnecessary attributes. Vakharia et al. [12] have compared five FS methods for fault diagnosis of ball bearing in rotating machinery, reporting that both the combination of Fisher score and SVM [22] and the combination of RELIEFF and artificial neural network (ANN) [23] have good accuracy. Additionally, Upadhyay et al. [24] have explored three methods to select informative features in wavelet domains. Specifically, they used the least square SVM and discovered that Fisher score has the highest discrimination ability for epilepsy detection.

This study performed an evaluation of FS methods, and a total of 8 filter methods, 1 wrapper method, and 1 embedded method were involved. Specifically, the evaluation was conducted in a proposed unified framework where features were ranked and incrementally added; RF was the classifier, and 4 metrics were used to assess the classification performance. Notably, the digital database for screening mammography (DDSM) [25] was investigated which contains 1104 benign and 980 malignant lesions. In the end, a test-retest study was concerned and the reliability of built models was discussed.

2. Methods

2.1. Data Collection. The DDSM is one of the largest databases for mammographic breast image analysis [25–27], which is available online (http://www.eng.usf.edu/cvprg/Mammography/Database.html). The database includes 12 volumes of normal cases, 16 volumes of benign cases, and 15 volumes of malignant mass lesion cases. Each case is represented by 6 to 10 files, i.e., an "ics" file, an overview 16-bit portable gray map (PGM) file, four image files compressed with lossless joint photographic experts group (LJPEG) encoding, and a zero to four overlay files.

Using the toolbox DDSM Utility (https://github.com/trane293/DDSMUtility) [28], a total of 2084 histologically verified breast lesions (1104 benign and 980 malignant lesions) and 4016 mammographic images were obtained. Full details on how to convert the dataset from an outdated image format (LJPEG) to a usable format (i.e., portable network graphic) and on how to extract these outlined regions of interest are described in the toolbox manual.

2.2. Lesion Representation. Previous studies have suggested computational and informative features for mammographic lesion representation [29, 30]. In this study, 18 features were used to characterize breast mass lesions among which 7 features (mean, median, standard deviation, maximum, minimum, kurtosis, and skewness) represent the statistical analysis of mass intensity, 8 features (area, perimeter, circularity, elongation, form, solidity, extent, and eccentricity) describe the lesion shape, and 3 features (contrast, correlation, and entropy) are derived from the texture analysis using the grey-level cooccurrence matrix (GLCM) [31]. Full information to these quantitative features can be referred to [32].

2.3. Feature Selection Methods. In total, 10 feature selection methods (8 filter methods, 1 wrapper method, and 1 embedded method) were evaluated. Specifically, there were 6 methods based on unsupervised learning and 4 methods based on supervised learning (Table 1).

Brief description of each method is as below

(a) Correlation-based feature selection (CFS) was used to quantify the relationship between feature vectors using Pearson's linear correlation coefficient [33]. It takes the minimal correlation coefficient of one feature vector to the other feature vectors as the score which represents the information redundancy. Finally, features were sorted according to the scores in ascending order

(b) Feature selection via eigenvector centrality (ECFS) [34] recasts the FS problem based on the affinity graph and the nodes in the graph present features. It estimates the importance of nodes through the indicator of eigenvector centrality (EC). And the purpose of EC is to quantify the importance of a feature with regard to the importance of its neighbors and these central nodes are ranked as candidate features

(c) Infinite latent feature selection (ILFS) [35] is a probabilistic latent FS approach that considers all the possible feature subsets. It further models feature "relevancy" through a generative process inspired by the probabilistic latent semantic analysis [36]. The mixing weights are derived to measure a graph of features, and a score of importance is provided by the weighted graph for each feature, which indicates the importance of the feature in relation to its neighboring features

(d) Laplacian score (LAPLACIAN) [37] evaluates the importance of a feature by its power of locality preserving. It constructs a nearest neighbor graph to model the local geometric structure, and it seeks the features that respect this graph structure

(e) Least absolute shrinkage and selection operator (LASSO) [38] performs feature selection and regularization simultaneously and thus, it can balance prediction accuracy and model interpretability. LASSO is L_1-constrained linear least squares fits, and the importance of each feature is weighted

(f) Feature selection using local learning-based clustering (LLCFS) [39] estimates the feature importance during the process of local learning-based clustering (LLC) [40] in an iterative manner. It associates a weight to each feature, while the weight is incorporated into the regularization of the LLC method by considering the relevance of each feature for the clustering

(g) RELIEFF [19] estimates the weight of each feature according to how well its value can differentiate between itself and its neighboring features [41]. Thus, if the difference in feature values is observed

Table 1: Feature selection methods.

ID	Acronym	Class	Learning strategy
A	CFS	Filter	Unsupervised
B	ECFS	Filter	Supervised
C	ILFS	Filter	Supervised
D	LAPLACIAN	Filter	Unsupervised
E	LASSO	Embedded	Supervised
F	LLCFS	Filter	Unsupervised
G	RELIEFF	Filter	Supervised
H	ROC	Filter	Unsupervised
I	UFSOL	Wrapper	Unsupervised
J	WILCOXON	Filter	Unsupervised

in a neighboring instance pair with the same class, its weight decreases; while if there are different classes, its weight increases

(h) ROC is an independent evaluation criterion [42] which is used to assess the significance of every feature in the separation of two labeled groups. It stands for the area between the empirical receiver operating characteristic (ROC) curve and the random classifier slope. Higher area value indicates better separation capacity

(i) Unsupervised feature selection with ordinal locality (UFSOL) [43] is a clustering-based approach. It proposes a triplet-induced loss function that captures the underlying ordinal locality of data instances. UFSOL can preserve the relative neighborhood proximities and contribute to the distance-based clustering

(j) Wilcoxon rank-sum test (WILCOXON) or Mann-Whitney *U* test is a nonparametric test [44]. It requires no assumption of normal distribution of feature values. The test provides the most accurate significance estimates, especially with small sample sizes and/or when the data do not approximate a normal distribution

Among these methods, 4 methods consider statistical analysis on differentiating each other features or on label classification (CFS, RELIEFF, ROC, and WILCOXON); 3 methods build a graph to map the relationship between features, and weights of features are quantified by the specific measure spaces (ECFS, ILFS, and LAPLACIAN); 2 methods concern data clustering (LLCFS and UFSOL) for feature weighting; and 1 method merges feature selection into a regularization problem to balance prediction accuracy and model interpretability (LASSO). During the procedure, FS methods put a weight to each feature and thus, these features can be ranked according to their weights from the most to the least important.

2.4. Performance Metrics. In this study, four metrics, the area under the curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE), were used to quantify the classification performance [45]. In particular, AUC presents the overall capacity of a model in lesion classification and it refers to the area under the ROC curve.

Based on histological verification, true positive (TP) is the number of positive cases that were correctly predicted as "positive," false negative (FN) represents the positive cases that were misclassified as "negative," true negative (TN) represents the true negative cases that were predicted correctly, while false positive (FP) is true negative cases that were predicted as "positive." ACC, SEN, and SPE can be formulated using the formula (1), (2), and (3), respectively.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}, \tag{1}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{3}$$

2.5. Experiment Design. Given 2084 lesion cases (1104 benign and 980 malignant lesions) of 4016 mammographic images, we took one image per lesion in the test study (a total of 1104 benign images and 980 malignant images) and the remaining images (1017 benign lesion images and 915 malignant lesion images) were used to retest the trained diagnostic models in the test study. Specifically, in the test study, 400 benign lesion images and 400 malignant lesion images were randomly picked for training and the other images were used for testing. The experiment was carried out 100 times, and performance metrics were reported on average.

RF is used as the classifier in this study. It is an ensemble learning method that has been widely applied for prediction, classification, and regression [20, 21, 46], and Strobl et al. utilized it to measure the variable importance [47]. The most important parameter in RF algorithm is the number of trees, and Oshio et al. stated that increasing the number of trees does not always mean the performance improvement [48]. Therefore, the number of trees is set as 10 and fewer trees indicates more generalizable of a trained model with regard to thousands of lesion cases in the DDSM database.

The unified framework is shown in Figure 1. It consists of feature ranking, incremental feature selection, RF optimization, and performance evaluation. Furthermore, feature ranking is based on the whole images in the study. In addition, after the RF-based model was built and evaluated on the testing samples, the model was further used to predict the malignance of the lesion images in the retest study. It is worth of note that parameters of FS methods are set as default.

2.6. Software Platform. Involved feature selection methods were implemented with MATLAB (MathWorks, Natick, MA, USA) where seven methods were from the Feature Selection Library [49], two methods (ROC and WILCOXON) were from the function *rankfeatures*, and one method (RELIEFF) was from the function *relieff*. Furthermore, the classifier RF was based on the function *randomForest* [50] in R (https://www.r-project.org/). The experiments were run on a personal laptop, and the laptop

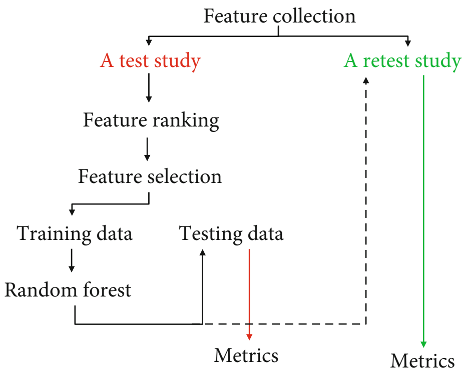


Figure 1: The proposed unified framework. It includes feature ranking, incremental feature selection, RF-based lesion classification, and performance evaluation, where features were precollected.

was equipped with dual Intel (R) Cores (TM) of 2.50 GHz and 8 GB DDR RAM. The implementation did not rely on any optimization or strategies for algorithm acceleration.

2.7. Statistical Analysis. Quantitative metrics were summarized as the mean ± standard deviation (SD) (MATLAB, MathWorks, Natick, MA, USA). Comparison between performance metrics is made with Wilcoxon rank-sum test or two sample t-tests when appropriate. All statistical tests are two sided, and p values less than 0.05 are defined as significant difference.

3. Results

3.1. Perceived Increase of AUC Values. Figure 2 shows that the AUC values increased when features were added for mass lesion representation (red lines). When using top 2 features, both ECFS and CFS achieved AUC values that were averagely larger than 0.70 and AUC values from other FS methods that were larger than 0.60. Yet, the AUC values from UFSOL and LLCFS were <0.60, and the values did not show any obvious improvement until top 6 and 5 features were integrated in breast lesion classification, respectively. Compared to the baseline of AUC equal to 0.85 (green lines), both ILFS and RELIEFF obtained higher values when at least 6 features were used, followed by CFS (7 features) and WILCOXON (8 features), and other FS methods that required 9 to 10 features. In addition, for each diagnostic model, the error-bar plot of AUC in the retest study overlapped quite well with the plot in the test study.

3.2. Result Summary. Table 2 summarizes the number of features and corresponding performance metrics when a model achieves its AUC surpassing the baseline with the least feature number. It was observed that half of the methods required 10 or more features. In particular, when the first-time model exceeded the baseline, its SEN was higher than 0.85, while its ACC and SPE were relatively lower, indicating the potential false positive.

Table 3 summarizes the metric values when top two features are used for lesion representation. It was found that ECFS and CFS achieve AUC larger than 0.70, while three out of other eight methods reach AUC less than 0.60. We also found that ECFS, CFS, and ILFS reach SPE values larger than 0.50, while other methods tend to misclassify benign lesions into malignant ones.

The feature selection results are shown in Table 4 where the top-most important features of each model are highlighted in red. Frequency analysis of these features indicates that the 8th feature and the 16th feature are selected eight times, followed by the 4th feature 7 times, while other features are equally used or less than 6 times.

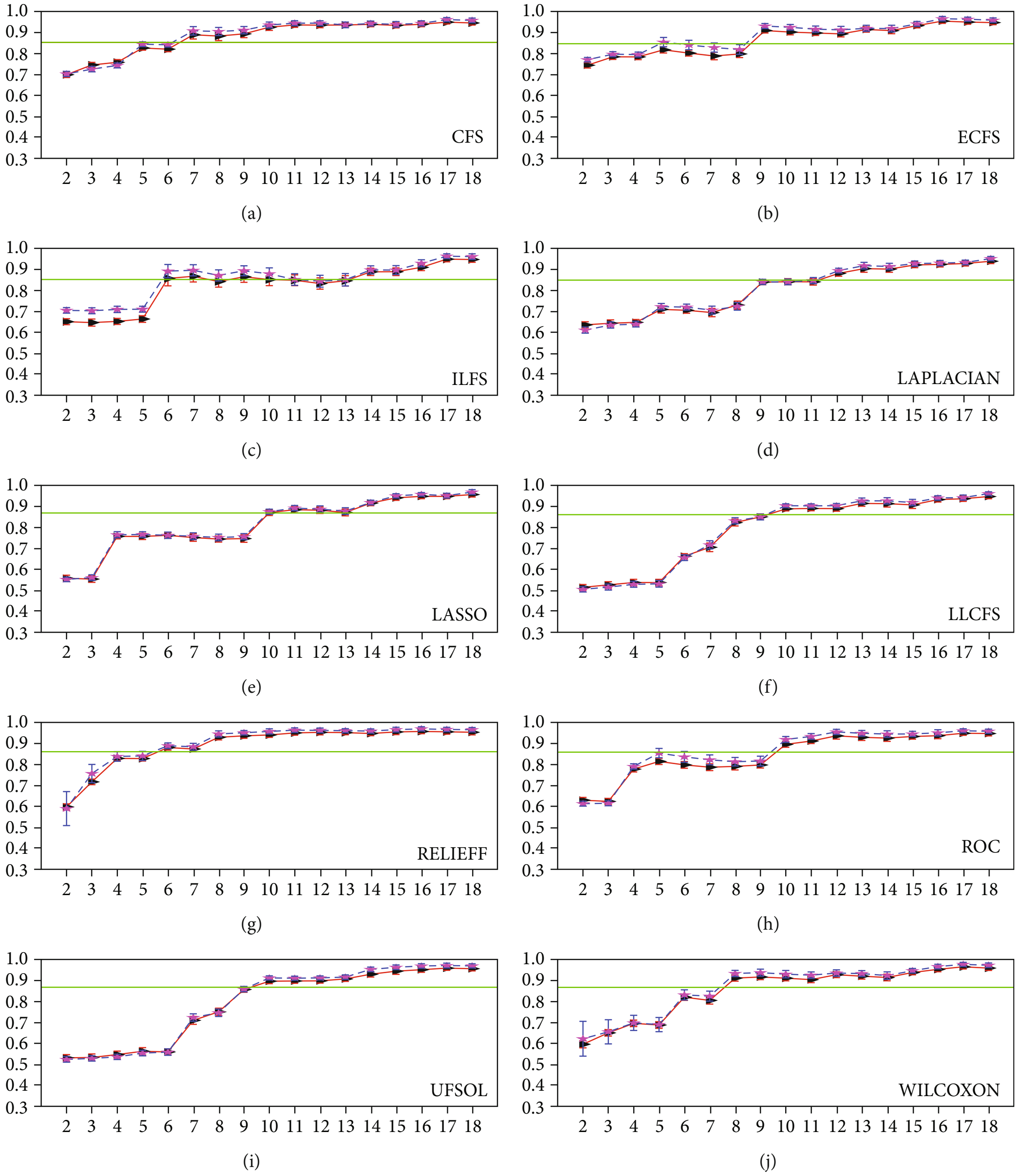
4. Discussion

This study evaluated 10 FS methods in a unified framework for mammographic breast cancer diagnosis where RF is used as the classifier. Besides, the reliability of each diagnosis model was verified. Experimental results suggested that CFS has the ability to retrieve generally discriminative features. Based on the features ranked by CFS, the classification performance keeps improving. In addition, the CFS-based model achieved the 2nd best performance when using top 2 features and it surpassed the baseline (AUC = 0.85) by using the top 7 features.

Some methods lead to unchanged or decreased performance at certain points when the number of features increases (Figure 2), which might be the selected features are redundant. These methods are ECFS, ILFS, LASSO, LLCFS, and ROC. In feature ranking, some methods omit the relationship between features. For instance, features *i_mean* and *i_median* (Appendix A) correlated well (Pearson's correlation coefficient, $p = 0.99$) and the two features are near each other in 8 out of 10 ranked feature lists (Table 4). Thus, it is helpful to remove the redundant features and continue to update diagnosis models in order to reach the optimal solution.

The use of a reasonable number of features is desirable in intelligent diagnosis since it implies a model lightweight computing; it is easy to interpret and can be generalized to other related applications. Investigation of top-ranked two features revealed that 7 out of 10 methods failed in distinguishing benign lesions from malignant ones (SPE < 0.5, Table 3). ECFS and CFS can achieve relatively good performance (AUC > 0.71, ACC > 0.63, SEN > 0.71, and SPE > 0.57). When the number of features increases, ILFS, RELIEFF, and CFS begin to exceed the baseline (Figure 2). On the other hand, except for AUC and SEN, other metrics have important roles since they allow for model evaluation from another perspectives. By comparing AUC, ACC, SEN, and SPE metrics, we found that most ACC and SPE values were lower than 0.80 when both AUC and SEN were larger than 0.85, which indicated that considerably benign lesions were misclassified and thereby, these patients would be exposed to unnecessary biopsies and would suffer from psychological anxiety.

Over recent years, FS has gained increasing attention. Notably, a series of models have been developed in radiomics [51–53]. Radiomics explores to represent one target from various perspectives where tens of thousands features



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j)

Figure 2: AUC. A baseline (green) of AUC equal to 0.85 is added to the plots. In each plot, the red solid line indicates the test result, while the blue dashed line shows the retest result. Besides, error bars are added. Please note that the figure can be enlarged to perceive details.

can be crafted. Consequently, the selection of these discriminative features is a crucial, indispensable, but challenging step. On the other hand, the efficiency of feature subsets is hard to compare due to number of reasons such as FS being data dependent, which means that different data splitting may lead to change in the feature weights. Moreover, different FS methods might lead to distinct results because of theoretical frameworks, and this study obtained ten different selection results (Table 4).

This study has several limitations. First, few features were considered. It is known that massive features can be handcrafted based on mass intensity, shape, and texture in various transformed domains [30, 51–53], while it might make FS become challenging if hundreds of thousands features are involved, in particular for high dimension but small sample data analysis [54]. Second, this study evaluated a total of 10 FS methods among which 8 methods belong to the filter method group. Since filter methods are independent of classifiers, it avoids classifier selection and thus, computes efficiently. On the other hand, if more wrapper and embedded methods are compared, the conclusion that CFS having better performance would be more strongly supported. However, it is worth noting that this imbalance of FS methods does not affect the use of the proposed framework.

Table 2: Performance comparison. The metric values in bold come from the test study, while the values in the line below are from the retest study with corresponding features and model.

	No.	AUC	ACC	SEN	SPE
CFS	7	**0.867 ± 0.023**	**0.733 ± 0.035**	**0.883 ± 0.018**	**0.793 ± 0.023**
		0.896 ± 0.020	0.724 ± 0.035	0.900 ± 0.018	0.806 ± 0.022
ECFS	9	**0.887 ± 0.018**	**0.739 ± 0.028**	**0.894 ± 0.011**	**0.806 ± 0.014**
		0.926 ± 0.013	0.717 ± 0.034	0.915 ± 0.012	0.816 ± 0.017
ILFS	6	**0.866 ± 0.028**	**0.678 ± 0.044**	**0.854 ± 0.030**	**0.763 ± 0.031**
		0.907 ± 0.025	0.665 ± 0.043	0.884 ± 0.027	0.779 ± 0.029
LAPLACIAN	12	**0.863 ± 0.018**	**0.730 ± 0.030**	**0.880 ± 0.013**	**0.790 ± 0.016**
		0.891 ± 0.013	0.716 ± 0.028	0.893 ± 0.011	0.799 ± 0.014
LASSO	10	**0.858 ± 0.020**	**0.685 ± 0.030**	**0.851 ± 0.013**	**0.763 ± 0.016**
		0.862 ± 0.019	0.692 ± 0.025	0.856 ± 0.011	0.772 ± 0.013
LLCFS	10	**0.855 ± 0.020**	**0.735 ± 0.027**	**0.876 ± 0.009**	**0.789 ± 0.013**
		0.887 ± 0.014	0.714 ± 0.025	0.891 ± 0.009	0.796 ± 0.012
RELIEFF	6	**0.855 ± 0.020**	**0.718 ± 0.026**	**0.868 ± 0.011**	**0.780 ± 0.013**
		0.880 ± 0.015	0.695 ± 0.037	0.876 ± 0.012	0.782 ± 0.019
ROC	10	**0.878 ± 0.019**	**0.728 ± 0.029**	**0.885 ± 0.013**	**0.796 ± 0.016**
		0.919 ± 0.012	0.706 ± 0.035	0.908 ± 0.013	0.807 ± 0.018
UFSOL	10	**0.858 ± 0.020**	**0.731 ± 0.028**	**0.877 ± 0.011**	**0.788 ± 0.013**
		0.889 ± 0.016	0.709 ± 0.029	0.892 ± 0.009	0.794 ± 0.014
WILCOXON	8	**0.887 ± 0.019**	**0.726 ± 0.027**	**0.890 ± 0.013**	**0.799 ± 0.015**
		0.925 ± 0.013	0.707 ± 0.036	0.910 ± 0.013	0.810 ± 0.019

Table 3: Performance comparison when using top two features for lesion representation.

	No.	AUC	ACC	SEN	SPE
CFS	2	**0.711** ± 0.012	**0.636** ± 0.013	**0.714** ± 0.027	**0.572** ± 0.030
		0.715 ± 0.011	0.642 ± 0.012	0.718 ± 0.019	0.573 ± 0.026
ECFS	2	**0.734** ± 0.013	**0.660** ± 0.012	**0.755** ± 0.026	**0.581** ± 0.024
		0.759 ± 0.010	0.677 ± 0.011	0.785 ± 0.018	0.579 ± 0.021
ILFS	2	**0.678** ± 0.012	**0.606** ± 0.012	**0.698** ± 0.023	**0.530** ± 0.026
		0.724 ± 0.011	0.635 ± 0.011	0.752 ± 0.016	0.529 ± 0.025
LAPLACIAN	2	**0.649** ± 0.014	**0.603** ± 0.012	**0.738** ± 0.025	**0.492** ± 0.024
		0.626 ± 0.014	0.590 ± 0.011	0.737 ± 0.023	0.458 ± 0.020
LASSO	2	**0.557** ± 0.014	**0.526** ± 0.013	**0.651** ± 0.025	**0.422** ± 0.028
		0.552 ± 0.010	0.525 ± 0.010	0.653 ± 0.023	0.410 ± 0.023
LLCFS	2	**0.517** ± 0.013	**0.499** ± 0.013	**0.645** ± 0.028	**0.379** ± 0.024
		0.507 ± 0.012	0.498 ± 0.011	0.648 ± 0.025	0.363 ± 0.025
RELIEFF	2	**0.611** ± 0.013	**0.568** ± 0.014	**0.689** ± 0.022	**0.486** ± 0.028
		0.604 ± 0.073	0.574 ± 0.066	0.668 ± 0.021	0.490 ± 0.129
ROC	2	**0.632** ± 0.013	**0.582** ± 0.013	**0.694** ± 0.025	**0.491** ± 0.027
		0.616 ± 0.011	0.571 ± 0.011	0.716 ± 0.021	0.440 ± 0.034
UFSOL	2	**0.543** ± 0.015	**0.514** ± 0.012	**0.654** ± 0.027	**0.399** ± 0.021
		0.527 ± 0.013	0.513 ± 0.011	0.652 ± 0.024	0.388 ± 0.023
WILCOXON	**2**	**0.605** ± 0.015	**0.563** ± 0.015	**0.686** ± 0.024	**0.461** ± 0.028
		0.629 ± 0.075	0.587 ± 0.069	0.679 ± 0.020	0.505 ± 0.133

Table 4: Feature selection results. The top-most important features that achieve AUC larger than 0.85 are in bold to each FS method.

	The most to the least important features																	
CFS	**16**	**7**	**14**	**3**	**11**	**5**	**15**	6	2	8	13	17	10	9	1	4	12	18
ECFS	**8**	**9**	**17**	**4**	**10**	**2**	**1**	**16**	**12**	3	14	6	13	15	7	11	5	18
ILFS	**11**	**14**	**18**	**5**	**3**	**15**	13	1	4	2	10	6	9	7	16	12	8	17
LAPLACIAN	**8**	**5**	**4**	**3**	**9**	**2**	**1**	**16**	**7**	**18**	**6**	**11**	15	10	13	14	17	12
LASSO	**17**	**18**	**15**	**13**	**6**	**16**	**4**	**1**	**2**	**8**	9	5	3	7	11	14	10	12
LLCFS	**3**	**5**	**4**	**2**	**1**	**8**	**9**	**7**	**16**	**11**	18	6	15	10	14	13	17	12
RELIEFF	**10**	**14**	**11**	**7**	**18**	**8**	4	12	3	9	13	17	16	6	15	5	1	2
ROC	**9**	**17**	**4**	**8**	**10**	**2**	**1**	**16**	**3**	**12**	11	15	6	14	13	18	7	5
UFSOL	**9**	**1**	**2**	**3**	**5**	**4**	**8**	**16**	**7**	**11**	18	6	17	12	15	10	13	14
WILCOXON	**10**	**16**	**9**	**17**	**4**	**12**	**6**	**8**	14	2	1	13	3	18	7	11	15	5

Third, RF performs as the classifier, since it is important in classification tasks due to its interpretability [21]. From the technical perspective, other MLCs, such as ANN and SVM, are also feasible [12, 17, 20, 21, 24, 30]. It is also desirable to investigate the effects of RF parameters on the lesion diagnosis. However, it might lead to massive result reports and thus, only the number of trees is empirically determined and other parameters are set as default. Last but not the least, how to choose a proper FS method is a long-term problem in the field of computer-aided diagnosis. It should be admitted that feature extraction, FS methods, and MLCs are closely related to the ultimate goal of breast cancer diagnosis. Depending on specific purposes, such as diagnosis accuracy, model simplicity, interpretability, and generalization capacity, the selection of features, FS methods, and MLCs is different. Fortunately, the proposed framework can be expanded to incorporate more features as radiomics, more FS methods, and MLCs for classification or diagnosis tasks. Therefore, it is promising that systematic and comprehensive analysis on additional mammographic databases could deepen our understanding of breast cancer diagnosis from mammographic images.

5. Conclusions

This study evaluated ten feature selection methods for breast cancer diagnosis based on the digital database for screening mammography, where the random forest served as the machine learning classifier. Different methods led to distinct feature ranking results, and the correlation-based feature selection method was found to have superior performance in general. The way to find discriminative features out of thousands of features is challenging but indispensable for intelligent diagnosis and thus, extra efforts should be made towards accurate and efficient feature selection.

Abbreviations

FS:	Feature selection
AUC:	The area under the curve
ACC:	Accuracy
SEN:	Sensitivity
SPE:	Specificity
MLC:	Machine learning classifier
SVM:	Support vector machine
RF:	Random forest
ANN:	Artificial neural network
DDSM:	Digital database for screening mammography
PGM:	Portable gray map
LJPEG:	Lossless joint photographic experts group
GLCM:	Grey-level cooccurrence matrix
CFS:	Correlation-based feature selection
ECFS:	Feature selection via eigenvector centrality
EC:	Eigenvector centrality
ILFS:	Infinite latent feature selection
LAPLACIAN:	Laplacian score
LASSO:	Least absolute shrinkage and selection operator
LLCFS:	Feature selection using local learning-based clustering
LLC:	Local learning-based clustering
ROC:	Receiver operating characteristic
UFSOL:	Unsupervised feature selection with ordinal locality
WILCOXON:	Wilcoxon rank-sum test
TP:	True positive
FN:	False negative
TN:	True negative
FP:	False positive
SD:	Standard deviation.

Data Availability

The data and toolboxes are available online. The data used to support the findings of this study are from http://www.eng.usf.edu/cvprg/Mammography/Database.html; the Feature Selection Library is https://www.mathworks.com/matlabcentral/fileexchange/56937-feature-selection-library; and the toolbox DDSM Utility from https://github.com/trane293/DDSMUtility is for data format transformation.

Disclosure

The funding source had no role in the design of this study and will not have any role during its execution, analyses, interpretation of the data, or decision to submit results.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

JL conceived the idea and drafted the manuscript; CJT collected the dataset and conducted experiments; XFX focused on data analysis and helped code implementation. All authors discussed the experimental results and proofread the final manuscript.

Acknowledgments

The authors would like to thank Dr. G. Roffo for the toolbox FSLib2018, Mr. A. Sharma for the toolbox DDSM Utility, and University of South Florida for the DDSM. Sincere thanks are also given to the editor and reviewers for their valuable advice that have helped to improve the paper quality and these students in the department who helped the data collection. This project is supported by the Tianjin Hospital of ITCWM Nankai Hospital and Tianjin Central Hospital of Obstetrics and Gynecology.

References

[1] S. Zhao, Y. Zhang, H. Xu, and T. Han, "Ensemble classification based on feature selection for environmental sound recognition," *Mathematical Problems in Engineering*, vol. 2019, Article ID 4318463, 7 pages, 2019.

[2] E. Adeli, X. Li, D. Kwon, Y. Zhang, and K. M. Pohl, "Logistic regression confined by cardinality-constrained sample and feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1713–1728, 2020.

[3] S. A. Mostafa, A. Mustapha, M. A. Mohammed et al., "Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease," *Cognitive Systems Research*, vol. 54, pp. 90–99, 2019.

[4] M. A. Khan, T. Akram, M. Sharif et al., "An implementation of normal distribution based segmentation and entropy controlled features selection for skin lesion detection and classification," *BMC Cancer*, vol. 18, no. 1, 2018.

[5] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Applied Soft Computing*, vol. 62, pp. 203–215, 2018.

[6] S. Tian, C. Wang, and H. Chang, "A longitudinal feature selection method identifies relevant genes to distinguish complicated injury and uncomplicated injury over time," *BMC Medical Informatics and Decision Making*, vol. 18, no. S5, 2018.

[7] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: a survey from the search perspective," *Methods*, vol. 111, pp. 21–31, 2016.

[8] B. H. Zheng, L. Z. Liu, Z. Z. Zhang et al., "Radiomics score: a potential prognostic imaging feature for postoperative survival of solitary HCC patients," *BMC Cancer*, vol. 18, no. 1, 2018.

[9] S. Tian, C. Wang, and B. Wang, "Incorporating pathway information into feature selection towards better performed gene signatures," *BioMed Research International*, vol. 2019, Article ID 2497509, 12 pages, 2019.

[10] X. Lun, M. Wang, Z. Yu, and Y. Hou, "Commercial video evaluation via low-level feature extraction and selection," *Advances in Multimedia*, vol. 2018, Article ID 2056381, 20 pages, 2018.

[11] P. Y. Lee, W. P. Loh, and J. F. Chin, "Feature selection in multimedia: the state-of-the-art review," *Image and Vision Computing*, vol. 67, pp. 29–42, 2017.

[12] V. Vakharia, V. K. Gupta, and P. K. Kankar, "A comparison of feature ranking techniques for fault diagnosis of ball bearing," *Soft Computing*, vol. 20, no. 4, pp. 1601–1619, 2016.

[13] J. Li, K. Cheng, S. Wang et al., "Feature Selection," *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, 2018.

[14] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 971–989, 2016.

[15] S. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 157–176, 2011.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[17] H. Wang, T. M. Khoshgoftaar, and K. Gao, "A comparative study of filter-based feature ranking techniques," in *2010 IEEE International Conference on Information Reuse & Integration*, pp. 43–48, Las Vegas, NV, USA, 2010.

[18] R. L. Plackett, "Karl Pearson and the chi-squared test," *International Statistical Review*, vol. 51, no. 1, pp. 59–72, 1983.

[19] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018.

[20] L. Ma, T. Fu, T. Blaschke et al., "Evaluation of feature selection methods for object-based land cover mapping of unmanned aerial vehicle imagery using random forest and support vector machine classifiers," *ISPRS International Journal of Geo-Information*, vol. 6, no. 2, 2017.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[23] G. G. Towell and J. W. Shavlik, "Knowledge-based artificial neural networks," *Artificial Intelligence*, vol. 70, no. 1-2, pp. 119–165, 1994.

[24] R. Upadhyay, P. K. Padhy, and P. K. Kankar, "A comparative study of feature ranking techniques for epileptic seizure detection using wavelet transform," *Computers and Electrical Engineering*, vol. 53, pp. 163–176, 2016.

[25] K. Bowyer, D. Kopans, W. P. Kegelmeyer et al., "The digital database for screening mammography," in *Third International Workshop on Digital Mammography*, vol. 58, Springer, Berlin, Heidelberg, 1996.

[26] C. Rose, D. Turi, A. Williams, K. Wolstencroft, and C. Taylor, *Web Services for the DDSM and Digital Mammography Research*, International Workshop on Digital Mammography. Springer, Berlin, Heidelberg, 2006.

[27] M. Benndorf, C. Herda, M. Langer, and E. Kotter, "Provision of the DDSM mammography metadata in an accessible format," *Medical Physics*, vol. 41, no. 5, article 051902, 2014.

[28] A. Sharma, “DDSM Utility,” 2015, https://github.com/trane293/DDSMUtility.

[29] N. P. Pérez, M. A. Guevara López, A. Silva, and I. Ramos, “Improving the Mann-Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography,” *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 19–31, 2015.

[30] D. C. Moura and M. A. Guevara López, “An evaluation of image descriptors combined with clinical data for breast cancer diagnosis,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 561–574, 2013.

[31] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[32] J. J. M. van Griethuysen, A. Fedorov, C. Parmar et al., “Computational radiomics system to decode the radiographic phenotype,” *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[33] D. J. Best and D. E. Roberts, “Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 3, pp. 377–379, 1975.

[34] G. Roffo and S. Melzi, “Feature selection via eigenvector centrality,” in *Proceedings of New Frontiers in Mining Complex Patterns*, pp. 1–12, Riva del Garda, Italy, 2016.

[35] G. Roffo, S. Melzi, U. Castellani, and A. Vinciarelli, “Infinite latent feature selection: a probabilistic latent graph-based ranking approach,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1398–1406, Venice, Italy, 2017.

[36] T. Hofmann, “Probabilistic latent semantic analysis,” in *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*, pp. 289–296, San Francisco, CA, United States, 1999.

[37] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” *Advances in Neural Information Processing Systems*, vol. 18, pp. 507–514, 2006.

[38] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[39] Hong Zeng and Yiu-ming Cheung, “Feature selection and kernel learning for local learning-based clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1532–1547, 2011.

[40] M. Wu and B. Schlkopf, “A local learning approach for clustering,” *Advances in neural information processing systems*, vol. 19, pp. 1529–1536, 2007.

[41] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, “Overcoming the myopia of inductive learning algorithms with RELIEFF,” *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[42] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Springer Science and Business Media, 2012.

[43] J. Guo, Y. Quo, X. Kong, and R. He, “Unsupervised feature selection with ordinal locality,” in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1213–1218, Hong Kong, China, 2017.

[44] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[45] Z. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, “A technical review of convolutional neural network-based mammographic breast cancer diagnosis,” *Computational and Mathematical Methods in Medicine*, vol. 2019, Article ID 6509357, 16 pages, 2019.

[46] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019.

[47] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: illustrations, sources and a solution,” *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[48] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How Many Trees in a Random Forest?,” in *In International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 154–168, Springer, Berlin, Heidelberg, 2012.

[49] G. Roffo, “Feature selection library (MATLAB toolbox),” 2016, https://arxiv.org/abs/1607.01327.

[50] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[51] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar et al., “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature Communications*, vol. 5, no. 1, 2014.

[52] V. Kumar, Y. Gu, S. Basu et al., “Radiomics: the process and the challenges,” *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[53] S. S. F. Yip and H. J. W. L. Aerts, “Applications and limitations of radiomics,” *Physics in Medicine and Biology*, vol. 61, no. 13, pp. R150–R166, 2016.

[54] X. Zhang, E. Zhang, and R. Li, “Optimized feature extraction by immune clonal selection algorithm,” in *2012 IEEE Congress on Evolutionary Computation*, pp. 1–6, Brisbane, QLD, Australia, 2012.

[55] L. Cehovin and Z. Bosnic, “Empirical evaluation of feature selection methods in classification,” *Intelligent data analysis*, vol. 14, no. 3, pp. 265–281, 2010.

[56] D. Dernoncourt, B. Hanczar, and J. D. Zucker, “Analysis of feature selection stability on high dimension and small sample data,” *Computational Statistics and Data Analysis*, vol. 71, no. 71, pp. 681–693, 2014.

Hindawi
BioMed Research International
Volume 2021, Article ID 8811056, 10 pages
https://doi.org/10.1155/2021/8811056

Research Article

Using Machine Learning to Unravel the Value of Radiographic Features for the Classification of Bone Tumors

Derun Pan,[1] Renyi Liu,[1] Bowen Zheng,[1] Jianxiang Yuan,[2] Hui Zeng,[1] Zilong He,[1] Zhendong Luo,[3] Genggeng Qin,[1] and Weiguo Chen[1]

[1] *Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong Province, China*
[2] *Department of Radiology, Foshan Hospital of TCM, Foshan, Guangdong Province, China*
[3] *Department of Radiology, University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong Province, China*

Correspondence should be addressed to Genggeng Qin; zealotq@smu.edu.cn and Weiguo Chen; chen1999@smu.edu.cn

Derun Pan and Renyi Liu contributed equally to this work.

Received 27 September 2020; Revised 5 November 2020; Accepted 3 March 2021; Published 15 March 2021

Academic Editor: Shaode Yu

Objectives. To build and validate random forest (RF) models for the classification of bone tumors based on the conventional radiographic features of the lesion and patients' clinical characteristics, and identify the most essential features for the classification of bone tumors. *Materials and Methods*. In this retrospective study, 796 patients (benign bone tumors: 412 cases, malignant bone tumors: 215 cases, intermediate bone tumors: 169 cases) with pathologically confirmed bone tumors from Nanfang Hospital of Southern Medical University, Foshan Hospital of TCM, and University of Hong Kong-Shenzhen Hospital were enrolled. RF models were built to classify tumors as benign, malignant, or intermediate based on conventional radiographic features and potentially relevant clinical characteristics extracted by three musculoskeletal radiologists with ten years of experience. SHapley Additive exPlanations (SHAP) was used to identify the most essential features for the classification of bone tumors. The diagnostic performance of the RF models was quantified using receiver operating characteristic (ROC) curves. *Results*. The features extracted by the three radiologists had a satisfactory agreement and the minimum intraclass correlation coefficient (ICC) was 0.761 (CI: 0.686-0.824, $P < .001$). The binary and tertiary models were built to classify tumors as benign, malignant, or intermediate based on the imaging and clinical features from 627 and 796 patients. The AUC of the binary (19 variables) and tertiary (22 variables) models were 0.97 and 0.94, respectively. The accuracy of binary and tertiary models were 94.71% and 82.77%, respectively. In descending order, the most important features influencing classification in the binary model were margin, cortex involvement, and the pattern of bone destruction, and the most important features in the tertiary model were margin, high-density components, and cortex involvement. *Conclusions*. This study developed interpretable models to classify bone tumors with great performance. These should allow radiographers to identify imaging features that are important for the classification of bone tumors in the clinical setting.

1. Introduction

The bone tumor is relatively rare, but the malignant bone tumor is the third leading cause of cancer-related death in individuals before 20 years old. In the United States, in 2020, an estimated 3,600 individuals (2,120 males, 1,480 females) will be diagnosed with primary malignant tumors of the bone and joints, and 1,720 individuals (1000 males, 720 females) will die from the disease [1].

The fourth edition of the World Health Organization (WHO) Classification of Tumours of Soft Tissue and Bone published in 2013 classifies bone tumors as benign, malignant, and intermediate [2]. Compared with the third edition, the most significant change is the addition of intermediate bone tumors. Intermediate bone tumors include the locally aggressive type and occasional metastatic type. Locally aggressive type often has a recurrence after resection, which is typical of osteoblastoma [2, 3]. Occasionally, metastatic type has the ability of distant metastasis, which is typically represented by giant cell tumors of bone [4]. However, the aggression and metastasis degree is lower than that of malignant bone tumors. Therefore, this classification method can better guide the formulation of clinical treatment plans. In clinical practice, bone tumor classification involves a comprehensive evaluation of a patient's demographics, medical history, and the lesion's imaging features [5]. There are significant differences in the treatment of different bone tumors; hence, the early classification of bone tumors helps guide therapy and improve patient management [6–9].

Conventional radiography is the preferred imaging modality for evaluating primary bone tumors [10]. Although the benefits of early classification of bone tumors are widely acknowledged, differentiating between bone tumor types can be difficult. Challenges include the variation in the imaging manifestation and their rarity, making it difficult for radiologists to make an accurate diagnosis [2]. Several studies have classified benign and malignant bone tumors based on patient characteristics such as age, gender, and imaging features such as tumor location, margins, periosteal reaction, and mineralization [11, 12]. Despite these efforts, no single radiographic criteria for bone tumor classification have been identified, increasing the risk for diagnostic error.

Machine learning refers to models designed to evaluate and make predictions about relationships between data [13, 14]. Classifying bone tumors using machine learning models based on predefined radiographic or clinical features may help radiologists differentiate between various bone tumors.

A random forest model is an ensemble classifier that consists of many decision trees [15]. The random forest model outputs the class voted by a majority of the individual trees or the mean individual tree prediction [16]. It generates an internal unbiased estimate of the generalization error in the forest building processes and uses a nodes' splitting process to estimate the essential variables [17]. Random forest models are highly predictive as classifiers when analyzing medical imaging data [18, 19].

We hypothesize that a random forest model with high predictive accuracy for bone tumor classification may benefit the clinical setting. This study's objectives were to (1) build and validate a random forest model to classify bone tumors based on the conventional radiographic features of the lesion and patients' clinical characteristics and (2) identify the most important conventional radiographic features for the bone tumor classification.

2. Materials and Method

This retrospective study was approved by the research ethics review board of Nanfang Hospital of Southern Medical University. The necessity to obtain written informed consent from included patients was waived. Data was collected by Nanfang Hospital of Southern Medical, Foshan Hospital of TCM, and University of Hong Kong-Shenzhen Hospital.

2.1. Study Population. The study collected 796 patients (26 ± 18 years) with pathologically confirmed bone tumors from Nanfang Hospital of Southern Medical University between 2014 and 2019, Foshan Hospital of TCM, and University of Hong Kong-Shenzhen Hospital between 2018 and 2019 as a data set. The inclusion criteria were as follows: (1) patients who underwent at least one preoperative conventional radiographic examination in one of the three academic medical centers between 2014 and 2019 and (2) patients who had a pathological diagnosis via biopsy. The exclusion criteria were as follows: (1) patients who relapse after surgery, (2) patients with poor quality preoperative conventional radiographic images, and (3) there is a foreign body in the conventional radiographic images.

For each included patient, the first preoperative conventional radiographic examination was defined as the index examination.

2.2. Conventional Radiography. All conventional radiographic images were collected from the picture archiving and communication system (PACS) of three hospitals. Anteroposterior and lateral views showing the bone tumor were obtained from each included patient.

2.3. Feature Analysis. Preoperative conventional radiographic features and potentially relevant clinical characteristics were extracted and compiled in a structured database by three musculoskeletal radiologists (with ten years of experience) without knowledge of pathological diagnoses. PACS was used to capture conventional radiographic features of each bone tumor, including location, margin, eccentric growth, expansive growth, sclerotic border, periosteal reaction, radiographic density, high-density components, the pattern of bone destruction, source, pathological fracture, and cortex involvement. The radiologists independently extracted features from the conventional radiographic images in DICOM format. Medical records were reviewed for patients' clinical characteristics, including erythrocyte sedimentation rate (ESR), age, gender, redness and hyperemia, swelling, warmth, pain, palpable mass, and dyskinesia (Table 1).

The radiologists independently scored each conventional radiographic feature, and scores were averaged across radiologists. The presence/absence of nominal features was scored on a scale from 0 to 1, where 0 indicated none of the radiologists had a positive opinion and 1 indicated all three radiologists had a positive opinion. For example, if 2 of 3 radiologists consider the margin of the bone tumor to be "sharp," whereas the remaining 1 of 3 radiologists considered it to be "ill-defined," the score was sharp = 0.67 (2/3) and ill-defined = 0.33 (1/3). Age and ESR were assigned numerical values.

2.4. Random Forest Classifier. Patients were randomly divided into a 70% training and validation data set and a 30% testing data set. A 6-fold cross-validation method was used to establish random forest models and verify the classification accuracy. The study used recursive feature elimination (RFE) to select features related to the classification during training, which enables feature interaction. RFE returns a ranking of all features by recursively training random forest models and removing the feature with the smallest ranking score. At each iteration, the feature's removal least affects the objective function. The iterations continued until the best performance of models was reached.

A binary model was built to classify tumors as benign or malignant based on the imaging and clinical data from 627

Table 1: Preoperative radiographic features and clinical characteristics with potential clinical importance for diagnosis.

Features	Feature class	Permissible value	ICC
Location*	Categorical	Upper tibia (a)/inferior femur (b)/upper humerus (c)/middle humerus (d)	0.954
Location	Categorical	Epiphysis (a)/metaphysis (b)/diaphysis (c)/not applicable (d)	0.854
Eccentric growth	Binary	Without (0)/with (1)	0.921
Expansive growth	Binary	Without (0)/with (1)	0.888
Margin	Binary	Sharp(0)/ill-defined (1)	0.832
Sclerotic border	Binary	Without (0)/with (1)	0.796
Periosteal reaction	Categorical	Without (a)/continuous (b)/interrupted (c)	0.899
Radiographic density	Categorical	Mixed (a)/low (b)/high (c)	0.863
High-density components	Categorical	Without (a)/calcification or ossification (b)/tumor bone (c)/unrecognizable (d)	0.761
Pattern of bone destruction	Categorical	Geographic (a)/moth-eaten (b)/permeated (c)/not applicable (d)	0.812
Source	Binary	Medullary (0)/cortical (1)	0.909
Pathological fracture	Binary	Without (0)/with (1)	0.888
Cortex involvement	Categorical	Complete cortex (a)/cortical expansion and thinning (b)/interrupted cortex (c)	0.870
Clinical data			
ESR	Numerical		—
Age	Numerical		—
Gender	Binary	Male (0)/female (1)	—
Redness and hyperemia	Binary	Without (0)/with (1)	—
Swelling	Binary	Without (0)/with (1)	—
Warmth	Binary	Without (0)/with (1)	—
Pain	Binary	Without (0)/with (1)	—
Palpable mass	Binary	Without (0)/with (1)	—
Dyskinesia	Binary	Without (0)/with (1)	—

Note: *The location details are shown in supplement section.

patients. The training and validation set included data from 438 patients. The test set included data from 189 patients. A tertiary model was built to classify tumors as benign, malignant, or intermediate based on the imaging and clinical data from 796 patients. The training and validation set consisted of data from 557 patients. The test set included data from 239 patients.

SHapley Additive exPlanations (SHAP) was used to describe the most important conventional radiographic features for the classification. The diagnostic performance of the random forest classifiers was evaluated in the test sets using area under curve (AUC), accuracy, sensitivity, and specificity.

2.5. Statistical Analysis. Statistical analysis was conducted using the SPSS version 20.0 software (SPSS, Chicago, Ill). Clinical variables were compared among patients with benign, malignant, and intermediate bone tumors using one-way analysis of variance (ANOVA). The intraclass correlation coefficient (ICC) was used to assess three radiologists' agreement who extracted radiographic features. The weights of all input variables were calculated during training and verification; the higher value of the weight indicates the greater importance. Statistical significance was set at $P < 0.05$.

3. Results

3.1. Study Population. The study enrolled 412 patients with benign bone tumors (23 ± 16 years), 215 patients with malignant bone tumors (33 ± 20 years), and 169 patients with intermediate bone tumors (24 ± 16 years). The most commonly benign, malignant, and intermediate bone tumors were osteochondroma (36.1%), osteosarcoma (45.5%), and giant cell tumor (38.5%), respectively.

For the binary classification model, the training and validation set ($n = 438$; 26 ± 18 years) consisted of 298 patients (68.0%) with a benign bone tumor and 140 (32.0%) patients with a malignant bone tumor. The test set ($n = 189$, mean age, 27 ± 18 years) consisted of 114 patients (60.3%) with a benign bone tumor and 75 (39.7%) patients with a malignant bone tumor. For the tertiary classification model, the training and validation set ($n = 557$; 26 ± 18 years) consisted of 289 (51.9%) patients with a benign bone tumor, 118 (21.2%) patients with an intermediate bone tumor, and 150 (26.9%) patients with a malignant bone tumor. The test set ($n = 239$; mean age, 26 ± 18 years) consisted of 123 (51.5%) patients with a benign bone tumor, 51 (21.3%) patients with an intermediate bone tumor, and 65 (27.2%) patients with a malignant bone tumor (Table 2). The details of the tertiary model's test set were shown in the supplement section (available here).

The clinical characteristics of the included patients stratified by bone tumor type (benign, intermediate, or malignant bone tumor) were summarized in Table 3. Patients with a malignant bone tumor were significantly older than those with a benign bone tumor (33 vs. 23 years old; $P < 0.001$). The pathological type of bone tumor was significantly associated with all clinical parameters examined except gender ($P > 0.05$).

Table 2: Patients characteristics: training, validation and test sets.

Characteristics	Binary model Training and validation set	Test set	Tertiary model Training and validation set	Test set
No. of patients	438	189	557	239
Age (y)*	26 ± 18	27 ± 18	26 ± 18	26 ± 18
ESR*	19.01 ± 22.20	20.47 ± 24.09	19.95 ± 23.16	20.97 ± 24.70
Pathological results				
Biopsy benign for bone tumor	298 (68.0)	114 (60.3)	289 (51.9)	123 (51.5)
Biopsy malignant for bone tumor	140 (32.0)	75 (39.7)	150 (26.9)	65 (27.2)
Biopsy intermediate for bone tumor	0	0	118 (21.2)	51 (21.3)

Note: unless otherwise indicated, data are numbers (%) of patient. *Data are means ± standard deviation.

Table 3: Clinical characteristics of the included patients stratified by benign, intermediate, or malignant bone tumor.

Clinical characteristics	All ($N = 796$)	Benign ($N = 412$)	Malignant ($N = 215$)	Intermediate ($N = 169$)	P value
Age	26 ± 18	23 ± 16	33 ± 20	24 ± 16	<0.001*
ESR	20.26 ± 23.63	12.25 ± 14.96	33.25 ± 28.24	23.25 ± 26.38	<0.001*
Male	496 (62.3)	253 (61.4)	135 (62.8)	108 (63.9)	0.841
Female	300 (37.7)	159 (38.6)	80 (37.2)	61 (36.1)	
Redness and hyperemia	20 (2.5)	5 (1.2)	13 (6.0)	2 (1.2)	0.018*
Without redness and hyperemia	776 (97.5)	407 (98.8)	202 (94.0)	167 (98.8)	
Swelling	211 (26.5)	79 (19.2)	85 (39.5)	47 (27.8)	<0.001*
Without swelling	585 (73.5)	333 (80.8)	130 (60.5)	122 (72.2)	
Warmth	92 (11.6)	12 (2.9)	59 (27.4)	21 (12.4)	<0.001*
Without warmth	704 (88.4)	400 (97.1)	156 (72.6)	148 (87.6)	
Pain	472 (59.3)	173 (42.0)	174 (80.9)	125 (74.0)	<0.001*
Without pain	324 (40.7)	239 (58.0)	41 (19.1)	44 (26.0)	
Palpable mass	275 (34.5)	169 (41.0)	73 (34.0)	33 (19.5)	<0.001*
Without palpable mass	521 (65.5)	243 (59.0)	142 (66.0)	136 (80.5)	
Dyskinesia	171 (21.5)	53 (12.9)	69 (32.1)	49 (29.0)	<0.001*
Without dyskinesia	625 (78.5)	359 (87.1)	146 (67.9)	120 (71.0)	

*$P < 0.05$.

3.2. Radiographic Features. The ICC agreement of three radiologists for feature extraction was high (ICC >0.75), with the lowest 0.761 for components (CI: 0.686-0.824, $P < .001$) and the highest 0.954 for location (CI: 0.936-0.967, $P < .001$), as per Table 1.

Examples of the conventional radiographic features of bone tumors and their scores from 3 patients are shown in Figure 1. Patient A was an 8-year-old female with a benign bone tumor. Patient B was a 34-year-old man with an intermediate bone tumor. Patient C was a 46-year-old man with a malignant bone tumor. Images were scored for the presence or absence of sharp vs. ill-defined margins, geographic vs. moth-eaten vs. permeated pattern of bone destruction, and with vs. without expansive growth.

3.3. Random Forest Models. Two random forest models were used to classify bone tumors based on imaging and clinical data (Figure 2). The binary classification model consisted of 15 random decision trees and the maximum tree depth was 10. The tertiary classification model consisted of 85 random decision trees and the maximum tree depth was 8.

The binary classification model classified bone tumors as benign or malignant. The 19 predictor variables included age, location, ESR, margin, cortex involvement, the pattern of bone destruction, high-density components, radiographic density, source, eccentric growth, gender, swelling, warmth, pain, dyskinesia, sclerotic border, location relationship with epiphysis, periosteal reaction, and pathological fracture.

The tertiary classification model classified bone tumors as benign, malignant, or intermediate. The 22 predictor variables included all the extracted conventional radiographic features and clinical characteristics.

In descending order of importance, the binary model features were as follows: margin, cortex involvement, the pattern of bone destruction, and high-density components. The important features for the tertiary model were as follows:

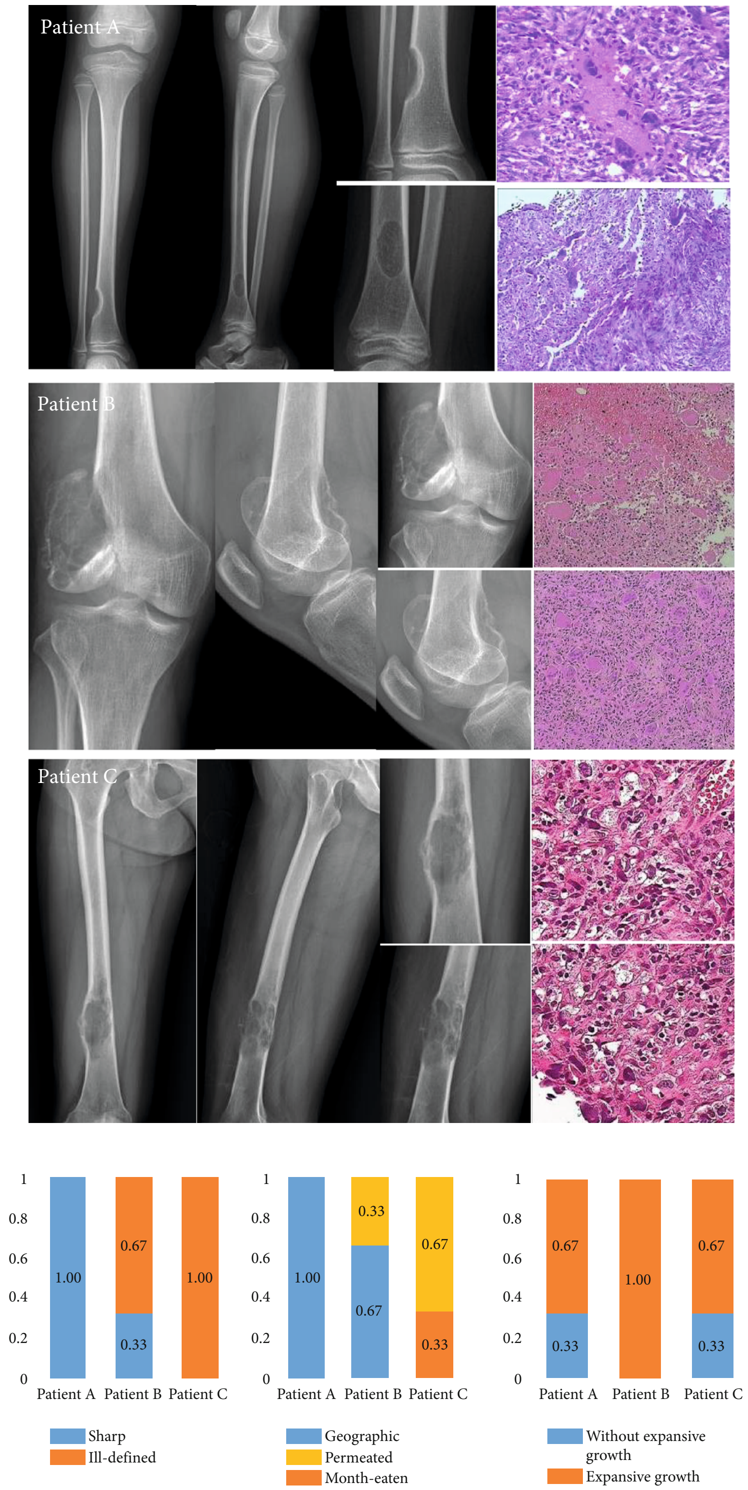


Figure 1: Examples of the features (upper panel) and scores depicting the presence or absence of sharp vs. ill-defined bone margins, geographic vs. moth-eaten vs. permeated pattern of bone destruction, and with vs. without expansive growth (lower panel) as seen on conventional radiographic images obtained from 3 patients. Patient A was an 8-year-old female with nonossifying fibroma. Patient B was a 34-year-old man with a giant cell tumor of bone, and Patient C was a 46-year-old woman with osteosarcoma.

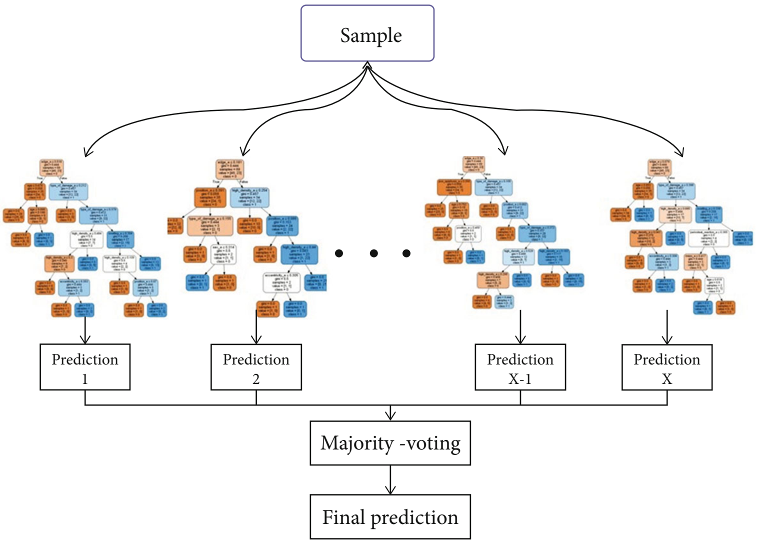


Figure 2: Flow chart of the random forest models. The binary classification model consisted of an ensemble of 15 random decision trees and maximum depth set to 10. The tertiary classification model consisted of an ensemble of 85 random decision trees and maximum depth set to 8. Output from all decision trees determines the final prediction.

margin, high-density components, cortex involvement, and pattern of bone destruction (Figure 3).

3.4. Random Forest Model Performance. The random forest models were tested for their ability to classify bone tumors as benign, malignant, or intermediate (Table 4). Overall, the binary classification model outperformed the tertiary classification model. For the binary classification model, AUC, accuracy, sensitivity, and specificity were 0.97, 94.71%, 93.33%, and 95.61%, respectively. For the tertiary classification model, AUC, accuracy, sensitivity and specificity were 0.95, 84.94%, 86.18%, and 83.62%, respectively, for predicting benign bone tumor; 0.98, 92.05%, 90.77%, and 92.53%, respectively, for predicting malignant bone tumor, and 0.89, 86.19%, 58.82%, and 93.62%, respectively, for predicting intermediate bone tumor. Figure 4 shows the receiver operating characteristic curves for the random forest models.

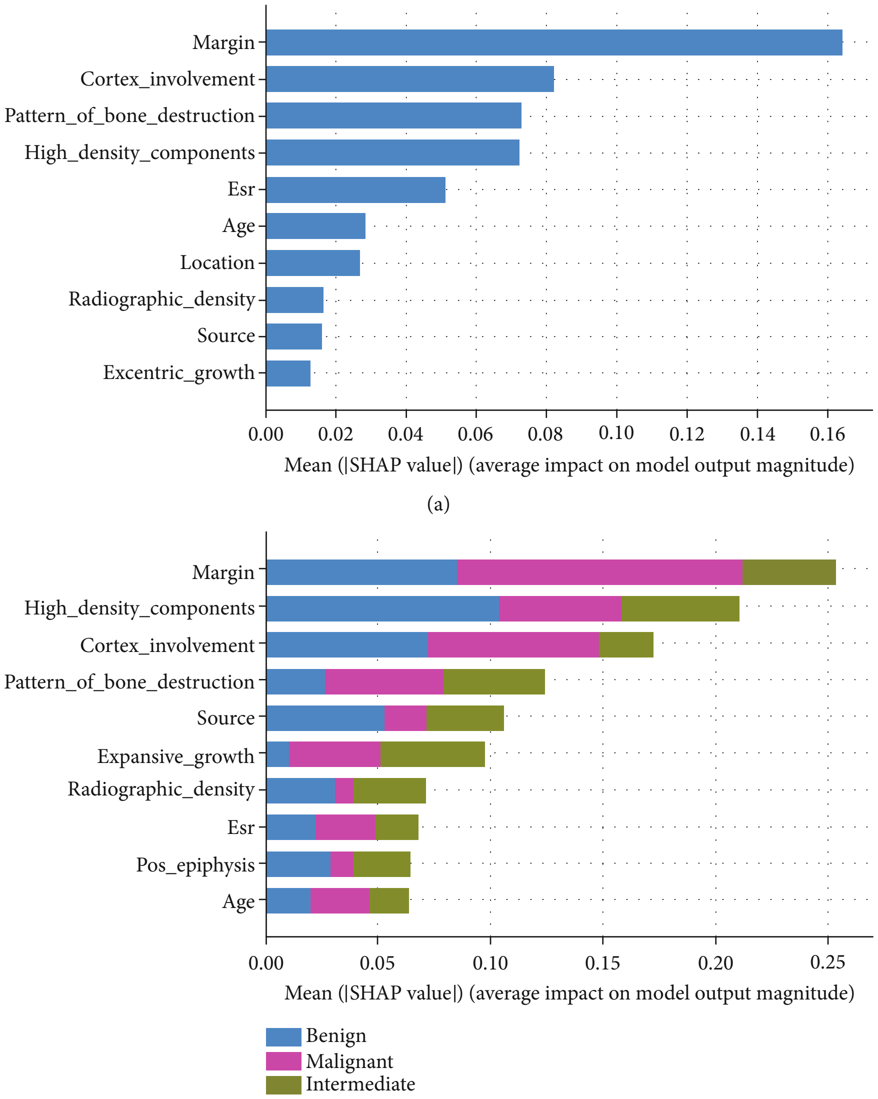
4. Discussion

This study built, validated, and tested random forest models for the bone tumors classification based on the lesion's conventional radiographic features and patients' clinical characteristics and identified the most important conventional radiographic features for bone tumors classification. A random forest model with high performance for bone tumors classification will have utility in the clinical setting.

In this study, the most important features influencing the binary classification model were margin, cortex involvement, pattern of bone destruction, and high-density components, indicating that malignant bone tumors were more destructive and aggressive than benign bone tumors. Consistent with these results, previous reports indicate that conventional radiographic features such as lesion margins, cortical destruction, presence and type of periosteal reaction, and matrix mineralization can be applied in differentiating benign from malignant bone tumors [20, 21]. However, these studies failed to quantify which feature was more important. Regarding imaging features, the margin is considered the most critical reflection of a primary bone tumor's malignant or benign nature. Malignant tumors typically manifest as ill-defined and indistinct margins with a broad transition zone between the tumor and normal bone, while benign tumors exhibit a sclerotic rim and a narrow transition zone. In terms of high-density components, malignant bone tumors such as osteosarcoma usually include more calcified and ossified components than benign bone tumors [20]. However, some malignant tumors, including Ewing sarcoma and plasmacytoma, did not show this feature in the present study.

As for the tertiary classification model, the most important features were margin, high-density components, cortex involvement, and pattern of bone destruction. Overall, these findings support the hypothesis that an interpretable model based on conventional radiographic features and clinical characteristics can be reliably applied to classify bone tumors in clinical practice.

The binary and tertiary classification models' performances were evaluated in the test sets using AUC value, accuracy, sensitivity, and specificity. The tertiary classification model relied on more features than the binary classification model to learn and predict, while the binary model was more accurate than the tertiary model. This may be because some imaging features of intermediate bone tumors are similar to those of benign or malignant bone tumors. For example, giant cell tumor of bone appears as an eccentric lytic lesion without marginal sclerosis and may have cortical destruction on radiography [22, 23], and eosinophilic granuloma of the bone appears as a moth-eaten lytic-bone lesion without marginal sclerosis, but with a continuous periosteal reaction [24]. Retrospective analysis of misclassified cases in this study revealed that 92.3% of misclassifications involved benign vs. intermediate bone tumors or malignant vs. intermediate bone tumors.



(a)

(b)

Figure 3: The most ten important features influencing the classification of bone tumors in the binary (a) and tertiary (b) models. The features are presented in descending order according to their absolute impact on the classification of bone tumor. The SHAP model takes into account all possible combinations of features in the presence/absence of a specific feature to evaluate its contribution to the prediction.

Table 4: The performance of random forest models.

	AUC	Sensitivity	Specificity	Accuracy	Overall accuracy	Micro AUC
Binary model	0.97	93.33%	95.61%	94.71%		
Tertiary model						
Benign	0.95	86.18%	83.62%	84.94%		
Malignant	0.98	90.77%	92.53%	92.05%	82.77%	0.94
Intermediate	0.89	58.82%	93.62%	86.19%		

Note: AUC: area under the receiver operating characteristic curve. All the results were obtained in the test set of two models.

Applying machine learning to classify bone tumors is scarce in the current study, probably because of bone tumor's rareness, variable location, and appearance, making data collection a challenge. Benndorf et al. built a pretest probabilistic (naive Bayes) classifier for primary malignant bone tumors based on the patient's age, sex, and tumor localization. Results from ten-fold cross-validation showed that the pretest probability of primary malignant bone tumor was correctly raised in 79.8% of cases [25]. Do et al. used a naive Bayes machine that processed 18 demographic and radiographic

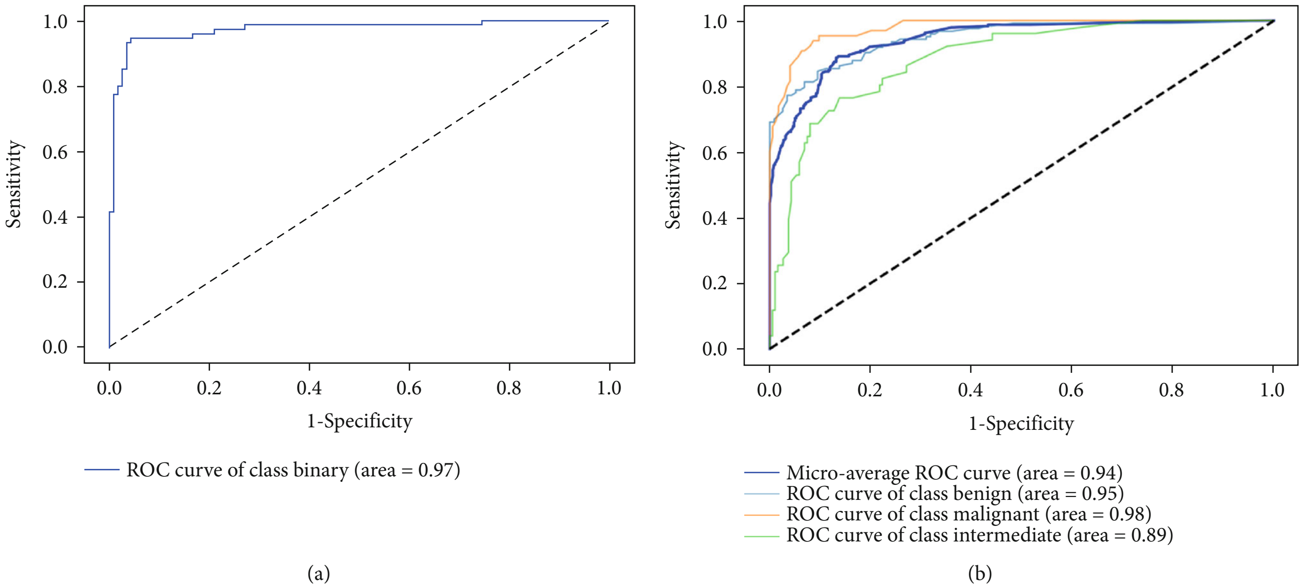


Figure 4: Receiver operating characteristic (ROC) curves of the binary (a) and tertiary (b) models.

features to evaluate primary and differential accuracy for the diagnosis of bone tumors. Primary accuracy was 62% and differential accuracy was 80% for the top 10 most common diagnoses [26]. In the present study, the binary and tertiary classification models' accuracy was 94.71% and 82.77%, indicating that these random forests outperformed previously reported models with superior accuracy. Unlike the previous study, this study evaluated model performance using AUC, which is more suitable for medical bias data.

The random forest model with reliable classification performance may assist radiologists in bone tumor diagnosis. Misdiagnosis and inappropriate treatment can also be reduced to a certain extent. It can improve the cure rate and prognosis of patients with bone tumors to a great extent eventually.

To the author's knowledge, the present study is the first to identify the most important conventional radiographic features for the bone tumor classification [27–29]. Thirteen conventional radiographic features were used to distinguish among benign, malignant, and intermediate bone tumors. Data from three medical centers were used to train, validate, and test the models, implying that the models are widely applicable across various clinical settings. This contrasts with other approaches based on image analysis, such as radiomics, which can be limited by different healthcare institutions' scanner parameters and image processing software [30, 31].

There are several limitations to this study. First, the classification models were based on conventional radiographic features without considering other imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI). Thus, some imaging features that are important for the classification may have been missed. However, conventional radiography is the preferred imaging modality for evaluating primary bone tumors. Therefore, models based on conventional radiographic features provide suitable and convenient solutions to guide clinical decision-making in bone tumor classification. Second, some patients' clinical characteristics were incomplete, and several specific biochemical markers of bone tumors, such as alkaline phosphatase, were not collected.

In conclusion, our study developed binary and tertiary models trained on a data set of linked conventional radiographic features and clinical characteristics to classify bone tumors, which obtained outstanding performance. Unlike previous studies, the SHapley Additive exPlanations was used to help radiologists, and other physicians recognize imaging features that are important for bone tumor classification. This approach may allow doctors to understand models easily so that they can integrate it into clinical practice to make precise diagnoses. In the future, the models may be enhanced by integrating CT and MRI features, potentially improving bone tumor classification and patient outcomes.

Data Availability

In order to protect the privacy of patients, the access to data is restricted.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Derun Pan and Renyi Liu Both contributed equally to this work.

Acknowledgments

This study has received funding from the Presidential Foundation of the National Key R&D Program of China (2019YFC0117301), National Key R&D Program of China (2019YFC0121903), Natural Science Foundation of Guangdong Province, China (2019A1515011168), Natural Science

Foundation of Guangdong Province, China (2018A0303130215), Clinical Research Project of Nanfang Hospital, Southern Medical University (2020CR003), and High Level-Hospital Program, Health Commission of Guangdong Province, China (HKUSZH201901026).

Supplementary Materials

Supplement Section: the locations represented by each letter. The type and number of tumors in the tertiary model's test set. *(Supplementary Materials)*

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Colorectal cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[2] C. D. M. Fletcher, J. A. Fletcher, and U. Krishnan, "WHO classification of tumours of soft tissue and bone," in *Pathology and Genetics of Tumours of Soft Tissue and Bone*, pp. 239–394, IARC Press, Lyon, 4th edition, 2013.

[3] F. Limaiem, D. W. Byerly, and R. Singh, *Osteoblastoma, In: StatPearls [Internet], Treasure Island (FL)*, StatPearls Publishing, 2021.

[4] Y. Yang, Z. Huang, X. Niu, H. Xu, Y. Li, and W. Liu, "Clinical characteristics and risk factors analysis of lung metastasis of benign giant cell tumor of bone," *Journal of Bone Oncology*, vol. 7, pp. 23–28, 2017.

[5] T. Woo, R. Lalam, V. Cassar-Pullicino et al., "Imaging of upper limb tumors and tumorlike pathology," *Radiologic Clinics of North America*, vol. 57, no. 5, pp. 1035–1050, 2019.

[6] H. Fritzsche, K. D. Schaser, and C. Hofbauer, "Benign tumours and tumour-like lesions of the bone : general treatment principles," *Orthopade*, vol. 46, no. 6, pp. 484–497, 2017.

[7] C. J. Gutowski, A. Basu-Mallick, and J. A. Abraham, "Management of bone sarcoma," *The Surgical Clinics of North America*, vol. 96, no. 5, pp. 1077–1106, 2016.

[8] Z. Luo, C. Ye, and H. X. Sang, "Osteosarcoma in the coracoid process that mimicked an osteochondroma: a case report," *Medicine*, vol. 96, no. 46, article e8608, 2017.

[9] P. F. Horstmann, W. H. Hettwer, and M. M. Petersen, "Treatment of benign and borderline bone tumors with combined curettage and bone defect reconstruction," *Journal of Orthopaedic Surgery*, vol. 26, no. 3, article 2309499018774929, 2018.

[10] I. N. Gemescu, K. M. Thierfelder, C. Rehnitz, and M. A. Weber, "Imaging features of bone tumors: conventional radiographs and MR imaging correlation," *Magnetic Resonance Imaging Clinics of North America*, vol. 27, no. 4, pp. 753–767, 2019.

[11] V. Vieth, "The importance of radiology in bone sarcoma diagnostics : initial and advanced diagnostics," *Orthopade*, vol. 48, no. 9, pp. 727–734, 2019.

[12] S. D. Yarmenitis, "Conventional radiology of bone and soft tissue tumors," in *Imaging in Clinical Oncology*, A. D. Gouliamos, J. Andreou, P. Kosmidis, A. D. Gouliamos, J. Andreou, and P. Kosmidis, Eds., pp. 83–88, Springer Milan, Milano, 2014.

[13] M. Kohli, L. M. Prevedello, R. W. Filice, and J. R. Geis, "Implementing machine learning in radiology practice and research," *AJR American Journal of Roentgenology*, vol. 208, no. 4, pp. 754–760, 2017.

[14] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[15] A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review," *Frontiers in Aging Neuroscience*, vol. 9, p. 329, 2017.

[16] J. Kruppa, Y. Liu, G. Biau et al., "Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory," *Biometrical Journal*, vol. 56, no. 4, pp. 534–563, 2014.

[17] M. Bahl, R. Barzilay, A. B. Yedidia, N. J. Locascio, L. Yu, and C. D. Lehman, "High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision," *Radiology*, vol. 286, no. 3, pp. 810–818, 2017.

[18] M. Montazeri, M. Montazeri, M. Montazeri, and A. Beigzadeh, "Machine learning models in breast cancer survival prediction," *Technology and Health Care*, vol. 24, no. 1, pp. 31–42, 2016.

[19] V. A. Zimmer, B. Glocker, N. Hahner et al., "Learning and combining image neighborhoods using random forests for neonatal brain disease classification," *Medical Image Analysis*, vol. 42, pp. 189–199, 2017.

[20] K. Mehta, M. P. McBee, D. C. Mihal, and E. B. England, "Radiographic analysis of bone tumors: a systematic approach," *Seminars in Roentgenology*, vol. 52, no. 4, pp. 194–208, 2017.

[21] D. H. Lee, J. M. Hills, M. I. Jordanov, and K. A. Jaffe, "Common tumors and tumor-like lesions of the shoulder," *The Journal of the American Academy of Orthopaedic Surgeons*, vol. 27, no. 7, pp. 236–245, 2019.

[22] N. F. Andrade, M. J. D. Teixeira, L. H. do Carmo Araújo, and C. E. B. Ponte, "Knee bone tumors: findings on conventional radiology," *Radiologia Brasileira*, vol. 49, no. 3, pp. 182–189, 2016.

[23] K. A. Raskin, J. H. Schwab, H. J. Mankin, D. S. Springfield, and F. J. Hornicek, "Giant cell tumor of bone," *JAAOS - Journal of the American Academy of Orthopaedic Surgeons*, vol. 21, no. 2, pp. 118–126, 2013.

[24] A. Angelini, A. F. Mavrogenis, E. Rimondi, G. Rossi, and P. Ruggieri, "Current concepts for the diagnosis and management of eosinophilic granuloma of bone," *Journal of Orthopaedics and Traumatology*, vol. 18, no. 2, pp. 83–90, 2017.

[25] M. Benndorf, J. Neubauer, M. Langer, and E. Kotter, "Bayesian pretest probability estimation for primary malignant bone tumors based on the surveillance, epidemiology and end results program (SEER) database," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 3, pp. 485–491, 2017.

[26] B. H. Do, C. Langlotz, and C. F. Beaulieu, "Bone tumor diagnosis using a naïve Bayesian model of demographic and radiographic features," *Journal of Digital Imaging*, vol. 30, no. 5, pp. 640–647, 2017.

[27] P. Caro-Domínguez and O. M. Navarro, "Bone tumors of the pediatric foot: imaging appearances," *Pediatric Radiology*, vol. 47, no. 6, pp. 739–749, 2017.

[28] O. Ahmed, D. D. Moore, and G. S. Stacy, "Imaging diagnosis of solitary tumors of the phalanges and metacarpals of the hand," *AJR. American Journal of Roentgenology*, vol. 205, no. 1, pp. 106–115, 2015.

[29] J. Panotopoulos, P. T. Funovics, and R. Windhager, "General diagnostic work-up for benign tumors of the musculoskeletal system," *Orthopade*, vol. 46, no. 6, pp. 473–476, 2017.

[30] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: a systematic review," *International Journal of Radiation Oncology·Biology·Physics*, vol. 102, no. 4, pp. 1143–1158, 2018.

[31] S. Rizzo, F. Botta, S. Raimondi et al., "Radiomics: the facts and the challenges of image analysis," *European Radiology Experimental*, vol. 2, no. 1, p. 36, 2018.

Hindawi
BioMed Research International
Volume 2021, Article ID 6685943, 15 pages
https://doi.org/10.1155/2021/6685943

Research Article

A Sparse Volume Reconstruction Method for Fetal Brain MRI Using Adaptive Kernel Regression

Qian Ni,[1] Yi Zhang,[2] Tiexiang Wen ,[3,4] and Ling Li[5]

[1]*Shenzhen Hospital of Guangzhou University of Chinese Medicine, Shenzhen, China*
[2]*Radiology & Vascular Surgery, Department of Radiology, Zhongda Hospital, Medical School, Southeast University, Nanjing, China*
[3]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*
[4]*National Innovation Center for Advanced Medical Devices, Shenzhen, China*
[5]*Suzhou Institute of Advanced Technology, Chinese Academy of Sciences, Suzhou, China*

Correspondence should be addressed to Tiexiang Wen; tx.wen@siat.ac.cn

Received 9 November 2020; Revised 25 January 2021; Accepted 18 February 2021; Published 8 March 2021

Academic Editor: Erlei Zhang

Slice-to-volume reconstruction (SVR) method can deal well with motion artifacts and provide high-quality 3D image data for fetal brain MRI. However, the problem of sparse sampling is not well addressed in the SVR method. In this paper, we mainly focus on the sparse volume reconstruction of fetal brain MRI from multiple stacks corrupted with motion artifacts. Based on the SVR framework, our approach includes the slice-to-volume 2D/3D registration, the point spread function- (PSF-) based volume update, and the adaptive kernel regression-based volume update. The adaptive kernel regression can deal well with the sparse sampling data and enhance the detailed preservation by capturing the local structure through covariance matrix. Experimental results performed on clinical data show that kernel regression results in statistical improvement of image quality for sparse sampling data with the parameter setting of the structure sensitivity 0.4, the steering kernel size of $7 \times 7 \times 7$ and steering smoothing bandwidth of 0.5. The computational performance of the proposed GPU-based method can be over 90 times faster than that on CPU.

1. Introduction

Magnetic resonance imaging (MRI) is an ideal diagnostic technique for researchers to investigate the development of the fetal brain [1]. Its advantages are the absence of ionizing radiation, the availability of different contrast options (T1-weighted, T2-weighted, and diffusion-weighted imaging), and the superior contrast of soft tissue compared with ultrasonography, and MRI is also a safe and noninvasive procedure for patients and fetuses [2–4]. For these reasons, MRI has been widely used to investigate the developing fetal brain in vivo [5]. For fetal brain MRI, the high-quality volume representation of 3D acquisition has significant clinical meaning [6]. By the observation of the reconstructed volume data, researchers can study the mechanism of brain development and maturation [7] and identify the fetal brain abnormality or potential injury [8, 9], such as brain tumors, vascular malformations, and posterior fossa abnormalities. Fetal brain MRI can provide abundant information about aid clinical management, prognostication, and counseling [10].

The duration of an examination is typically 45 to 60 minutes for fetal brain MRI [1]. One major problem of fetal brain MRI is motion artifacts caused by fetal and maternal motion, because of the long acquisition times of 3D MRI scanning. Maternal motion may be avoided by some measures, but fetal motion is usually fast and unpredictable, especially for the younger fetus. Thus, it is still challenging to reconstruct high-fidelity image for fetal brain MRI due to the presence of fetal motion. For fetal motion, different strategies can be adopted to reduce the motion artifacts on MRI [11]. The first strategy tries to prevent the motion occurring during the examination, such as maternal sedation. The second one tries to quicken the data sampling speed. The faster the acquisition techniques for fetal brain MRI are, the lower the motion occurs. For example, the single-shot fast spin echo (SSFSE) T2-weighted imaging can acquire a slice at 1-second speed [12]. On the other hand, sparse data sampling technique can be applied to shorten the time of data acquisition. The last strategy tries to reconstruct high-quality image through advanced postprocessing motion detection and correction algorithms, such as the SVR method [13].

For the SVR framework [14], it includes the following steps to reduce the fetal motion and reconstruct the high-quality 3D result: motion identification and exclusion step, registration step, reconstruction step, and regularization step. For the motion identification and exclusion step, we should estimate the amount of motion and exclude the slices with large amount of motion corruption. Early reconstruction approaches need to manually exclude the motion corrupted slices. The intersection-based motion correction approach can automatically detect and reject motion corrupted and incorrect registration slices by the abnormal level of their mean squared intensity difference with respect to all other intersecting slices [15]. In [16], Kainz et al. have proposed an approach to automatically estimate the amount of motion based on the low-rank decomposition for linearly correlated image slices [17]. Using this approach, we can reject stacks with large motion and choose the stack with the least motion as the template to prepare for the registration step. Registration step can be utilized to correct the motion between slice and the reconstructed volume. Rousseau et al. [18] combined the 2D/3D registration with the PSF to achieve the 3D reconstruction. PSF [14] is a mathematical function to model the actual appearance of data points in physical space. By PSF, we can physically correct estimation of the image acquisition process. Subsequently, the SVR method was modified to improve the robustness of the 2D/3D registration [19]. For the reconstruction step, superresolution methods [20, 21] are utilized to reconstruct the 3D volume. In [22], Gholipour and Warfield combined the superresolution method with slice-to-volume registration to reduce the burring effect. Because the motion identification and exclusion steps can exclude the slice of which the motion amount is greater than the threshold, the amounts of the slightly corrupted slices are still preserved for reconstruction. Using the robust superresolution volume reconstruction method [23], the weight of slightly corrupted and misaligned slices would be reduced to minimize the effect of motion. During the process of superresolution reconstruction, maximum likelihood estimation (MLE) is treated as an optimum solution to estimate the point's value [24]. To get better results, we should minimize the difference between the estimated slices and the acquired slices. Since the minimization only depends on the acquired samples, the estimation in the MLE framework is ill-posed and inaccurate when the samples are sparse [23]. The regularization step is used to solve the overfitting problem, and it can reduce image noise and registration errors. In [25], Charbonnier et al. proposed a deterministic edge-preserving regularization method to deal with image. However, this method makes it difficult to avoid the smoothing of edges. Adaptive regularization techniques can be employed to reduce the smoothing effects of regularization [26]. In [27], Rousseau et al. took advantage of total variation regulation to extend the superresolution reconstruction method.

The general SVR framework with the superresolution reconstruction method has been developed in [28]. One important way to alleviate fetal motion is to quicken the data acquisition time by the sparse data sampling technique. However, the traditional SVR method could not deal well with the sparse sampling problem and cannot provide high-quality image. In this paper, we utilize the SVR method with adaptive kernel regression to cope with the sparse volume reconstruction with minimum motion artifacts under the condition of sparse data acquisition. The key improvements compared to previous works are as follows: firstly, we make use of the sparse samples to get faster speed of data acquisition in fetal brain MRI. Next, the adaptive kernel regression-based reconstruction method [29] with robust statistics calculation [24] can reconstruct high-quality volume under the condition of sparse sampling. In general, our comprehensive reconstruction method for fetal brain MRI mainly includes slice-to-volume registration, the robust statistics calculation, the PSF-based volume update, and adaptive kernel regression-based volume update.

The rest of the paper is organized as follows. The detailed methodology is discussed in Section 2. We design the actual implementation of the algorithm in Section 3. Section 4 involves the experiment results and compares with those of superresolution methods. In this section, we also discuss how to determine the optimal values of related parameters using GPU-based fast reconstruction. Finally, we make a brief conclusion in Section 5.

2. Methods

2.1. Model of Data Acquisition and Motion Estimation. During data acquisition of fetal brain MRI, we collected several stacks of 2D slices in different orientations. Because of the fetal motion, the movement could be observed between these slices. Assume that the acquired k misaligned 2D slices are $I_j \in \mathbf{R}^{n\times h}$, $j = 1, \cdots, k$, and the corresponding sparse 2D slices are $I_j^s \in \mathbf{R}^{n\times h}$, $j = 1, \cdots, k$. During the slice acquisitions of MRI, the inhomogeneity of the magnetic field B_j, $j = 1, \cdots, k$, affects the intensities of the slices and the scaling factor S_j', $j = 1, \cdots, k$, is potentially different for each acquired slices. In [30], the logarithmic transformation was chosen to make the bias additive. However, field in-homogeneities are known to be multiplicative. Differently, we use the multiplicative bias field to form the multiplicative exponential model which replaces the logarithmic model. So the scaled and bias corrected slice I_j' can be modeled as

$$\begin{aligned} I_j^s &= \operatorname{sparse}\left(I_j\right), \\ \operatorname{vec}\left(I_j'\right) &= S_j' \cdot \exp\left(-B_j\right) \cdot \operatorname{vec}\left(I_j^s\right), \end{aligned} \tag{1}$$

where I_j^s is the sparsely sampled slice coming from the sparse operator sparse(•), vec(•) is the vectorization operator that transforms a m-pixel ($m = n \times h$) image $\mathbf{R}^{n\times h}$ into a vector of intensity values $\mathbf{R}^m$. The corresponding k-aligned 2D ground-truth slices are $I_j^* \in \mathbf{R}^{n\times h}$, $j = 1, \cdots, k$. The relationship between corrected slices I_j' and the ground-truth slices I_j^* can be denoted as follows:

$$\mathrm{vec}\left(I_j'\right) = \theta_j \cdot \mathrm{vec}\left(I_j^*\right) + \mathrm{vec}\left(e_j\right), \quad j = 1, \cdots, k, \tag{2}$$

where e_j is the motion error, and θ_j denotes the unknown motion transformation parameter of slice I_j^*. Then, we can define the following data matrix:

$$\begin{aligned} D &= \left[\mathrm{vec}\left(I_1'\right); \cdots; \mathrm{vec}\left(I_k'\right)\right] \in \mathbf{R}^{m\times k}, \\ X &= \left[\mathrm{vec}(I_1^*); \cdots; \mathrm{vec}(I_k^*)\right] \in \mathbf{R}^{m\times k}, \\ E &= \left[\mathrm{vec}(e_1); \cdots; \mathrm{vec}(e_k)\right] \in \mathbf{R}^{m\times k}, \\ T_{\mathrm{total}} &= \left[\theta_1; \cdots; \theta_k\right] \in \mathbf{R}^{m\times k}. \end{aligned} \tag{3}$$

where D, X, E, and T_{total} denote the observed data matrix, reconstructed data matrix, motion error matrix, and the rigid transformation matrix. Given these definitions, the observed data matrix D can be described as $D = T_{\mathrm{total}} \bullet X + E$. The motion error matrix E is mainly caused by misaligned slices. The misaligned slices can cause the inaccurate reconstructed volume, and we want to exclude the stack which has many misaligned slices. However, we cannot directly calculate the amount of stack motions for the observed data matrix D, but a low-rank approximation D^* as surrogate estimate can be used to evaluate the stack motion indirectly [16]. It has been shown that D^* provides the best approximation to D [31]. The difference value between D^* and D measures the motion error E. The smaller difference value indicates that the stack has fewer motions. To provide the low-rank approximation, the singular value decomposition is used to decompose the data matrix D as $D_{m\times k} = U_{m\times k} S_{k\times k} V_{k\times k}^T$. The singular value decomposition of D produces three matrices U, S, and V. U and V are both orthogonal matrices, and S is the diagonal matrix containing the singular values on the diagonal. And the singular value decomposition of D^* is the first r singular values of the original matrix D, i.e., $D_{m\times k}^* = U_{m\times r}^* S_{r\times r}^* V_{r\times k}^{*T}$, $r = 1, \cdots, k$. U^* and V^* are the first r columns of U and V, and S^* is the top left $r \times r$ submatrix of S. The relative error based on the Frobenius norm $\|D - D^*\|$ is used to measure the approximation between D^* and D, i.e. $\delta_r = \|D - D_r^*\|/\|D\|$. For the different values of $r = 1, \cdots, k$, we can find the minimal rank r for each stack that satisfies the given threshold β, i.e., $\arg\min_r\{\delta_r < \beta\}$. Combining δ_r and r, the surrogate estimate for the amount of motion is given by $\mu_r = \delta_r \bullet r$.

Based on the low-rank decomposition method, we can choose one stack with minimal motion as the target template and first perform the 3D rigid volumetric registration between the target template and the other stacks (stack to template registration). During the first registration, we can get the corresponding rigid global transformation matrix T_{global}. Then, second, the 3D rigid volumetric registration between the reconstructed volume and all slices (slice to reconstructed volume registration) can produce local transformation matrix T_{local}. The prerequisite for two registrations is that all stacks and reconstructed volume should be mapped to the world coordinates. Thus, we need to define two transformations to map each pixel in the 2D slice and each voxel in the reconstructed volume to a continuous location in the world coordinates. The first one is world transformation $W_{\mathrm{s}} = [\theta_1^w, \cdots, \theta_k^w]$ that transforms the discrete coordinates of a pixel $p_{\mathrm{s}} = [i, j, 0, 1]^T \in I_j^{\mathrm{s}}$ in the acquired slice to the continuous local world coordinates. The second one is world transformation $W_r = [\theta_1^{w'}, \cdots, \theta_k^{w'}]$ that transforms the discrete coordinates of a voxel $p_{\mathrm{r}} = [x, y, z, 1]^T \in X$ in the reconstructed volume to the continuous local world coordinates. Meanwhile, the mapping and registrations can be combined and formulated as Equation (4). Thus, Figure 1 illustrates the whole transformation process from the pixels in the sparse slice to voxels in the 3D reconstructed volume.

$$p_{\mathrm{r}} = \left[W_{\mathrm{r}}^{-1} \cdot T_{\mathrm{total}} \cdot W_{\mathrm{s}}\right] \cdot p_{\mathrm{s}} = \left[W_{\mathrm{r}}^{-1} \cdot \left(T_{\mathrm{global}} \cdot T_{\mathrm{local}}\right) \cdot W_{\mathrm{s}}\right] \cdot p_{\mathrm{s}}. \tag{4}$$

2.2. PSF-Based Volume Update. To model the actual appearance of sampling data points in physical space, the point spread functions (PSFs) are used to make the exact estimation for every voxel value in the reconstructed target volume. For the MRI ssFSE sequence in this paper, the exact shape of the PSF has been measured using a phantom and rotating imaging encoding gradient in [14]. The resulting shapes of the PSF in in-plane and in through-slice are given by a sinc function and the slice profile, respectively. Since the ideal rectangle profile has the very dense and inefficient spatial sampling, Kuklisova-Murgasova et al. [28] have proposed to use the 3D Gaussian function with the full width at half maximum (FWHM) equal to the slice thickness as an approximation for the sinc function. The PSF function based on 3D Gaussian profile is used to approximately model the SSFSE sequence and is expressed as follows:

$$\mathrm{PSF}_{\mathrm{G}} = \exp\left(\frac{-dx^2}{2\sigma_x^2} + \frac{-dy^2}{2\sigma_y^2} + \frac{-dz^2}{2\sigma_z^2}\right), \tag{5}$$

where dx, dy, and dz are the offsets from the center of a reconstructed voxel, σ_x and σ_y are the full width at half maximum (FWHM) in the in-plane x - and y -directions, and the σ_z equals to the slice thickness in the through-plane direction. For each pixel in the sampled slice, the $\mathrm{PSF}_{\mathrm{G}}$ is applied to obtain the corresponding PSF coefficient matrix. Since every sampling pixel (i.e., p_{s}) does not perfectly align itself with the reconstructed voxel (i.e., p_{r}), one p_{s} contributes to more than one p_{r}. To model this, every voxel is sampled

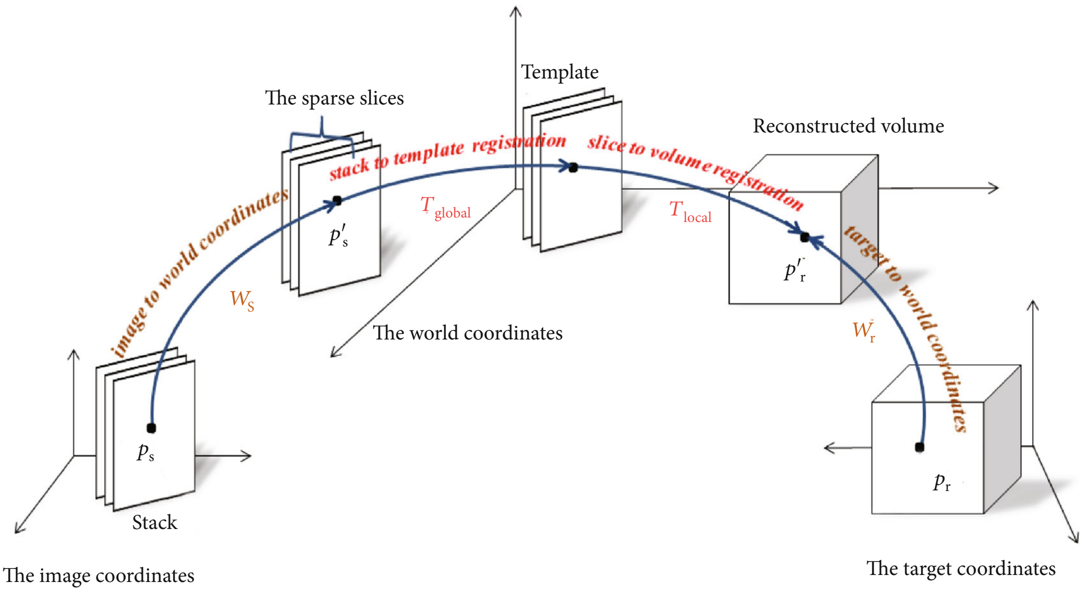


Figure 1: The illustration of the whole transformation process from pixels p_{s} to voxel p_{r}.

around its local surrounding neighbor in the reconstructed volume to make sure that it has at least one corresponding pixel in the acquired slices. Then, the PSF coefficients are used to weigh the pixel's contribution during the*n*th iteration.

$$p_{\text{r}} = \left\lfloor \left(W_{\text{r}}^{-1}\cdot T_{\text{total}}\cdot W\right)\cdot p_{\text{s}}\right\rfloor, \widetilde{p_{\text{s}}} = \left(W_{\text{r}}^{-1}\cdot T_{\text{total}}\cdot W\right)^{-1} p_{\text{r}},$$
$$\text{X}\left(p_{\text{r}}^{n+1}\right) = \text{PSF}\left(p_{\text{s}} - \widetilde{p_{\text{s}}}\right)\cdot S'_j\cdot \exp\left(-B_j\right)\cdot I_j^i(p_{\text{s}}) + X\left(p_{\text{r}}^{n}\right), \tag{6}$$

where $\lfloor\bullet\rfloor$ is the operation that finds the nearest voxel center in the space of the reconstructed volume. The reconstructed volume X is updated iteratively through the PSF-based data sampling model, and every voxel of X is filled at an arbitrarily chosen voxel size.

2.3. Robust Outlier Removal. Once the target volume is updated based on the Gaussian PSF, the simulated slices $I^{\text{ss}} = [I_1^{\text{ss}}, \cdots, I_k^{\text{ss}}] \in \mathbf{R}^{n\times h}$ can be generated from the updated reconstructed volume. Then, the misaligned error e^* between the corrected acquired sparse slices I' and simulated slices I^{ss} can be computed as

$$E(e^*) = I'(p_{\text{s}}) - I^{\text{ss}}(p_{\text{s}}). \tag{7}$$

In [28], an EM model-based robust statistics approach was proposed to classify each slice pixel into two classes: inliers and outliers. Specially, the probability density function (PDF) for the inlier class is modeled as a zero-mean Gaussian distribution with variance σ^2: $E \sim N(0,\sigma^2)$, and the PDF for the outlier class is modeled as a uniform distribution with constant density, which is a reciprocal of the range $[a,b]$ $: E \sim U(a,b)$. Then, the likelihood of the observing error e^* can be expressed as

$$P(e^* \mid \sigma, c) = c\cdot N_\sigma(e^*) + (1-c)\cdot U, \tag{8}$$

where c is a mixing proportion of inliers representing the correctly matched voxels. Then, the posterior probability of a voxel being an inlier can be computed using the expectation step as

$$p_{ij} = \frac{c\cdot N_\sigma\left(e_{ij}^*\right)}{c\cdot N_\sigma\left(e_{ij}^*\right) + (1-c)\cdot U}. \tag{9}$$

The variables σ and c are updated by the following maximization step:

$$\sigma = \sqrt{\frac{\sum p_{ij}\cdot\left(e_{ij}^*\right)^2}{\sum p_{ij}}},$$
$$c = \frac{\sum p_{ij}}{\sum N_j}, \tag{10}$$

where N is the number of the pixels in the slice. By constantly iterating, we can get the best parameters σ and c. The inlier probability can be used to weigh the PSF-based volume update. By the same way, each slice is classified into inlier and outlier as well using the EM algorithm. The probability of an inlier slice is defined as $p_j^{\text{slice}} = \sqrt{\sum_i p_{ij}^2 / N_j}$. The slices inferred to be an outlier are excluded from the PSF-based volume update to remove artifacts of motion corruption and misregistration.

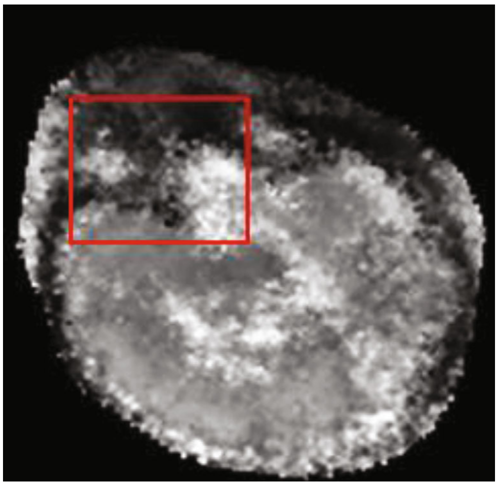


Figure 2: The reconstructed volume after PSF-based volume update.

The purpose of the outlier removal is to make the framework more robust by rejecting the outlier slices. The outlier removal module is adopted directly from the cited previous work [16], where the accuracy of the motion recognition and outlier removal has been evaluated in detail by simulating the slice motion at a variety of amplitudes and comparing the known motion amplitude to the surrogate measure provided through rank approximation. They have shown that there was strong correlation between the amplitude of the known motion and the values of μ_{r} derived from the stack data matrices.

2.4. Steering Kernel Regression-Based Volume Update. For sparse reconstruction, it is experimentally found that the reconstructed volume still remains unallocated or inaccurate voxels after PSF-based volume update and the reconstructed result is noise as shown in Figure 2.

In [32], the kernel regression can make better nonparametric estimation for the empty pixels. In this paper, the steering kernel regression approach [29] is introduced to update the voxels for the previous sparse volume data. The model for the kernel regression is expressed as

$$Y_i = r(X_i) + \varepsilon_i, \quad i = 1, \cdots, M, \tag{11}$$

where $r(\bullet)$ is the function of kernel regression, $X_i = (x_i, y_i, z_i)$ is the 3D coordinate of the voxel, ε_i is a zero-mean Gaussian noise with variance σ_0^2 as $X \sim N(0, \sigma_0^2)$, and Y_i is the voxel after PSF-based Gaussian volume update.

Assuming that the voxel X_i is close to the known voxel X in the reconstructed volume, we have the following approximation for $r(X_i)$ using the N-term-order Taylor series:

$$\begin{aligned} r(X_i) &\approx r(X) + \{\nabla r(X)\}^T (X_i - X) \\ &\quad + \frac{1}{2!}(X_i - X)^T \{Hr(X)\}(X_i - X) + \cdots \\ &= \beta_0 + \boldsymbol{\beta}_1^T (X_i - X) + \boldsymbol{\beta}_2^T \mathrm{vech} \\ &\quad \cdot \left\{ (X_i - X)(X_i - X)^T \right\} + \cdots, \end{aligned} \tag{12}$$

where ∇ and H are, respectively, the gradient (3×1) and Hessian (3×3) operators; $\beta_0 = r(X)$, which is the voxel value of interest; and the vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are defined as

$$\boldsymbol{\beta}_1 = \left[G_x, G_y, G_z\right]^T = \left[\frac{\partial r(X)}{\partial x}, \frac{\partial r(X)}{\partial y}, \frac{\partial r(X)}{\partial z}\right]^T, \tag{13}$$

$$\begin{aligned} \boldsymbol{\beta}_2 = \frac{1}{2}\Bigg[& \frac{\partial^2 r(X)}{\partial x^2}, 2\frac{\partial^2 r(X)}{\partial x \partial y}, 2\frac{\partial^2 r(X)}{\partial x \partial z}, \\ & \frac{\partial^2 r(X)}{\partial y^2}, 2\frac{\partial^2 r(X)}{\partial y \partial z}, \frac{\partial^2 r(X)}{\partial z^2} \Bigg]^T. \end{aligned} \tag{14}$$

$\mathrm{vech}(\bullet)$ is the half-vectorization operator that transforms the upper triangular portion of a symmetric matrix into a column-stacked vector, i.e.,

$$\mathrm{vech}\left(\begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} \right) = \begin{bmatrix} a & b & c & d & e & f \end{bmatrix}^T. \tag{15}$$

Based on the least-squares formula, we can optimize Equation (12) as

$$\begin{aligned} \min_{\{\boldsymbol{\beta}_n\}_{n=0}^{N}} \sum_{i=1}^{L} & \left[Y_i - \beta_0 - \boldsymbol{\beta}_1 (X_i - X) \right. \\ & \left. - \boldsymbol{\beta}_2 (X_i - X)^2 - \cdots \right]^2 \cdot \frac{1}{h} K\left(\frac{X_i - X}{h}\right), \end{aligned} \tag{16}$$

where L is the number of known voxels within the neighborhood window, $K(\bullet)$ is the distance-weighted kernel function which penalizes distance away from the local position, and h is the smoothing parameter that controls the strength of the penalty. The kernel function is chosen as the exponential function, Gaussian function, or other functions which satisfy the following conditions:

$$\begin{aligned} \int t K(t) dt &= 0, \\ \int t^2 K(t) dt &= c. \end{aligned} \tag{17}$$

For the computation simplicity, the Gaussian-based

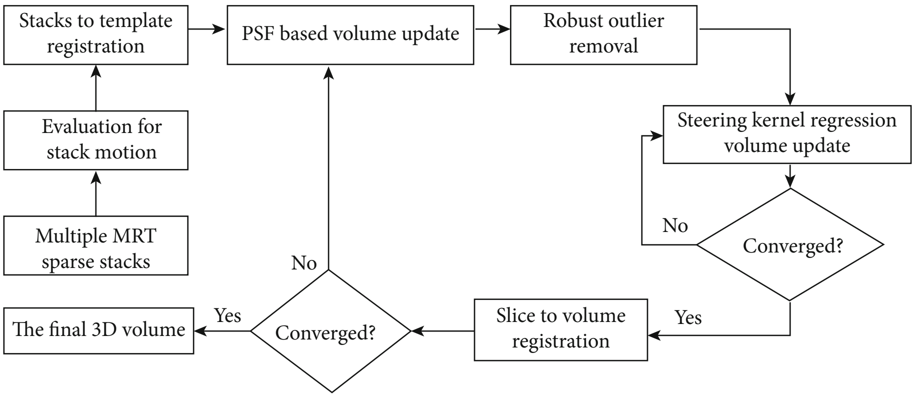

Stacks to template registration
PSF based volume update
Robust outlier removal
Steering kernel regression volume update
Evaluation for stack motion
Multiple MRT sparse stacks
No
Converged?
No
Yes
The final 3D volume
Converged?
Slice to volume registration
Yes

Figure 3: Flowchart of the proposed algorithm.

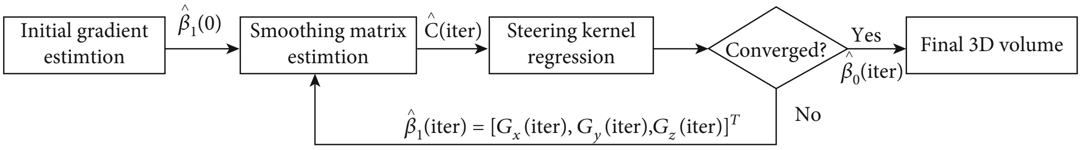
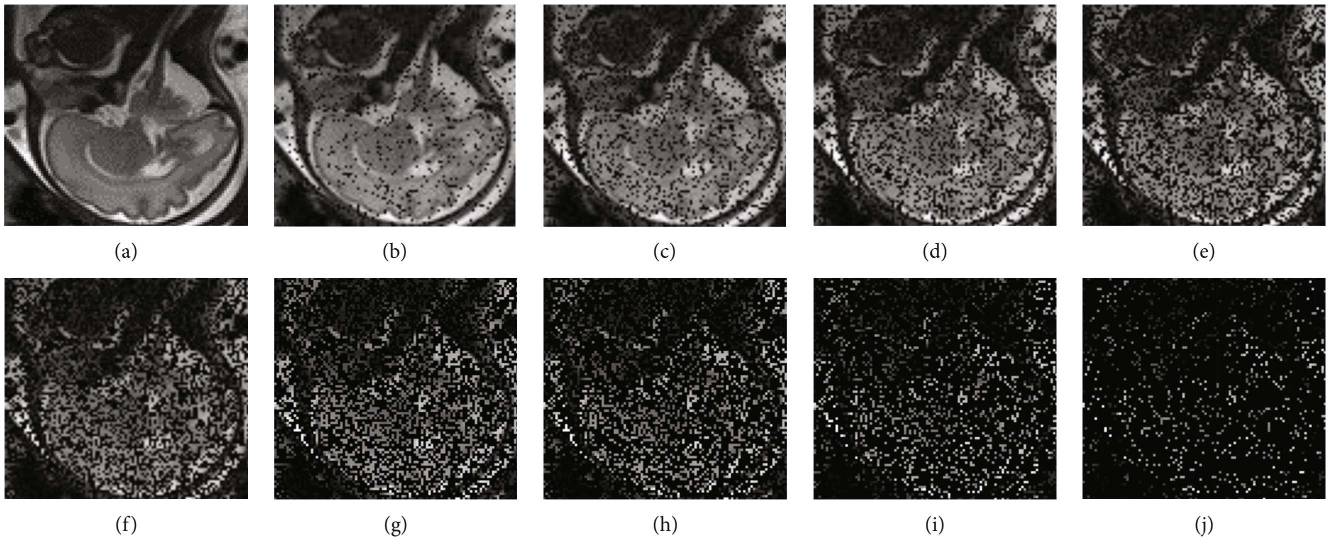

Initial gradient estimtion
$\hat{\beta}_1(0)$
Smoothing matrix estimtion
$\hat{C}(\text{iter})$
Steering kernel regression
Converged?
Yes
$\hat{\beta}_0(\text{iter})$
Final 3D volume
No
$\hat{\beta}_1(\text{iter}) = [G_x(\text{iter}), G_y(\text{iter}), G_z(\text{iter})]^T$

Figure 4: The iterative steering kernel regression.


(a) (b) (c) (d) (e)

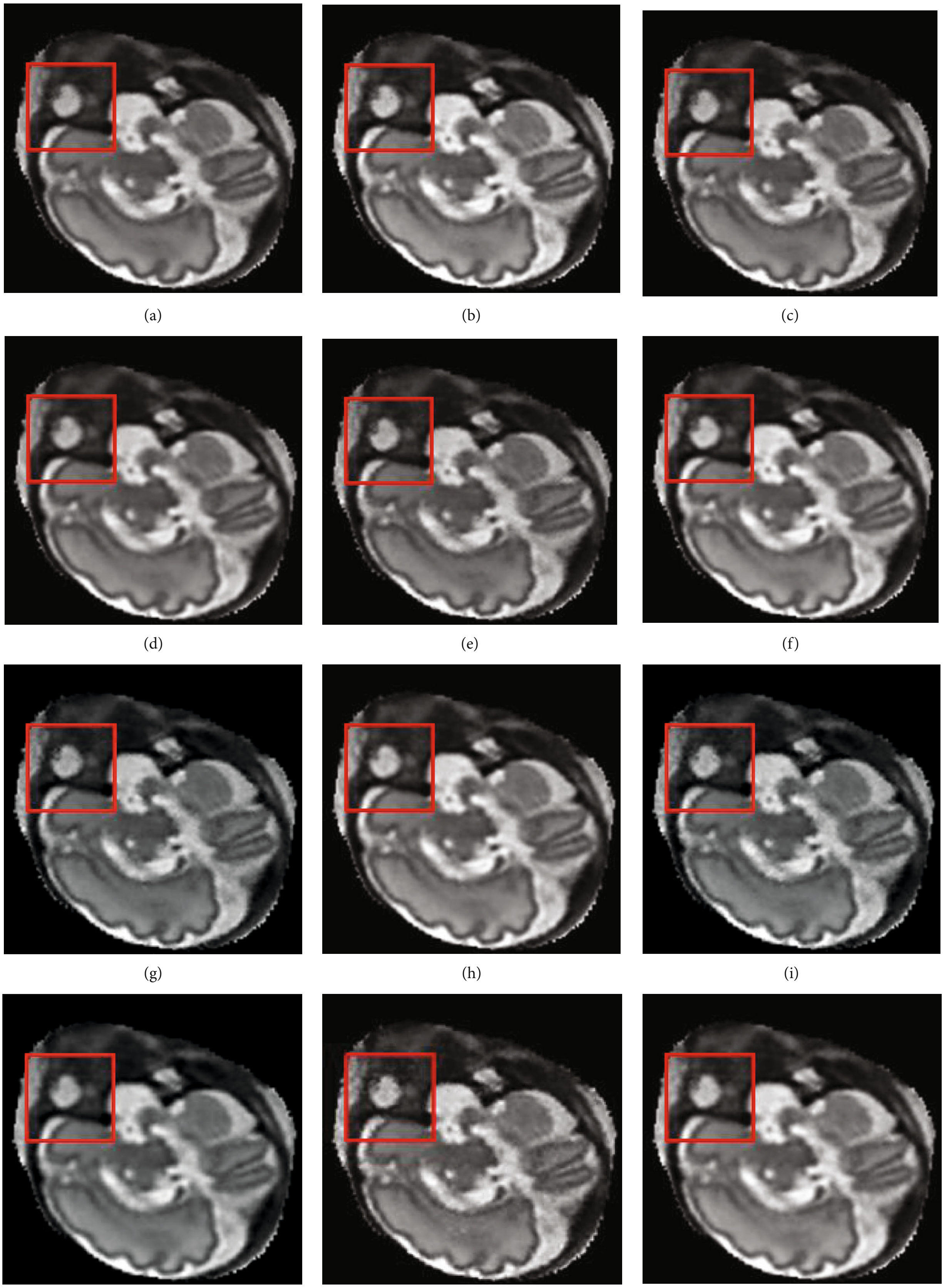
(f) (g) (h) (i) (j)

Figure 5: The original and spare slices: (a) the typical 30th original slice; (b–j) the corresponding simulated sparse slice by removing once every 10% proportion pixels ranging from 10% to 90%.

kernel function is chosen in the steering kernel regression [33]. The steering kernel adapts locally to image structures (e.g., edges, flat, and texture areas), which are captured by the kernel footprint. For example, the kernel footprint is large in the flat areas, elongated in edge areas, and compact in texture areas. The 3D steering kernel function takes from

$$K_s(X_i - X) = \frac{\sqrt{\det(\mathbf{C}_i)}}{2\pi h^2} \exp\left\{-\frac{1}{2h^2}\left\|\mathbf{C}_i^{1/2}(X_i - X)\right\|_2^2\right\}, \tag{18}$$

where $\|\bullet\|_2^2$ is the L_2 norm and $\mathbf{C}_i$ is the symmetric covariance matrix. Since the local image structure is highly related to the gradient covariance, we can make the data-dependent



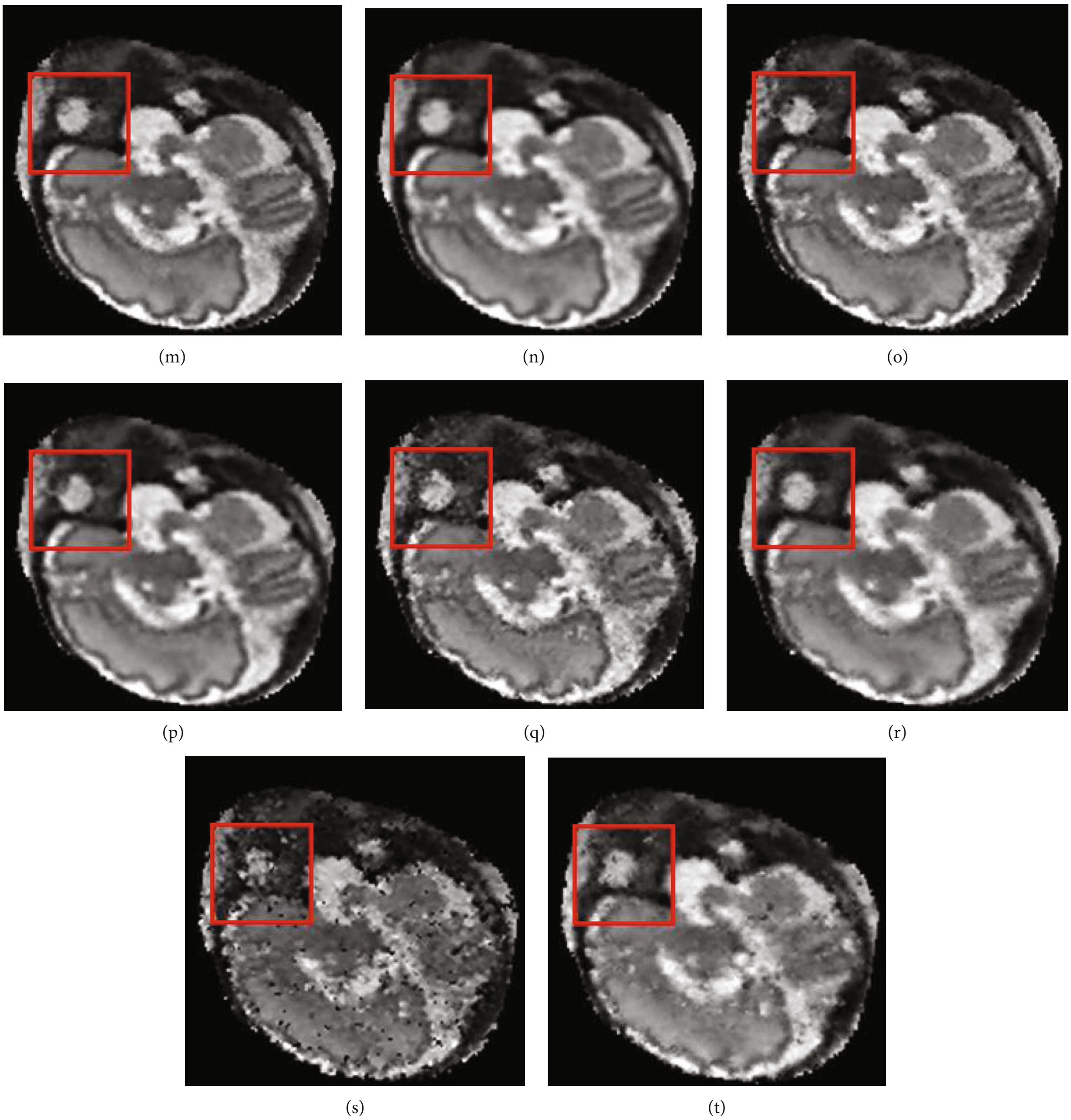
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 6: Continued.


(m)

(n)

(o)

(p)

(q)

(r)

(s)

(t)

Figure 6: Reconstruction results of different data removal ratio by Kainz et al. method (2015) and our proposed method. (a), (c), (e), (g), (i), (k), (m), (o), (q), (s) are the reconstructed results by Kainz et al. method for the sparsely sampled dataset with once every 10% data removal ratio ranging from 0% to 90% respectively. (b), (d), (f), (h), (j), (l), (n), (p), (r), (t) are the reconstructed results by the proposed methodfor the sparsely sampled dataset with once every 10% data removal ratio ranging from 0% to 90%, respectively. The red rectangle points to the obvious difference, which appears as artifacts in the reconstructed image if no steering kernel regression volume updated is used.

covariance matrix estimation utilizing the local edge gradients:

$$\widehat{\mathbf{C}}_i \approx \begin{bmatrix} \sum_{X_i \epsilon w} G_x(X_i)G_x(X_i) & \sum_{X_i \epsilon w} G_x(X_i)G_y(X_i) & \sum_{X_i \epsilon w} G_x(X_i)G_z(X_i) \\ \sum_{X_i \epsilon w} G_x(X_i)G_y(X_i) & \sum_{X_i \epsilon w} G_y(X_i)G_y(X_i) & \sum_{X_i \epsilon w} G_y(X_i)G_z(X_i) \\ \sum_{X_i \epsilon w} G_x(X_i)G_z(X_i) & \sum_{X_i \epsilon w} G_y(X_i)G_z(X_i) & \sum_{X_i \epsilon w} G_z(X_i)G_z(X_i) \end{bmatrix}, \quad (19)$$

Table 1: The RMSE and SSIM value comparison of fetal brain reconstruction with different removal proportions, respectively.

Different removal proportions	RMSE Kainz et al.'s method (2015)	RMSE Our method	MSSIM Kainz et al.'s method (2015)	MSSIM Our method
0%	19.096	19.096	1	1
10%	32.578	29.120	0.9752	0.9759
20%	38.043	33.947	0.9690	0.9691
30%	43.171	36.790	0.9591	0.9608
40%	49.194	41.027	0.9478	0.9458
50%	55.894	45.480	0.9202	0.9366
60%	67.053	53.305	0.9063	0.9271
70%	81.522	62.964	0.8667	0.8993
80%	111.917	78.886	0.7862	0.8449
90%	180.483	112.249	0.6338	0.7407

where w is a local analysis window and $G_x(\bullet)$, $G_y(\bullet)$, and $G_z(\bullet)$ are the gradients along the x-, y-, and z-directions.

Equation (16) can be expressed in the matrix form as

$$\widehat{\mathbf{B}} = \min_{\mathbf{b}} \|\mathbf{Y} - \mathbf{X}_{\mathbf{F}}\mathbf{B}\|_{\mathbf{W}}^2 = \min_{\mathbf{b}} (\mathbf{Y} - \mathbf{X}_{\mathbf{F}}\mathbf{B})^T \mathbf{W} (\mathbf{Y} - \mathbf{X}_{\mathbf{F}}\mathbf{B}), \quad (20)$$

where $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_L]^T$ is the vector set of all known voxels, $\mathbf{B} = [\beta_0, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_N^T]^T$ is the vector set of all estimated parameters, $\mathbf{W} = \mathrm{diag}\,[K_s(X_0 - X), K_s(X_1 - X), \cdots, K_s(X_L - X)]$ is the diagonal matrix whose elements on the diagonal are the value of $K_s(\bullet)$, and the other elements are zero. According to the least-squares method, we have the following solution:

$$\widehat{\mathbf{B}} = \left(\mathbf{X}_{\mathrm{F}}^T \mathbf{W} \mathbf{X}_{\mathrm{F}}\right)^{-1} \mathbf{X}_{\mathrm{F}}^T \mathbf{W} \mathbf{Y}, \quad (21)$$

where $\mathbf{B} = [\beta_0, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_N^T]^T$, $\widehat{r}(X_i) = \widehat{\beta}_0 = \mathbf{e}_1^T (\mathbf{X}_{\mathrm{F}}^T \mathbf{W} \mathbf{X}_{\mathrm{F}})^{-1} \mathbf{X}_{\mathrm{F}}^T \mathbf{W} \mathbf{Y}$ is the voxel value estimated by the steering kernel regression, $\widehat{\boldsymbol{\beta}}_1 = [G_x, G_y, G_z]^T$ is applied for computing the symmetric covariance matrix $\widehat{\mathbf{C}}_{i+1}$ iteratively, and $\mathbf{X}_{\mathrm{F}}$ is a coordinate matrix expressed as follows:

$$\mathbf{X}_{\mathrm{F}} = \begin{bmatrix} 1 & (X_0 - X)^T & \mathrm{vech}^T\left\{(X_0 - X)(X_0 - X)^T\right\} & \cdots \\ 1 & (X_1 - X)^T & \mathrm{vech}^T\left\{(X_1 - X)(X_1 - X)^T\right\} & \cdots \\ . & . & . & \cdots \\ . & . & . & \cdots \\ 1 & (X_L - X)^T & \mathrm{vech}^T\left\{(X_L - X)(X_L - X)^T\right\} & \cdots \end{bmatrix}. \quad (22)$$

Once the reconstructed volume is updated based on the steering kernel regression, we update the simulated slices I^{ss} and the misaligned error $E(e^*)$ according to Equation (7). To remove artifacts caused by motion corruption and misregistration and enhance image edges, we further update the reconstructed volume using the following equation:

$$X\left(p_{\mathrm{r}}^{n+1}\right) = \mathrm{PSF}\left(p_{\mathrm{s}} - \widetilde{p_{\mathrm{s}}}\right) \cdot p_j^{\mathrm{slice}} \cdot p_{ij} \cdot E\left(e^*\right) + X\left(p_{\mathrm{r}}^n\right). \quad (23)$$

3. Implementation

The experiment computer is equipped with Intel Core i5 2.6 GHz CPU, and the operating system is Windows 7 64 bit. We have implemented the proposed algorithm using the Microsoft Visual Studio 2012 and Image Registration Toolkit (IRTK) software package which includes many useful methods to do registration, transformation, and other image processing. In this section, we discuss the key implementation details. The diagram of the total algorithm is expressed in Figure 3.

The first step is to evaluate the stack motion according to the method of low-rank decomposition. We estimate the amount of the stack motion by the surrogate $\mu_r = \delta_r \bullet r$ and choose the stack with the minimum amount of stack motion as the template. The second step is to perform the global registration, which calculates the matrix of global transformation T_{global} from the other stacks to the template. The third step is the iterative registration-based volume reconstruction, which consists of the outer registration step and the inner reconstruction step. The outer loop step includes the PSF-based volume update, robust outlier removal, steering kernel regression-based volume update, and slice to volume registration. The PSF-based volume update step makes the initial estimation of the reconstructed volume based on Equation (6). Then, the simulated slices are created and used for the robust misaligned error calculation between the simulated slices and the acquired slices as described in Section 2.3. The robust statistic calculation achieves the classification of outlier slices and inlier slices. The outlier slices are excluded to remove artifacts of motion corruption and misregistration. The slice to volume registration is to calculate the local transformation T_{local} from slices to reconstructed volume. The whole transformation process is described by Equation (4). The volume update based on the adaptive steering kernel regression is aimed at reconstructing the accurate volume iteratively as shown in Figure 4. The initial gradients $\widehat{\boldsymbol{\beta}}_1(0) = [G_x(0), G_y(0), G_z(0)]^T$ are estimated by the classical kernel regression. Then, the gradient information is used to calculate the covariance smoothing matrix $\widehat{\mathbf{C}}(\mathrm{iter})$ (i.e., Equation (19)). We use smoothing matrix to update the voxel value $\widehat{\beta}_0(\mathrm{iter})$ and its corresponding gradients $\widehat{\boldsymbol{\beta}}_1(\mathrm{iter})$ according to Equation (21), respectively. To obtain a more reliable voxel estimation, the process is iterated three times in our experiment.

4. Experimental Results and Evaluation

4.1. Evaluation of Image Quality. In the experimental evaluation, we used the datasets from the fetal MRI datasets [16], which were acquired by a Philips Achieva 3 T MR scanner. During the experiment, the volunteers were lying at a 20° tilt on the left side to avoid the pressure on the inferior vena cava.

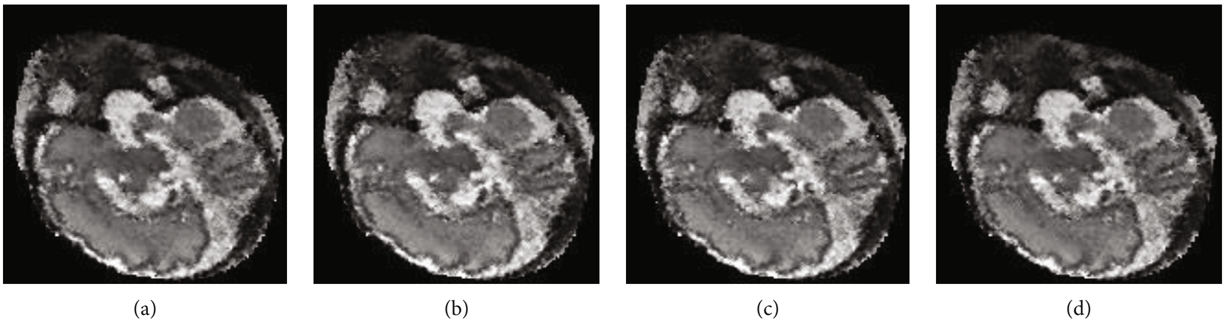
Table 2: The running time comparison of the adaptive kernel regression method for fetal brain reconstruction based on CPU and GPU, respectively.

Processor	Single-threaded CPU	Multithreaded CPU	GPU	Single-threaded CPU vs. GPU	Multithreaded CPU vs. GPU
Gradient information (s)	416.960	108.762	5.047	82.62	21.55
Covariance smoothing matrix (s)	46.082	48.902	0.580	79.45	84.31
Steering kernel regression (s)	1402.727	887.629	15.091	92.95	58.82
Total time (s)	1865.769	1045.293	20.718	90.06	50.45

Table 3: The RMSE and MSSIM values and running time comparison of fetal brain reconstruction with different window sizes ranging from $3 \times 3 \times 3$ to $9 \times 9 \times 9$, respectively.

Window size w	$w = 3$	$w = 5$	$w = 7$	$w = 9$
RMSE	125.061	125.869	129.298	126.929
MSSIM	0.6796	0.6663	0.6606	0.6763
TIME (s)	$7.056 = (4.810 + 0.534 + 1.712)$	$10.268 = (4.822 + 3.714 + 1.732)$	$124.678 = (4.801 + 118.211^* + 1.666)$	$222.353 = (4.798 + 215.888^* + 1.667)$

Note: TIME denotes the time caused only by running the adaptive kernel regression method. $T = (A + B + C)$: A is the time to calculate the gradient information. B is the time to calculate the covariance smoothing matrix. C is the time to calculate steering kernel regression function. T is the sum of A, B, and C. $*$ indicates that CPU is chosen as the running processor for the covariance matrix calculation due to the limitation of GPU kernel memory for the large window size.



(a) (b) (c) (d)

Figure 7: Reconstructed results of the MRI data with different window sizes w: (a) $w = 3$, (b) $w = 5$, (c) $w = 7$, and (d) $w = 9$.

The volunteer's womb was scanned with single-shot fast spin echo (SSFSE) T2-weighted sequence. Three stacks of images from axial, coronal, and sagittal orientation are used to construct the final high-resolution volume. To obtain the sparse stacks, we randomly remove different proportions of pixels of the stack once every 10% proportion ranging from 10% to 90%. The different removal proportions control the removal number of pixels. The typical 30th slice of the collected stack and its corresponding simulated spare slices are illustrated in Figure 5.

For different data removal ratios, the sparse stacks are used to reconstruct the high-resolution 3D fetal brain MRI volume with the method of Kainz et al. [16] (SVR with super-resolution) and our proposed method. Figure 6 shows the reconstructed results by Kainz et al.'s method and the proposed method for the sparsely sampled dataset with once every 10% data removal ratio ranging from 0% to 90%, respectively. In Figure 6, we can observe that as the removal ratio increases, the reconstructed results by Kainz et al. method have much more noise for the sparse sampled dataset compared with our proposed method. On the other hand, the proposed method is capable of reconstructing high-resolution images without obvious artifacts even for the 90% data removal ratio.

Table 4: The RMSE and MSSIM value comparison of fetal brain reconstruction with different structure sensitivities α ranging from 0.1 to 0.5, respectively.

Structure sensitivity	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
RMSE	125.06	112.10	99.927	91.073	97.556
MSSIM	0.6823	0.7117	0.7584	0.7712	0.7523

For the sake of quantitative evaluation, the image quality assessment index of root mean square error (RMSE) [9] and mean structure similarity (MSSIM) [34] is introduced to quantitatively assess the algorithms under different removal ratios. The RMSE score can be computed by the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(z(i) - g(i)\right)^2}, \tag{24}$$

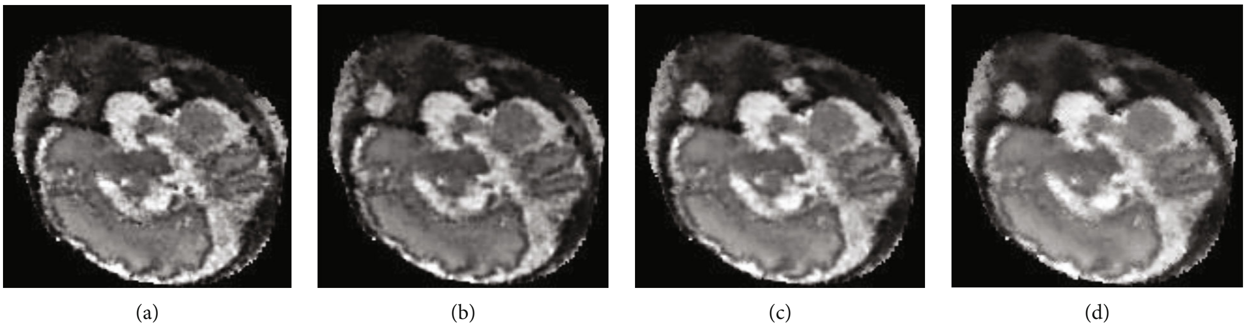


Figure 8: Reconstructed results of the MRI data with different structure sensitivities α: (a) $\alpha = 0.2$, (b) $\alpha = 0.3$, (c) $\alpha = 0.4$, and (d) $\alpha = 0.5$.

where $z(\bullet)$ is the reconstructed result, $g(\bullet)$ is the ground-truth volume, and N is the number of voxels. A good reconstruction method is capable of estimating the removal data very close to the original data. Given $z(\bullet)$ and $g(\bullet)$, a low RMSE value represents that the estimated result is satisfying while a high RMSE means that the interpolation accuracy is poor.

The structure similarity (SSIM) index explores the structural information for image quality assessment based on the main idea that the pixels have strong interdependency when they are spatially close. The SSIM metric is calculated based on the intensity, contrast, and structure and is computed as

$$\mathrm{SSIM}(z, g) = \frac{\left(2\mu_z\mu_g + c_1\right)\left(2\sigma_{zg} + c_2\right)}{\left(\mu_z^2 + \mu_g^2 + c_1\right)\left(\sigma_z^2 + \sigma_g^2 + c_2\right)}, \tag{25}$$

where μ_z, μ_g, σ_z, σ_g, and σ_{zg} denote the mean, variance, and covariance on square window, which moves pixel by pixel in images $z(i)$ and $g(i)$, respectively. The two variables $c_1 = k_1 L$ and $c_2 = k_2 L$ are used to stabilize the division with weak denominator. Here, L is the dynamic range of pixel value (e.g., 255 for 8-bit grayscale image), with $k_1 = 0.01$ and $k_1 = 0.03$ by default. Since the SSIM metric is calculated on various windows of a volume image, the mean SSIM (MSSIM) index is used in this experiment to assess the overall image quality:

$$\mathrm{MSSIM}(z, g) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{SSIM}(z_i, g_i), \tag{26}$$

where M is the number of local windows in the image. $\mathrm{MSSIM}(z, g) \in [0, 1]$; the higher MSSIM indicates better structural similarity between two images.

For the clinical datasets, it is impractical to obtain the ground-truth volume in advance. For the sake of fair comparison among different methods, the quantitative evaluation is performed based on an average reconstructed volume. We first use the original stacks without data removal to reconstruct a complete volume by Kainz et al.'s method (2015) and our method (e.g., Figures 6(a) and 6(b)), respectively. Both volumes are adopted to create an average volume as the ground truth. Table 1 shows the quantitative results of the RMSE and MSSIM values with different data removal ratios for each method. As can be seen, the results of Kainz et al.'s method produce the highest RMSE scores and lowest scores for all sampling rates. Both the high RMSE value and low MSSIM value for Kainz et al.'s method indicate poor image quality because of the artifacts and noise. For all levels of sampling rate, the proposed method performs better than the Kainz et al.'s method. More importantly, both of the difference of the RMSE and MSSIM index between Kainz et al.'s method and our method increase while the data removal ratio increases, indicating that our method outperforms much more compared with the Kainz et al.'s method when the data removal ratio increases.

Table 5: The RMSE and MSSIM value comparison of fetal brain reconstruction with the regularization parameter λ ranging from 0.1 to 2.0, respectively.

Regularization parameter λ	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$	$\lambda = 2.0$
RMSE	91.073	91.447	91.276	91.566	91.071
MSSIM	0.7720	0.7710	0.7708	0.7684	0.7732

4.2. Evaluation of Computational Efficiency. Our approach is capable of reconstructing the accurate volume from the highly sparse sampling dataset, but it requires largely computational burden as well due to the iterative kernel regression estimation. To reduce the long processing time of the adaptive kernel regression, the proposed method is accelerated by the GPU-based parallel implementation based on the NVIDIA GeForce GTX 1080 and CUDA 8.0 libraries. In the experiment, we make the evaluation of the computational efficiency of the adaptive kernel regression method, including the computation of the gradient information, the covariance smoothing matrix, and the steering kernel regression. The computational efficiency of the other modules (i.e., motion estimation, stack-to-template registration, PSF-based volume update, robust outlier removal, and slice-to-volume registration) has been evaluated in detail in [16]. The comparisons are based on the single-threaded CPU, multithreaded CPU, and GPU for the dataset of 80% data removal ratio under the parameter setting as the kernel size $k_c = 5$ and the smoothing parameter $h_c = 2.0$ in the initial gradient estimation step based on the classical kernel regression, the steering kernel size $k_s = 7$ and the steering smoothing parameter h_s

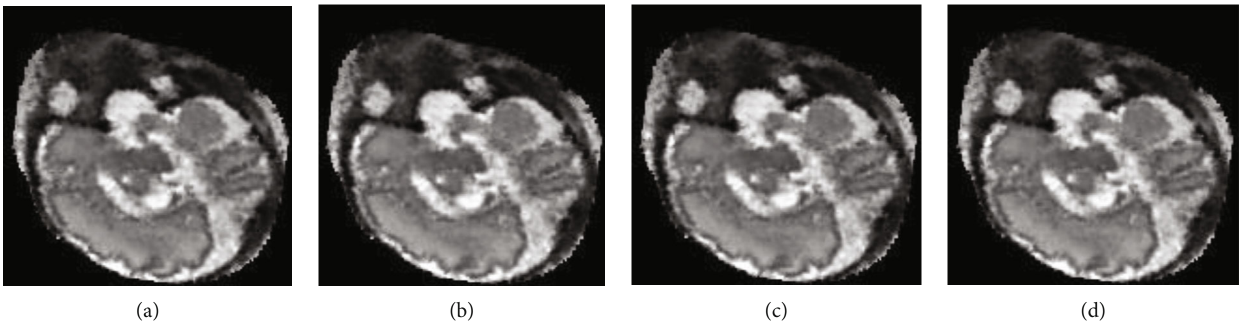


Figure 9: Reconstructed results of the MRI data with different regularization parameters λ: (a) $\lambda = 0.5$, (b) $\lambda = 1.0$, (c) $\lambda = 1.5$, and (d) $\lambda = 2.0$.

Table 6: The RMSE and MSSIM values and running time comparison of fetal brain reconstruction with different steering kernel sizes k_s ranging from $3 \times 3 \times 3$ to $9 \times 9 \times 9$, respectively.

Steering kernel size	$k_s = 3$	$k_s = 5$	$k_s = 7$	$k_s = 9$
RMSE	91.07	79.554	79.005	81.502
MSSIM	0.7791	0.7973	0.8054	0.7781
TIME (s)	7.907 = (5.407 + 0.581 + 1.919)	11.995 = (5.404 + 0.581 + 6.010)	21.148 = (5.402 + 0.581 + 15.165)	37.333 = (5.408 + 0.581 + 31.344)

Note: TIME denotes the time caused only by running the adaptive kernel regression method. $T = (A + B + C)$: A is the time to calculate the gradient information. B is the time to calculate the covariance smoothing matrix. C is the time to calculate the steering kernel regression function. T is the sum of A, B, and C.

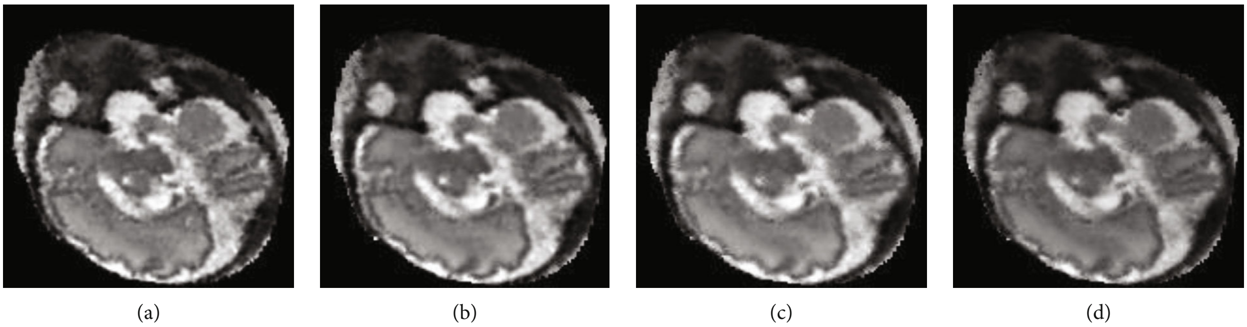


Figure 10: Reconstructed results of the MRI data with different kernel window sizes: (a) $k_s = 3$, (b) $k_s = 5$, (c) $k_s = 7$, and (d) $k_s = 9$.

Table 7: The RMSE and MSSIM value comparison of fetal brain reconstruction with different steering smoothing parameters h_s ranged from 0.1 to 2.5, respectively.

Steering smoothing parameter	$h_s = 0.1$	$h_s = 0.5$	$h_s = 1.0$	$h_s = 1.5$	$h_s = 2.0$
RMSE	79.005	78.886	79.230	79.159	82.87
MSSIM	0.7927	0.8092	0.7933	0.7944	0.7946

= 0.5 in the steering kernel regression step, and the window size $w = 3$, the regularization parameter $\lambda = 2.0$, and the structure sensitivity $\alpha = 0.4$. The practical running time for the proposed method is shown in Table 2. For single-threaded CPU, the running time of the adaptive kernel regression method is 1865.769 s, which includes the computation of the gradient information in 416.96 s, the covariance smoothing matrix in 46.082 s, and the steering kernel regression in 1402.727 s. For the multithreaded CPU, we use 4 threads to run the adaptive kernel regression method and its computational time is 1045.293 s, indicating less improvement compared with the single-threaded CPUs. The running time of the GPU implementation is 20.718 s in total. From Table 2, it can be observed that the GPU-based processing time has significantly decreased by 98.89% and 98.02%, compared with the single-threaded CPU and the multithreaded CPU, respectively.

4.3. The Choice of the Adaptive Kernel Regression Parameters.

There are seven parameters which can be adjusted to affect the reconstructed image quality for the proposed method. These parameters include the kernel size k_c and the smoothing parameter (i.e., the kernel bandwidth) h_c in the initial gradient estimation step based on the classical kernel regression, the steering kernel size k_s and the steering smoothing parameter h_s in the steering kernel regression step, and the window size w, the regularization parameter λ, and the structure sensitivity α ($0 \leq \alpha \leq 0.5$) in the covariance matrix estimation step. In our method, k_c and h_c are related with the initial calculation of gradient information and have a negligible effect in the experiment. For the adaptive sparse

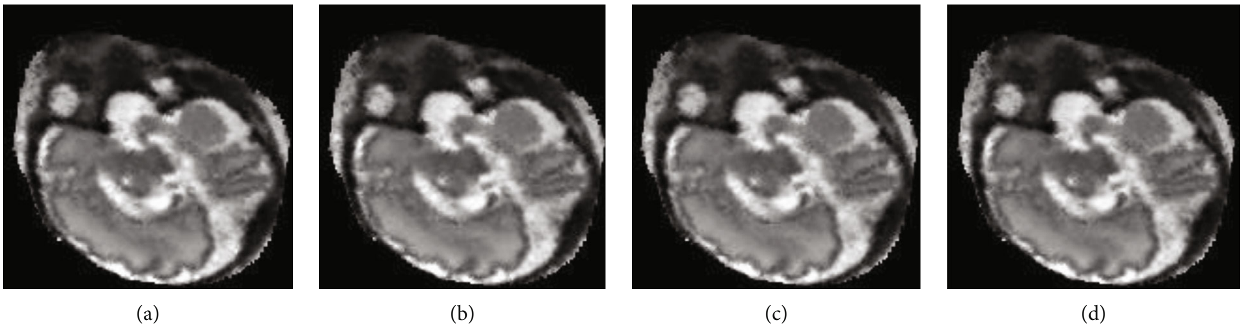


Figure 11: Reconstructed results of the MRI data with different steering smoothing parameters: (a) $h_s = 0.5$, (b) $h_s = 1.0$, (c) $h_s = 1.5$, and (d) $h_s = 2.0$.

reconstruction, covariance matrix estimation and steering kernel estimation are the two of the important steps and their parameters (i.e., w, α, λ, k_s, and h_s) play an important role in the volume reconstruction and deserve much more investigation.

With the help of GPU-based fast implementation, we firstly adjust the parameters (i.e., w, α, and λ) of the covariance matrix estimation one by one. The window size w decides how many neighbor points in the gradient matrix are taken for the estimation of the covariance matrix. Table 3 shows the RMSE and MSSIM values and the running time for different window sizes. Both of the RMSE and MSSIM values differ slightly, indicating that the window size has a negligible influence on the reconstructed image quality, as shown in Figure 7. However, the running time increases with the increase of window size. It can be observed that the window size of $w = 3$ is chosen because of its faster implementation and lower RMSE value.

Table 4 shows the RMSE and MSSIM values influenced by the structure sensitivity parameter α, and the lowest RMSE value and highest MSSIM value are obtained for the structure sensitivity $\alpha = 0.4$ indicating the best performance of the algorithm. Figure 8 shows the corresponding reconstructed images for different α values. As can be seen, the result with large structure sensitivity (e.g., $\alpha = 0.5$) results in oversmoothing image, while small structure sensitivity (e.g., $\alpha = 0.2$) overemphasizes the image edges. The experiment shows that the structure sensitivity α has a significant influence on the reconstructed volume.

Under different regularization parameter settings, the RMSE and MSSIM measurements of the reconstructed results are calculated and shown in Table 5. The illustrative results are further shown in Figure 9. The regularization parameter λ is used to suppress the noise. However, the regularization parameter has negligible influence on the reconstructed image quality in the experiments.

The next group parameters (i.e., k_s and h_s) come from the steering kernel regression for the adaptive voxel value estimation. The kernel window size k_s has a great impact on the processing time for the kernel regression-based algorithm under different data removal proportions. When the kernel window increases, the estimation of each voxel involves more nearby pixels and leads to more computation [32]. The smaller the kernel window size is, the faster our algorithm runs. On the other hand, if the size of the kernel window is too small, we could obtain the fault result, because there are not enough samples to make the current voxel estimation, especially for large data removal proportion. The larger the data removal proportion is, the sparser the sampled data will be. The RMSE and MSSIM index and processing time measurement of the reconstructed results under different kernel window sizes are shown in Table 6. With the increase of the kernel window size, the running time of steering kernel regression function is becoming longer. The corresponding images of different steering kernel sizes are shown in Figure 10. The proper kernel window size (i.e., $7 \times 7 \times 7$) produces a trade-off between the processing time and the reconstruction accuracy under different removal proportions.

Table 7 shows the RMSE and MSSIM values and running time with different steering smoothing parameters h_s. As can be seen, the results with the steering smoothing parameter (i.e., $h_s = 0.5$) achieve the lowest RMSE value and highest MSSIM value among these settings. The reconstructed results produced by different steering smoothing parameters are shown in Figure 11. In [33], it has been given that the steering smoothing parameter indicates the "footprint" of the kernel function. The large footprint of the kernel function could reduce the noise but at the cost of oversmoothing details, while small footprints are desirable to preserve the edges. In the experiment, the footprint setting $h_s = 0.5$ is chosen for reaching a trade-off between the noise reduction and edge preservation. Finally, all parameters of the adaptive steering kernel regression algorithm are determined as follows: the window size $w = 3$, the regularization parameter $\lambda = 2.0$, the structure sensitivity $\alpha = 0.4$, the steering kernel size $k_s = 7$, and the steering smoothing parameter $h_s = 0.5$. Under such parameter setting, the RMSE value decreases from 126.47 to 78.89, indicating the quality improvement by 37.62%.

5. Conclusion

In this paper, we proposed an adaptive reconstruction method to deal with the sparse sampling dataset for fetal brain MRI. Our method combines the latest SVR framework, including the stack motion evaluation, PSF-based volume update, robust outlier removal, slice-to-volume registration, and the proposed adaptive kernel regression-based volume update. Compared with the existing SVR framework, our method has advantages of sparse volume reconstruction and is capable of reconstructing superresolution image even for 80%~90% data removal. With the capability of sparse reconstruction, the data sampling time can be greatly shortened and thus, the motion artifacts can be reduced indirectly. To accelerate the voxel estimation, we use the CUDA to implement the steering kernel regression approach. For the proposed method, the running times of GPU-based implementation are speeded up to 90x than that of the CPU. The GPU-based parallel implementation of the proposed method is more practical to meet the requirements of fetal brain MRI. Meanwhile, we make the detailed investigation on the choice of parameters for the adaptive kernel regression-based volume reconstruction with the help of GPU-based fast implementation. To summarize, the structure sensitivity α and the steering kernel window size k_s are two of the important parameters on sparse kernel regression volume reconstruction. Meanwhile, the kernel window size has a strong relationship with the running time. Larger window size requires longer processing time. Overall, our approach is used to reconstruct superresolution image from the highly sparse sampled dataset of fetal brain MRI corrupted with motion artifacts. One of its potential applications includes other motion organ MRI reconstruction, such as the heart MRI with the heart beating motion artifacts.

Data Availability

The test data was downloaded from the publicly available dataset on GitHub (https://github.com/bkainz/fetalReconstruction.git).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qian Ni and Yi Zhang contributed equally to this work and should be considered co-first authors.

Acknowledgments

This study was financed partially by the National Key R&D Program of China (No. 2018YFA0704102), the National Natural Science Foundation of China (Nos. 81827805, 61401451), and the Shenzhen Key R&D Program (No. JCYJ20200109114812361).

References

[1] O. A. Glenn, “Normal development of the fetal brain by MRI,” *Seminars in Perinatology*, vol. 33, no. 4, pp. 208–219, 2009.

[2] D. Bulas and A. Egloff, “Benefits and risks of MRI in pregnancy,” *Seminars in Perinatology*, vol. 37, no. 5, pp. 301–304, 2013.

[3] S. C. O'Connor, V. J. Rooks, and A. B. Smith, “Magnetic resonance imaging of the fetal central nervous system, head, neck, and chest,” *Seminars in Ultrasound, CT and MRI*, vol. 33, no. 1, pp. 86–101, 2012.

[4] L. M. Tee, E. Y. Kan, J. C. Cheung, and W. Leung, “Magnetic resonance imaging of the fetal brain,” *Hong Kong Medical Journal*, vol. 22, no. 3, pp. 270–278, 2016.

[5] C. Limperopoulos and C. Clouchoux, “Advancing fetal brain MRI: targets for the future,” *Seminars in Perinatology*, vol. 33, no. 4, pp. 289–298, 2009.

[6] M. A. Rutherford, “Magnetic resonance imaging of the fetal brain,” *Current Opinion in Obstetrics & Gynecology*, vol. 21, no. 2, pp. 180–186, 2009.

[7] D. Prayer, G. Kasprian, E. Krampl et al., “MRI of normal fetal brain development,” *European Journal of Radiology*, vol. 57, no. 2, pp. 199–216, 2006.

[8] N. Girard, C. Raybaud, D. Gambarelli, and D. Figarella-Branger, “Fetal brain MR imaging,” *Magnetic Resonance Imaging Clinics of North America*, vol. 9, no. 1, pp. 19–56, vii, 2001.

[9] V. Merzoug, S. Ferey, C. Andre, A. Gelot, and C. Adamsbaum, “Magnetic resonance imaging of the fetal brain,” *Journal of Neuroradiology*, vol. 29, no. 2, pp. 76–90, 2002.

[10] S. N. Saleem, “Fetal magnetic resonance imaging (MRI): a tool for a better understanding of normal and abnormal brain development,” *Journal of Child Neurology*, vol. 28, no. 7, pp. 890–908, 2013.

[11] C. Malamateniou, S. J. Malik, J. M. Allsop et al., “Motion-compensation techniques in neonatal and fetal MR imaging,” *American Journal of Neuroradiology*, vol. 34, no. 6, pp. 1124–1136, 2013.

[12] D. Prayer, P. C. Brugger, and L. Prayer, “Fetal MRI: techniques and protocols,” *Pediatric Radiology*, vol. 34, no. 9, pp. 685–693, 2004.

[13] F. Rousseau, O. Glenn, B. Iordanova et al., “A novel approach to high resolution fetal brain MR imaging,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 548–555, Springer-Verlag, 2005.

[14] S. Z. Jiang, H. Xue, A. Glover, M. Rutherford, D. Rueckert, and J. V. Hajnal, “MRI of moving subjects using multislice snapshot images with volume reconstruction (SVR): application to fetal, neonatal, and adult brain studies,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 7, pp. 967–980, 2007.

[15] K. Kim, P. Habas, F. Rousseau, O. Glenn, A. Barkovich, and C. Studholme, “Intersection based motion correction of multislice MRI for 3-D in utero fetal brain image formation,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 146–158, 2010.

[16] B. Kainz, M. Steinberger, W. Wein et al., “Fast volume reconstruction from motion corrupted stacks of 2D slices,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 9, pp. 1901–1913, 2015.

[17] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, “RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.

[18] F. Rousseau, A. Glenn, B. Iordanova et al., “Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images,” *Academic Radiology*, vol. 13, no. 9, pp. 1072–1081, 2006.

[19] A. Bertelsen, P. Aljabar, H. Xue et al., “Improved slice to volume reconstruction of the fetal brain for automated cortex segmentation,” in *Proceedings of the International Society for*

Magnetic Resonance in Medicine, p. 3437, Honolulu, Hawai'i, 2009.

[20] H. Greenspan, *Super-resolution in medical imaging*, vol. 52, no. 1, 2009Oxford University Press, 2009.

[21] P. Milanfar, *Super-Resolution Imaging*, CRC Press, 2007.

[22] A. Gholipour and S. K. Warfield, "Super-resolution reconstruction of fetal brain MRI," in *MICCAI Workshop on Image Analysis for the Developing Brain(IADB)*, London, UK, 2009.

[23] A. Gholipour, J. A. Estroff, and S. K. Warfield, "Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain MRI," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1739–1758, 2010.

[24] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, Wiley, 2nd edition, 2009.

[25] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, 1997.

[26] G. Peyre, "A review of adaptive image representations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 896–911, 2011.

[27] F. Rousseau, K. Kim, C. Studholme, M. Koob, and J.-L. Dietemann, "On super-resolution for fetal brain MRI," in *Medical Image Computing and Computer Assisted Intervention*, pp. 355–362, Springer-Verlag, 2010.

[28] M. Kuklisova-murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, "Reconstruction of fetal brain MRI with intensity matching and complete outlier removal," *Medical Image Analysis*, vol. 16, no. 8, pp. 1550–1564, 2012.

[29] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, 2007.

[30] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of MR images of the brain," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 885–896, 1999.

[31] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[32] T. X. Wen, L. Li, Q. S. Zhu et al., "GPU-accelerated kernel regression reconstruction for freehand 3D ultrasound imaging," *Ultrasonic Imaging*, vol. 39, no. 4, pp. 240–259, 2017.

[33] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 1958–1975, 2009.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

Hindawi
BioMed Research International
Volume 2021, Article ID 8840835, 9 pages
https://doi.org/10.1155/2021/8840835

Research Article

An Interpretable Model-Based Prediction of Severity and Crucial Factors in Patients with COVID-19

Bowen Zheng,[1] Yong Cai,[2] Fengxia Zeng,[1] Min Lin,[3] Jun Zheng,[3] Weiguo Chen,[1] Genggeng Qin,[1] and Yi Guo[4]

[1] *Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong 510515, China*
[2] *Department of CT, Maoming People's Hospital, Maoming, Guangdong 525000, China*
[3] *Department of Radiology, Honghu People's Hospital, Honghu, Hubei 433220, China*
[4] *Department of Medical Services Section, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong 510515, China*

Correspondence should be addressed to Genggeng Qin; zealotq@smu.edu.cn and Yi Guo; 52415229@qq.com

Received 16 September 2020; Revised 1 November 2020; Accepted 4 February 2021; Published 4 March 2021

Academic Editor: Shaode Yu

This study established an interpretable machine learning model to predict the severity of coronavirus disease 2019 (COVID-19) and output the most crucial deterioration factors. Clinical information, laboratory tests, and chest computed tomography (CT) scans at admission were collected. Two experienced radiologists reviewed the scans for the patterns, distribution, and CT scores of lung abnormalities. Six machine learning models were established to predict the severity of COVID-19. After parameter tuning and performance comparison, the optimal model was explained using Shapley Additive explanations to output the crucial factors. This study enrolled and classified 198 patients into mild ($n = 162$; 46.93 ± 14.49 years old) and severe ($n = 36$; 60.97 ± 15.91 years old) groups. The severe group had a higher temperature ($37.42 \pm 0.99^{\circ}$C vs. $36.75 \pm 0.66^{\circ}$C), CT score at admission, neutrophil count, and neutrophil-to-lymphocyte ratio than the mild group. The XGBoost model ranked first among all models, with an AUC, sensitivity, and specificity of 0.924, 90.91%, and 97.96%, respectively. The early stage of chest CT, total CT score of the percentage of lung involvement, and age were the top three contributors to the prediction of the deterioration of XGBoost. A higher total score on chest CT had a more significant impact on the prediction. In conclusion, the XGBoost model to predict the severity of COVID-19 achieved excellent performance and output the essential factors in the deterioration process, which may help with early clinical intervention, improve prognosis, and reduce mortality.

1. Introduction

Coronavirus disease 2019 (COVID-19), pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a highly infectious respiratory disease with a variable incubation period ranging from 1 to 14 days, and people are generally vulnerable to the virus.

Reverse transcription-polymerase chain reaction (RT-PCR) for SARS-CoV-2 is the standard for diagnosing COVID-19. However, RT-PCR takes 1–2 days to complete and may report false-negative results. Some areas even faced a shortage of RT-PCR testing kits [1, 2]. Under these circumstances, chest computed tomography (CT) played a vital role in detecting and assessing patients with COVID-19, especially in detecting patients with COVID-19 in the early stage [3].

According to clinical presentation, patients with COVID-19 were classified into four categories: mild type, moderate type, severe type, and critical type [4]. Most patients were classified as the mild type and moderate type with mild symptoms, whereas a small group of patients may experience acute respiratory distress syndrome (ARDS), septic shock, coagulation dysfunction, and multiple organ failure. These patients required ventilators and extracorporeal membrane oxygenation during an expensive treatment and had a high death rate [5]. Previous researchers showed that up to 5.0% of the patients were admitted to the intensive care unit (ICU), 2.3% of the patients needed invasive mechanical ventilation, and 1.4% of patients died eventually [6]. It is unclear why some patients develop into severe or critical cases, while others only get mild or no symptoms.

The crucial factors in the deterioration process remain unknown.

Early identification of severity and crucial factors are of great value, prompting early clinical intervention and preventing deterioration of patients' condition. However, it is hard for the doctor to identify those patients under the human limitation on information processing. Hence, artificial intelligence has been widely applied in the medical domain, enabling radiologists to make full use of data, including imaging information, and explore the images' biological nature. Since the initial outbreak, attempts have been made to detect COVID-19 using chest CT.

In this study, we established a machine learning model, combining clinical information, laboratory tests, and chest CT features for early prediction of the severity and crucial factors of patients with COVID-19. Our model may help identify patients who require early clinical intervention to improve prognosis and reduce mortality.

2. Materials and Methods

2.1. Study Participants. This retrospective study evaluated de-identified data and involved no potential risk to the patients. Therefore, the institutional review board waived the requirement of obtaining written informed consent. This study included patients with COVID-19, as confirmed by RT-PCR, admitted to the People's Hospital of Honghu and Honghu Xiaotangshan Hospital from January 1 to March 27, 2020. The inclusion criteria were as follows: (a) a positive RT-PCR result for SARS-CoV-2 infection, (b) patients who underwent a chest CT scan and laboratory tests at admission in the two hospitals mentioned above, and (c) no other viral infection or serious complication. The exclusion criteria were as follows: (a) patients who underwent a chest CT scan and laboratory tests in other hospitals and (b) patients whose chest CT images showed no lesion in the lungs.

Patients' triage, sex, age, symptoms, pre-existing diseases, the temperature at admission, and laboratory tests, such as white blood cell (WBC), neutrophil, and lymphocyte counts, were collected. Patients with COVID-19 were classified into four categories [4]: (1) The mild type includes those who have mild clinical symptoms and no pneumonia manifestations found in imaging. (2) The moderate type includes the patients who have symptoms such as fever and respiratory tract symptoms with pneumonia manifestations seen on imaging. (3) The severe type fulfilled the following criteria: respiratory frequency ≥ 30/minute, blood oxygen saturation $\leq 93\%$, arterial partial pressure of oxygen (PaO_2)/oxygen concentration (FiO_2) ratio < 300, and lung infiltrates $> 50\%$ within 24–48 hours. (4) The critical type meets any of the following criteria: occurrence of respiratory failure requiring mechanical ventilation and the presence of shock and other organ failures that require monitoring and treatment in the ICU.

In this study, all patients were classified into four clinical types according to the criteria mentioned above during treatment. The mild type was excluded because of no pneumonia manifestations found in imaging. The moderate type was classified into the mild group. Concerning the rareness of the critical type, the severe type and critical type were classified into the severe group in this study (Figure 1).

2.2. Imaging Techniques. Chest CT scanning (Go Now, Siemens Healthcare, Germany; GE optima 680, GE Healthcare, USA) was performed at the end of full inspiration in the supine position. The images were acquired and reconstructed with 80–130 kV tube voltage and automatic tube current modulation (up to 400 mA). The slice thicknesses were 0.6 mm (GE optima CT680) and 1.5 mm (Go Now), respectively. The lung window setting was at a window level of -600 Hounsfield units (HU) and a window width of 1500 HU. The scanning range was from the apex to the lung base.

2.3. Image Interpretation. All chest CT images were reviewed by two radiologists with over five years of clinical experience in the respiratory system independently. Any disagreement was resolved by discussion and consensus. The following aspects were reviewed for each patient: (1) stage (early stage, progress stage, or restoration stage); (2) distribution (subpleural, scatter, or diffuse) and shape (nodular, patchy, or large patchy); (3) number of lung lobes involved; (4) presence of ground-glass opacity (GGO); (5) presence of consolidation, fibrotic lesions, reticular shadow, crazy paving pattern, air bronchogram, pleural effusion, pleural thickening, and mediastinal lymphadenopathy (axil diameter > 10 mm); and (6) CT scores of the percentage of lung involved [7, 8]. Each lobe was evaluated for the percentage involved on a scale of 0–4 (0: 0% involvement, 1: <25% involvement, 2: 25%–50% involvement, 3: 50%–75% involvement, and 4: ≥75% involvement). The total score on the chest CT was the summation of all five lobes. The maximum possible score was 20.

2.4. Statistical Analysis. Statistical analyses were performed using SPSS (version 26.0). Continuous variables are expressed as means and standard deviations and compared by an independent-sample t-test; categorical variables are expressed as counts and frequencies (%) and compared using Fisher's exact test between the mild and severe groups. Statistical significance was set at $p < 0.05$. The area under the curve (AUC) of different models was compared by the DeLong test using MedCalc (version 19.4.1).

2.5. Interpretable Machine Learning Model Building. A dataset was built, including clinical information, laboratory tests, and chest CT features, from 198 patients with COVID-19, as confirmed by RT-PCR. The machine learning model was established using Python 3.7. We randomly split the dataset into a 70% training and validation set and a 30% test set. All quantitative features were normalized to the range of 0 to 1. The categorical features were transformed into a one-hot numerical array. Six machine learning models, including logistic regression (LR), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), support vector machine (SVM), and eXtreme gradient boosting (XGBoost), were built based on the features after preprocessing. After parameter tuning, the model's performance was assessed using the AUC. The receiver operating characteristic (ROC) curve of

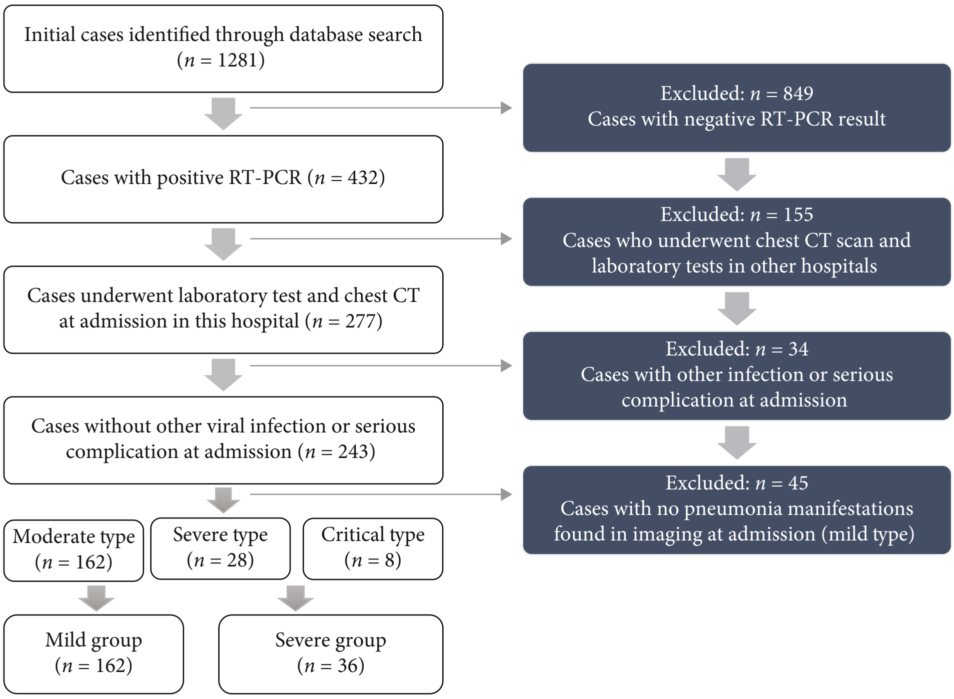

Initial cases identified through database search (n = 1281)
Excluded: n = 849
Cases with negative RT-PCR result
Cases with positive RT-PCR (n = 432)
Excluded: n = 155
Cases who underwent chest CT scan and laboratory tests in other hospitals
Cases underwent laboratory test and chest CT at admission in this hospital (n = 277)
Excluded: n = 34
Cases with other infection or serious complication at admission
Cases without other viral infection or serious complication at admission (n = 243)
Excluded: n = 45
Cases with no pneumonia manifestations found in imaging at admission (mild type)
Moderate type (n = 162)
Severe type (n = 28)
Critical type (n = 8)
Mild group (n = 162)
Severe group (n = 36)

FIGURE 1: Flow diagram of patient enrollment.

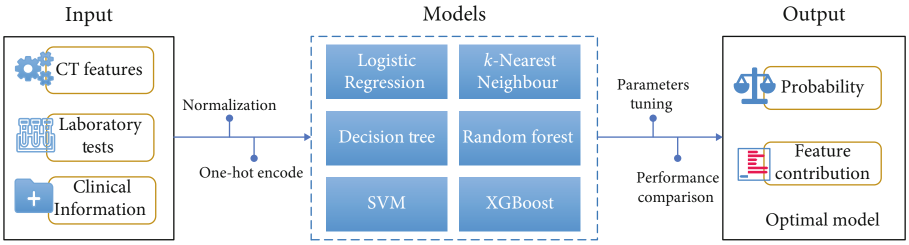


FIGURE 2: Illustration of the modeling framework.

each model was further evaluated using DeLong's test on MedCalc (Figure 2).

Based on Shapley values from coalitional game theory, Shapley Additive explanations (SHAP) were used to explain the model [9, 10]. The SHAP explains the model prediction by computing each feature's contribution individually or jointly to the prediction. With kernelSHAP, treeSHAP, and deepKernal subclasses, SHAP can explain any machine learning model's output.

3. Results

3.1. Statistical Analysis. This study enrolled 198 patients (mild group: 162 cases and severe group: 36 cases), including 80 males and 118 females. The average age of the mild (46.93 ± 14.49 years) and severe (60.97 ± 15.91 years) groups was significantly different. Patients in the mild group were admitted to the hospital 10.40 ± 5.58 days after the onset, which is longer than that in the severe group (8.00 ± 4.88 days, $p = 0.038$). However, the temperature of patients in the severe group was higher than that of those in the mild group (37.42 ± 0.99°C vs. 36.75 ± 0.66°C). Fever, cough, shortness of breath, and dyspnea were significant features associated with the severe group. In terms of basic diseases, 22.22% (8/36) and 6.79% (11/162) of patients in the severe and mild groups, respectively, had high blood pressure ($p = 0.008$) (Table 1).

There were 9.35 ± 7.44 and 6.44 ± 4.08 days between the first CT scan and onset of chest CT features in the mild and the severe groups, respectively. However, the total CT score and the number of different lobes involved in the severe group were significantly higher than those in the mild group. Patients with diffuse (23/36, 63.89%) and large patchy (18/36, 50.00%) appearances were likely to deteriorate. In contrast, patients with diffuse location and patchy shape of the mild group were 35.80% and 81.48%, respectively. Moreover, 80.6% of severe group patients showed lung lesions that had invaded five lobes at admission, compared to 39.5% of

Table 1: Demographic, clinical characteristics, and laboratory tests of the patients.

	Mild group ($n = 162$)	Severe group ($n = 36$)	p
Age (years)			
Mean (SD)	46.93 ± 14.49	60.97 ± 15.91	<0.001
Range	17-81	28-86	
Median age	46	64.50	
Gender			0.513
Male	67 (41.36%)	13 (36.11%)	
Female	95 (58.64%)	23 (63.89%)	
Signs and symptoms at admission			
Days from onset (days)	10.40 ± 5.58	8 ± 4.88	0.038
Temperature (°C)	36.75 ± 0.66	37.42 ± 0.99	<0.001
Fever*	119 (73.46%)	27 (75.00%)	<0.001
Cough*	96 (59.29%)	23 (63.89%)	<0.001
Fatigue	34 (20.99%)	10 (27.78%)	0.352
Shortness of breath*	16 (9.88%)	7 (19.44%)	0.140
Chest tightness*	13 (8.02%)	3 (8.33%)	1
Dyspnea*	6 (3.70%)	7 (19.44%)	0.002
Fear of cold*	6 (3.70%)	5 (13.89%)	0.026
Diarrhea*	8 (4.94%)	1 (2.78%)	1
Headache*	8 (4.94%)	1 (2.78%)	1
Dizziness	4 (2.47%)	3 (8.33%)	0.105
Palpitation*	1 (0.62%)	3 (8.33%)	0.018
Preexisting disease			
Hypertension*	11 (6.79%)	8 (22.22%)	0.008
Diabetes*	6 (3.70%)	4 (11.11%)	0.077
CAD*	5 (3.09%)	2 (5.56%)	0.356
Lung cancer*	0	1 (2.78%)	0.176
Myocardial infarction*	0	1 (2.78%)	0.176
Cerebral infarction*	0	1 (2.78%)	0.176
Tuberculosis*	0	1 (2.78%)	0.176
Laboratory tests			
WBC ($\times 10^9$/L)	5.53 ± 2.30	7.11 ± 3.53	0.014
Neutrophil ($\times 10^9$/L)	3.61 ± 2.10	5.80 ± 3.50	0.001
Neutrophil ratio (%)	62.48 ± 13.15	76.15 ± 12.11	<0.001
Lymphocyte ($\times 10^9$/L)	1.40 ± 0.50	0.99 ± 0.47	<0.001
Lymphocyte ratio (%)	27.33 ± 10.07	16.80 ± 9.71	<0.001
NLR	3.04 ± 2.75	8.12 ± 9.69	0.004

CAD: coronary artery disease; WBC: white blood cell; NLR: neutrophil-to-lymphocyte ratio. *Fisher's exact test.

the mild patients ($p = 0.001$). The manifestations of pleural effusion, consolidation, crazy paving, and air bronchogram played an essential role in predicting COVID-19 deterioration, indicating that these patients were more likely to develop into severe and critically ill patients (Table 2).

As for laboratory tests, the severe group had a higher WBC count, neutrophil count, and neutrophil ratio and a lower lymphocyte count and lymphocyte ratio than the mild group. Furthermore, the neutrophil-to-lymphocyte ratio (NLR) in the severe group was significantly higher than that in the mild group (8.12 ± 9.69 vs. 3.04 ± 2.75) (Table 1).

3.2. Machine Learning Model Performance and Interpretability. A dataset was built, including enrolled patients' clinical information, laboratory tests, and chest CT features. We randomly split the dataset into a 70% training

Table 2: Chest CT features of the patients.

	Mild group ($n = 162$)	Severe group ($n = 36$)	p
Stage			0.208
Early stage	44 (27.16%)	10 (27.78%)	
Progress stage	105 (64.81%)	26 (72.22%)	
Restoration stage	13 (8.02%)	0	
Location			0.002
Subpleural	50 (30.86%)	2 (5.56%)	
Scatter	54 (33.33%)	11 (30.56%)	
Diffuse	58 (35.80%)	23 (63.89%)	
Shape			<0.001
Nodular	11 (6.79%)	1 (2.78%)	
Patchy	132 (81.48%)	17 (47.22%)	
Large patchy	19 (11.73%)	18 (50.00%)	
Number of lobes involved			<0.001
1	22 (13.58%)	0	
2	22 (13.58%)	1 (2.78%)	
3	20 (12.35%)	1 (2.78%)	
4	34 (20.99%)	5 (13.89%)	
5	64 (39.51%)	29 (80.56%)	
Image manifestations			
Pleural effusion	1 (0.62%)	4 (11.11%)	0.004
Fibrosis	64 (39.51%)	15 (41.67%)	0.811
Consolidation	85 (52.47%)	28 (77.78%)	0.006
Reticular shadow	95 (58.64%)	34 (94.44%)	<0.001
Crazy paving	9 (5.56%)	15 (41.67%)	<0.001
Air bronchogram	55 (33.95%)	26 (72.22%)	<0.001
Pleural thickening	62 (38.27%)	24 (66.67%)	0.002
Lymphadenovarix	10 (6.17%)	4 (11.11%)	0.493
GGO	162 (100.00%)	36 (100.00%)	—
Nodules	68 (41.98%)	19 (52.78%)	0.211
Quantitative features			
CT from onset (days)	9.36 ± 7.44	6.44 ± 4.08	0.002
Total score	4.24 ± 2.54	8.50 ± 4.44	<0.001
UOR	0.75 ± 0.65	1.75 ± 1.23	<0.001
MOR	0.62 ± 0.66	1.36 ± 0.90	<0.001
IOR	1.10 ± 0.68	2.00 ± 1.20	<0.001
UOL	0.73 ± 0.59	1.53 ± 0.97	<0.001
IOL	1.04 ± 0.67	1.86 ± 1.13	<0.001

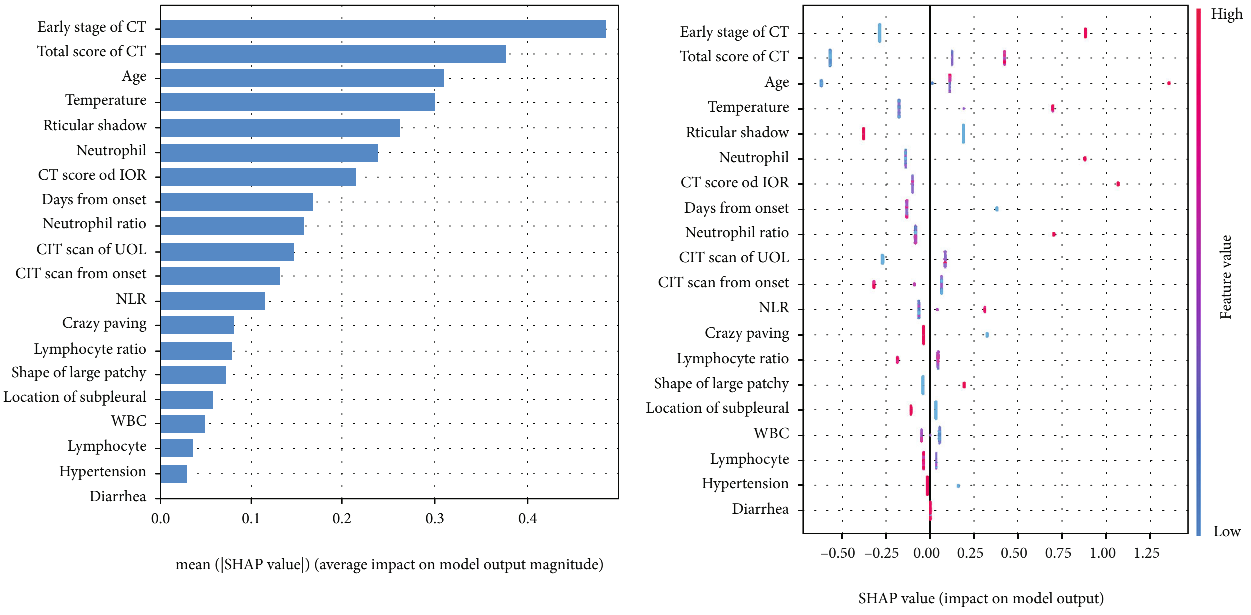
GGO: ground-glass opacity; UOR: upper lobe of right lung; MOR: middle lobe of right lung; IOR: inferior lobe of right lung; UOL: upper lobe of left lung; IOL: inferior lobe of left lung.

and validation set (138 cases, 113 in the mild group and 25 in the severe group) and a 30% test set (60 cases, 49 in the mild group and 11 in the severe group). Six machine learning models were built, validated, and tested based on the dataset. The performance of the models is reported in Table 3. Five of the six models showed a good fit, except for the DT model with an AUC of 0.707 (95% confidence interval (CI) (0.575, 0.817), $p = 0.0097$). The AUC of XGBoost ranked first for all models, with an AUC of 0.924 (95% CI (0.826, 0.976), $p < 0.0001$). XGBoost achieved 90.91% sensitivity (95% CI (58.7%, 99.8%)) and 97.96% specificity (95% CI (89.10%, 99.90%)). The RF model achieved a 0.907 AUC (95% CI (0.804, 0.967), $p < 0.0001$), 90.91% sensitivity (95% CI (58.7%, 99.8%)), and 95.92% specificity (95% CI (80.4%, 97.7%)). The KNN model obtained a 100% sensitivity (95% CI (71.5%, 100.00%)); however, KNN had a 0.857 AUC (95% CI (0.743, 0.934), $p < 0.0001$) and 61.22% specificity (95% CI (46.2%, 74.8%)). The difference in AUCs between

Table 3: The AUC, sensitivity, and specificity comparisons.

	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	p
LR	0.891 (0.783, 0.956)	90.91 (58.7, 99.8)	93.88 (83.1, 98.7)	0.1306
KNN	0.857 (0.743, 0.934)	100.00 (71.5, 100.0)	61.22 (46.2, 74.8)	0.2844
DT	0.707 (0.575, 0.817)	45.45 (16.7, 76.6)	95.92 (86.0, 99.5)	0.0095
RF	0.907 (0.804, 0.967)	90.91 (58.7, 99.8)	95.92 (86.0, 99.5)	0.1915
SVM	0.892 (0.785, 0.958)	90.91 (58.7, 99.8)	91.84 (80.4, 97.7)	0.2006
XGBoost	0.924 (0.826, 0.976)	90.91 (58.7, 99.8)	97.96 (89.1, 99.9)	—

Two-sided p values were calculated by comparing AUC for the XGBoost model with the other models. AUC comparisons were evaluated using the DeLong test; LR: logistic regression; KNN: k-nearest neighbor; DT: decision tree; RF: random forest; SVM: support vector machine; XGBoost: eXtreme gradient boosting.



(a)

(b)

Figure 3: The contribution of various features to the prediction model. The features are listed in descending order according to their contribution to the prediction of a patient becoming severe or critically ill. (a) The importance of features measured by the mean absolute Shapley values according to their contribution. (b) The combination of feature importance and feature effects. The color shows the value of the features from high to low. The horizontal location shows whether the effect of that value caused a higher or lower prediction. Each point is a Shapley value for a feature and an instance.

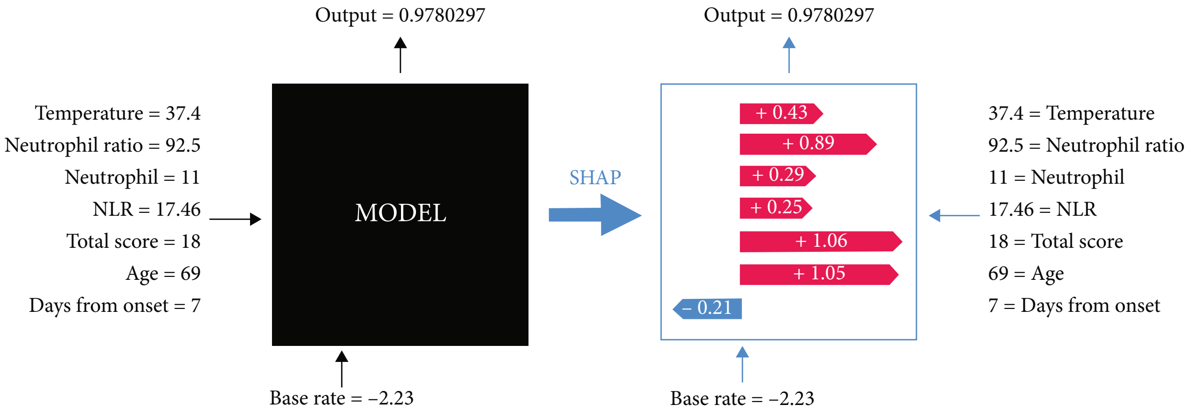


Figure 4: With the help of an interpretable module, we can know how the machine learning model concludes each individual. A 69-year-old patient was predicted to be deteriorating with a possibility of 0.978 (97.8%). The days from symptom onset to hospital admission was seven days, and the temperature at admission was 37.4°C. The neutrophil was 11×10^9/L, with a neutrophil ratio of 92.5% and an NLR of 17.46.

the XGBoost and RF models was not statistically significant ($p = 0.192$). The sensitivity of the two models remained the same; however, XGBoost had higher specificity. Although the AUC between XGBoost and LR, KNN, and RF showed no statistical difference, XGBoost acquired the highest Youden index, sensitivity, and specificity. In general, XGBoost was the best model in this dataset.

We further explored the interpretability of XGBoost using the TreeExplainer of SHAP [11]. Figure 3(a) shows the top 19 features that influenced the severe group prediction in descending order. The early stage of chest CT, total CT score of the percentage of lung involvement, and age were the top three contributors to the prediction of deterioration (Figure 3(a)). Patients in the early stage of chest CT at admission were more likely to deteriorate. Moreover, a higher chest CT total score meant that a broader area of the lung was involved; the patients had an increased risk of becoming severe or critically ill (Figure 3(b)). Specifically, injury to the inferior lobe of the right lung (IOR) and upper lobe of the left lung (UOL) had a more significant impact on the prediction than the other lobes.

The high neutrophil count, neutrophil ratio, and NLR were also useful in predicting severe and critically ill patients. We can take one step further to explore the feature contribution in individual predictions. The model outputs the probability of a patient becoming severe or critically ill, followed by the specific weight of contribution in the single prediction. Figure 4 shows an example of a SHAP. While the conventional machine learning model merely outputs the prediction, SHAP was able to show the details of how AI concluded.

4. Discussion

The universal manifestation of COVID-19, such as GGO, has low specificity, making it difficult to distinguish COVID-19 from other types of pneumonia solely based on chest CT appearance [12, 13]. It would be even harder, more time-consuming, and often unfeasible for radiologists to assess the disease severity based on the lobar extent, type of pulmonary opacities, clinical information, and laboratory tests, especially in urgent situations or high demand [8, 14, 15]. Since the COVID-19 outbreak, attempts using AI have been made to integrate the information from molecular, medical, and epidemiological scales [16, 17]. The cluster computing power of AI can help with early and improved disease detection and diagnosis, treatment monitoring, and contact tracing of infected individuals, which may help predict the future course of COVID-19 [18]. Moreover, AI can help with designing and developing vaccines and drugs [19–21]. This study took a step further and established six machine learning models to predict COVID-19 patients' prognosis; XGBoost ranked first in performance.

Homayounieh et al. [22] performed multiple logistic regression tests combined with the radiomics of chest CT, clinical information, and laboratory tests on 115 RT-PCR positive patients to predict the possibility of ICU admission, i.e., severe patients. They achieved a 0.84 AUC (95% CI (0.78, 0.85), $p < 0.02$). In comparison, the XGBoost model showed a 0.924 AUC (95% CI (0.826, 0.976), $p < 0.0001$), 90.91% sensitivity (95% CI (58.7%, 99.8%)), and 97.96% specificity (95% CI (89.10%, 99.90%)) based on the clinical information, laboratory tests, and chest CT features. Another issue with AI applications is interpretability. Most AI-predicted models are a "black box"; that is, it is not possible to know further details about each feature's contribution towards model prediction, an important issue with AI applications in clinical settings. Therefore, we established an interpretable XGBoost-based module called SHAP.

This interpretable module outputs the contribution of important features. Patients with features on the list have a higher possibility of deteriorating to severe or critically ill condition. In this cohort, the early stage of chest CT manifestation made the most significant contribution to the prediction, followed by the total score of chest CT and age. Lesions in the severe and critically ill patients seem to be more extensive than mild cases, meaning a higher total score of chest CT and presence of diffused patchy and large patchy appearances on the CT image. Similar to MERS-CoV, patients in the severe group were usually older than those in the mild group, indicating that the elderly tends to develop severe or critical forms of COVID-19, possibly due to comorbidities such as hypertension and underlying immune response [23]. Fever was a typical symptom of COVID-19, and those with a higher temperature at admission were more likely to worsen in the future. The cough was another common symptom, whereas fatigue, shortness of breath, and dyspnea were more common in the severe group, which is consistent with previous research [24, 25]. Furthermore, higher neutrophil count, neutrophil ratio, and NLR ratio increased the possibility of deterioration. Lymphocytopenia is a characteristic of COVID-19 [26]. The virus proliferates in the respiratory system, causing a series of immune responses, leading to changes in lymphocytes and other immune cells [25]. The lower lymphocyte count and lymphocyte ratio, higher WBC and neutrophil counts, and higher neutrophil ratio and NLR may be related to the severity and mortality rate of COVID-19 [27]. Similar to the days from onset to admission, the days from symptom onset to the first CT scan for the severe group were shorter than those for the mild group, meaning that the initial symptoms were serious, resulting in early hospital presentation. In contrast, the lesions appeared to be more extensive in the severe group, suggesting the rapid progression of COVID-19 in these patients. It is worth noting that the more extensive injuries in the IOR and UOL, the more significant their contribution to the deterioration.

With the interpretable machine learning model's application, the medical institutions could identify the potential severe type and critical type patients, hence applying the main observation since admission. Once the crucial factors change during treatment, the doctors could take the early clinical intervention to stop deterioration in the early stage.

Our study has some limitations. First, the small sample size and differences in the number of mild and severe patients may have affected the statistical power of our study. In this study, we applied stratified sampling in data segmentation to reduce the influence brought by imbalanced numbers. Second, the prognostic prediction model may be further improved by combining chest CT radiomics or deep learning models. The application of radiomics and deep learning models may eliminate subjective bias and improve performance. Attempts have been made in a previous study on the detection, outcome, and prognosis prediction of COVID-19 [2, 28, 29]. Third, this was a retrospective study, indicating uncontrollable data loss in the collection, such as procalcitonin and C-reactive protein. In order to ensure a sufficient data size, we had to give up some laboratory results, which may have decreased the performance of the model. Given the limited scale and data, the established XGBoost model requires further clinical validation.

In conclusion, this study established an interpretable machine learning model based on the XGBoost algorithm combined with clinical information, laboratory tests, and chest CT features, aimed at predicting the possibility of COVID-19 patients becoming severe and critically ill, which achieved excellent performance. Furthermore, we explored the most important features in the deterioration process using the interpretable SHAP module, which enabled us to determine the factors that put the patients at risk of developing ARDS and dying from respiratory failure and take necessary clinical interventions to improve the patient prognosis and reduce mortality among the severe and critically ill patients.

Data Availability

All data used to support the findings of this study are restricted by the Ethics Committee of Honghu People's Hospital in order to protect patient privacy.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Bowen Zheng and Yong Cai contributed equally to this work.

Acknowledgments

We would like to thank Editage (https://www.editage.com) for English language editing. This study was funded by the National Key Research and Development Program of China (2019YFC0121903), National Key Research and Development Program of China (2019YFC0117301), Natural Science Fund of Guangdong Province (2018A0303130215), Natural Science Fund of Guangdong Province (2019A1515011168), and High-level Hospital Construction Research Project of Maoming People's Hospital (Yueweihan (2018) 413).

References

[1] S. Kundu, H. Elhalawani, J. W. Gichoya, and C. E. Kahn, "How might AI and chest imaging help unravel COVID-19's mysteries?," *Radiology: Artificial Intelligence*, vol. 2, no. 3, 2020.

[2] X. Mei, H.-C. Lee, K.-y. Diao et al., "Artificial intelligence-enabled rapid diagnosis of patients with COVID-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[3] M. Chung, A. Bernheim, X. Mei et al., "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[4] Y.-H. Jin, Evidence-Based Medicine Chapter of China International Exchange and Promotive Association for Medical and Health Care (CPAM), Q.-Y. Zhan et al., "Chemoprophylaxis, diagnosis, treatments, and discharge management of COVID-19: an evidence-based clinical practice guideline (updated version)," *Military Medical Research*, vol. 7, no. 1, p. 41, 2020.

[5] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China," *JAMA*, vol. 323, no. 13, pp. 1239–1242, 2020.

[6] W.-j. Guan, Z.-y. Ni, Y. Hu et al., "Clinical characteristics of coronavirus disease 2019 in China," *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[7] G. C. Ooi, P. L. Khong, N. L. Müller et al., "Severe acute respiratory syndrome: temporal lung changes at thin-section CT in 30 Patients," *Radiology*, vol. 230, no. 3, pp. 836–844, 2004.

[8] A. Bernheim, X. Mei, M. Huang et al., "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," *Radiology*, vol. 295, no. 3, p. 200463, 2020.

[9] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, https://arxiv.org/abs/1802.03888.

[10] S. Lundberg and L. S-I, "A unified approach to interpreting model predictions," 2017, https://arxiv.org/abs/1705.07874.

[11] S. Lundberg, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[12] H. X. Bai, B. Hsieh, Z. Xiong et al., "Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT," *Radiology*, vol. 296, no. 2, pp. E46–e54, 2020.

[13] H. Choi, X. Qi, S. H. Yoon et al., "Erratum: extension of coronavirus disease 2019 (COVID-19) on chest CT and implications for chest radiograph interpretation," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, article e204001, 2020.

[14] K. Li, J. Wu, F. Wu et al., "The clinical and chest CT features associated with severe and critical COVID-19 pneumonia," *Investigative Radiology*, vol. 55, no. 6, pp. 327–331, 2020.

[15] Y. Wang, C. Dong, Y. Hu et al., "Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: a longitudinal study," *Radiology*, vol. 296, no. 2, pp. E55–e64, 2020.

[16] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz, "Mapping the landscape of artificial intelligence applications against COVID-19," *Journal of Artificial Intelligence Research*, vol. 69, pp. 807–845, 2020.

[17] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 4, pp. 337–339, 2020.

[18] A. Haleem, R. Vaishya, M. Javaid, and I. H. Khan, "Artificial intelligence (AI) applications in orthopaedics: an innovative technology to embrace," *Journal of Clinical Orthopaedics and Trauma*, vol. 11, Supplement 1, pp. S80–S81, 2020.

[19] R. Gupta, A. Ghosh, A. K. Singh, and A. Misra, "Clinical considerations for patients with diabetes in times of COVID-19 epidemic," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 14, no. 3, pp. 211-212, 2020.

[20] E. Fast, R. B. Altman, and B. Chen, "Potential T-cell and B-cell epitopes of 2019-nCoV," 2020, https://www.biorxiv.org/content/10.1101/2020.02.19.955484v1.abstract.

[21] E. Ong, M. U. Wong, A. Huffman, and Y. He, "COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning," *Frontiers in Immunology*, vol. 11, p. 1581, 2020.

[22] F. Homayounieh, R. Babaei, H. K. Mobin et al., "Computed tomography radiomics can predict disease severity and outcome in coronavirus disease 2019 pneumonia," *Journal of Computer Assisted Tomography*, vol. 44, no. 5, pp. 640–646, 2020.

[23] A. Badawi and S. G. Ryoo, "Prevalence of comorbidities in the Middle East respiratory syndrome coronavirus (MERS-CoV): a systematic review and meta-analysis," *International Journal of Infectious Diseases*, vol. 49, pp. 129–133, 2016.

[24] M. Yu, D. Xu, L. Lan et al., "Thin-section chest CT imaging of COVID-19 pneumonia: A comparison between patients with mild and severe disease," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, 2020.

[25] C. Huang, Y. Wang, X. Li et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[26] H. J. Koo, S. Lim, J. Choe, S.-H. Choi, H. Sung, and K.-H. Do, "Radiographic and CT features of viral pneumonia," *Radiographics*, vol. 38, no. 3, pp. 719–739, 2018.

[27] M. A. Matthay, L. B. Ware, and G. A. Zimmerman, "The acute respiratory distress syndrome," *The Journal of Clinical Investigation*, vol. 122, no. 8, pp. 2731–2740, 2012.

[28] D. Singh, V. Kumar, and M. K. Vaishali, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 39, no. 7, pp. 1379–1389, 2020.

[29] L. Huang, R. Han, T. Ai et al., "Serial quantitative chest CT assessment of COVID-19: deep-learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, article e200075, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 6679603, 8 pages
https://doi.org/10.1155/2021/6679603

Research Article

Automated Segmentation Method for Low Field 3D Stomach MRI Using Transferred Learning Image Enhancement Network

Luguang Huang,[1] Mengbin Li,[1] Shuiping Gou,[2,3] Xiaopeng Zhang,[2] and Kun Jiang[1]

[1] *Xijing Hospital of the Fourth Military Medical University, Xian, Shaanxi, China*
[2] *School of Artificial Intelligent, Xidian University, Xian, Shaanxi, China*
[3] *Intelligent Medical Imaging Big Data Frontier Research Center, Xidian University, Xian, Shaanxi, China*

Correspondence should be addressed to Mengbin Li; limbin@fmmu.edu.cn and Shuiping Gou; shpgou@mail.xidian.edu.cn

Received 13 November 2020; Revised 22 December 2020; Accepted 19 January 2021; Published 11 February 2021

Academic Editor: Zhiguo Zhou

Accurate segmentation of abdominal organs has always been a difficult problem, especially for organs with cavities. And MRI-guided radiotherapy is particularly attractive for abdominal targets compared with low CT contrast. But in the limit of radiotherapy environment, only low field MRI segmentation can be used for stomach location, tracking, and treatment planning. In clinical applications, the existing 3D segmentation network model is trained by the low field MRI, and the segmentation result cannot be used in radiotherapy plan since the bad segmentation performance. Another way is that historical high field intensity MR images are directly used for data expansion to network learning; there will be a domain shift problem. How to use different domain images to improve the segmentation accuracy of deep neural network? A 3D low field MRI stomach segmentation method based on transfer learning image enhancement is proposed in this paper. In this method, Cycle Generative Adversarial Network (CycleGAN) is used to construct and learn the mapping relationship between high and low field intensity MRI and to overcome domain shift. Then, the image generated by the high field intensity MRI through the CycleGAN network is with transferred information as the extended data. The low field MRI combines these extended datasets to form the training data for training the 3D Res-Unet segmentation network. Furthermore, the convolution layer, batch normalization layer, and Relu layer together were replaced with a residual module to relieve the gradient disappearance of the neural network. The experimental results show that the Dice coefficient is 2.5 percent better than the baseline method. The over segmentation and under segmentation are reduced by 0.7 and 5.5 percent, respectively. And the sensitivity is improved by 6.4 percent.

1. Introduction

Image-guided radiotherapy has become the mainstream of radiotherapy for gastric cancer, and it is very important to refer to the precise contour of target organs in the process of image-guided radiotherapy. CT has low contrast for soft tissue, so it is very difficult to locate and trace accurately abdominal organs. Dynamic Magnetic Resonance Imaging (dMRI) has the flexibility to image in the orientations most relevant to organ and tumor motion and for a prolonged duration without ionizing radiation. Therefore, MRI-guided radiotherapy is particularly attractive for abdominal targets. The normal MRI can provide a high spatial resolution anatomy and morphology proton distribution information of organs to get accurately tumor contour [1–3]. But it cannot meet the requirements of image-guided radiotherapy since high field interferes radiotherapy equipment. In order to realize the real-time MRI-guided radiotherapy, American View-Ray company developed the MRIdian system with a magnetic field intensity of 0.35 T, which makes the peak signal-to-noise ratio (PSNR) and spatial resolution of the images are very low.

At present, the low field MRI-guided radiotherapy is very few, which leads to the serious shortage of low field MR images. If high field MRI are directly used for data expansion training, domain shift will reduce greatly the segmentation performance of the existing deep learning model. Inspired by transfer learning and CycleGAN model, one way of meeting clinical application is to make high field intensity MRI transfer to low field intensity MRI to expand training data of the deep neural network [4].

Recently, many segmentation models based on UNet [5, 6] have been proposed in the last few years, like HDenseUNet [7], nnUnet res-Unet [8, 9], and LW-HCN, in which res-Unet achieved state-of-the-art performance on segmentation tasks in a different medical dataset. But these networks (Unets) are based on single domain segmentation task. And few reports were on the multidomain segmentation problem. To reduce the appearance gap cross-image modalities, a generative adversarial network (GAN) has been proposed to generate an image following a distribution. Nie et al. [10] introduced the GAN in medical image synthesis. Tanner et al. proposed a GAN for MR-CT deformable registration. Zhu et al. [11, 12] proposed the Cycle Generative Adversarial Network (CycleGAN) for the image translation from different domains. Compared with the GAN, it can be efficiently trained by the unpaired image data. This advantage can benefit to the cross-domain medical image registration. In particular, for the cross-modal medical images with big appearance and morphological gaps, CycleGAN can be introduced to map relieve domain shift.

In this paper, an automated segmentation method for low field 3D stomach MRI using transferred learning image enhancement network (TLLASN) is proposed. In the TLLASN model, our proposed multiscale 3D CycleGAN method is used to map the relationship between high and low field intensity MRI images, which overcomes domain shift between high and low field intensity images. And res-u-net is as the basic network, and then the convolution layer, batch normalization layer, and Relu layer together were replaced with a residual module. Furthermore, Dice loss function is selected to deal with the label sample unbalance in order to improve the segmentation ability of the proposed algorithm.

2. Dataset and Preprocessing

The experimental dataset included the low field MRI that were taken from 14 patients and the high field MRI that were taken from 9 patients in tumor radiation from the University of California, Los Angeles. The parameters of low field intensity MRI data acquisition are as follows: the thickness of scanning layer is 2-5 mm, the resolution of each scanning layer is $1.5 \times 1.5\,\text{mm}^2$, and the magnetic field intensity of scanning surface is 0.35 T. The parameters of high field intensity MRI data acquisition are as follows: the thickness of scanning layer is 2-5 mm, the resolution of each scanning layer is $1.5 \times 1.5\,\text{mm}^2$, and the magnetic field intensity of scanning surface is 3 T. In order to ensure the accuracy of data labeling, all the data are labeled by a radiologist. We randomly selected the images of four groups of patients from the data of 14 patients as the test set, and the rest of the images as the training set for deep network training and selected 20% of the training set as the validation set.

In the course of magnetic resonance imaging, because of the change of magnetic field, MRI scan often shows intensity inhomogeneity, and the same tissue has also bright inhomogeneous change in vision. This change is called the bias field. Because the change of signal intensity is not due to anatomical differences, bias field will bring many problems to the subsequent image processing, which will aggravate the class imbalance and affect the image segmentation. Therefore, three preprocess strategies are made as follows:

(Step 1) Bias field correction: we use N4 bias field correction technology [13] to correct the image by extracting the bias field to ensure that each tissue type has the same intensity in a single image, as is shown in Figure 1.

(Step 2) Image resampling: all data were resampled with SimpleITK toolkit, and the resolution was uniformly sampled to $1.5 \times 1.5 \times 3$.

(Step 3) Image cutting: (1) 2500 seed points were randomly scattered in the inner region of the whole 3D MRI image. (2) A $64 \times 64 \times 32$ image patch is cut out centered on the selected point. (3) The label of each image is processed in the same way as (1) and (2). (4) Detect the number of pixels in the image patch cut out of the label one by one. If the number is more than 5, the image patch is retained; otherwise, the image patch and the corresponding MRI image patch are deleted together.

(Step 4) Data normalization: the gray value of the image needs be normalized to the [0, 1], when the full convolution neural network is used for image segmentation. The following normalization formula is taken on all images:

$$\widehat{X} = \frac{X - \min(X)}{\max(X) - \min(X)}. \tag{1}$$

3. Algorithm

The lack of low field intensity stomach MRI data and the large amount of 3D segmentation network parameters lead to overfit of the model. Zhang and his group have proved that the traditional data expansion is effective in reducing over fit and improving the generalization performance, especially without a large label training set. Therefore, expanding the existing low field MRI is the best way, such as rotating, flipping, and shearing the images. However, the traditional data enhancement methods lead to a high correlation between the expanded image and the original image, so the improvement of segmentation accuracy is limited. Another way of image expansion is to enhance the image by synthesizing the data with the same distribution as the target domain. The synthesized data does not come from the target domain directly and was obtained by a transferred image in different domains, which contains abundant anatomical and topological information, and is a good supplement to the target and image.

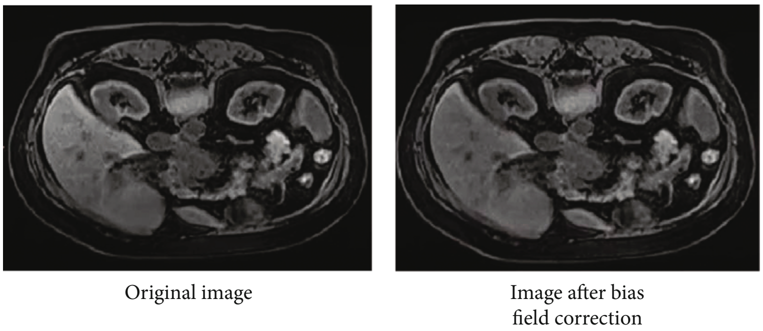


Figure 1: Image comparison before and after bias field correction.

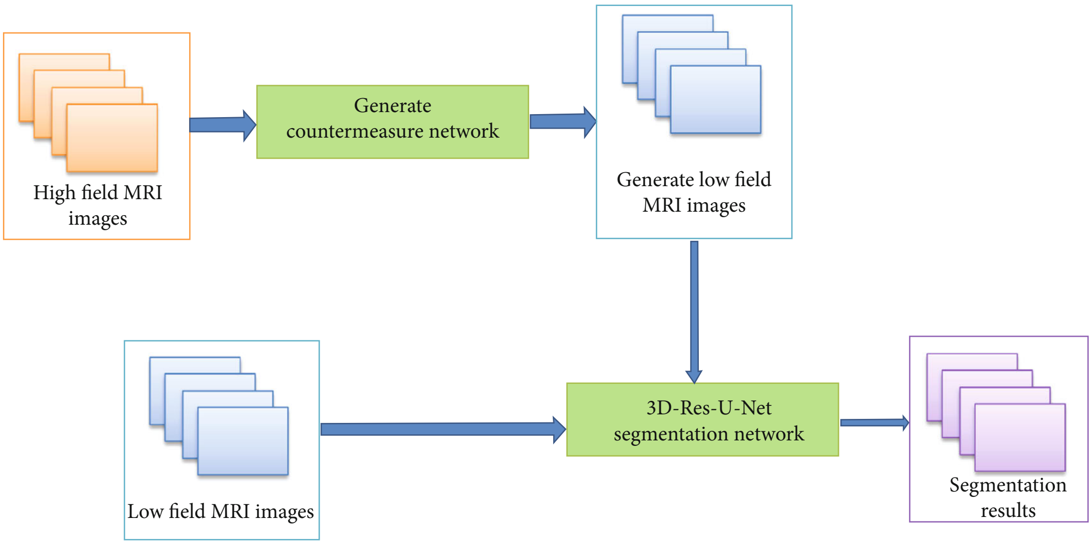

High field MRI images
Generate countermeasure network
Generate low field MRI images
Low field MRI images
3D-Res-U-Net segmentation network
Segmentation results

Figure 2: The architecture of the overall model.

Chartsias and his group [14] used CycleGAN to generate paired synthetic MRI and corresponding myocardial masks from paired CT slices and their myocardial segmentation masks. The authors based on CycleGAN image synthesis module, because it neither needs to match CT and MR image nor demands these images belonging to the same patient. Once the synthetic data was generated, synthetic MRI and original MRI are used to train the myocardial segmentation model and the segmentation performance increased 15% compared with the training of myocardial segmentation models only using real MRI.

Inspired by the study, the most direct solution is to enlarge the existing low field intensity MRI to improve the segmentation model. We expand training data by style transfer high field intensity MRI for network training in this paper, so that the trained model has good generalization ability. Cyclic Generation Adversarial Network can get the fake low field intensity MRI to map relationship between high and low field intensity MRI. Then, the fake low field intensity images generated by high field intensity magnetic resonance images through CycleGAN network are used as the extended data with transferred information. The low field MRI combines the extended dataset to train the 3D Res-Unet segmentation network, so as to overcome the problem of domain shift between high and low field intensity images.

3.1. Proposed Model. This paper proposes a low field 3D MRI segmentation model based on transfer learning image enhancement, which is mainly composed of two parts: one is the high and low field intensity MRI image transfer network based on 3D CycleGAN model to map relationship, and the other is the stomach segmentation network based on 3D res-u-net. The model structure is shown in Figure 2.

3.2. Model Optimization. The model structure of the 3D CycleGAN used is shown in Figure 3. There are five optimization strategies for the network to modify the traditional CycleGAN. First, two-dimensional convolution in CycleGAN is replaced by three-dimensional convolution. Secondly, in the generator part, the encoder decoder structure, i.e., the U-shaped structure, is adopted, and jump connection is added in the convolution layer corresponding to the encoder and decoder to fuse features of different scales. Thirdly, considering a certain correlation between the image patches of MRI, in the discriminator part, referring to the idea of PatchGan, the final output is $8 \times 8 \times 4$ matrix, and

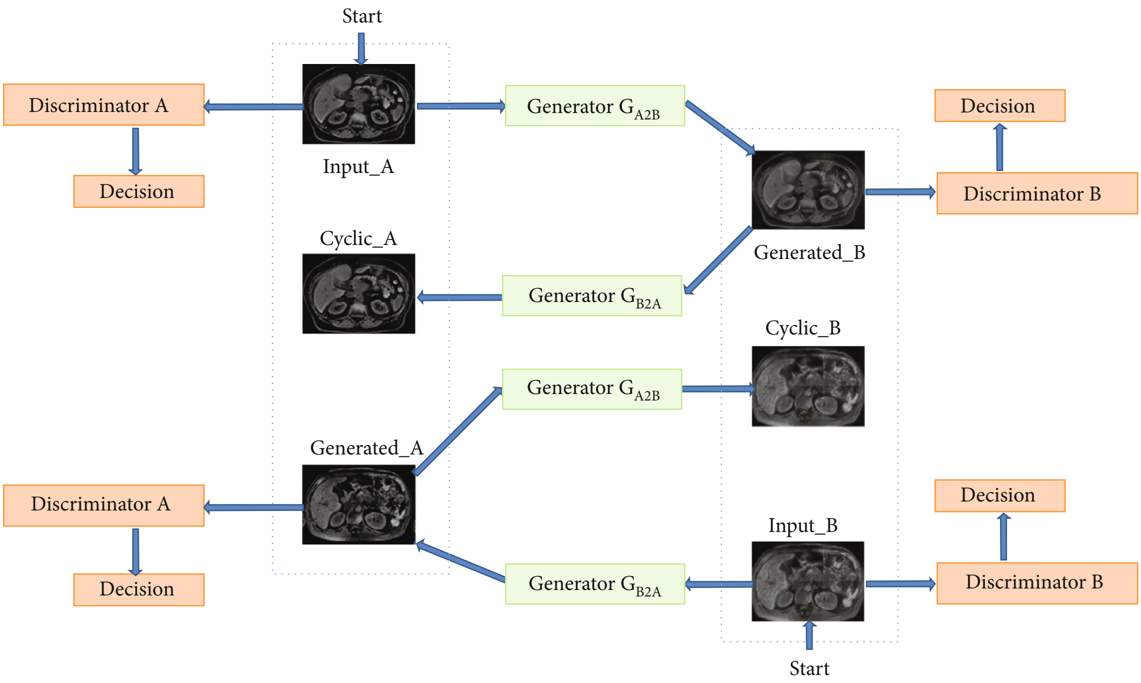


FIGURE 3: Structure diagram of high and low field intensity image transfer model.

then, the mean value of the matrix was calculated as true or false output. Fourthly, the stomach tag and MRI are used as the input of generator and discrimination, which makes the network pay more attention to the gastric region. Fifthly, label smoothing is used to reduce the label of real image from 1 to 0.9, which can avoid overconfidence of discriminator and improve the stability of model training.

There are three optimization to res-u-net the network. Firstly, the convolution layer, batch normalization layer, and Relu layer in 3D res-u-net are replaced together with a residual module, which increases the fitting ability of the network and alleviates the gradient disappearance of deep neural network. In order to reduce the parameters of the network without reducing the fitting ability of the network, a module with $1 \times 1 \times 1$ convolution kernel is added before and after each convolution kernel is $3 \times 3 \times 3$ modules, and its modified structure is shown in Figure 4.

Secondly, Dice loss is used as the loss function to solve the imbalance problem of positive and negative samples. The expression is rewrite as follows:

$$L_{DC} = 1 - DC(A, B). \qquad (2)$$

Thirdly, the convolution neural network generally requires that the input image size is fixed. For different size images, it is necessary to cut them to adapt to the input size of the network. In order to make the network adaptive to segmentation of any size images and reduce the memory consumption, the patch $(64 \times 64 \times 32)$ is used to train the network here. In the test phase, the window is used to segment the image for patch prediction. The overlap between image patches is maintained $(32 \times 32 \times 16)$, and the average value of the prediction results of overlapped parts is taken to reduce the patching effect and improve the accuracy of segmentation.

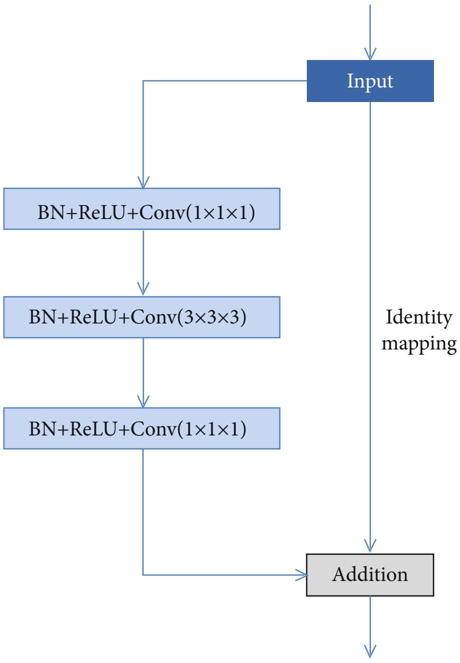


FIGURE 4: Bottleneck residual module.

3.3. Evaluation Metrics. We used Dice coefficient (DC), sensitivity (SEN), specificity (SPE), Hausdorff distance (Haus), over segmentation rate (OR), and under segmentation rate (UR) to quantitatively analyze the segmentation results. Dice coefficient is used to measure the coincidence between the

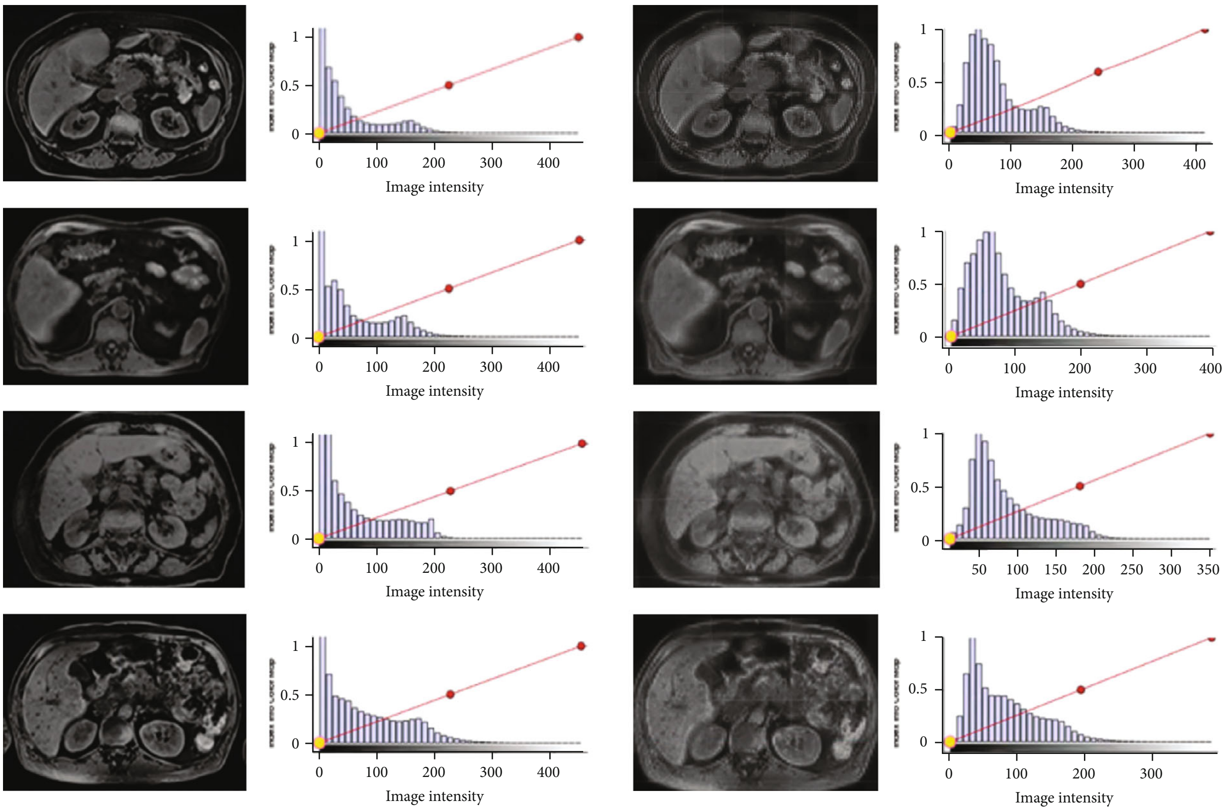


Figure 5: High field intensity MRI, axial image, and gray histogram of MRI were generated.

segmentation results and the gold standard. The larger the value, the higher the coincidence degree. It is more sensitive to the internal filling of segmentation results.

The calculation formula of Dice coefficient is as follows:

$$\mathrm{DC} = 2 \times \frac{|X \cap Y|}{|X \cup Y|}. \tag{3}$$

The sensitivity and specificity are calculated as follows:

$$\begin{aligned}\mathrm{Sen} &= \frac{|X \cap Y|}{Y},\\ \mathrm{Spe} &= \frac{|X^c \cap Y^c|}{Y^c}.\end{aligned} \tag{4}$$

Hausdorff distance is the maximum distance between the segmentation result and the nearest point in the gold standard. The smaller the value is, the higher the similarity is, and it is more sensitive to the boundary of segmentation results. The formula is as follows:

$$\begin{aligned}d_H(X, Y) &= \max |d_{XY}, d_{YX}|\\ &= \max \left\{\max_{x\in X} \min_{y\in Y} d(x, y), \max_{y\in Y} \min_{x\in X} d(x, y)\right\},\end{aligned} \tag{5}$$

where $d(x, y)$ is the Euclidean distance between the pixel in the segmentation result and the pixel in the gold standard.

The over segmentation ratio refers to the ratio of pixels whose segmentation results are beyond the gold standard. The calculation formula is as follows:

$$\mathrm{OR} = \frac{O_S}{R_S + O_S}, \tag{6}$$

where O_S is the number of pixels that should not have been included in the segmentation results, but actually are in the segmentation results, and R_S is the number of pixels in the segmentation results that coincide with the gold standard.

The under segmentation ratio refers to the ratio of pixels missing in the segmentation result within the gold standard. The calculation formula is as follows:

$$\mathrm{UR} = \frac{U_S}{R_S + O_S}, \tag{7}$$

where U_S is the number of pixels that should have been included in the segmentation results, but are not actually in the segmentation results.

3.4. Implementation Details. Our experiments were carried out using Keras with Ten-sorflow, whose backend is Python 3.5, and used Nvidia Ge-Force RTX2080, Cuda 9.0, and Cudnn v7.3.1 toolkit for parallel acceleration. The hardware configuration of the computer is 4.0 GHz Intel Core i7-4790k CPU. During optimization, Adam is used to optimize the generator and discriminator of cyclic generation countermeasure network. The weight of momentum is set to 0.5. The

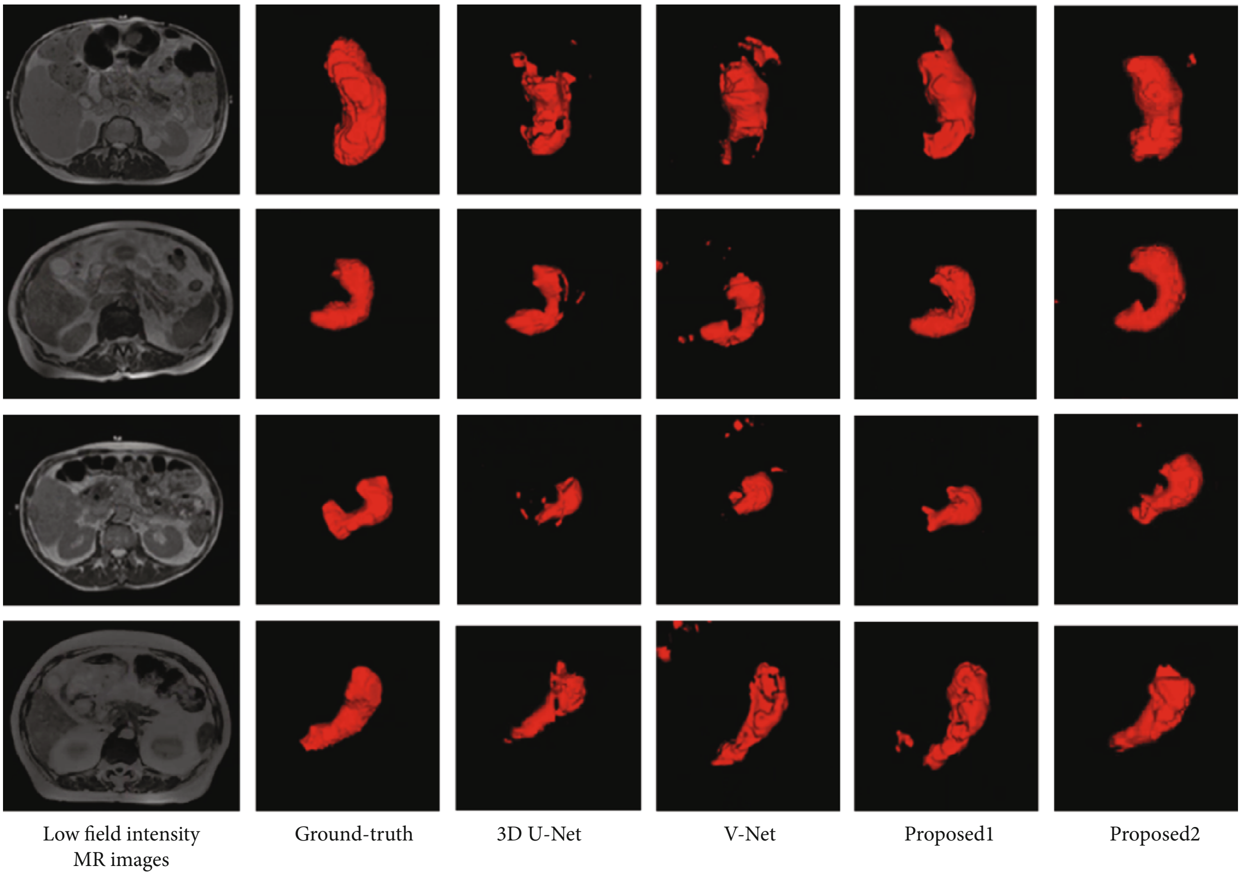


Figure 6: Image enhancement based on transfer learning in low field MR stomach segmentation.

learning rate of the generator and discriminator is set to 0.0001 and 0.0004, respectively. In the segmentation network, the size of input image patch is 64 × 64 × 32, the batch size is set to 4, the optimizer uses Adam as the optimizer, the learning rate is 0.0001, and early stop is used to prevent over fitting.

4. Results

Figure 5 shows the axial and gray histogram of some high field MRI and the axial and gray histogram of the generated MRI images. The first column of each row is the axial image of the original high field intensity MRI, the second column is the gray histogram of the original high field intensity MRI, the third column is the fake image generated by CycleGAN, and the fourth column is the gray histogram corresponding to the transformed image. It can be seen that the histogram distribution of the transformed image is similar and follows the same distribution. From the perspective of anatomical structure, the position and shape of each organ in the image did not change, but the gray distribution was different, which is equivalent to the style migration, so the high field MRI label can be used as the label of the transferred image for segmentation network training.

In order to analyze the segmentation of different experiments more intuitively, the segmentation results are 3D reconstructed. Figure 6 shows the low field MRI images and their 3D segmentation style of different patients using 4 deep networks, in which "Proposed1" means the segmentation method for traditional image enhance the low field intensity MRI by flipping and rotating and "Proposed2" means the

Table 1: Comparison of Dice coefficient (DC) index of four segmentation methods.

Method	#patient 1	2	3	4	Mean
3D U-net	0.562	0.704	0.648	0.349	0.566
V-net	0.613	0.759	0.597	0.493	0.616
Proposed1	0.684	0.865	0.681	0.532	0.690
Proposed2	0.654	0.874	0.730	0.631	0.722

Table 2: Comparison of Hausdorff distance index of four segmentation methods.

Method	#patient 1	2	3	4	Mean
3D U-net	10.05	10.1	7	7	8.54
V-net	9	8.1	6.71	8.60	8.10
Proposed1	9	7.87	6.71	8.60	8.04
Proposed2	9	7.87	6.4	8.60	7.97

Table 3: Sensitivity index comparison of four segmentation methods.

Method	#patient 1	2	3	4	Mean
3D U-net	0.395	0.568	0.495	0.212	0.417
V-net	0.472	0.697	0.539	0.396	0.526
Proposed1	0.535	0.811	0.551	0.509	0.602
Proposed2	0.497	0.814	0.600	0.465	0.594

Table 4: Comparison of specificity indexes of four segmentation methods in test set.

Method	#patient 1	2	3	4	Mean
3D U-net	0.999	0.999	0.999	0.999	0.999
V-net	0.999	0.999	0.999	0.999	0.999
Proposed1	0.999	0.999	0.999	0.999	0.999
Proposed2	0.999	0.999	0.999	0.999	0.999

Table 5: Comparison of over segmentation rate of four experimental methods in test set.

Method	#patient 1	2	3	4	Mean
3D U-net	0.011	0.044	0.031	0.003	0.023
V-net	0.063	0.122	0.211	0.172	0.142
Proposed1	0.029	0.064	0.064	0.025	0.045
Proposed2	0.023	0.044	0.043	0.008	0.029

segmentation method for traditional image enhance combining with transfer learning.

Compared with the segmentation results of 3D u-net and v-net, we can see that 3D u-net can segment the general stomach shape; however, under segmentation is more serious. V-net has improved to the under segmentation, but it is over segmentation. Compared with v-net and Proposed1, it can be seen that after the transfer learning image enhancement, the high field intensity MRI is transformed into the pseudo low field intensity MRI with the similar distribution as the low field intensity MRI by using CycleGAN, which increases the diversity of training samples and improves the segmentation performance significantly. Although "Proposed1" segmentation results also have partial over segmentation, the stomach region is smoother than that of 3D u-net and v-net. Compared with the results of the other three methods, "Proposed2" is a little over segmentation and under segmentation regions. To sum up, "Proposed2" has more completed region consistency and clear contour, which is closest to the ground truth.

Table 1 and Table 2 show the comparison of Dice coefficient and Hausdorff distance of segmentation results of different algorithms. From the point of view of Hausdorff distance, the indexes of different algorithms are almost the same, but the average Hausdorff distance of the proposed algorithm on the test set is optimal. From the perspective of Dice coefficient, there are different degrees of improvement in each test sample by using image enhancement strategy. The result of "Proposed2" shows that the use of high field intensity MRI transformed by CycleGAN network increases the diversity of training samples, alleviates the over fitting phenomenon to a certain extent, and improves the generalization ability of the network.

Table 6: Comparison of four segmentation methods.

Method	#patient 1	2	3	4	Mean
3D U-net	0.598	0.413	0.489	0.785	0.571
V-net	0.494	0.266	0.364	0.499	0.406
Proposed1	0.491	0.180	0.382	0.530	0.396
Proposed2	0.451	0.174	0.419	0.478	0.381

Table 7: Comparison of segmentation results of different image enhancement methods.

Method	Metric	#patient 1	2	3	4	Mean
Traditional_based	DC	0.643	0.853	0.621	0.544	0.665
	Haus	9	8.062	6.708	8.485	8.064
	Sen	0.494	0.781	0.475	0.403	0.538
	Spe	0.999	0.999	0.999	0.999	0.999
	OR	0.039	0.047	0.052	0.073	0.053
	UR	0.486	0.208	0.497	0.553	0.436
CycleGAN_based	DC	0.684	0.865	0.681	0.532	0.690
	Haus	9	7.874	6.708	8	7.895
	Sen	0.535	0.814	0.551	0.509	0.602
	Spe	0.999	0.999	0.999	0.999	0.999
	OR	0.029	0.064	0.064	0.026	0.046
	UR	0.451	0.174	0.419	0.478	0.381

In order to analyze the under segmentation and over segmentation of 4 methods, we use indexes of sensitivity, specificity, over segmentation rate, and under segmentation rate for comparison, as shown in Tables 3–6. In Table 5, the over segmentation rate of the 4 methods is generally low, which shows that the over segmentation is not obvious in the segmentation results. This is consistent with the high specificity index in Table 4. The over segmentation rate of the algorithm we proposed is the lowest. Although the under segmentation rate of patient 3 in method 4 was slightly higher than that in method 2, its sensitivity was 1% higher. On the whole, the indexes of method 4 are better than those of the other three methods, which means that the segmentation model based on transfer learning image enhancement is effective for stomach segmentation of low field intensity MRI images.

Table 7 shows the comparison of segmentation results between the traditional data enhancement method and the combining image enhancement method. Traditional_based is to enhance the low field intensity MRI by flipping and rotating, CycleGAN_based is to enhance an image by using the trained CycleGAN network, which transformed the high field intensity MRI image into a pseudo low field intensity MRI image as the low field intensity MRI. From the perspective of segmentation index, each segmentation index of the CycleGAN_based is better than that of Traditional_based. Dice coefficient of CycleGAN_based is 2.5 percent higher than that of Traditional_based, over segmentation rate and less segmentation rate of CycleGAN_based are 0.7 and 5.5 lower percent than those of Traditional_based, respectively, and sensitivity of CycleGAN_based is higher 6.4 percent than that of Traditional_based.

5. Conclusion

The stomach is a kind of cavity organ in the abdomen, which is easy to deform and has uneven gray distribution. Moreover, low field intensity stomach MRI are noisy and lacking of data, which increases the difficulty of stomach segmentation in 3D images. TLLASN is proposed to cope with these problems. CycleGAN can get the fake low field intensity MRI to reduce the difference between high and low field intensity MRI. In other words, domain adaption between high and low field intensity images is achieved. In this study, the fake low field intensity images transferred information to train 3D Res-Unet segmentation network. High field intensity MRI is used to expand training data by style transfer for network training in this paper, so that the trained model has good generalization ability. The experimental results show that the automated segmentation method for low field 3D stomach MRI using transferred learning image enhancement network effectively increases the amount and diversity of training data and achieves good segmentation results.

Data Availability

Data used to support the findings of this study are included within the article

Conflicts of Interest

The authors declare that there are no conflicts of interest that could be perceived as prejudicing the impartiality of the researched reported.

Authors' Contributions

Luguang Huang and Mengbin Li contributed equally to this work.

Acknowledgments

This study is supported by the Natural Science Foundation of Shaanxi Province of China under Grant Nos. 2019ZDLGY03-02-02 and 2017JM8026, the Research Plan of Improving Public Scientific Quality in Shaanxi Province (No. E219360001), the Fundamental Research Funds for the Central Universities (No. JC2001), and Focus on Developing of Shaanxi Province under Grant No. 2020GY-049.

References

[1] R. S. Desikan, F. Ségonne, B. Fischl et al., "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[2] J. N. Giedd, J. Blumenthal, N. O. Jeffries et al., "Brain development during childhood and adolescence: a longitudinal MRI study," *Nature Neuroscience*, vol. 2, no. 10, pp. 861–863, 1999.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[4] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[5] Y. Chen, L. Xing, L. Yu, H. P. Bagshaw, M. K. Buyyounouski, and B. Han, "Automatic intraprostat-ic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch Unet," *Medical Physics*, vol. 47, no. 12, pp. 6421–6429, 2020.

[6] Y. Cao, S. Liu, Y. Peng, and J. Li, "DenseUNet: densely connected UNet for electron microscopy image segmentation," *IET Image Processing*, vol. 14, no. 12, pp. 2682–2689, 2020.

[7] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: a novel multiphoton *in vivo* cellular image segmentation model based on a convolutional neural network," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 6, pp. 1275–1285, 2020.

[8] K. Cao and X. Zhang, "An improved res-UNet model for tree species classification using airborne high-resolution Images," *Remote Sensing*, vol. 12, no. 7, 2020.

[9] X. Liu, Y. Zhang, H. Jing, L. Wang, and S. Zhao, "Ore image segmentation method using U-Net and Res_Unet convolutional networks," *RSC Advances*, vol. 10, no. 16, pp. 9396–9406, 2020.

[10] D. Nie, R. Trullo, J. Lian et al., "Medical image syn-thesis with deep convolutional adversarial networks," *IEEE Transactions onBiomedicalEngineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Un-paired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE Interna-tional Conference on Computer Vission*, pp. 2223–2232, Venice, Italy, 2017.

[12] O. Bernard, A. Lalande, C. Zotti et al., "Deep learning techniques for au-tomatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[13] N. J. Tustison, B. B. Avants, P. A. Cook et al., "N4ITK: improved N3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[14] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *Simulation and Synthesis in Medical Imaging*, vol. 10557, pp. 3–13, Springer, 2017.

Hindawi
BioMed Research International
Volume 2021, Article ID 6668510, 8 pages
https://doi.org/10.1155/2021/6668510

Research Article

To Explore MR Imaging Radiomics for the Differentiation of Orbital Lymphoma and IgG4-Related Ophthalmic Disease

Ye Yuan,[1,2] **Guangyu Chu**,[1] **Tingting Gong**,[1] **Lianze Du**,[1] **Lizhi Xie**,[2] **Qinghai Yuan**,[1] **and Qinghe Han**[1]

[1] *Department of Radiology, The Second Hospital of Jilin University, Changchun 130041, China*
[2] *GE Healthcare, MR Research China, Beijing 100176, China*

Correspondence should be addressed to Qinghe Han; hanqinghehe@126.com

Received 4 November 2020; Revised 29 December 2020; Accepted 23 January 2021; Published 8 February 2021

Academic Editor: Erlei Zhang

Among orbital lymphoproliferative disorders, about 55% of diagnosed cancerous tumors are orbital lymphomas, and nearly 50% of benign cases are immunoglobulin G4-related ophthalmic disease (IgG4-ROD). However, due to nonspecific characteristics, the differentiation of the two diseases is challenging. In this study, conventional magnetic resonance imaging-based radiomics approaches were explored for clinical recognition of orbital lymphomas and IgG4-ROD. We investigated the value of radiomics features of axial T1- (T1WI-) and T2-weighted (T2WI), contrast-enhanced T1WI in axial (CE-T1WI) and coronal (CE-T1WI-cor) planes, and 78 patients (orbital lymphoma, 36; IgG4-ROD, 42) were retrospectively reviewed. The mass lesions were manually annotated and represented with 99 features. The performance of elastic net-based radiomics models using single or multiple modalities with or without feature selection was compared. The demographic features showed orbital lymphoma patients were significantly older than IgG4-ROD patients ($p < 0.01$), and most of the patients were male (72% in the orbital lymphoma group vs. 23% in the IgG4-ROD group; $p = 0.03$). The MR imaging findings revealed orbital lymphomas were mostly unilateral (81%, $p = 0.02$) and wrapped eyeballs or optic nerves frequently (78%, $p = 0.02$). In addition, orbital lymphomas showed isointense in T1WI (100%, $p < 0.01$), and IgG4-ROD was isointense (60%, $p < 0.01$) or hyperintense (40%, $p < 0.01$) in T1WI with well-defined shape (64%, $p < 0.01$). The experimental comparison indicated that using CE-T1WI radiomics features achieved superior results, and the features in combination with CE-T1WI-cor features and the feature preselection method could further improve the classification performance. In conclusion, this study comparatively analyzed orbital lymphoma and IgG4-ROD from demographic features, MR imaging findings, and radiomics features. It might deepen our understanding and benefit disease management.

1. Introduction

Orbital lymphoproliferative disorders (OLPDs) consist of a broad range of benign and malignant tumors [1, 2]. Among diagnosed cancerous tumors, nearly 55% of cases are orbital lymphomas [3], while luckily, most orbital lymphomas are primary, low-grade, and amendable to low-dose radiotherapy [1–3]. To improve the diagnosis performance, many studies explored to figure out some discriminative characteristics. Eckardt et al. evaluated the diagnostic approach in 11 orbital lymphoma patients and found that orbital swelling, pain, and motility impairment were the leading clinical symptoms [4]. Another study observed the proptosis, eyelid lesions, decreased visual acuity, and optic nerve compression in 26 cases with orbital lymphoma [5]. Moreover, Priego et al. described different orbital lymphoma patterns at diagnosis and follow-up in 19 cases, and superior-lateral quadrant and extraconal location were predominantly observed on imaging scans [6]. The patterns were further confirmed by Jin et al. who evaluated the computed tomography (CT) imaging and magnetic resonance imaging (MRI) features of primary orbital lymphoma to establish a differential diagnosis in 14 cases, reporting that periorbital preseptal tissues were mainly involved in the upper lateral quadrant of the orbit [7]. They also suggested that MRI may be very useful for assessing the location, configuration, inner structure, and characteristic manifestations of orbital lymphomas [7]. However, these symptoms were either qualitative such as laterality or nonspecific such as decreased visual acuity; thus, they could not have a wider application [1–8].

It is also figured out that nearly 50% of benign OLPD cases are immunoglobulin G4-related ophthalmic disease (IgG4-ROD) [1, 9]. IgG4-ROD is an inflammatory disease of unknown etiology, which can be treated using corticosteroid therapy [1, 2]. Typical IgG4-ROD is characterized by painless enlarging masses over the lacrimal gland with or without proptosis. Bilateral disease is common but not necessarily symmetrical; visual acuity is usually not impaired. Besides the lacrimal gland, IgG4-ROD has been reported in various orbit tissues, including muscle, fat, eyelid, and nerve [9]. A short summary of related publications indicated that the signs and symptoms of IgG4-ROD included chronic noninflammatory lid swelling and proptosis. Moreover, patients often had a history of allergic disease and increased serum levels of IgG4, IgE, and hypergammaglobulinemia [10]. In addition, a study comparing both IgG4-ROD and non-IgG4-ROD European patients revealed that infraorbital nerve enlargement was frequently presented in IgG4-ROD patients [11].

Except for conventional MR images, diffusion-weighted imaging (DWI) has been extensively explored over the years. The apparent diffusion coefficient (ADC) values have been revealed useful in diagnosing OLPDs [12, 13]. Haradome et al. observed that the mean ADC of orbital lymphomas was significantly lower than that of benign OLPDs ($p < 0.01$). In addition, an optimal cutoff of ADC values could yield a superior prediction of orbital lymphoma, and the prediction was even better than that using the contrast-enhancement ratio of lesions [1]. Xu et al. also found significantly lower ADC ($p < 0.001$) in malignant OLPDs when compared to benign ones, and a receiver operating characteristic curve analysis indicated ADC alone could achieve an optimal sensitivity in the classification of benign and malignant OLPDs [2]. In addition, ElKhamary et al. suggested that median ADC was significantly different between benign and malignant OLPDs, and an ideal threshold of ADC values benefited the classification of diffuse orbital masses [14]. Notably, Lecler et al. [15] and Maldonado et al. [16] also reported similar results.

CT is another useful imaging approach for analyzing OLPDs. Jin et al. [7] found that isodense soft tissue masses characterized primary orbital lymphoma with clear demarcation on CT images; the lesions showed homogeneously marked enhancement when contrast medium was used. Simon et al. [17] discovered that benign lesions were more likely hyperdense or hypodense on CT in comparison with inflammatory and malignant tumors. Briscoe et al. [5] suggested that bone changes were more common on CT images when orbital lymphomas were suspected. Thus, combining CT and MR imaging could be useful for accurate diagnosis of OLPDs.

Preoperative identification of orbital lymphoma and IgG4-ROD facilitates disease management, treatment planning, and health care [1–3]. Yet, due to nonspecific presenting signs and symptoms and lack of qualitative findings, diagnosis is still somehow challenging. For diagnosis, a biopsy is routinely utilized in clinics. However, considering the tumor's specific location, i.e., orbital lesions, a biopsy is difficult, and thus, may lead to misdiagnosis, mistreatment, and even missed diagnosis [9].

Radiomics has been widely explored for intelligent diagnosis [18–20]. It extracts quantitative features from medical images using advanced algorithms [21–23], and the features are further mined for disease diagnosis and cancer staging [24–27]. However, to the best of our knowledge, no machine learning-based radiomics models have yet been designed for orbital lymphoma and IgG4-ROD. Since previous studies suggested that MR imaging is a promising tool to accurately visualize the location, shape, and internal structure of orbital lymphoma [1, 2, 7, 11]; in this study, we assessed the value of conventional MR images in machine learning-based radiomics approaches for clinical identification of orbital lymphoma and IgG4-ROD.

2. Materials and Methods

2.1. Patients and Data Collection. This retrospective study was approved by the institutional review board of the Second Hospital of Jilin University, and written informed consent from patients was waived. Through a review of our hospital database, 36 cases of orbital lymphoma and 42 cases of IgG4-ROD were identified. All patients were historically confirmed by surgical biopsy between March 2013 and September 2018. It should be noted that all patients received MR imaging before the surgical biopsy.

Histopathologic features were used for pathologic diagnosis. Orbital lymphoma was diagnosed using flow cytometric and gene rearrangement analysis. IgG4-ROD was identified according to the immunohistochemical staining results, which require the number of IgG4-positive plasma cells more than 50 cells per high-power field samples, the ratio of IgG4-positive plasma cells over IgG-positive plasma cells >40%, and serum IgG4 concentration of 1.35 g/L [28].

All diagnosed patients were without a history of previous treatment or surgery. They had no history of orbital diseases or other tumors. All imaging was performed using a 3.0-T MR equipment (GE MR 750) with imaging parameters as in Table 1. Axial fast spin-echo (FSE) T1-weighted (T1WI) and T2-weighted (T2WI) images, contrast-enhanced T1WI in the axial (CE-T1WI) and coronal (CE-T1WI-cor) planes were acquired using Gd-DTPA (dose: 0.1 mmol/kg; and injection rate: 2.0 ml/s).

2.2. Manual Annotation and Feature Extraction. Mass lesions were manually outlined by using the ITK-SNAP software (version 3.8.0). Two board-certified radiologists with 6 and 10 years of experience in head and neck imaging performed the manual annotation together and were blinded to clinical information and histologic diagnosis. If consensus was not reached, the annotation was further arbitrated by a senior radiologist with 16 years of experience to ensure the annotation quality.

Manual annotation and feature extraction were performed as follows: MR images of one patient were imported

Table 1: MR imaging parameters on the 3.0-T scanner.

	TR (ms)	TE (ms)	Slice thickness (mm)	Slice gap (mm)	Matrix size	Field of view ([mm, mm])
T1W1	515	17	3.0	0.3	[512, 512]	[15, 15]
T2WI	2000	85	3.0	0.3	[512, 512]	[15, 15]
CE-T1WI	463	8.5	3.0	0.3	[512, 512]	[15, 15]
CE-T1WI-cor	650	8.8	3.0	0.3	[512, 512]	[15, 15]

into the ITK-SNAP. Then, the radiologists performed the image analysis from the laterality (left/right/bilateral) and the shape of the margins (well-defined or ill-defined) to figure out obvious lesion boundaries. If the lesion boundaries were ambiguous, MR images from the four imaging sequences were displayed for observation, and CE-T1WI and CE-T1WI-cor were set as the baseline. After discussion, the consensus was reached, and lesion delineation was performed slice by slice. Specifically, the delineation was made from the head to the feet direction to avoid bone structures and eyeball regions. When eye muscles and/or optic nerves were involved, eye muscles and organ tissues were outlined if necessary, as the lesion was our point of interest.

Two representative examples are shown in Figure 1. The top row represents a case of orbital lymphoma, and the bottom row shows a case of IgG4-ROD. From left to right is one slice of T1W1, T2W1, CE-T1WI, and CE-T1WI-cor image in addition to the mask of volume region of interest. In each slice, the region in red lines represents the mass lesion.

Annotated tumors were quantified using a public package Pyradiomics (version 3.0), and a total of 99 features were computed. Among the features, 14 were for shape description, 18 were from first-order histogram analysis, 22 were from gray-level cooccurrence matrix (GLCM) analysis, 14 were from gray-level run-length matrix (GLRLM) features, 16 were from gray-level size zone matrix (GLSZM) analysis, and 15 were from gray-level differential matrix (GLDM) analysis. These features have been applied for lesion representation, radiomics, and intelligent diagnosis [29].

2.3. Disease Differentiation. Figure 2 shows the workflow of disease differentiation using elastic net fitting [30]. First, a data set was divided into a training set and a testing set by random splitting. The Wilcoxon rank-sum test was optionally used to figure out these statistically significant features by comparing the two groups of data samples. Consequently, the default parameters of the elastic net were tuned, finally generating a trained model. At the testing stage, the trained elastic net was evaluated via a testing data set, and its performance was assessed. The rectangle with a dashed line indicated a comparison study to investigate the effect of the Wilcoxon rank-sum test in disease diagnosis.

2.4. Experimental Design. This study investigated the effect of single modality, multiple modalities, and preselection of important features on disease classification performance. Single modality data sets included T1WI, T2WI, CE-T1WI, and CE-T1WI-cor; multiple modality data sets comprise different combinations of single modality data (T1WI + T2WI + CE-T1WI, T1WI + T2WI + CE-T1WI-cor, CE-T1WI + CE-T1WI-cor, and T1WI + T2WI + CE-T1WI + CE-T1WI-cor). In addition, the effect of selecting statistically important features using a nonparametric test of Wilcoxon rank-sum test on disease diagnosis was observed.

The elastic net has been widely used in feature selection, regularized regression, and data classification [30]. It linearly combined both L_1 and L_2 penalties using a parameter α to overcome some limitations of the least absolute shrinkage and selection operator (LASSO) [31]. In this study, the elastic net was used for feature selection and classification ($\alpha = 0.75$). First, 80% of data samples were randomly selected for training the elastic net model, and 10-folder cross-validation was used for automatic parameter tuning. Next, the trained elastic net model was tested on the testing samples. Then, the prediction performance was evaluated using four metrics, including the area under the curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE) [32]. In addition, the procedure was repeated 100 times, and the performance metrics were averaged. The whole procedure was implemented with MATLAB2018a (MathWorks, USA) and the elastic net using the embedded function "lasso.m."

2.5. Statistical Analysis. The group differences were assessed by a two-tailed t-test or Pearson's chi-squared test based on the SPSS software (version 25.0, IBM Corp., Armonk, NY). p value <0.05 was considered statistically significant.

3. Results

3.1. Patient Characteristics and Tumor Distribution. Table 2 shows patient characteristics and tumor distribution between the two groups. Significant differences were found between groups. Patients with orbital lymphoma were 9 years older than patients with IgG4-ROD. Moreover, most patients with malignant tumors were male (26/36, 72%). In the IgG4-ROD group, 10/42 (23%) were male. Yet, no statistical difference in gender was found between the two groups. In addition, most orbital lymphomas were unilaterally involved (29/36, 81%), while IgG4-RODs were equally unilateral and bilateral.

Table 3 summarizes MR features between the two groups. Significant differences were found in 3 attributes. First, the shape of margins of IgG4-ROD lesions was well-defined (27/42, 64%) in comparison with that of orbital lymphomas (14/36, 39%). Second, the lesions were more frequently wrapped around eyeballs and/or optic nerves in patients with orbital lymphomas (28/36 (78%)) compared to those with IgG4-RODs (22/42 (52%)). Third, in T1WI images, orbital lymphoma was perceived as isointense (36/36, 100%), while

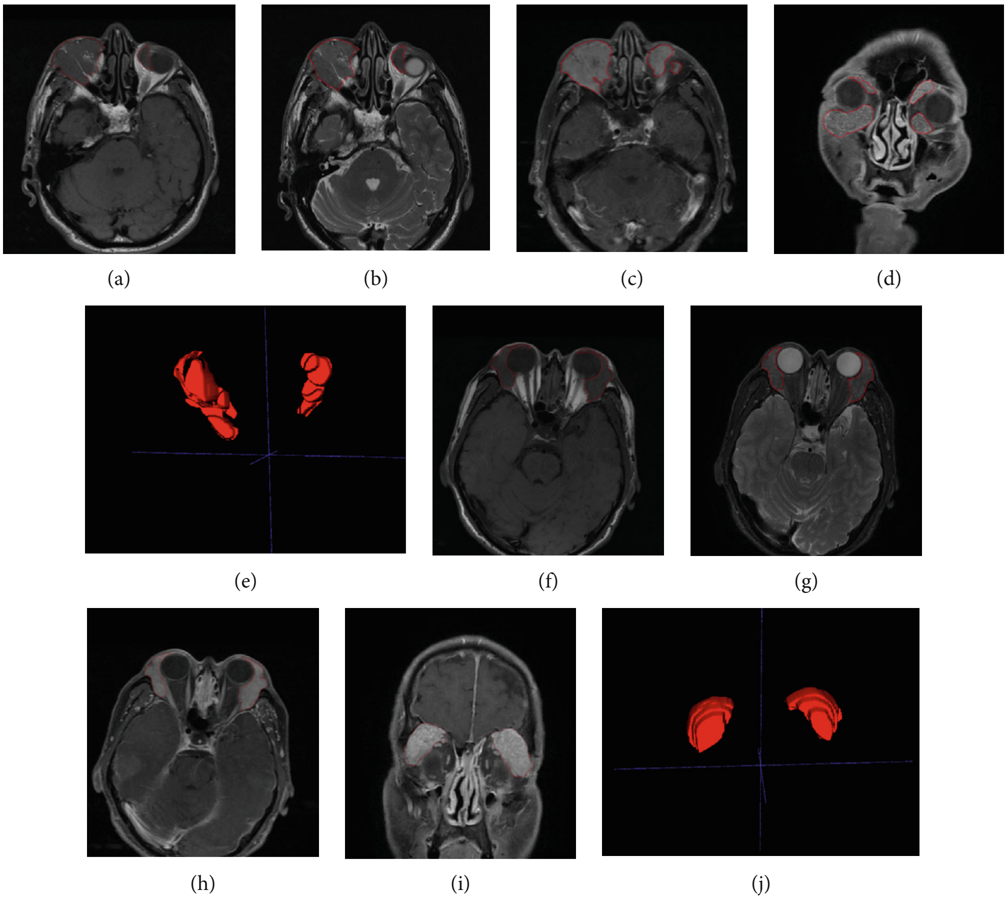


Figure 1: Two representative cases. The top row shows a 60-year male patient with orbital lymphoma, and the bottom row shows a 60-year female patient with IgG4-ROD. In each case, one image of T1WI, T2WI, CE-T1WI, and CE-T1WI-cor and the volume mask are shown from left to right. Mass lesions are the region in red lines. Note that images are cropped for display purposes.

IgG4-ROD as isointense (25/42, 60%) or hyperintense (17/42, 40%). We also found that some patients in both groups had flow void sign and in-homogeneity in lesion regions. Moreover, most orbital lymphomas (26/36, 72%) and IgG4-RODs (31/42, 74%) were perceived as hypointense signals in T2WI images.

3.2. Parameter Optimization. Figure 3 shows the automated optimization λ when training samples were fitted by elastic net using 10-folder cross-validation (CV). The x-axis indicates the change of λ value, and the y-axis corresponds to the mean square error (MSE). In addition, the green dotted line locates the λ with minimum CV error, and the solid blue line points to the minimum CV error plus one standard deviation (SE). In this study, a larger λ was used when the MSE was within one SE of the minimum one for the consideration of model reliability.

3.3. Performance Using Single versus Multiple Modality Data. Table 4 summarizes the performance when using single or multiple modality data for disease classification. The best result was obtained when using CE-T1WI, followed by CE-T1WI-cor. Both T1WI and T2WI caused poor SPE (<0.50), while T1WI led to a fair AUC value (<0.60). The application of multimodality increased the diagnosis results. The addition of CE-T1WI-cor increased the AUC and SPE by 5% and 9%, respectively. However, adding T1WI and T2WI to the combination of CE-T1WI + CE-T1WI-cor did not improve the classification performance.

3.4. Performance with Feature Preselection. The results with feature preselection are shown in Table 4. By comparing both Table 4 and Table 5, we found that feature preselection improved the combination of CE-T1WI and CE-T1WI-cor ($p < 0.02$) and benefited single- (such as CE-T1WI and CE-T1WI-cor) and other multiple modality data (such as T1WI + T2WI + CE-T1WI) based disease differentiation.

3.5. Feature Analysis. Wilcoxon rank-sum test indicated that 13, 18, 75, and 40 features were with statistical significance ($p < 0.05$) corresponding to T1WI, T2WI, CE-T1WI, and CE-T1WI-cor. In disease classification, elastic net further verified that 1, 5, 6, and 4 features were frequently selected (>50 times) between the two groups of patients on T1WI, T2WI, CE-T1WI, and CE-T2WI-cor, respectively. It is worth noting the elastic net model with the superior performance required 6 features, among which 5 were computed from CE-T1WI (1 shape feature, major axis length; 2 GLCM features, correlation and autocorrelation; 1 GLDM feature, large dependence high gray-level emphasis; 1 GLRM feature, long-

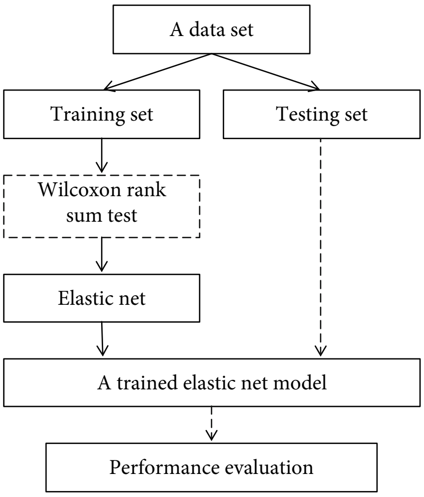


Figure 2: The procedure of disease diagnosis. It includes data splitting, identification of significant features, elastic net-based feature selection, disease diagnosis, and performance assessment.

Table 2: Patient characteristics and tumor distribution.

	Orbital lymphoma ($n = 36$)	IgG4-ROD ($n = 42$)	p value	K value
Age (years)			< 0.01	
Mean ± std	64.89 ± 10.30	55.21 ± 13.88		
Range	[38, 84]	[25, 78]		
Gender			0.03	4.85
Male	26	20		
Female	10	22		
Laterality			0.02	7.53
Left	18	11		
Right	11	11		
Bilateral	7	20		

Table 3: Perceived MR imaging features.

	Orbital lymphoma ($n = 36$)	IgG4-ROD ($n = 42$)	p value	K value
Margin			< 0.01	37.05
Well-defined	14	27		
Ill-defined	22	15		
Local spread of eyeball or optic nerve			0.02	5.43
Yes	28	22		
No	8	20		
Extraocular muscles involved			0.12	2.47
Yes	15	25		
No	21	17		
Flow void sign present on T2WI			0.23	1.44
Yes	14	11		
No	22	31		
Signal intensity on T1WI			< 0.01	18.63
Low	0	0		
Iso	36	25		
High	0	17		
Signal intensity on T2WI			0.59	1.04
Low	26	31		
Iso	9	8		
High	1	3		
Homogeneity			0.75	0.10
Yes	29	35		
No	7	7		

run high gray-level emphasis), and 1 GLDM feature (large dependence high gray-level emphasis) from CE-T2WI-cor. In addition, frequently selected features were all from post-contrast T1WI images, and the GLDM feature was highlighted.

4. Discussion

This study investigated demographic characteristics, MR imaging features, and radiomics models of orbital lymphoma and IgG4-ROD, thus aiming to facilitate preoperative diagnosis of these two different tumor types. Seventy-eight patients were retrospectively reviewed, and mass lesions were manually annotated. Clinical characteristics, MR findings, and the performance of single and multimodality data with and without feature preselection were analyzed.

Demographic characteristics revealed that orbital lymphoma patients were significantly older than IgG4-ROD patients. This has also been previously reported by other studies that examined the difference between orbital lymphoma and other diseases, such as benign OLPDs [1, 2], pseudotumor [33], and lymphoma subtypes [34]. Thus, the patient's age should be considered when performing a diagnosis. Moreover, we discovered that most patients with orbital lymphoma (72%) were male, yet there was no significant difference between patients with orbital lymphoma and those with IgG4-ROD, which is consistent with data published by Olsen and Steffen [34] and inconsistent with some other studies [1, 2, 33]. Therefore, the predominance of male patients in orbital lymphoma requires to be further investigated by future clinical studies.

MR imaging features suggested orbital lymphomas had unilateral involvement compared to benign OLPDs, which was consistent with previous data [1, 2, 6–8, 33, 34]. Moreover, orbital lymphomas were frequently located around organs, such as eyeballs, and compress optic nerves, which might explain the decreased visual acuity, eye irritation, excessive tearing, and pain in the eye in these patients [6, 8, 34]. In addition, when comparing the signal intensity with that of the cerebral cortex, orbital lymphomas showed isointense in T1WI and hypointense signals in T2WI. At the same time, IgG4-RODs had iso- or hyperintense signals in T1WI

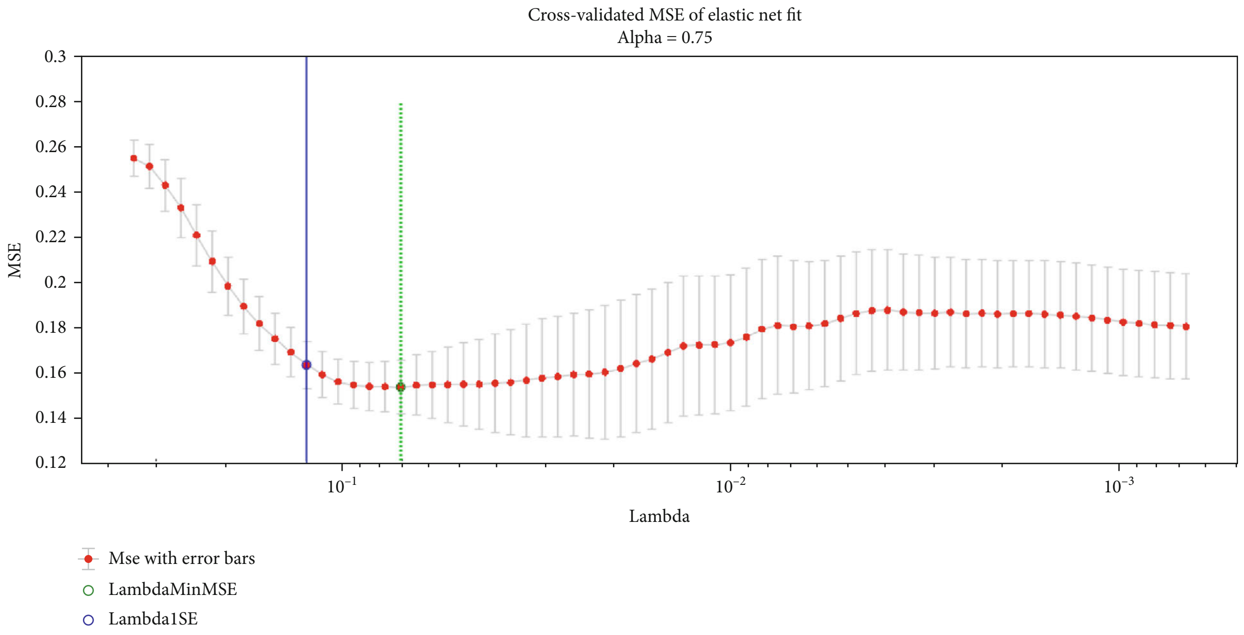


Figure 3: Automated optimization of the parameter λ when the training samples are fitted by elastic net using 10-folder cross-validation (CV). The x-axis shows the change of λ, and the y-axis indicates the mean square error (MSE). The green dotted line locates the λ with minimum CV error, and the solid blue line points to the minimum CV error plus one standard deviation (SE).

Table 4: Disease classification using single or multiple modality data.

	AUC	ACC	SEN	SPE
T1WI	0.54 ± 0.10	0.53 ± 0.13	0.79 ± 0.16	0.29 ± 0.20
T2WI	0.63 ± 0.12	0.62 ± 0.13	0.79 ± 0.11	0.46 ± 0.18
CE-T1WI	0.74 ± 0.10	0.74 ± 0.11	0.81 ± 0.16	0.67 ± 0.16
CE-T1WI-cor	0.72 ± 0.10	0.72 ± 0.11	0.83 ± 0.14	0.61 ± 0.21
T1WI + T2WI + CE-T1WI	0.70 ± 0.12	0.70 ± 0.12	0.79 ± 0.10	0.62 ± 0.14
T1WI + T2WI + CE-T1WI-cor	0.71 ± 0.12	0.69 ± 0.14	0.85 ± 0.10	0.58 ± 0.16
T1WI + T2WI + CE-T1WI + CE-T1WI-cor	0.78 ± 0.10	0.77 ± 0.10	0.82 ± 0.14	0.74 ± 0.17
CE-T1WI + CE-T1WI-cor	0.79 ± 0.11	0.78 ± 0.11	0.82 ± 0.15	0.76 ± 0.19

and hypointense signals in T2WI. As to the shape of margins, most IgG4-RODs were well-defined, which was verified by prior disease classification [7]. However, these MR findings, nonspecific, qualitative, and subjective, could be found between orbital lymphoma and other non-IgG4-ROD. Thus, these nondiscriminative features might require other advanced imaging modalities for deeper understanding.

Experimental results highlighted the importance of CE-T1WI for disease classification. CE-T1WI achieves superior performance, and in combination with CE-T1WI-cor and preselection of features, it could further improve the diagnostic performance. Contrast-enhanced T1-weighted MR imaging was highlighted in this study. Six discriminative features (5 from CE-T1WI and 1 from CE-T1WI-cor) were retrieved. As these features were quantitative and meaningful, they can help understand the machine learning-based radiomics models. On the other hand, two studies explored machine learning methods for the quantitative analysis of ocular adnexal lymphoma and idiopathic orbital inflammation. Guo et al. discovered that five features (4 from CE-T1WI and 1 from T2WI) achieved a larger AUC (> 0.70) [35]. Hou and his colleagues found bag-of-words features from CE-T1WI may significantly outperform the features from no-enhanced MR images [36]. In general, both studies indirectly provided support for our findings, suggesting that contrast-enhanced MR imaging may improve the differentiation between orbital lymphoma and IgG4-ROD.

To our knowledge, this is the first study that aimed at building a machine learning model for the differentiation of orbital lymphoma and IgG4-ROD. The elastic net is the backbone of the proposed radiomics model. It retrieves informative features for data representation and also acts as the classifier for disease prediction. When analyzing the performance of single- and multimodal data, CE-T1WI resulted as the most informative. To reduce the feature number, improve the prediction performance, and enhance the model interpretability, feature preselection via statistical comparison was conducted, and a handful of features were identified.

Table 5: Disease classification with feature preselection.

Wilcoxon rank-sum test	AUC	ACC	SEN	SPE
T1WI	0.57 ± 0.09	0.55 ± 0.11	0.74 ± 0.16	0.39 ± 0.19
T2WI	0.59 ± 0.12	0.58 ± 0.13	0.80 ± 0.14	0.38 ± 0.18
CE-T1WI	0.73 ± 0.11	0.73 ± 0.11	0.78 ± 0.15	0.68 ± 0.19
CE-T1WI-cor	0.74 ± 0.10	0.73 ± 0.11	0.86 ± 0.12	0.61 ± 0.21
T1WI + T2WI + CE-T1WI	0.75 ± 0.11	0.74 ± 0.11	0.80 ± 0.11	0.69 ± 0.12
T1WI + T2WI + CE-T1WI-cor	0.70 ± 0.10	0.69 ± 0.12	0.83 ± 0.10	0.58 ± 0.16
T1WI + T2WI + CE-T1WI + CE-T1WI-cor	0.78 ± 0.11	0.78 ± 0.10	0.83 ± 0.13	0.73 ± 0.18
CE-T1WI + CE-T1WI-cor	0.82 ± 0.09	0.81 ± 0.09	0.84 ± 0.12	0.79 ± 0.14

The present study proposed a radiomics model, which revealed the importance of CE-T1WI in the classification and might further be used to screen and diagnose eye diseases.

This study also has a few limitations. First, T1WI, T2WI, and CE-T1WI are conventional MR imaging modalities, yet other modalities, such as DWI and CT, and some other parameters, such as contrast-enhancement ratio and ADC, should also be considered. Second, a limited number of features were collected for tumor description; more features should be collected to quantify mass lesions from various perspectives. Third, this study applied elastic net for feature selection and disease diagnosis. Several other approaches, such as feature ranking methods [37], can be used for feature selection in this binary classification task. Finally, the sample size was small, and large-scale studies are required to confirm these findings.

5. Conclusion

In the present study, several quantitative MR features were identified as relevant for differentiation of orbital lymphoma and IgG4-ROD. The machine learning-based radiomics model verified that contrast-enhanced T1 MR imaging was discriminative in disease classification. The next step is to incorporate other modalities and advanced techniques to further explore the differences between diseases.

Data Availability

The MR images that support this study's findings are restricted by the Medical Ethics Committee of the Second Hospital of Jilin University to protect patient privacy. Requests for access to the data sets or the radiomics features can be made to the corresponding author Qinghe Han (hanqinghehe@126.com).

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Ye Yuan and Guangyu Chu contributed equally to this work and co-first authors.

Acknowledgments

This study is supported by the Jilin Province Education Department (JJKH20201062KJ) and the Department of Science and Technology of Jilin Province (20200404200YY).

References

[1] K. Haradome, H. Haradome, Y. Usui et al., "Orbital lymphoproliferative disorders (OLPDs): value of MR imaging for differentiating orbital lymphoma from benign OPLDs," *American Journal of Neuroradiology*, vol. 35, no. 10, pp. 1976–1982, 2014.

[2] X. Xu, H. Hu, H. Liu et al., "Benign and malignant orbital lymphoproliferative disorders: differentiating using multiparametric MRI at 3.0T," *Journal of Magnetic Resonance Imaging*, vol. 45, no. 1, pp. 167–176, 2017.

[3] G. E. Valvassori, S. S. Sabnis, R. F. Mafee, M. S. Brown, and A. Putterman, "Imaging of orbital lymphoproliferative disorders," *Radiologic Clinics of North America*, vol. 37, no. 1, pp. 135–150, 1999.

[4] A. M. Eckardt, J. Lemound, M. Rana, and N.-C. Gellrich, "Orbital lymphoma: diagnostic approach and treatment outcome," *World journal of surgical oncology*, vol. 11, no. 1, p. 73, 2013.

[5] D. Briscoe, C. Safieh, Y. Ton, H. Shapiro, E. I. Assia, and D. Kidron, "Characteristics of orbital lymphoma: a clinicopathological study of 26 cases," *International Ophthalmology*, vol. 38, no. 1, pp. 271–277, 2018.

[6] G. Priego, C. Majos, F. Climent, and A. Muntane, "Orbital lymphoma: imaging features and differential diagnosis," *Insights Into Imaging*, vol. 3, no. 4, pp. 337–344, 2012.

[7] C. W. Jin, N. Rana, Y. Wang, S. H. Ma, M. Li, and M. Zhang, "CT and MRI features of primary orbital lymphoma: review of 14 cases," *Asian Journal of Medical Sciences*, vol. 1, no. 2, pp. 87–90, 2010.

[8] G. Gerbino, P. Boffano, R. Benech et al., "Orbital lymphomas: clinical and radiological features," *Journal of Cranio-Maxillofacial Surgery*, vol. 42, no. 5, pp. 508–512, 2014.

[9] W.-K. Yu, C.-C. Tsai, S.-C. Kao, and C. J.-L. Liu, "Immunoglobulin G4-related ophthalmic disease," *Taiwan journal of ophthalmology*, vol. 8, no. 1, pp. 9–14, 2018.

[10] T. Kubota and M. Suzuko, "Orbital IgG4-related disease: clinical features and diagnosis," *ISRN rheumatology*, vol. 2012, 5 pages, 2012.

[11] J. Soussan, R. Deschamps, J. C. Sadik et al., "Infraorbital nerve involvement on magnetic resonance imaging in European patients with IgG4-related ophthalmic disease: a specific sign," *European Radiology*, vol. 27, no. 4, pp. 1335–1343, 2017.

[12] A. R. Sepahdari, L. S. Politi, V. K. Aakalu, H. J. Kim, and A. A. K. A. Razek, "Diffusion-weighted imaging of orbital masses: multi-institutional data support a 2-ADC threshold model to categorize lesions as benign, malignant, or indeterminate," *American Journal of Neuroradiology*, vol. 35, no. 1, pp. 170–175, 2014.

[13] A. R. Sepahdari, R. Kapur, V. K. Aakalu, J. P. Villablanca, and M. F. Mafee, "Diffusion-weighted imaging of malignant ocular masses: initial results and directions for further study," *American Journal of Neuroradiology*, vol. 33, no. 2, pp. 314–319, 2012.

[14] S. M. ElKhamary, A. Galindo-Ferreiro, L. AlGhafri, R. Khandekar, and S. A. Schellini, "Characterization of diffuse orbital mass using apparent diffusion coefficient in 3-tesla MRI," *European Journal of Radiology Open*, vol. 5, pp. 52–57, 2018.

[15] A. Lecler, L. Duron, M. Zmuda et al., "Intravoxel incoherent motion (IVIM) 3 T MRI for orbital lesion characterization," *European Radiology*, vol. 31, no. 1, pp. 1–10, 2020.

[16] F. R. Maldonado, J. P. Princich, L. Micheletti, M. S. Toronchik, J. I. Erripa, and C. Rugilo, "Quantitative characterization of extraocular orbital lesions in children using diffusion-weighted imaging," *Pediatric Radiology*, vol. 51, no. 1, pp. 1–9, 2020.

[17] G. J. Ben Simon, C. C. Annunziata, J. Fink, P. Villablanca, J. D. McCann, and R. A. Goldberg, "Rethinking orbital imaging: establishing guidelines for interpreting orbital imaging studies and evaluating their predictive value in patients with orbital tumors," *Ophthalmology*, vol. 112, no. 12, pp. 2196–2207, 2005.

[18] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[19] S. S. F. Yip and J. W. L. A. Hugo, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, pp. R150–R166, 2016.

[20] V. Kumar, Y. Gu, S. Basu et al., "Radiomics: the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[21] J. Lao, Y. Chen, Z.-C. Li et al., "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme," *Scientific Reports*, vol. 7, no. 1, pp. 1–8, 2017.

[22] S. Yu, L. L. Liu, Z. Y. Wang, G. Z. Dai, and Y. Q. Xie, "Transferring deep neural networks for the differentiation of mammographic breast lesions," *SCIENCE CHINA Technological Sciences*, vol. 62, no. 3, pp. 441–447, 2019.

[23] X. Chen, M. Zeng, Y. Tong et al., "Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis," *Bio Med Research International*, vol. 2020, 2020.

[24] Z.-C. Li, G. Zhai, J. Zhang et al., "Differentiation of clear cell and non-clear cell renal cell carcinomas by all-relevant radiomics features from multiphase CT: a VHL mutation perspective," *European Radiology*, vol. 29, no. 8, pp. 3996–4007, 2019.

[25] J. Dong, M. Yu, Y. Miao et al., "Differential diagnosis of solitary fibrous tumor/hemangiopericytoma and angiomatous meningioma using three-dimensional magnetic resonance imaging texture feature model," *BioMed Research International*, vol. 2020, 8 pages, 2020.

[26] F. Valdora, N. Houssami, F. Rossi, M. Calabrese, and A. S. Tagliafico, "Rapid review: radiomics and breast cancer," *Breast Cancer Research and Treatment*, vol. 169, no. 2, pp. 217–229, 2018.

[27] Y. Guan, P. Wang, Q. Wang et al., "Separability of acute cerebral infarction lesions in CT based radiomics: toward artificial intelligence-assisted diagnosis," *BioMed Research International*, vol. 2020, 8 pages, 2020.

[28] H. Goto, Japanese Study Group for IgG4-Related Ophthalmic Disease, M. Takahira, and A. Azumi, "Diagnostic criteria for IgG4-related ophthalmic disease," *Japanese Journal of Ophthalmology*, vol. 59, no. 1, pp. 1–7, 2015.

[29] J. J. M. van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[30] H. Zou and H. Trevor, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[32] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A technical review of convolutional neural network-based mammographic breast cancer diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2019, 16 pages, 2019.

[33] J. Ren, Y. Yuan, Y. Wu, and X. Tao, "Differentiation of orbital lymphoma and idiopathic orbital inflammatory pseudotumor: combined diagnostic value of conventional MRI and histogram analysis of ADC maps," *BMC Medical Imaging*, vol. 18, no. 1, p. 6, 2018.

[34] T. G. Olsen and H. Steffen, "Orbital lymphoma," *Survey of Ophthalmology*, vol. 64, no. 1, pp. 45–66, 2019.

[35] J. Guo, Z. Liu, C. Shen et al., "MR-based radiomics signature in differentiating ocular adnexal lymphoma from idiopathic orbital inflammation," *European Radiology*, vol. 28, no. 9, pp. 3872–3881, 2018.

[36] Y. Hou, X. Xie, J. Chen et al., "Bag-of-features-based radiomics for differentiation of ocular adnexal lymphoma and idiopathic orbital inflammation from contrast-enhanced MRI," *European Radiology*, vol. 31, no. 1, pp. 1–10, 2020.

[37] Z. Zhang, X. Liang, W. Qin, S. Yu, and Y. Xie, "matFR: a MATLAB toolbox for feature ranking," *Bioinformatics*, vol. 36, no. 19, pp. 4968-4969, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 6621894, 9 pages
https://doi.org/10.1155/2021/6621894

Research Article

Development and Validation of a Radiomics Nomogram for Prognosis Prediction of Patients with Acute Paraquat Poisoning: A Retrospective Cohort Study

Shan Lu, Duo Gao, Yanling Wang, Xuran Feng, Yongzhi Zhang, Ling Li, and Zuojun Geng

Department of Medical Imaging, The Second Hospital of Hebei Medical University, Shijiazhuang 050000, China

Correspondence should be addressed to Zuojun Geng; 1980756261@qq.com

Received 10 November 2020; Revised 6 January 2021; Accepted 25 January 2021; Published 2 February 2021

Academic Editor: Zhiguo Zhou

Objective. To evaluate the efficiency of a radiomics model in predicting the prognosis of patients with acute paraquat poisoning (APP). *Materials and Methods*. Chest computed tomography images and clinical data of 80 patients with APP were obtained from November 2014 to October 2017, which were randomly assigned to a primary group and a validation group by a ratio of 7 : 3, and then the radiomics features were extracted from the whole lung. Principal component analysis (PCA) and least absolute shrinkage and selection operator (LASSO) regression were used to select the features and establish the radiomics signature (Rad-score). Multivariate logistic regression analysis was used to establish a radiomics prediction model incorporating the Rad-score and clinical risk factors; the model was represented by nomogram. The performance of the nomogram was confirmed by its discrimination and calibration. *Result*. The area under the ROC curve of operation was 0.942 and 0.865, respectively, in the primary and validation datasets. The sensitivity and specificity were 0.864 and 0.914 and 0.778 and 0.929, and the prediction accuracy rates were 89.5% and 87%, respectively. Predictors included in the individualized predictive nomograms include the Rad-score, blood paraquat concentration, creatine kinase, and serum creatinine. The AUC of the nomogram was 0.973 and 0.944 in the primary and validation datasets, and the sensitivity and specificity were 0.943 and 0.955, respectively, in the primary dataset and 0.889 and 0.929 in the validation dataset, and the prediction accuracy was 94.7% and 91.3%, respectively. *Conclusion*. The radiomics nomogram incorporates the radiomics signature and hematological laboratory data, which can be conveniently used to facilitate the individualized prediction of the prognosis of APP patients.

1. Introduction

Although some countries have banned the use of paraquat (PQ), paraquat can still be obtained on the market by other forms of preparations. In recent years, the incidence rate of paraquat poisoning is still high in some areas of China, and paraquat poisoning has become the first cause of death of poisoning. After ingestion, PQ is rapidly absorbed and distributed to the lung, liver, kidney, and muscle, and if left untreated, the accumulation of PQ can cause fulminant multiple organ failure, including pulmonary edema and heart, kidney, and liver failure [1], with a mortality rate of up to 50%~90% [2]. However, there is still no effective antidote. PQ mainly accumulates in the lung, where it is retained even when blood levels start to decrease, resulting in a free radical build-up that triggers inflammatory responses and leading to lung fibrosis [3, 4]. Lung damage and respiratory failure are common causes of death [5, 6]. Although many studies suggested that the lung injury caused by paraquat is irreversible, a case study, in fact, by Lee et al. [7] showed that lung damage may not be irreversible if treated in time. Thus, evaluation of lesions in the lungs and their severity at the early stage of poisoning may be crucial to guide the clinical adjustment of the treatment plan and improve patient outcomes.

Chest computed tomography (CT) has been demonstrated to be useful in detecting early lung lesions and

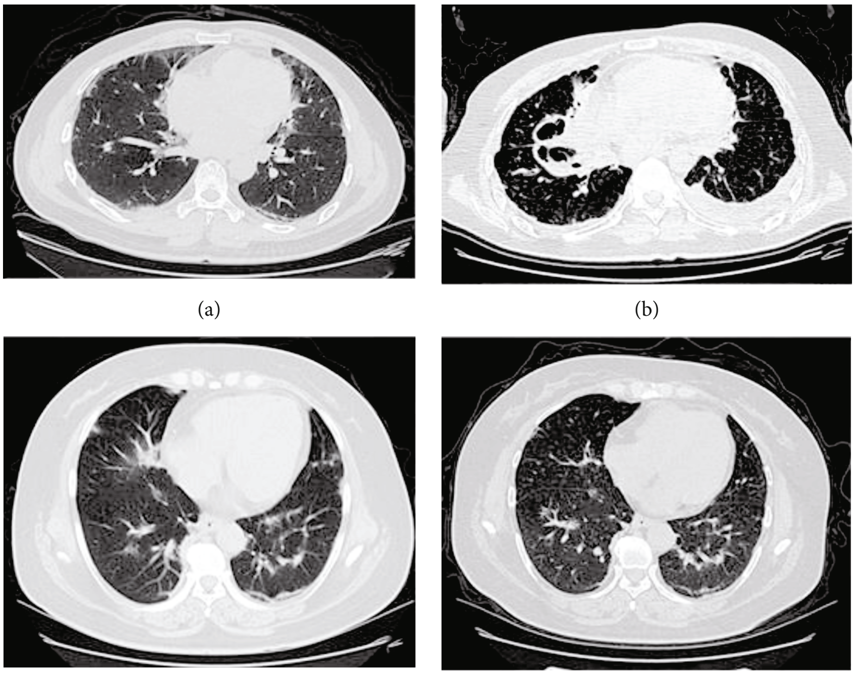


Figure 1: (a, b) CT images of paraquat poisoning in a 53-year-old man at the beginning and 2 months later. (c, d) CT images of paraquat poisoning in a 40-year-old woman at the beginning and 2 months later.

assessing long-term damage in PQ-poisoned survivors [1]. In the 1990s, Im et al. [8] and Lee et al. [4] described the radiologic high-resolution CT (HRCT) manifestations of PQ-induced pulmonary damage, with special emphasis on the sequential changes, but without quantitative studies. Recently, the number of injured lung segments and the volume or area ratio of gross glass density shadow (GGO) found in CT examination in patients with acute PQ poisoning have been used to predict the prognosis of PQ poisoning [5, 6, 9]; although these studies obtained a certain accuracy rate, their observation object was limited to a single injury sign, which lacked estimation of the total lung injury and ignored a large part of the CT image information.

Although studies have shown that many blood laboratory indicators can also be used to predict the prognosis of patients with PQ poisoning [10–17], most of these studies used only one or several indicators that were almost lung nonspecific, which cannot effectively reflect the major causes of death of APP patients: injury of the lung. A comprehensive predictive model, which combines CT lung injury signs and blood laboratory indicators, to evaluate multisystem injury or functional failure is yet to be developed. Previous studies have also shown that objective and quantitative imaging descriptors could potentially be used as prognostic or predictive biomarkers. The combined analysis of a panel of biomarkers, rather than individual analyses, as a signature is the most promising approach that is powerful enough to change clinical management [18–20]. As is shown in Figure 1, the two patients had similar lung damage at the initial stage of poisoning. However, CT examination showed that the severity of pulmonary disease was different after 2 months of follow-up so as the prognosis. Radiomics, as one of the most representative methods, is the process of the conversion of medical images into high-dimensional, mineable data via high-throughput extraction of quantitative features, followed by subsequent data analysis for decision support, which has been demonstrated useful in many kinds of focal lesions [21, 22]. However, to our best knowledge, rare radiomics applications for diffuse lesions were reported yet.

Therefore, the purpose of this study is to explore the feasibility of radiomics for the study of diffuse inflammatory diseases and to develop and validate a nomogram based on CT radiomics features and clinical prognostic risk factors for predicting the prognosis of patients with APP.

2. Patients and Methods

This study was approved by the Hospital Ethics Committee, and the requirement for written informed consent was waived.

2.1. Participants. Initial clinical baseline data and CT examination images of acute paraquat-poisoned patients, who were admitted to the emergency department from November 2014 to October 2017 and received individualized comprehensive treatment (Data Supplement (available here)), were collected. The patient screening process is shown in Figure 2. Data Supplement presents the inclusion and exclusion criteria.

The initial clinical baseline data of the poisoned patients included the following: age, gender, PQC, and blood routine and biochemical indicators within 24 hours, which included white blood cell count (WBC), high-sensitivity C-reactive protein (hsCRP), lactate dehydrogenase (LDH), creatine kinase isoenzyme (CK-MB), alanine aminotransferase

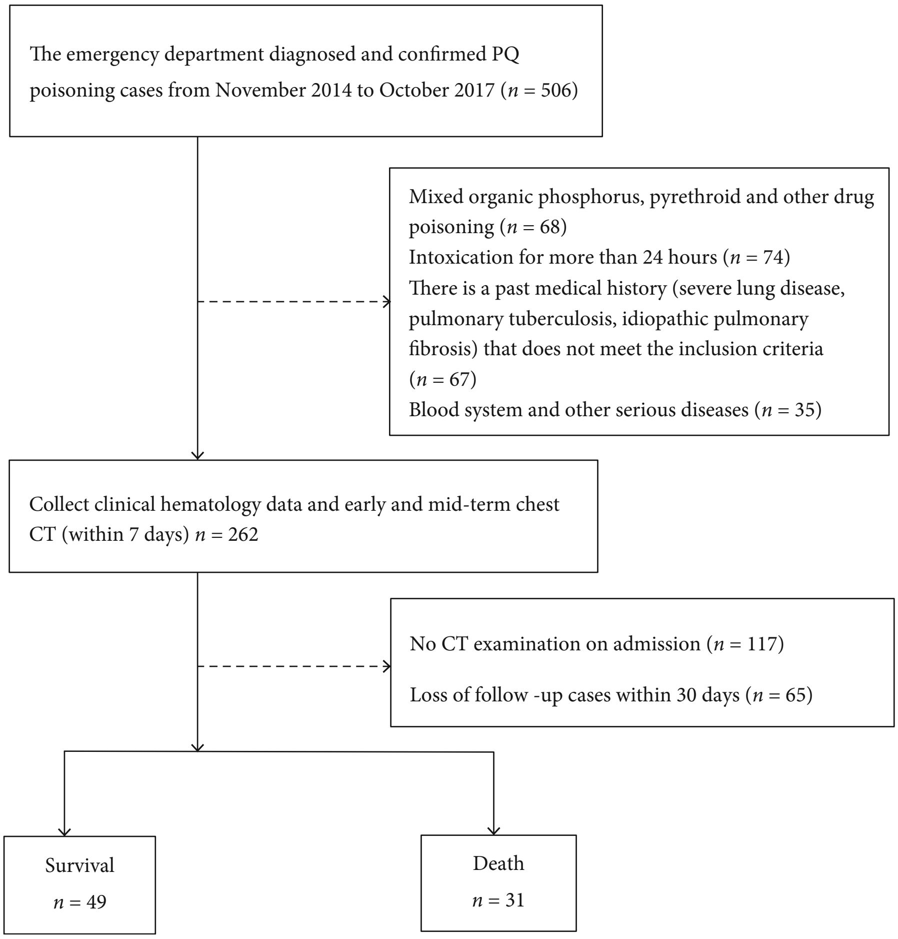

The emergency department diagnosed and confirmed PQ poisoning cases from November 2014 to October 2017 (n = 506)
Mixed organic phosphorus, pyrethroid and other drug poisoning (n = 68)
Intoxication for more than 24 hours (n = 74)
There is a past medical history (severe lung disease, pulmonary tuberculosis, idiopathic pulmonary fibrosis) that does not meet the inclusion criteria (n = 67)
Blood system and other serious diseases (n = 35)
Collect clinical hematology data and early and mid-term chest CT (within 7 days) n = 262
No CT examination on admission (n = 117)
Loss of follow -up cases within 30 days (n = 65)
Survival n = 49
Death n = 31

Figure 2: Flow chart of study enrollment.

(ALT), aspartate aminotransferase (AST), albumin (ALB), urea, creatinine (Cr), amylase (AMY), and glucose (GLU), a total of 12 indicators.

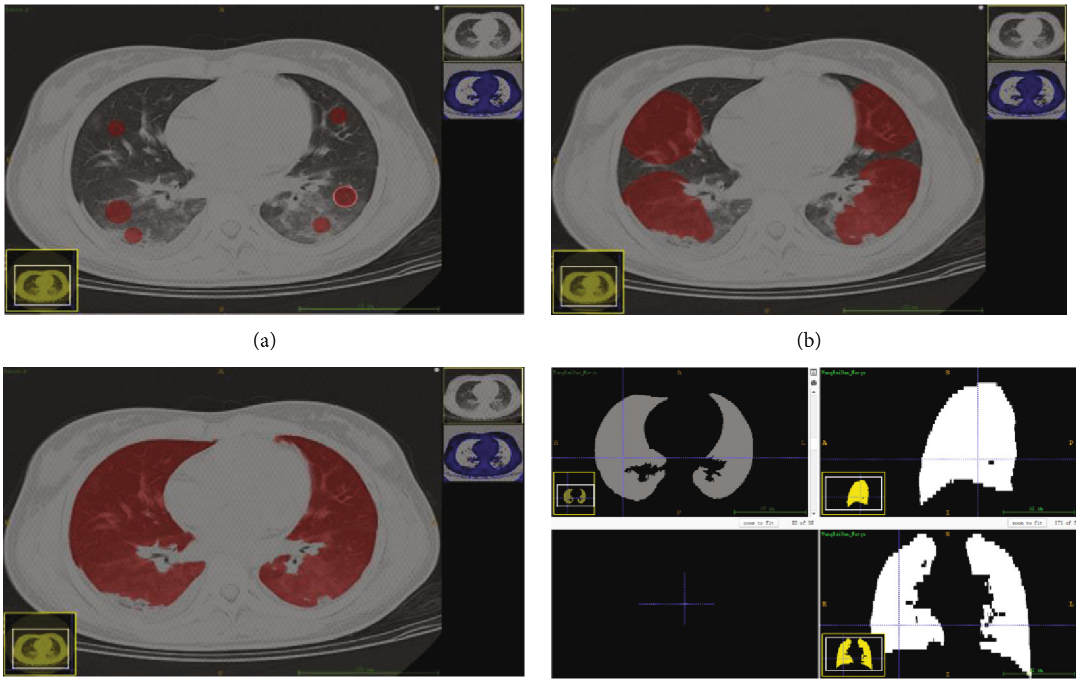
Finally, 80 patients were included in the study. According to the follow-up outcome of 30 days after PQ ingestion, patients were divided into the survival group (>30 days) and the death group. All 80 patients were randomly divided into two groups according to a ratio of 7 : 3.

2.2. CT Image Acquisition. Chest CT examinations were performed using a GE LightSpeed/16-slice scanner. CT scanning parameters were the same as those of the chest: 120 kV, 100 mA, 5 mm thickness and slice interval, and standard lung window (window width, 1500 HU; window level, -700 HU) were selected. Within 7 days after taking PQ, a chest CT examination was performed every average of 3 days.

2.3. Image Segmentation: ROI Drawing Methods and Modification Criteria. We used the region growing method in the ITK-SNAP software (version 3.6.0, https://www.itksnap.org) to sketch the whole lung as the ROI, which was then manually modified by two physicians with licensed physician qualifications. The interobserver correlation coefficients (ICCs) were used to assess the agreement of radiomics features by two-level radiologists. Data Supplement presents the ROI drawing methods and modification criteria in Figure 3.

The region growing method is mainly divided into three steps. (1) The seed points were selected from the seed area that can represent the extraction area, and the seed was a small area including a couple of pixels. (2) Determine the criteria for region growing and measure whether the pixels adjacent to the seed point meet the criteria. The standards outlined in this study were as follows: the lower threshold was -1200 HU, and the upper threshold was -100 HU. (3) Stop growing [18]. After region growing, the boundary between the apex of the lung and the edge of the lung needed manual modification.

The criteria for manual modification were as follows. (1) In the boundary between the chest wall and the lung, the lesion-free areas were automatically outlined without modification; those areas with lesions (but the lesions were not totally included in the ROI) were manually modified. (2) If the demarcation of lung atelectasis caused by pleural effusion was unclear, automatic delineation of results was used without manual modification. (3) For the higher density of lung lesions, such as cords and nodules, which were not covered in the ROI, manual delineation was applied. (4) For lung lesions that were not included in the ROI automatically, manual delineation was applied. (5) For the vascular and



(a)

(b)

(c)

(d)

Figure 3: Image segmentation diagram. (a) Seed points were selected in three higher density lesions, lower density lesions, and normal lung tissue in both lungs. (b) The seed points began to grow. (c) Growth of seed points completed basically. (d) ROI obtained after manual modification.

bronchi, the ROI contained no main and leaf bronchi; if the segmental and inferior bronchi were connected to pixels that were distinguishable by the naked eye, we did not sketch them into the ROI. Otherwise, we sketched them into the ROI; the small scattered bronchus of lungs was contained in the ROI. (6) For those lesions with a poor borderline in the hilum, the principle was not missing lesions as far as possible. (7) For the apex and bottom of the lung, slices without lung tissue were removed manually; slices with lung tissue but only scattered pixels in the border were included; we modified the ROI to the edge of the lung tissue manually.

2.4. Radiomics Feature Extraction. Analysis Kit software (GE Healthcare, Life Sciences, China) was utilized to extract the radiomics features. A total of 385 radiomics features, including 42 histogram features, 154 grey-level cooccurrence matrix (GLCM) features, 180 run-length matrix (RLM) features, and 11 grey-level zone size matrix (GLZSM) features, were extracted from the ROI. Details of the radiomics feature extraction methodology and the individual parameters can be found in the Data Supplement. The interobserver correlation coefficient (ICC) between two radiologists' agreement is 0.823 (0.762 to 0.971, 95% CI).

2.5. Feature Selection and Radiomics Signature Building. The principal component analysis (PCA) and the least absolute shrinkage and selection operator (LASSO) method were used to select the most useful predictive features from the primary cohort. A radiomics signature (here we called the Rad-score) was calculated for each patient via a linear combination of selected features that were weighted by their respective coefficients.

2.6. Radiomics Signature Validation. We evaluated the ability of the Rad-score to differentiate survival and death in the primary cohort and then validated it in the validation cohort. Sensitivity, specificity, and AUC (area under the ROC curve) were used to evaluate the diagnostic efficiency. The diagnostic accuracy rate was shown as a color bar chart.

2.7. Development of an Individualized Prediction Model. Statistical analysis and ROC curve analysis were performed for each initial clinical baseline data, and backward logistic regression was used to select clinical risk factors to be included in the nomogram.

An individualized prediction model was established based on the primary dataset by incorporating the radiomics signature with the clinical risk factors. And it was presented with a radiomics nomogram so as to provide the clinicians with a quantitative tool to predict prognosis. Calibration curves were plotted to assess the calibration of the radiomics nomogram. Decision curve analysis (DCA) was conducted to determine the clinical usefulness of the radiomics nomogram by quantifying the net benefits at different threshold probabilities in the testing dataset.

2.8. Statistical Analysis. Statistical analyses were performed by using SPSS 21.0.$P < 0.05$ was considered statistically significant. A chi-squared test was used for the comparison of count data. Measurement data were compared by using the independent-samples t-test if the data satisfied the normal distribution, otherwise by using the Mann-Whitney *U* test. Measurement data were generally expressed as the mean ± standard deviation or the median and the interquartile range according to whether satisfying normal distribution.

Feature selection and model building were conducted with R software (version 3.3.2; http://www.Rproject.org).

3. Results

3.1. Clinical Risk Factor Selection. The statistical test results of demography and initial blood laboratory data and are shown in Table 1. The results showed that PQA, PQC, WBC, CK-MB, LDH, Cr, and GLU were statistically significant among the survival and death groups ($P < 0.05$), and the ROC curve showed AUC of PQA, PQC, WBC, CK-MB, and Cr were all above 0.7. Finally, PQC, CK-MB, and Cr were selected by backward logistic regression to be included in the nomogram.

Table 1: The demography and initial blood laboratory.

Factors		Survival group	Death group	*P* value	AUC
Age (years)		34.29 ± 10.98	37.61 ± 14.38	0.468	—
Gender	Male	25 (51%)	18(58.1%)	—	—
	Female	24 (49%)	13(41.9%)	—	—
PQC (mL)		3.84 (3.12)	8.60 (4.93)	0.000	0.804
WBC ($\times 10^9$/L)		10.60 (4.15)	14.90 (8.30)	0.000	0.759
hsCRP (mg/L)		1.05 (3.78)	1.90 (6.70)	0.410	0.564
LDH (U/L)		205.50 (54.15)	228.55 (58.68)	0.031	0.698
CK-MB (U/L)		17.00 (5.83)	25.15 (15.23)	0.000	0.759
ALT (U/L)		16.55 (13.53)	19.15 (11.68)	0.253	0.562
AST (U/L)		19.95 (7.08)	22.35 (9.88)	0.277	0.550
ALB (g/L)		46.30 (4.53)	45.50 (8.95)	0.445	0.483
K (mmol/L)		3.71 (0.45)	3.56 (0.63)	0.091	0.387
Urea (mmol/L)		3.90 (2.40)	5.05 (2.10)	0.078	0.633
Cr (μmol/L)		63.00 (19.20)	86.10 (41.00)	0.000	0.732
AMY (U/L)		109.50 (164.65)	92.00 (137.00)	0.698	0.468
GLU (mmol/L)		6.38 (1.69)	7.05 (2.24)	0.024	0.699

3.2. Feature Selection and Radiomics Signature Building. Among the 385 original features from the primary dataset extracted, 23 constant terms were deleted first; 8 features with a cumulative variance contribution rate of 95% were retained after PCA (Appendix Figure A1 is given in the Data Supplement). The seven most relevant features were finally selected using LASSO, which gave the minimum mean classification error of cross-validation (Figures 4(a) and 4(b)).

3.3. Diagnostic Validation of the Radiomics Signature. ROC curves were plotted to evaluate the diagnostic efficiency of the logistic regression models (Figure 5(a)). The accuracy of the Rad-score is shown in Table 2. Distributions of the Rad-score and prognosis status in the primary and validation cohorts are given in the Data Supplement Appendix Figure A2.

3.4. Development of Individualized Prediction Comprehensive Models. Incorporated clinical factors included PQC, CK-MB, and Cr with the Rad-score; using multivariable logistic regression analysis, an individualized prediction model was built and is shown as a nomogram in Figure 6. The ROC curves were plotted to evaluate the diagnostic efficiency of the comprehensive model and are shown in Figure 5(b) and Table 2.

3.5. Clinical Use. The calibration curves of the primary dataset and validation dataset showed good agreement between prediction probability and real probability (Figure 7(a)). The decision curve showed that if the threshold probability of a patient or doctor is >10%, using the Rad-score to predict the prognosis of the patients adds more benefit than either the treat-all-patients scheme or the treat-none scheme. If the threshold probability exceeds 30%, the nomogram combining the Rad-score and clinical risk factors will be the best choice to maximize the net benefit (Figure 7(b)).

4. Discussion

Our study results revealed 385 radiomics features of pulmonary CT images, and we reduced them to 7 potential predictors and established the radiomics signature. The AUC of the primary dataset and validation dataset, respectively, were 0.942 (95% CI 0.886-0.997) and 0.865 (95% CI 0.658-1), and the sensitivity and specificity, respectively, were 0.864 and 0.914 and 0.778 and 0.929. The prediction accuracy of primary and validation datasets was 89.5% and 87%, respectively, which showed that the Rad-score had a good performance in the prediction of patient prognosis.

In previous studies about prognosis based on the pulmonary CT, Zhang et al. [5] found significantly fewer involved lung segments, or the presenting lesions were observed in baseline CT images (average admission 2.4 days) from the survivor group than the nonsurvivor group, indicating a smaller baseline disease extent in surviving patients. In their study, the sensitivity and specificity to predict prognosis were 72.2% and 28.6%, respectively, and the AUC was 0.767 (95% CI 0.656-0.878), based on the number of injured lung segments in the baseline CT examination. Their sensitivity and specificity were not very high for patient prognostic evaluation. Kim et al. [9] calculated the ratio of the sum of the areas of GGO at five levels (the top of the aortic arch, AP window, LUL bronchus, right inferior vein, and the top of the left diaphragm, respectively) and the sum of the area of the total lungs at the respective levels of pulmonary HRCT images 7 days after PQ ingestion, thinking that the area of GGO in the lung was an additional useful predictor for survival, especially when the PQ level was low. Kang et al. [6] calculated the maximum GGO volume ratio to the whole lung within the first 5 days after intoxication and showed that the AUC was 0.871 (95% CI 0.857-0.884), the sensitivity was 85.4%, the specificity was 89.3%, and the diagnostic accuracy was 87.6%. However, their study lacked independent validation; thus, the reliability of the obtained results needed to be further studied. Early lung injury of PQ intoxication mainly manifested as alveolitis, which was often shown as GGO and consolidation in pulmonary CT images. Therefore,

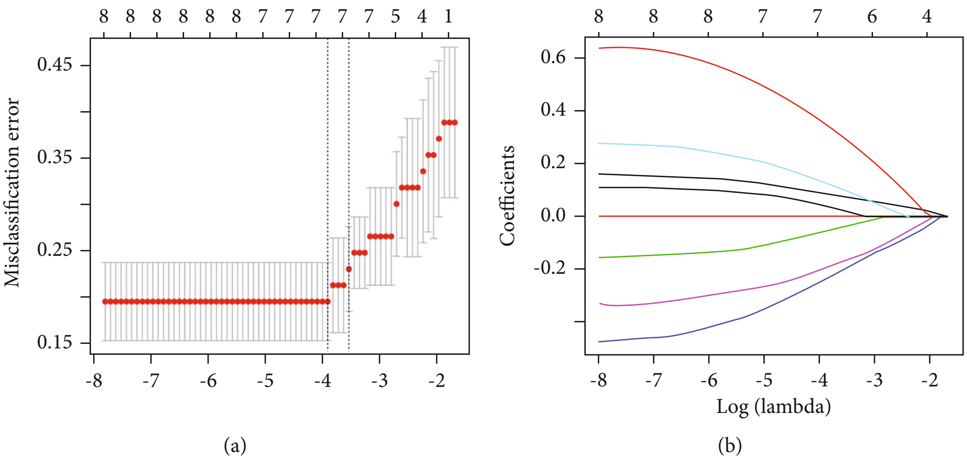


Figure 4: The number of features that LASSO selected after cross-validation. The underlined part is the value of log (lambda) and the number of features when the misclassification error is minimum.

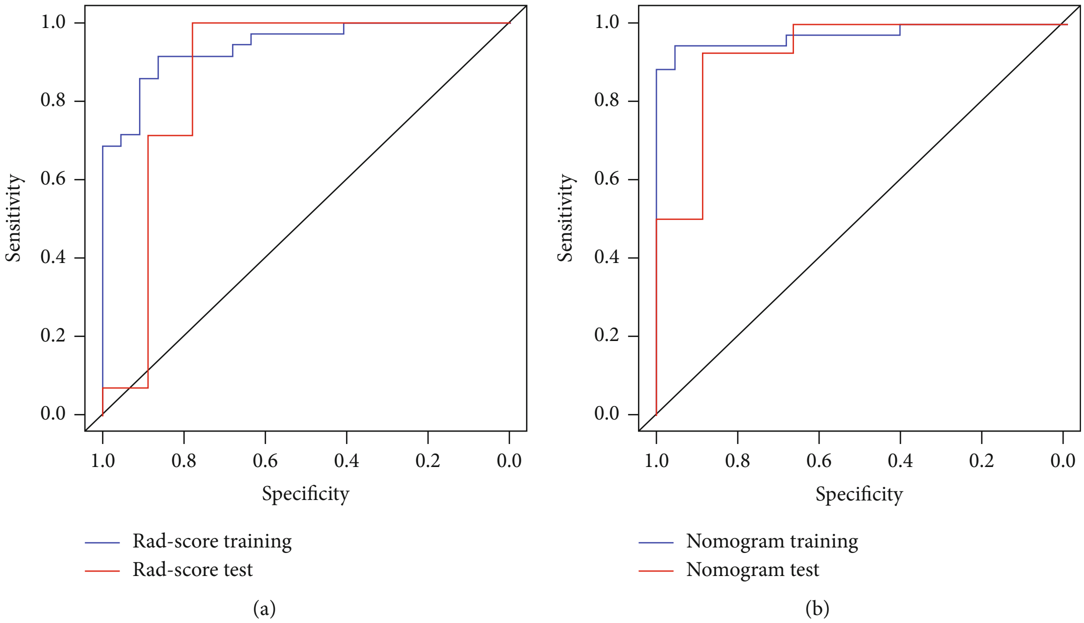


Figure 5: The ROC curve for the Rad-score and nomogram of the primary dataset and validation dataset.

Table 2: The accuracy of the Rad-score.

Information	Rad-score		Nomogram	
	Train dataset	Validation dataset	Train dataset	Validation dataset
AUC (95%)	0.942 (0.886-0.997)	0.865 (0.658-1)	0.973 (0.936-1)	0.944 (0.844-1)
ACC	0.895	0.87	0.947	0.913
Specificity	0.914	0.929	0.955	0.929
Sensitivity	0.864	0.778	0.943	0.889
Threshold	0.358	0.358	0.607	0.607

GGOs could reflect a certain extent of lung injury. The relatively accurate results of previous studies proved that the range of lung injury was an important factor for patient prognosis. However, the number of injured lung segments, GGO area ratio or volume ratio, could not completely reflect the extent of lung injury involving the whole lung and neglected other lung injuries that were not easily quantified, such as the thickening of bronchovascular bundles. In addition, all the GGO lesions in their study were manually delineated, resulting in large errors and poor consistency; and regarding calculating the area ratio or volume ratio, lesions and whole lungs needed to be delineated twice or even repeatedly examined.

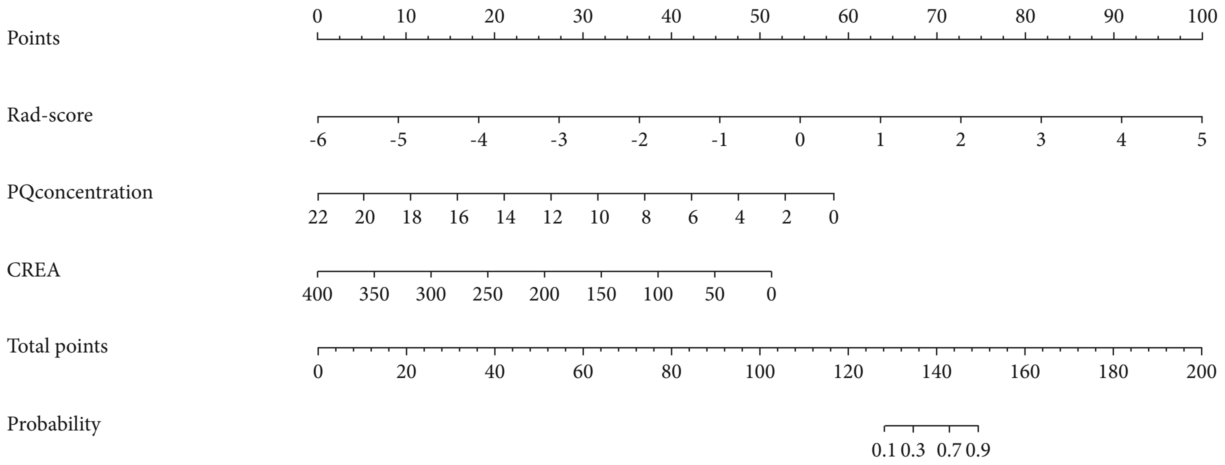


Figure 6: Radiomics nomogram was developed in the primary dataset.

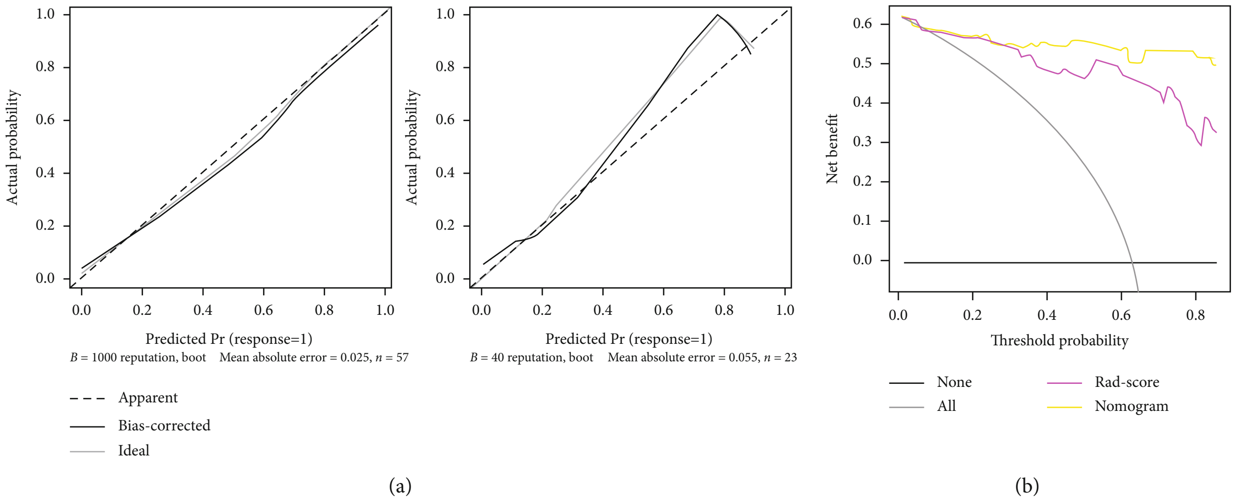


Figure 7: (a) Calibration curves of the radiomics nomogram. (b) Decision curve analysis for the radiomics nomogram.

Not only was the work inefficient, but it also further increased the error.

In this study, the region growing method was used to semiautomatically delineate the ROI. The whole lung was selected as the ROI of the CT images that the lung injury reached the peak (mainly 2-4-day images). Not only did it cover all the signs of lung injury we observed, but it was also easier to study ubiquitous lesions that are difficult to quantify, such as the thickening of the bronchovascular bundle. This provided a comprehensive measure of the extent and severity of lung injury, which would not ignore the microstructure changes that were invisible to naked eyes. Moreover, the more injury signs were observed in the same image, the more rapidly the lung injury developed, so the whole lung was selected as the ROI and was more scientific and rigorous.

In the early stage of lung injury caused by PQ poisoning, CT image mainly manifested as lung texture enhancement, GGO or consolidation, and was mainly distributed under the pleura. The features of density, range, and distribution of the above lung injuries may be the response of microstructural changes, including cell morphological changes and apoptosis, alveolar rupture and alveolar collapse, vascular basement membrane rupture, fibroblast precursor proliferation, and Clara cell migration [3, 23–27]. The Rad-score calculated based on the radiomics features that were extracted from CT images can effectively distinguish the different prognoses of the patients; thus, we guess that the radiomics features, such as the first-order histogram features and texture features, not only reflected the visible injuries by the naked eyes but also suggested the changes of the lung microstructure.

Among the laboratory data obtained at presentation, the levels of potassium, protein, arterial pH, $PaCO_2$, bicarbonate, albumin, amylase, AST, BUN, creatinine, and glucose were significantly related with prognosis by univariable analysis in a previous study [17]. However, among many similar studies, the strength of the correlation of various indicators with prognosis was different, which may be explained by the different equipment used, the follow-up time of prognosis, and patients' specificity. Our results showed that the PQC, CK-MB, and Cr were significantly different between the survival group and the death group. A large number of studies [28, 29] showed that the PQC was significantly associated with the prognosis of APP patients; our results also proved this point of view but, unfortunately, did not reach the same high correlation of prognosis compared with previous studies. PQ itself had direct nephrotoxicity; renal failure also impaired the excretion of PQ through the kidney; therefore, renal function injury may have a significant contribution to the mortality of APP [3]; the increase of Cr could suggest kidney injury [30]. CK-MB is the most specific and common indicator in the diagnosis of myocardial and skeletal muscle diseases, and a previous study that examined skeletal muscles obtained in both the biopsy and the autopsy of APP patients revealed extensive degeneration and fibrosis [3].

Compared with the single Rad-score, the nomogram that combined the clinical risk factors improved sensitivity, specificity, AUC, and diagnostic accuracy. The possible reason was that the CT image radiomics features mainly reflected the lung injury; by adding the clinical risk factors, the nomogram could reflect the damage of PQ to other tissues such as the liver, kidney, and muscle, so the performance of the model can be improved. However, the contribution of clinical risk factors was still lower than the radiomics signature, which indicated that the lung injury was the main prognostic factor in the early stage of poisoning.

In the previous studies about the mortality of APP patients, more attention was focused on lung nonspecific indicators. Many blood laboratory indicators were demonstrated to be useful in predicting the prognosis of patients with PQ poisoning [11, 14, 28]. These studies suggested that various laboratory indicators were related to prognosis in different degrees, but they all lacked independent validation. In a recent study [31], among 103 APP patients, aspartate aminotransferase, prothrombin time, prothrombin activity, total bilirubin, direct bilirubin, indirect bilirubin, alanine aminotransferase, urea nitrogen, and creatinine were found to be the most highly correlated indices in PQ poisoning and showed statistical significance ($P < 0.05$) in predicting PQ poisoning prognosis. Based on the above indicators, they established the grey wolf optimization-extreme learning machine (GWO-ELM) model. And the 10-fold cross-validation achieved a prediction accuracy of 81.45%, sensitivity of 81.24%, and specificity of 90.48%, respectively. Although the single-clinical factor model or multiclinical factor prediction model reached a certain accuracy, they were still lower than the prediction results of the Rad-score clinical model. This may be explained by two reasons; firstly, the baseline clinical data cannot specifically reveal the lung damage, which was the main cause of death; secondly, the data collection time was too early to fully reflect the damage of PQ toxicity to various organs. It was expected that lung CT images contained complementary and interchangeable information compared to other indexes, such as demographics, pathology, blood biomarkers, and genomics; combining the information would improve individualized treatment selection and monitoring [32].

This study has several limitations. Firstly, when choosing the ROI, mediastinal emphysema or pneumothorax and pleural effusion were not included; the main reason is that these signs may conceal the damage caused by PQ to the lung tissue, but the previous studies [6, 33] showed the appearance of mediastinal emphysema or pneumothorax, which suggested that the prognosis is bad and the mortality is high, so these signs' value of prognosis should be further studied. Secondly, in this study, the clinical risk factors of prognosis are not rich, such as urine PQ concentration, and arterial blood gas analysis was not included in the study, which was mainly restricted by hospital conditions. Whether there are significant differences in these clinical factors between the two groups and whether they can increase the performance of the prediction model need to be further discussed. Lastly, the relatively small sample number is another limitation of our study, which may have brought some deviation to the result, so it is necessary to make a further multicenter validation with a large number of samples in the future.

5. Conclusion

This study presents a radiomics nomogram that incorporates both the radiomics signature and the clinical risk factors and can be conveniently used to facilitate the individualized prediction of prognosis in patients with paraquat poisoning. Our study also proved that radiomics can also be applied to nontumor and diffuse diseases.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Shan Lu and Duo Gao contributed equally to this work.

Acknowledgments

The authors sincerely thank Ph.D. Jialiang Ren for data statistics.

Supplementary Materials

Supplementary Materials provides the individualized comprehensive treatment plan, the inclusion and exclusion criteria, and the radiomics feature extraction methodology. *(Supplementary Materials)*

References

[1] I. B. Gawarammana and N. A. Buckley, "Medical management of paraquat ingestion," *British Journal of Clinical Pharmacology*, vol. 72, no. 5, pp. 745–757, 2011.

[2] L. Senarathna, M. Eddleston, M. F. Wilks et al., "Prediction of outcome after paraquat poisoning by measurement of the plasma paraquat concentration," *monthly journal of the Association of Physicians*, vol. 102, no. 4, pp. 251–259, 2009.

[3] R. J. Dinis-Oliveira, J. A. Duarte, A. Sánchez-Navarro, F. Remião, M. L. Bastos, and F. Carvalho, "Paraquat poisonings: mechanisms of lung toxicity, clinical features, and

treatment," *Critical Reviews in Toxicology*, vol. 38, no. 1, pp. 13–71, 2008.

[4] S. H. Lee, K. S. Lee, J. M. Ahn, S. H. Kim, and S. Y. Hong, "Paraquat poisoning of the lung: thin-section CT findings," *Radiology*, vol. 195, no. 1, pp. 271–274, 1995.

[5] H. Zhang, P. Liu, P. Qiao et al., "CT imaging as a prognostic indicator for patients with pulmonary injury from acute paraquat poisoning," *British Journal of Radiolog*, vol. 86, no. 1026, pp. 2013–2035, 2013.

[6] X. Kang, D. Y. Hu, C. B. Li et al., "The volume ratio of ground glass opacity in early lung CT predicts mortality in acute paraquat poisoning," *PLoS One*, vol. 10, no. 4, article e0121691, 2015.

[7] E.-Y. Lee, K.-Y. Hwang, J.-O. Yang, and S.-Y. Hong, "Predictors of survival after acute paraquat poisoning," *Toxicology and Industrial Health*, vol. 18, pp. 201–206, 2006.

[8] J. G. Im, K. S. Lee, M. C. Han, S. J. Kim, and I. O. Kim, "Paraquat poisoning: findings on chest radiography and CT in 42 patients," *American Journal of Roentgenology*, vol. 157, no. 4, pp. 697–701, 1991.

[9] Y. T. Kim, S. S. Jou, H. S. Lee et al., "The area of ground glass opacities of the lungs as a predictive factor in acute paraquat intoxication," *Journal of Korean Medical Science*, vol. 24, no. 4, pp. 636–640, 2009.

[10] Z. Q. Liu, H. S. Wang, and Y. Gu, "Hypokalemia is a biochemical signal of poor prognosis for acute paraquat poisoning within 4 hours," *Internal and Emergency Medicine*, vol. 12, no. 6, pp. 837–843, 2017.

[11] D.-C. Zhou, H. Zhang, Z.-M. Luo, Q.-X. Zhu, and C.-F. Zhou, "Prognostic value of hematological parameters in patients with paraquat poisoning," *Scientific Reports*, vol. 6, article 36235, 2016.

[12] Y. Li, M. Wang, Y. Gao et al., "Abnormal pancreatic enzymes and their prognostic role after acute paraquat poisoning," *Scientific Reports*, vol. 5, article 17299, 2015.

[13] X. W. Liu, T. Ma, B. Qu, Y. Ji, and Z. Liu, "Prognostic value of initial arterial lactate level and lactate metabolic clearance rate in patients with acute paraquat poisoning," *American Journal of Emergency Medicine*, vol. 31, no. 8, pp. 1230–1235, 2013.

[14] S. Y. Hong, D. H. Yang, and K. Y. Wang, "Associations between laboratory parameters and outcome of paraquat poisoning," *Toxicology Letters*, vol. 118, no. 1-2, pp. 53–59, 2000.

[15] H. Changbao and Z. Xigang, "Prognostic significance of arterial blood gas analysis in the early evaluation of paraquat poisoning patients," *Clinical Toxicology*, vol. 49, pp. 734–738, 2011.

[16] D. M. Roberts, M. F. Wilks, M. S. Roberts et al., "Changes in the concentrations of creatinine, cystatin C and NGAL in patients with acute paraquat self-poisoning," *Toxicology Letters*, vol. 202, no. 1, pp. 69–74, 2011.

[17] B. J. Chun, J. M. Moon, and H. H. Ryu, "Prognostic significance of initial laboratory parameters and plasma paraquat concentration in patients with paraquat poisoning," *Annals of Emergency Medicine*, vol. 4, no. 51, p. 472, 2018.

[18] M. D. Kuo and N. Jamshidi, "Behind the numbers: decoding molecular phenotypes with radiogenomics-guiding principles and technical considerations," *Radiology*, vol. 270, no. 2, pp. 320–325, 2014.

[19] O. Gevaert, J. Xu, C. D. Hoang et al., "Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results," *Radiology*, vol. 264, no. 2, pp. 387–396, 2012.

[20] M. A. Mazurowski, "Radiogenomics: what it is and why it is important," *Journal of the American College of Radiology*, vol. 12, no. 8, pp. 862–866, 2015.

[21] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, article 4006, 2014.

[22] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[23] R. J. Dinis-Oliveira, H. Pontes, M. L. Bastos, F. Remião, J. A. Duarte, and F. Carvalho, "An effective antidote for paraquat poisonings: the treatment with lysine acetylsalicylate," *Toxicology*, vol. 255, no. 3, pp. 187–193, 2009.

[24] R. J. Dinis-Oliveira, F. Remião, J. A. Duarte et al., "P-Glycoprotein induction: an antidotal pathway for paraquat-induced lung toxicity," *Free Radical Biology & Medicine*, vol. 41, no. 8, pp. 1213–1224, 2006.

[25] P. Smith and D. Heath, "The ultrastructure and time sequence of the early stages of paraquat lung in rats," *The Journal of Pathology*, vol. 114, no. 4, pp. 177–184, 1974.

[26] L. C. Dearden, R. D. Fairshter, J. T. Morrison, A. F. Wilson, and M. Brundage, "Ultrastructural evidence of pulmonary capillary endothelial damage from paraquat," *Toxicology*, vol. 24, no. 3-4, pp. 211–222, 1982.

[27] H. Kadoya, K. Ikeda, K. Nakatani, S. Seki, and K. Kaneda, "Cellular prion protein expression in non-ciliated epithelial cells (Clara cells) of proliferating bronchioles during bleomycin-induced pulmonary fibrosis in hamster," *Osaka City Medical Journal*, vol. 47, no. 1, pp. 23–32, 2001.

[28] L. F. Hu, Y. H. Tang, and G. L. Hong, "Prognosis and survival analysis of paraquat poisoned patients based on improved HPLC-UV method," *Journal of Pharmacological and Toxicological Methods*, vol. 80, pp. 75–81, 2016.

[29] H.-W. Gil, M.-S. Kang, J.-O. Yang, E.-Y. Lee, and S.-Y. Hong, "Association between plasma paraquat level and outcome of paraquat poisoning in 375 paraquat poisoning patients," *Clinical Toxicology*, vol. 46, pp. 515–518, 2009.

[30] K. W. Chen, M. H. Wu, J. J. Huang, and C. Y. Yu, "Bilateral spontaneous pneumothoraces, pneumopericardium, pneumomediastinum, and subcutaneous emphysema: a rare presentation of paraquat intoxication," *Annals of Emergency Medicine*, vol. 23, no. 5, pp. 1132–1134, 1994.

[31] L. Hu, H. Li, Z. Cai et al., "A new machine-learning method to prognosticate paraquat poisoned patients by combining coagulation, liver, and kidney indices," *PLoS One*, vol. 12, no. 10, article e0186427, 2017.

[32] P. Lambin, S. F. Petit, H. J. W. L. Aerts et al., "The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer," *Radiotherapy and Oncology*, vol. 96, pp. 145–152, 2017.

[33] T. J. Lin, D. Z. Hung, H. T. Yen, J. Ger, and J. F. Deng, "Survival of paraquat intoxication complicated with mediastinal emphysema: a case report," *Chinese Medical Journal*, vol. 54, pp. 363–367, 1994.

Hindawi
BioMed Research International
Volume 2021, Article ID 1235314, 10 pages
https://doi.org/10.1155/2021/1235314

Research Article

Qualitative and Quantitative MRI Analysis in IDH1 Genotype Prediction of Lower-Grade Gliomas: A Machine Learning Approach

Mengqiu Cao,[1] Shiteng Suo,[1,2] Xiao Zhang,[3] Xiaoqing Wang,[1] Jianrong Xu,[1] Wei Yang,[4] and Yan Zhou[1]

[1]*Department of Radiology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China*
[2]*Biomedical Instrument Institute, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China*
[3]*Zhuhai Precision Medical Center, Zhuhai People's Hospital (Zhuhai Hospital Affiliated with Jinan University), Zhuhai 519000, China*
[4]*Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China*

Correspondence should be addressed to Wei Yang; weiyanggm@gmail.com and Yan Zhou; clare1475@yeah.net

Received 6 August 2020; Revised 30 December 2020; Accepted 9 January 2021; Published 23 January 2021

Academic Editor: Zhiguo Zhou

Purpose. Preoperative prediction of isocitrate dehydrogenase 1 (IDH1) mutation in lower-grade gliomas (LGGs) is crucial for clinical decision-making. This study aimed to examine the predictive value of a machine learning approach using qualitative and quantitative MRI features to identify the IDH1 mutation in LGGs. *Materials and Methods*. A total of 102 LGG patients were allocated to training ($n = 67$) and validation ($n = 35$) cohorts and were subject to Visually Accessible Rembrandt Images (VASARI) feature extraction (23 features) from conventional multimodal MRI and radiomics feature extraction (56 features) from apparent diffusion coefficient maps. Feature selection was conducted using the maximum Relevance Minimum Redundancy method and 0.632+ bootstrap method. A machine learning model to predict IDH1 mutation was then established using a random forest classifier. The predictive performance was evaluated using receiver operating characteristic (ROC) curves. *Results*. After feature selection, the top 5 VASARI features were enhancement quality, deep white matter invasion, tumor location, proportion of necrosis, and T1/FLAIR ratio, and the top 10 radiomics features included 3 histogram features, 3 gray-level run-length matrix features, and 3 gray-level size zone matrix features and one shape feature. Using the optimal VASARI or radiomics feature sets for IDH1 prediction, the trained model achieved an area under the ROC curve (AUC) of 0.779 ± 0.001 or 0.849 ± 0.008 on the validation cohort, respectively. The fusion model that integrated outputs of both optimal VASARI and radiomics models improved the AUC to 0.879. *Conclusion*. The proposed machine learning approach using VASARI and radiomics features can predict IDH1 mutation in LGGs.

1. Introduction

Diffuse lower-grade gliomas (LGGs; World Health Organization (WHO) grade II or III) are infiltrative neoplasms which account for about 33%-45% of all adult gliomas [1, 2]. Although LGGs are usually less aggressive with better treatment response and prolonged prognosis compared with glioblastomas (WHO grade IV), many cases eventually progress to glioblastoma. Previous studies have shown that the high tumor heterogeneity in clinical behavior depends on genetics more than histology [1–3]. Therefore, the 2016 WHO classification of Tumors of the Central Nervous System integrates molecular biomarkers with histology for glioma diagnosis [4].

Isocitrate dehydrogenase (IDH) is one of the most important molecular biomarkers in gliomagenesis. In the

2016 WHO classification scheme, IDH mutation status serves as the first molecular determinant beyond histology, and accordingly, LGG is classified into IDH-mutant and IDH-wildtype entities [4]. Patients with an IDH-mutated glioma have a longer survival duration than those with an IDH-wildtype tumor. Recent evidence has also suggested that IDH may be a potential therapeutic target in IDH-mutant gliomas [5]. Therefore, preoperative prediction of IDH mutation status is crucial for prognosis and therapeutic decision-making.

MRI can facilitate glioma diagnosis in a noninvasive manner [6, 7]. Qualitative MRI analysis still remains the basis in imaging diagnosis. For interpretation accuracy and consistency, Visually Accessible Rembrandt Images (VASARI) lexicon based on conventional MRI has been proposed to describe the features and guidelines. Previous studies have shown the biological or clinical relevance of the VASARI features in gliomas. For example, Zhou et al. [6] reported that VASARI features including proportion of necrosis and lesion size were associated with IDH1 mutation status.

Quantitative MRI has emerged as a promising tool in the evaluation of gliomas as it can provide information on tumor functionality. Apparent diffusion coefficient (ADC) calculated from diffusion-weighted imaging (DWI) is one of the most clinically useful quantitative measurements [8–10]. Radiomics, a recently developed high-throughput approach, can add value to the routine MRI to a greater extent by extracting and mining a large number of imaging traits [11]. Growing evidence has revealed the feasibility and clinical implications of radiomics in the characterization of glioma phenotypes [6, 12].

We hypothesized that the use of both qualitative and quantitative MRI features could facilitate better IDH genotype discrimination. In this study, we aimed to develop a machine learning approach based on VASARI and ADC radiomics features to characterize the IDH1 mutation status in LGGs.

2. Materials and Methods

2.1. Subjects. This retrospective study was approved by the local institutional review board with a waiver of the written informed consent from patients. Patients were identified by searching the database of our institution for radiologic and histopathologic records from January 2015 to December 2018. The inclusion criteria for the study patients were as follows: (a) histologically proven LGG; (b) available IDH1 mutation records; (c) complete preoperative MRI data including native T1- and T2-weighted imaging (T1W and T2W); T2 fluid attenuation inversion recovery (FLAIR), DWI, and postcontrast T1W; and (d) sufficient image quality. Patients who had received treatment for glioma prior to MRI were excluded. Finally, 102 LGG patients (60 men and 42 women; age range, 18-77 years; mean age, 45.3 ± 16.3 years) were included for the subsequent analyses. Subjects were randomly divided into two subsets, a training cohort ($n = 67$) and a validation cohort ($n = 35$).

2.2. MRI. Images were acquired using a 3 Tesla MRI system (Signa HDxt; GE Medical Systems, Milwaukee, Wis, USA) with an eight-channel head coil. The protocol included native T1W, T2W, FLAIR, and DWI in the axial plane and postcontrast T1W in three orthogonal planes. Postcontrast imaging was achieved with intravenous administration of 0.1 mmol/kg dose of gadopentetate dimeglumine (Magnevist; Bayer Healthcare, Berlin, Germany). In all native sequences, the same asymmetric field of view (260×260 mm^2), section thickness (5 mm), and intersection gap (20%) were used. DWI was performed before the injection of contrast material with repetition time = 4850 ms, echo time = 74 ms, acquisition matrix = 160×160, b value = 0 and 1000 sec/mm^2, and number of averages = 2.

2.3. Feature Extraction. For qualitative image analysis, readings were performed on all sequences with a Digital Imaging and Communications in Medicine viewer (RadiAnt DICOM Viewer; Poznan, Poland) by two neuroradiologists (Mengqiu Cao and Yan Zhou, with 6 and 19 years of experience in neurological MRI interpretation, respectively) in consensus. Each tumor was scored according to the VASARI lexicon, which consists of 23 imaging traits related to the morphology of brain tumors. Detailed descriptions of the VASARI feature set are available in Supplementary Table S1.

For quantitative ADC analysis, segmentation of the tumor area was first manually performed using the 3D Slicer software (version 4.7; https://www.slicer.org). The tumor area was defined as the abnormal hyperintensity area on FLAIR images. The volume of interest (VOI) was generated by including all consecutive image sections containing tumor areas. Independent analysis of the segmentation labels (from 30 randomly selected subjects in the training set) by two neuroradiologists was conducted to evaluate the interobserver reliability of the segmentation. The Dice similarity coefficient (DSC) [13] was measured over the two labels per case from the two neuroradiologists. A DSC value of 0 indicates no overlap and a value of 1 corresponds to exact overlap. After registering ADC maps to FLAIR images, VOI was propagated to ADC maps. A total of 56 radiomics features were then extracted from the volumetric ADC data including 3 shape features, 13 first-order histogram features, 9 gray-level co-occurrence matrix (GLCM) features, 13 gray-level run-length matrix (GLRLM) features, 13 gray-level size zone matrix (GLSZM) features, and 5 neighborhood gray-tone difference matrix (NGTDM) features [14]. Before the feature selection process, all the radiomics features were normalized to the range of [0, 1] for standardization, so that features of different orders of magnitude could be reasonably compared. Feature extraction was performed using the Matlab software (version 2016a; MathWorks, Natick, Mass, USA). Detailed calculations of the radiomics features are provided in Supplementary Table S2.

2.4. Feature Selection. Our study adopted a two-step feature selection scheme to identify the most predictive variables. First, the maximum Relevance Minimum Redundancy (mRMR) method was used to select features that had the maximal mutual information with respect to the target class (maximum relevance) and minimal mutual information with respect to each other (minimum redundancy). Second, the

0.632+ bootstrap method and the area under the receiver operating characteristic curve (AUC) metric were used to explore the features with optimal discrimination performance on the training data set [14]. A random forest classifier was chosen as a statistical model in this process. According to the AUC metric, the top 5 VASARI and 10 radiomics features were finally selected for further predictive model building.

2.5. Machine Learning-Based Prediction. Predictive models of different orders (1–5 for VASARI features and 1–10 for radiomics features) were constructed separately on the optimal combinations of VASARI and radiomics features. Random forest classifiers were trained on the training cohort. The prediction performance was evaluated with the 0.632+ bootstrap AUC method. Sensitivity, specificity, accuracy, and AUC were calculated for each condition.

The random forest prediction models were then validated on the validation cohort. Further, the fusion model from the optimal VASARI model and radiomics model was obtained by integrating the predicted probability of both models. The weight value of fusion of the two models was set according to the weighted average fusion strategy, that was, 0.5. When analyzing a new case, we separately calculated the prediction probability of VASARI and radiomics models and, then, averaged the two values as the final prediction probability. To demonstrate the complementary roles of VASARI and radiomics features in the fusion model, the correlation analysis was performed using the Pearson correlation coefficient. The prediction performance of the fusion machine learning model was evaluated. The influence of common clinical variables including age and gender on the prediction performance was also tested.

The flowchart of the experimental design of the machine learning approach is illustrated in Figure 1. All the machine learning algorithms were implemented using the Matlab software.

2.6. Statistical Analysis. Comparison of categorical characteristics between groups was performed with the chi-square test or Fisher's exact test and comparison of continuous characteristics with Student's t-test. Receiver operating characteristic (ROC) curves were generated on the basis of the classification results of random forest models. Results with P values less than 0.05 were considered to indicate a significant difference. All the statistical analyses were performed using the Matlab software and IBM SPSS Statistics software (version 21; SPSS, Chicago, Ill, USA).

3. Results

3.1. Patient Characteristics. Of all the 102 LGG patients, 61 (59.8%) were diagnosed as WHO grade II glioma and 41 (40.2%) with WHO grade III. Among them, 50 (49%) and 52 (51%) patients were confirmed with IDH1-mutant and IDH1-wildtype LGG, respectively. Patient characteristics of the whole cohort, the training cohort, and the validation cohort were summarized in Table 1. No significant difference in age, gender, WHO grade, or IDH1 mutation status was noted between the training and validation cohorts ($P > 0.05$).

3.2. Interobserver Reliability of Segmentation. Interobserver reliability analysis of the manual segmentation showed good agreement between the neuroradiologists, with a DSC score of 0.879 ± 0.046. A representative case showing the interobserver reliability of segmentation is illustrated in Figure 2.

3.3. IDH1 Mutation Prediction with VASARI Features. After feature selection, the top 5 VASARI features were enhancement quality, deep white matter invasion, tumor location, proportion of necrosis, and T1/FLAIR ratio (Table 2). Prediction models with orders 1 to 5 were generated by incorporating the above optimal features. On the training cohort, the highest AUC of 0.827 ± 0.031 was reached, with a sensitivity of 0.671 ± 0.058 and a specificity of 0.712 ± 0.049, respectively. Using the optimal feature set (the single enhancement quality feature), the trained model achieved an AUC of 0.779 ± 0.001 on the validation cohort, with a sensitivity of 0.718 ± 0.070, a specificity of 0.733 ± 0.100, and an accuracy of 0.726 ± 0.017, respectively. Representative cases of IDH1-mutant and IDH1-wildtype LGGs are shown in Figures 3 and 4.

3.4. IDH1 Mutation Prediction with Radiomics Features. In ADC radiomics analysis, the top 10 quantitative features were listed in Table 2. On the training cohort, the highest AUC of 0.849 ± 0.027 was reached, with a sensitivity of 0.790 ± 0.038 and a specificity of 0.770 ± 0.043, respectively. Using the optimal feature set (all the 10 features), the trained model achieved an AUC of 0.849 ± 0.008 on the validation cohort, with a sensitivity of 0.724 ± 0.035, a specificity of 0.761 ± 0.017, and an accuracy of 0.743 ± 0.022, respectively.

3.5. IDH1 Mutation Prediction with a Fusion Model with Optimal VASARI and Radiomics Features. The fusion model was constructed with the optimal VASARI model (enhancement quality) and radiomics model (the top 10 radiomics features). Results of the Pearson correlation analysis showed that these two types of features remained very low correlation (Figure 5), demonstrating their complementary roles in the fusion model. The fusion model improved the AUC to 0.879, with a sensitivity of 0.765, a specificity of 0.778, and an accuracy of 0.771, respectively. ROC curves of the optimal VASARI model, radiomics model, and the fusion model with VASARI and radiomics features are illustrated in Figure 6. The inclusion of clinical variables including age and gender to the model did not benefit the prediction performance (AUC = 0.859).

4. Discussion

In this study, the machine learning algorithm was used to explore the predictive value of VASARI features based on preoperative conventional MRI images and the radiomics features based on ADC maps in IDH1 genotyping of LGG patients. The results obtained by random forest classifiers showed that the AUCs were 0.779 and 0.849 on the optimal VASARI and radiomics feature sets, respectively, and the

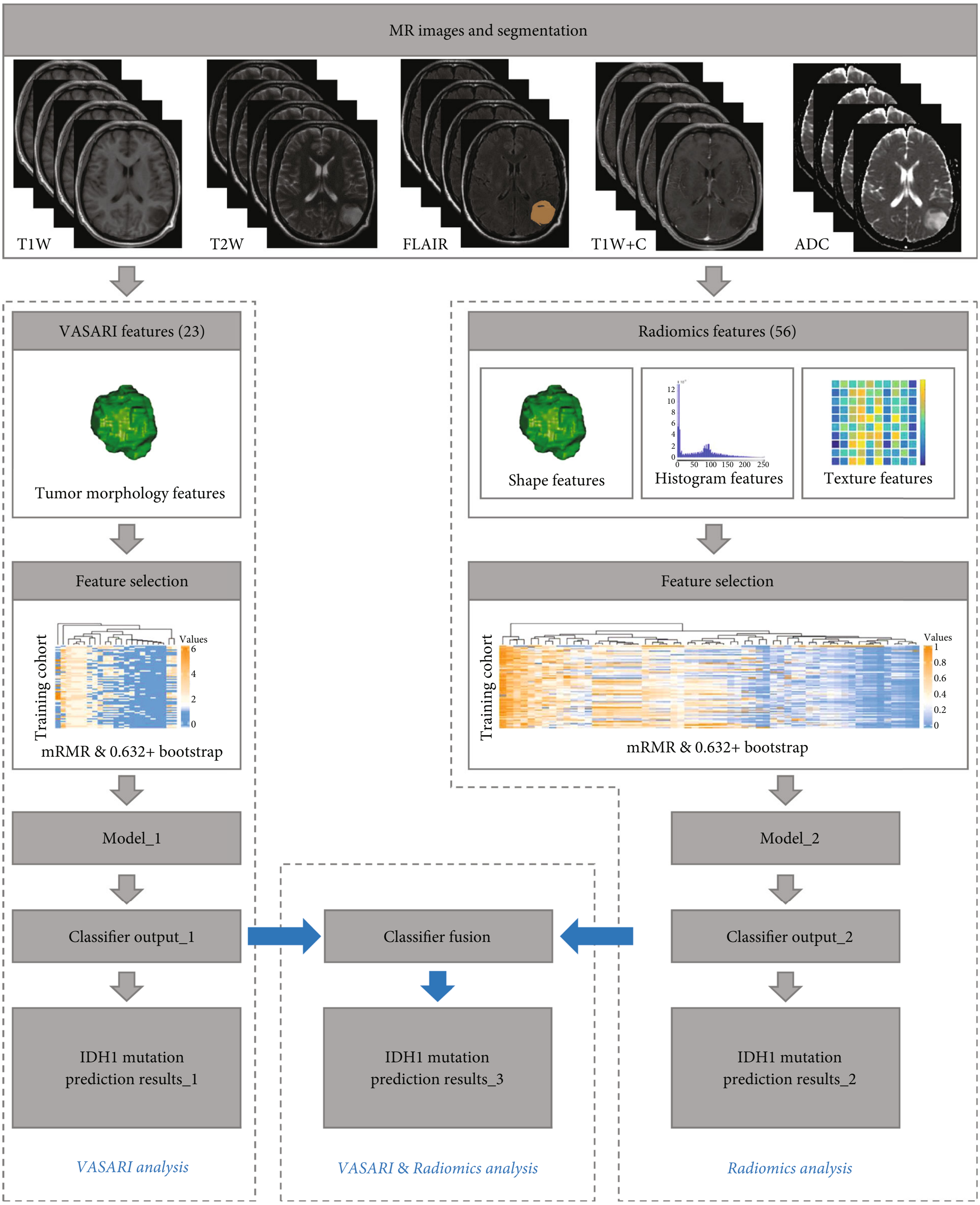


Figure 1: The flowchart of the experimental design of the machine learning approach.

fusion model with both feature sets achieved an improved AUC of 0.879 on the validation.

MRI is one of the essential methods for preoperative glioma diagnosis. Different imaging sequences can reveal different characteristics of tumor texture, blood supply, border, edema, hemorrhage, etc., and these characteristics are extremely important for the final diagnosis. The VASARI lexicon extracts features from routine MRI and provides

Table 1: Patient characteristics.

Characteristic	Whole cohort ($n = 102$)	Training cohort ($n = 67$)	Validation cohort ($n = 35$)	P value*
Age (years)†	45.3 ± 16.3	45.7 ± 17.1	44.6 ± 14.9	0.75
Gender				
Male	60 (58.8%)	38 (56.7%)	22 (62.9%)	0.55
Female	42 (41.2%)	29 (43.3%)	13 (37.1%)	
WHO grade				
II	61 (59.8%)	44 (65.7%)	17 (48.6%)	0.10
III	41 (40.2%)	23 (34.3%)	18 (51.4%)	
IDH1 status				
Mutant	50 (49.0%)	33 (49.3%)	17 (48.6%)	0.95
Wildtype	52 (51.0%)	34 (50.7%)	18 (51.4%)	

Unless otherwise specified, data are counts (percentages). WHO: World Health Organization; IDH 1: isocitrate dehydrogenase 1. †Data are means ± standard deviations. *P value was obtained by comparing each variable between training and validation cohorts.

standardized visual grading of MRI findings. In our study, enhancement quality was the most significant one for IDH1 mutation prediction among all VASARI features. IDH1-wildtype LGGs tended to represent a higher degree of contrast enhancement on the postcontrast T1W images compared with IDH1-mutant LGGs, which is consistent with previous studies [15–17]. Kickingereder et al. [18] found that IDH1-wildtype gliomas showed increased HIF1A activation, thus leading to a transcriptome signature induced by upregulating vasculo- and angiogenesis-related signaling pathways. Increase in proangiogenic molecules could result in more contrast agent uptake and more marked contrast enhancement on postcontrast T1W images. Besides enhancement quality, other VASARI features of strong predictive power for IDH1 mutation status included deep white matter invasion, tumor location, proportion of necrosis, and T1/FLAIR ratio. These findings are in line with those from previous studies [6, 15, 19, 20]. Among these features, tumor location in the frontal lobe in IDH1-mutant gliomas has been reported by many investigators in existing literature [21]. The frontal lobe predominance of IDH1-mutant gliomas may be because this type of tumors probably originates from glial progenitors in the forebrain subventricular zone [22]. VASARI-based random forest classifier showed an AUC of 0.779 on validation in predicting IDH1 mutation in LGGs, similar to the result reported by Park et al., who constructed a multivariable model with an AUC of 0.778 [20].

Radiomics is a method to extract quantitative features that are difficult to detect by human eyes from medical images and to use data mining and machine learning algorithms for diagnostic decision-making. In this study, radiomics analysis of ADC maps was conducted by extracting 57 quantitative features and subsequently building a prediction model with 10 optimal features. Given that the choice of classifier depends on the specific task as well as disease type, thus, comparative experiments were conducted, and ultimately random forest was chosen with the best performance for IDH1 prediction. The prediction performance on the independent validation set using different classifiers is shown in Supplementary Figure S1. Our optimal radiomics model achieved an AUC of 0.849 for IDH1 prediction in LGGs. ADC was used for radiomics analysis in our study, since ADC has been established as the most commonly used quantitative MRI metric, thus enabling first-order statistical features comparable between individuals. Previous studies have shown the benefit of ADC first-order statistical features in identifying IDH1 genotypes [23–25]. Our study further demonstrated the added value of ADC high-order radiomics features to first-order features for this purpose. Additionally, radiomics on other MRI modalities has also been investigated in terms of its relationship with IDH1 mutation status. Zhou et al. [6] found that random forest analysis of T2W-based texture features could predict IDH1 mutation status in LGGs with an AUC of 0.86, a sensitivity of 0.75, and a specificity of 0.89. By performing radiomics analysis on FLAIR images, Yu et al. [26] reported AUCs of 0.86 and 0.79 on the training and validation cohorts, respectively, in IDH1 prediction of LGGs. Interestingly, these results are consistent with ours on ADC radiomics analysis.

The major strength in our study design was the model building using both qualitative semantic and quantitative radiomics features, which were usually separately investigated in some previous studies [20, 27]. Results showed that the fusion model that integrated outputs of the optimal VASARI model and ADC-based radiomics model improved the AUC to 0.879 in IDH1 genotype prediction of LGGs, indicating that the fusion model was superior to the model using a single type of features. These findings suggest that radiomics analysis may add value to routine qualitative image analysis for IDH1 classification. Similarly, a recent study [28] also showed that the VASARI feature combined with ADC texture analysis could improve the accuracy of IDH1 mutation detection in anaplastic gliomas. In this study, although the mean age of patients with IDH1-wildtype LGG was higher than that of patients with IDH1-mutant LGG (47.3 years vs. 43.2 years), there was no statistical difference between the two groups ($P = 0.201$, independent sample t -test). Therefore, the inclusion of age factor in the final model failed to improve the accuracy of LGG IDH1 genotype identification.

Recently, with its rapid advancement in various fields within the past few years, deep learning has gained particular attention in the radiology domain. For example, Chang et al. [29] has used a deep learning method implemented with convolutional neural networks to classify genetic mutations in gliomas and a high accuracy of 0.94 in IDH mutation prediction was reached. Deep learning is advantageous in that it does not need human-derived feature extraction or prior feature selection [29]. However, big data are essential for a robust training process. A head-to-head comparison between conventional machine learning and deep learning methods is warranted in the future.

Apart from the intrinsic limitations of any retrospective study, several other limitations are discussed as follows. First, the cases were collected from a single center, and the patient population was relatively small. Further validation on diverse

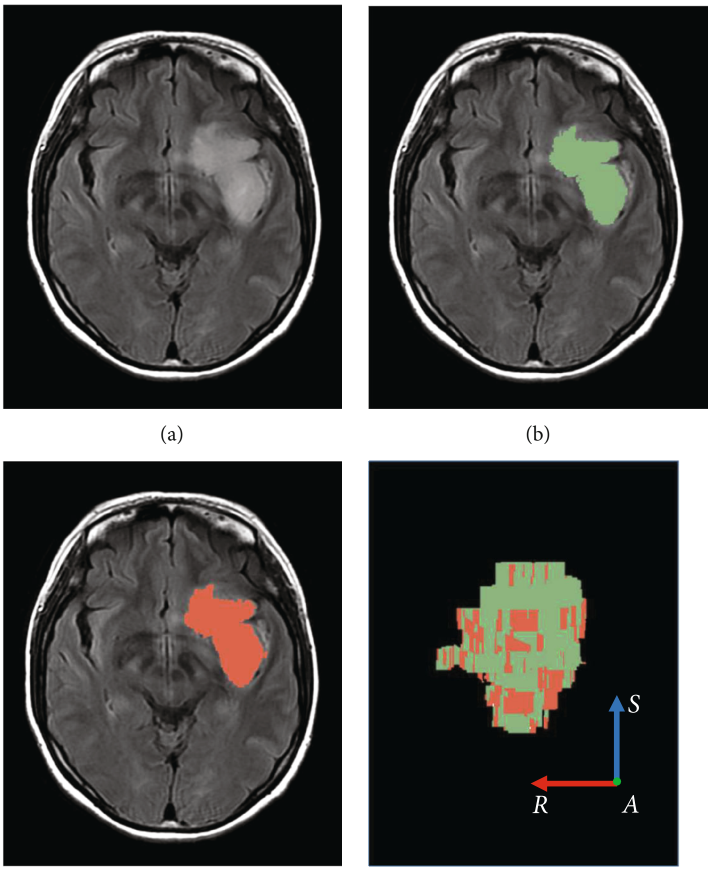


Figure 2: Interobserver reliability of contours between the two neuroradiologists. (a) One original section of the volumetric data. (b) Contour delineated by the first neuroradiologist. (c) Contour delineated by the second neuroradiologist. (d) Overlaid 3D volume rendering image (AP view).

Table 2: List of selected VASARI and radiomics features.

Feature selection Top 5 VASARI features	AUC value	Top 10 radiomics features	AUC value
Enhancement quality	0.752	GLRLM short run low gray-level emphasis	0.756
Deep white matter invasion	0.738	GLRLM low gray-level run emphasis	0.682
Tumor location	0.684	GLRLM run-length variance	0.678
Proportion of necrosis	0.682	Histogram minimum	0.677
T1/FLAIR ratio	0.632	Eccentricity	0.662
		GLSZM large zone high gray-level emphasis	0.641
		GLSZM low gray-level zone emphasis	0.628
		Histogram energy	0.616
		Histogram standard deviation	0.612
		GLSZM zone-size nonuniformity	0.607

VASARI: Visually Accessible Rembrandt Images; AUC: area under the receiver operating characteristic curve; GLRLM: gray-level run-length matrix; GLSZM: gray-level size zone matrix.

large data sets acquired from multiple vendors and across different centers is needed. Second, the numbers of included IDH1-mutant and IDH1-wildtype patients were similar (50 : 52), which did not reflect the actual prevalence of IDH mutation in LGG (around 80%) [3]. However, a balanced sampling could contribute to the model training process. Third, radiomics analysis was not performed on other routine MRI modalities. Routine MRI data were used to extract

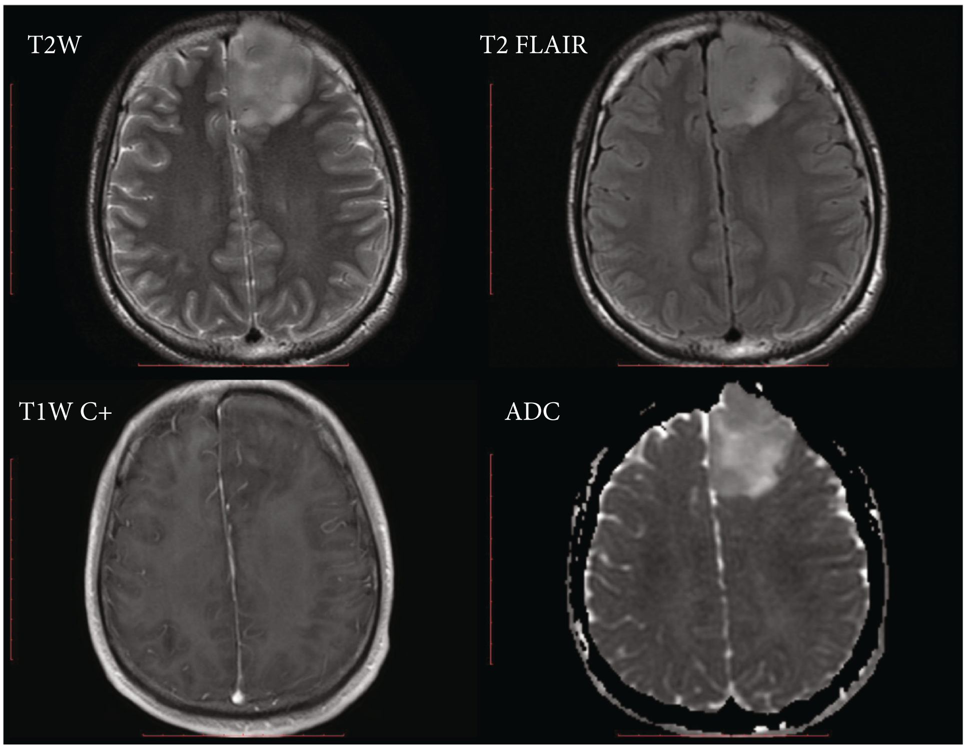


Figure 3: A 26-year-old man with an IDH1-mutant glioma (diffuse astrocytoma, WHO grade II). The tumor is located in the frontal lobe with no contrast enhancement, no deep white matter invasion, no necrosis, and an expansive tumor behavior (T1~FLAIR).

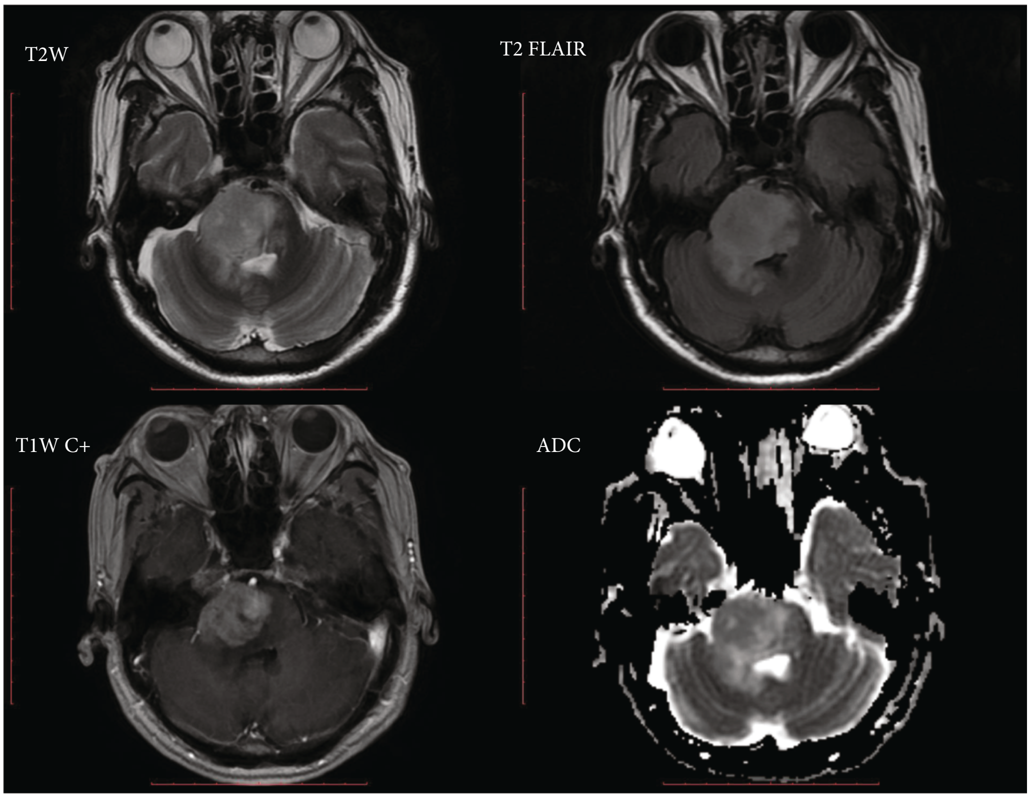


Figure 4: A 65-year-old man with an IDH1-wildtype glioma (diffuse astrocytoma, WHO grade II). The tumor is located in the brainstem with marked contrast enhancement, deep white matter invasion, a necrosis proportion of <33%, and a mixed tumor behavior (T1<FLAIR).

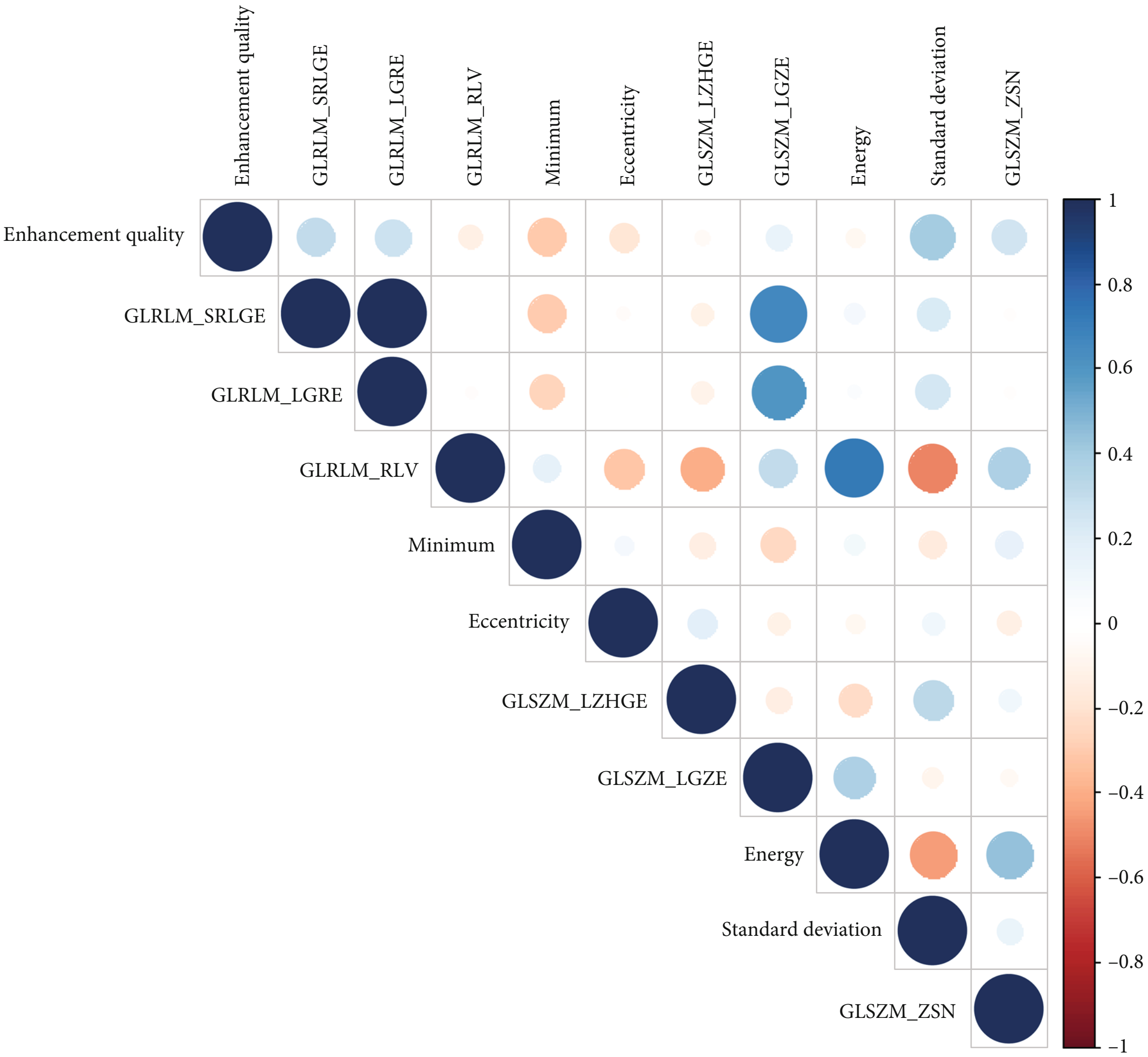


Figure 5: Correlation between VASARI and radiomics features in the fusion model.

semantic features, as is performed in clinical routine. However, the results on ADC maps were consistent with those on T2W images or FLAIR images reported before [6, 26]. Advanced MRI techniques such as perfusion-weighted imaging and magnetization transfer imaging were also not adopted for radiomics analysis. The inclusion of advanced MRI modalities could provide more comprehensive functional and metabolic information and should be considered in further studies. Fourth, interobserver agreement of image segmentation was evaluated in our study. However, interobserver agreement of features was not analyzed, although it has proven to be satisfactory for both VASARI and radiomics features in previous studies [6, 20]. Fifth, considering the small sample size of our study, we did not perform weight optimization in order to avoid overfitting of the training data. Although this weighting method may lose a little performance improvement (not always), we believe that the fusion results would be more robust, especially for new data, without performance bias. It can be seen from the results that our weighted average fusion strategy played a positive role in guiding the overall forecast performance. Last, according to the 2016 WHO classification of Tumors of the Central Nervous System, 1p/19q codeletion is also an important prognostic marker in molecular diagnosis of LGGs [4]. In the study, 1p/19q codeletion status was not evaluated because this information was not available on most subjects due to the retrospective nature.

5. Conclusion

In conclusion, preoperative MRI VASARI features and ADC radiomics features can effectively predict IDH1 mutation status in LGG, and the fusion model integrating both predictive features shows even better prediction performance. The proposed image-based machine learning approach may provide an alternative to the conventional workflow for the noninvasive identification of IDH1 genotypes. However, these findings should be validated in large multicenter data sets in future studies.

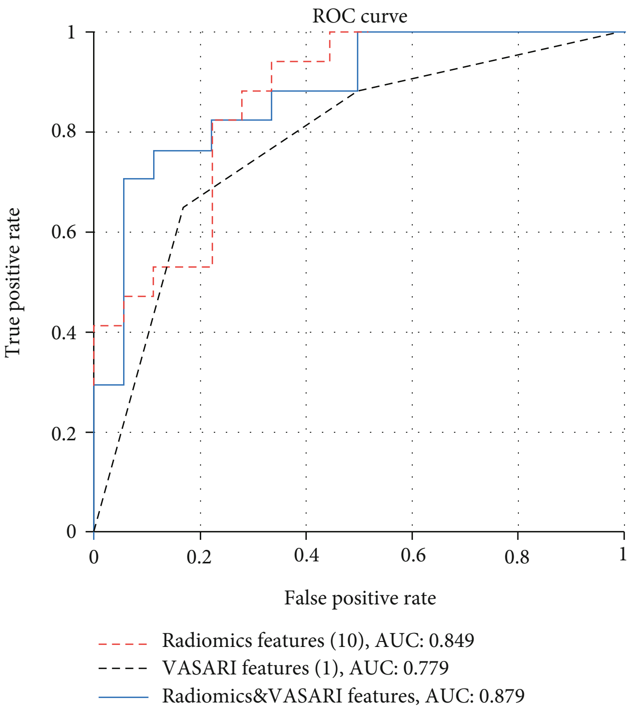


Figure 6: ROC curves of the constructed VASARI model, radiomics model, and the fusion model. The fusion prediction model improved the AUC to 0.879 on the validation dataset.

Data Availability

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

Disclosure

The abstract of this paper was presented at the 2020 ISMRM & SMRT Virtual Conference & Exhibition, as an oral presentation.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Mengqiu Cao, Shiteng Suo, and Xiao Zhang contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 81701642, 81501458, and 81571650), Shanghai Science and Technology Committee (grant numbers 17411964300, 18511102901, and 17441902700), and Shanghai "Rising Stars of Medical Talent" Youth Development Program (Youth Medical Talents—Medical Imaging Practitioner Program) (SHWRS(2020)_087).

Supplementary Materials

Table S1: summary of VASARI features. Table S2: summary of radiomics features. Figure S1: ROC curves show the prediction performance for IDH1 mutation on the independent validation set using different classifiers. Random forest classifier showed the best prediction performance with an area under the ROC curve of 0.849. RF: random forest; SVM: support vector machine; LDA: linear discriminant analysis; KNN: k-nearest neighbor. *(Supplementary Materials)*

References

[1] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[2] H. Suzuki, K. Aoki, K. Chiba et al., "Mutational landscape and clonal architecture in grade II and III gliomas," *Nature Genetics*, vol. 47, no. 5, pp. 458–468, 2015.

[3] D. J. Brat, R. G. Verhaak, K. D. Aldape et al., "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *The New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015.

[4] D. N. Louis, A. Perry, G. Reifenberger et al., "The 2016 World Health Organization classification of tumors of the central nervous system: a summary," *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, 2016.

[5] S. Pellegatta, L. Valletta, C. Corbetta et al., "Effective immuno-targeting of the IDH1 mutation R132H in a murine model of intracranial glioma," *Acta Neuropathologica Communications*, vol. 3, no. 1, p. 4, 2015.

[6] H. Zhou, M. Vallieres, H. X. Bai et al., "MRI features predict survival and molecular markers in diffuse lower-grade gliomas," *Neuro-Oncology*, vol. 19, no. 6, pp. 862–870, 2017.

[7] M. Cao, W. Ding, X. Han et al., "Brain T1ρ mapping for grading and IDH1 gene mutation detection of gliomas: a preliminary study," *Journal of Neuro-Oncology*, vol. 141, no. 1, pp. 245–252, 2019.

[8] J. A. Brunberg, T. L. Chenevert, P. E. McKeever et al., "In vivo MR determination of water diffusion coefficients and diffusion anisotropy: correlation with structural alteration in gliomas of the cerebral hemispheres," *AJNR. American Journal of Neuroradiology*, vol. 16, no. 2, pp. 361–371, 1995.

[9] Y. Kang, S. H. Choi, Y. J. Kim et al., "Gliomas: histogram analysis of apparent diffusion coefficient maps with standard- or high-b-value diffusion-weighted MR imaging–correlation with tumor grade," *Radiology*, vol. 261, no. 3, pp. 882–890, 2011.

[10] M. Cao, S. Suo, X. Han et al., "Application of a simplified method for estimating perfusion derived from diffusion-weighted MR imaging in glioma grading," *Frontiers in Aging Neuroscience*, vol. 9, p. 432, 2017.

[11] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[12] C. F. Lu, F. T. Hsu, K. L. Hsieh et al., "Machine learning-based radiomics for molecular subtyping of gliomas," *Clinical Cancer Research*, vol. 24, no. 18, pp. 4429–4436, 2018.

[13] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[14] M. Vallieres, C. R. Freeman, S. R. Skamene, and I. El Naqa, "A radiomics model from joint FDG-PET and MRI texture

features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Physics in Medicine and Biology*, vol. 60, no. 14, pp. 5471–5496, 2015.

[15] S. Qi, L. Yu, H. Li et al., "Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms," *Oncology Letters*, vol. 7, no. 6, pp. 1895–1902, 2014.

[16] S. Nakae, K. Murayama, H. Sasaki et al., "Prediction of genetic subgroups in adult supra tentorial gliomas by pre- and intraoperative parameters," *Journal of Neuro-Oncology*, vol. 131, no. 2, pp. 403–412, 2017.

[17] H. Arita, M. Kinoshita, A. Kawaguchi et al., "Lesion location implemented magnetic resonance imaging radiomics for predicting IDH and TERT promoter mutations in grade II/III gliomas," *Scientific Reports*, vol. 8, no. 1, p. 11773, 2018.

[18] P. Kickingereder, F. Sahm, A. Radbruch et al., "_IDH_ mutation status is associated with a distinct hypoxia/angiogenesis transcriptome signature which is non-invasively predictable with rCBV imaging in human glioma," *Scientific Reports*, vol. 5, no. 1, p. ???, 2015.

[19] P. Metellus, B. Coulibaly, C. Colin et al., "Absence of IDH mutation identifies a novel radiologic and molecular subtype of WHO grade II gliomas with dismal prognosis," *Acta Neuropathologica*, vol. 120, no. 6, pp. 719–729, 2010.

[20] Y. W. Park, K. Han, S. S. Ahn et al., "Prediction of IDH1-mutation and 1p/19q-codeletion status using preoperative MR imaging phenotypes in lower grade gliomas," *AJNR. American Journal of Neuroradiology*, vol. 39, no. 1, pp. 37–42, 2018.

[21] C. H. Suh, H. S. Kim, S. C. Jung, C. G. Choi, and S. J. Kim, "Imaging prediction of isocitrate dehydrogenase (IDH) mutation in patients with glioma: a systemic review and meta-analysis," *European Radiology*, vol. 29, no. 2, pp. 745–758, 2019.

[22] A. Lai, S. Kharbanda, W. B. Pope et al., "Evidence for sequenced molecular evolution of IDH1 mutant glioblastoma from a distinct cell of origin," *Journal of Clinical Oncology*, vol. 29, no. 34, pp. 4482–4490, 2011.

[23] S. Lee, S. H. Choi, I. Ryoo et al., "Evaluation of the microenvironmental heterogeneity in high-grade gliomas with IDH1/2 gene mutation using histogram analysis of diffusion-weighted imaging and dynamic-susceptibility contrast perfusion imaging," *Journal of Neuro-Oncology*, vol. 121, no. 1, pp. 141–150, 2015.

[24] Y. W. Park, K. Han, S. S. Ahn et al., "Whole-tumor histogram and texture analyses of DTI for evaluation of IDH1-mutation and 1p/19q-codeletion status in World Health Organization grade II gliomas," *AJNR. American Journal of Neuroradiology*, vol. 39, no. 4, pp. 693–698, 2018.

[25] C. De Looze, A. Beausang, J. Cryan et al., "Machine learning: a useful radiological adjunct in determination of a newly diagnosed glioma's grade and IDH status," *Journal of Neuro-Oncology*, vol. 139, no. 2, pp. 491–499, 2018.

[26] J. Yu, Z. Shi, Y. Lian et al., "Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma," *European Radiology*, vol. 27, no. 8, pp. 3509–3522, 2017.

[27] X. Zhang, Q. Tian, Y.-X. Wu et al., "IDH mutation assessment of glioma using texture features of multimodal MR images," *Medical imaging 2017: computer-aided diagnosis—proceedings of SPIE*, S. G. Armato and N. A. Petrick, Eds., no. article 101341S, 2017SPIE, Bellingham, WA, 2017.

[28] C. Q. Su, S. S. Lu, M. D. Zhou, H. Shen, H. B. Shi, and X. N. Hong, "Combined texture analysis of diffusion-weighted imaging with conventional MRI for non-invasive assessment of IDH1 mutation in anaplastic gliomas," *Clinical Radiology*, vol. 74, no. 2, pp. 154–160, 2019.

[29] P. Chang, J. Grinband, B. D. Weinberg et al., "Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas," *AJNR. American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1201–1207, 2018.

Hindawi
BioMed Research International
Volume 2021, Article ID 6683931, 7 pages
https://doi.org/10.1155/2021/6683931

Research Article

Automated Classification and Segmentation in Colorectal Images Based on Self-Paced Transfer Network

Yao Yao,[1,2] Shuiping Gou,[1] Ru Tian,[1] Xiangrong Zhang,[1] and Shuixiang He[3]

[1] *School of Artificial Intelligence, Xidian University, Xi'an, Shanxi 710071, China*
[2] *School of Information Engineering, Hangzhou Vocational and Technical College, Hangzhou, Zhejiang 310018, China*
[3] *Department of Gastroenterology, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shanxi 710071, China*

Correspondence should be addressed to Shuiping Gou; shpgou@mail.xidian.edu.cn and Shuixiang He; hesx123@126.com

Received 13 November 2020; Revised 22 December 2020; Accepted 6 January 2021; Published 21 January 2021

Academic Editor: Zhiguo Zhou

Colorectal imaging improves on diagnosis of colorectal diseases by providing colorectal images. Manual diagnosis of colorectal disease is labor-intensive and time-consuming. In this paper, we present a method for automatic colorectal disease classification and segmentation. Because of label unbalanced and difficult colorectal data, the classification based on self-paced transfer VGG network (STVGG) is proposed. ImageNet pretraining network parameters are transferred to VGG network with training colorectal data to acquire good initial network performance. And self-paced learning is used to optimize the network so that the classification performance of label unbalanced and difficult samples is improved. In order to assist the colonoscopist to accurately determine whether the polyp needs surgical resection, feature of trained STVGG model is shared to Unet segmentation network as the encoder part and to avoid repeat learning of polyp segmentation model. The experimental results on 3061 colorectal images illustrated that the proposed method obtained higher classification accuracy (96%) and segmentation performance compared with a few other methods. The polyp can be segmented accurately from around tissues by the proposed method. The segmentation results underpin the potential of deep learning methods for assisting colonoscopist in identifying polyps and enabling timely resection of these polyps at an early stage.

1. Introduction

The International Agency for Research on Cancer released research data on global cancer status in 2018. The report reported the incidence and mortality of 36 types of tumors in 185 countries around the world, comprehensively. The data showed that the incidence of colorectal cancer ranked third (10.2%), and the mortality rate ranked second (9.2%) [1]. As we known, the mortality rate of colorectal cancer can be reduced significantly by early removal of polyps [2] which can be found according to the early screening. Colorectal polyp, a benign disease, has specific imaging characteristics such as shape or surface structure and color [3]. Colorectal colonoscopy is the main method of diagnosing intestinal diseases. With a great number of colorectal images, the microscopic examination presents labor-intensive and time-consuming problems [4]. In addition, the pathological diagnosis of colonoscopy biopsy samples is prone to deviations due to individual pathologists' experience and knowledge [5]. The accuracy of diagnosis depends on the experience of the microscopy doctor, and the difference in diagnosis accuracy between experienced doctors and less experienced doctors is greater than 10%. Therefore, it is necessary to distinguish polyps from normal tissue and tumor using colorectal optical images.

However, it is difficult for the diagnosis of colorectal optical images. Firstly, the low light and interference of liquid often result in poor imaging quality of colorectal images. Secondly, the edges of normal tissue and polyp types are blurred. It causes the classification accuracy of normal tissue, polyp, and tumor to be low. Thirdly, the individual differences of polyps are mainly manifested in shape, color, and surface contour for polyps. And colorectal polyps are heterogeneous resulted that the segmentation of polyp becomes challenging.

Previous research showed that deep learning has given good results in medical images processing, such as tumor detection, classification, segmentation, retrieval, and prediction, especially for diagnosis and treatment of the brain [6, 7], breast [8], lung [9, 10], gastric [11], prostate cancers [12, 13], and histopathology [14]. Meanwhile, endoscopy-assisted diagnosis has also made some progress using deep learning, especially in colorectal endoscopy. There are two types for colorectal image detection: pathology and optical colonoscopy images. Here is the introduction to the image diagnosis progress.

For the pathology colorectal images, the recent advancement of deep learning is adapted. Thakur et al. [15] reviewed the development of an AI system in CRC pathology image analysis using deep learning. Korbar et al. [16] proposed an automatic image-understanding method to help pathologists with histopathological characterization and diagnosis of colorectal polyps. Sena et al. [17] propose a deep learning approach to recognize four different stages of cancerous tissue development. Lizuka et al. [18] trained convolutional neural networks (CNNs) and recurrent neural networks (RNNs) on biopsy histopathology whole-slide images (WSIs) of stomach and colon. *For the optical colorectal images, there are lots of researches on* detection and segmentation of colorectal polyps. Some methods take into account time series: Urban et al. [19] used deep learning to localize and identify polyps in real time with 96% accuracy in screening colonoscopy. Klare et al. [20] proposed the APDS with which the colonoscopy system of the video stream is captured by a frame-grabber device in HD. Wang et al. [21] used real-time automatic detection system to increase colonoscopic polyp and adenoma detection rates; some methods take into account spatial information: Li et al. [22] used a fully convolutional neural network structure for segmenting colorectal polyps. Yang et al. [23] developed convolutional neural network (CNN) models which automatically categorized colorectal lesions into several stages ranging from nonneoplastic lesions to advanced CRC with conventional white-light colonoscopy images. Zhang et al. [24] developed a fully automatic algorithm to detect and classify hyperplastic and adenomatous colorectal polyps. Others are from the semantic information: Wickstrom and Kampffmeyer [25] proposed a novel method for estimating the uncertainty associated with important features in the input and demonstrated how interpretability and uncertainty can be modeled for semantic segmentation of colorectal polyps. The above colorectal image processing methods using deep learning have achieved good performance.

Based on the above analysis of colonic *pathology and optical* colorectal image literature, deep learning methods are proposed on detection or segmentation of colorectal polyps. However, unlike the recent research based on single task, our method takes into account multitask: colorectal image classification and polyp image segmentation. In the proposed STVGG, transfer learning and self-paced learning are used to solve the unbalance and the difficult sample learning. STVGG transfers ImageNet network parameters to VGG network and calculates the loss value of each image in the forward propagation with the age parameter. In addition, the trained STVGG model of colorectal classification is shared to Unet segmentation model to deal with distinguishing polyp and normal tissues.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing. A total of 50 patients were examined under colonoscopies, and images were collected from the anorectal department of a hospital in Shaanxi Province, China, under ethical approval. Three experienced endoscopists were invited to classify the normal tissue, polyp, and tumor, and the ground truth was acquired. The data preprocessing was as follows:

Firstly, *data filtering*: uncleaned or unclear colorectal images were removed. After image filtering, the set of endoscopic images consisted of 487, 1374, and 1200 images with normal tissue, polyp, and tumor, respectively, taken under either white light (WL) or narrow band imaging (NBI) endoscopy.

Secondly, *dataset split*: the data were divided into training set, validation set, and test set according to the ratio of $2:1:2$.

Thirdly, *data argumentation*: the argumentation methods were rotation, flip, translation, and cropping. The training set and validation set were argumentized by four times

Finally, *data resizing*: the data was resized to $440 \times 440 \times 3$ to maintain the integrity of the intestinal wall.

2.2. Automatic Classification in Colorectal Endoscopy Based on STVGG. Because the performance of training network is poor by using colorectal images directly, Network pretrained on ImageNet is introduced to obtain a good classification result. Meanwhile, polyp areas are more difficult to be classified than normal tissue and tumor. This paper introduces self-paced regularization items to assign different sample weights for training samples. Self-paced learning injects the difficulty metric into the optimization model and updates the model parameters based on the current sample ordering and the metric based on the learning effect. It obtains a new round of difficulty ordering of samples and finally achieves the purpose of adaptive sample ordering.

In our method, in order to fully use data of ImageNet, the parameters of C_1 and C_2 from pretrained VGG19 model on ImageNet are transferred to STVGG. And the practical colorectal images are used as training data to update other layer parameters of the STVGG model. The self-paced learning algorithm is introduced to STVGG for dealing with those difficult and unbalance samples to improve classification performance. The overview of STVGG classification method is shown in Figure 1, where C_i represents the i^{th} convolutional patch, F represents the fully connected layer, and G represents the global average pooling layer. In this study, the first fully connected layer F_6 of VGG19 is replaced with the global average pooling layer G_6 to reduce the amount of model parameters and prevent overfitting and to get the pretrained model with the parameters of C_1 and C_2 layers.

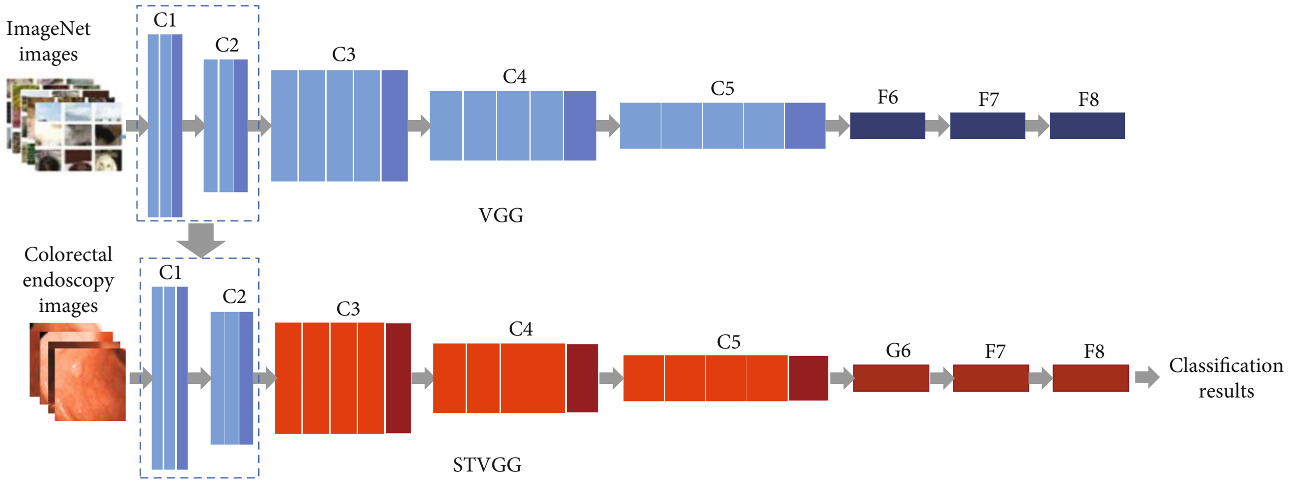


Figure 1: The overview of STVGG classification method.

The cross-entropy is selected as loss function $l(y_i, g(\mathbf{X}_i))$. Forward propagation is used to calculate the loss l in STVGG network:

$$l(y_i, g(\mathbf{X}_i)) = -\sum_{j=1}^{c} I\{y_i = j\} \log\left(\frac{e^{\boldsymbol{\omega}_j^T z_i^l}}{\sum_{k=1}^{c} e^{\boldsymbol{\omega}_k^T z_i^l}}\right). \qquad (1)$$

In the Eq. (1), c is the number of disease categories in the colorectal endoscopy dataset, and $\boldsymbol{\omega}_j^T$ is the weight parameter of the j^{th} output node. z_i^l represents the output vector in the last fully connected layer of the network. $I\{y_i = j\} \in \{0, 1\}$, when the predicted result of the sample is consistent with the label $I = 1$, otherwise $I = 0$.

Furthermore, self-paced learning is used to modify the STVGG network. The network objective function is rewritten as follows:

$$\min_{\omega,b} E_{\text{spl}}(w, b) = \min_{\omega,b}\left[\frac{1}{n}\sum_{i=1}^{n} v_i \cdot l(y_i, g(X_i)) + f(v_i, \lambda)\right] \qquad (2)$$

Parameters w and b are the weight and bias of the STVGG network, respectively, $\mathbf{v} = [v_1 \cdots v_i \cdots v_n]$ represents the weight of n samples, and $f(v_i, \lambda)$ is the binary self-paced regular term defined in Eq. (3).

$$f^H(v_{\text{i}}, \lambda) = -\lambda v_i \,; v_i * (l, \lambda) = \begin{cases} 1 \text{ if } l < \lambda, \\ 0 \text{ if } l \geq \lambda. \end{cases} \qquad (3)$$

w, b, and $\mathbf{v}$ of the STVGG network are optimized by iteration until the model converges to get a good classification network. Flowchart of STVGG algorithm is shown in Algorithm 1.

2.3. Automatic Segmentation in Polyp Image. After classification task is completed, the parameters $C_1 - C_5$ of the trained STVGG in the colorectal endoscopy classification task is shared to the segmentation task as the code part, while the Unet network framework is used in colorectal endoscopy segmentation task. And the decoding part of the original Unet is also adjusted with the corresponding encode part. Compared with the original Unet, the channel number of downsampling in the last layer is not increased for the proposed model. The framework of our segmentation model is shown in Figure 2.

Each rectangular box corresponds to a multichannel feature map. The number on the left side of the rectangular box indicates the size of each channel of the feature map. The number at the top of the rectangle indicates the channel number in the feature map. The blue, red, green, and purple arrows indicate the convolution operation with a convolution kernel size of 3 × 3, the max pooling with stride of 2 × 2, and the upsampling and the convolution with a convolution kernel size of 1 × 1, respectively. The gray arrow indicates that the feature map of the encoding part is cropped and copied with the feature map of the decoding part.

2.4. Comparison with Other Methods. The selection of comparison methods is based on the baseline VGG model adding some training strategies, and the specific strategies are as follows:

(1) VGG19 with transfer learning strategies (VGG+TL)

The parameters $C_1 - C_2$ in VGG19 are transferred from ImageNet network to extract low-level features well shared with natural images in colorectal endoscopy images.

(2) VGG19 with the strategy of structure retention color normalization (VGG+SRCN)

The data are collected from different periods, different patients, and equipment in different periods. Therefore, SRCN strategy is used so that color features of processed image tend to be consistent and reduce intraclass differences.

(3) VGG19 with strategy of spatial pyramid pooling (VGG+SPP)

SPP [26] layer on the VGG19 network is adopted, and it obtains a fixed length feature vector to aggregate the features and avoid geometric distortion in feature maps.

```
Flowchart of STVGG algorithm.
Input: Training set $\mathbf{D} = \{(\mathbf{x}_i, y_i), i = 1 \cdots n\}$, $\mathbf{x}_i$ represents the $i^{\text{th}}$ training data, and $y_i$ is the $i^{\text{th}}$ data label.
Initialization parameter: "age parameter"$\lambda$, a suitable initial value is given according to the approximate value range of the presample training error value; initialize the sample weight vector $\mathbf{v}$.
Model training parameter settings: total number of training iterations epoch, minimum batch size for training and verification, initial learning rate during model training $\alpha$, and decay rate of learning rate $\varphi$; update increment of age parameters $k$, $k > 0$.
a) Calculate network weights $\mathbf{w}$ and bias $\mathbf{b}$ by Eq. (2)
b) Calculate and update loss function $l(y_i, g(\mathbf{X}_i))$
c) Calculate self-paced regular term $f(v_i, \lambda)$ and update weight vector $\mathbf{v}$
d) Calculate and update $\min_{\omega,b} E_{\text{spl}}(w, b)$
e) Update age parameters $\lambda$ and learning rate $\alpha$, $\lambda = \lambda + k$, $\alpha = a \cdot \varphi$, $\varphi < 1$
f) Repeat steps a to e until the number of iterations epoch = 0
output: network weights $\mathbf{w}$ and bias $\mathbf{b}$
```

Algorithm 1.

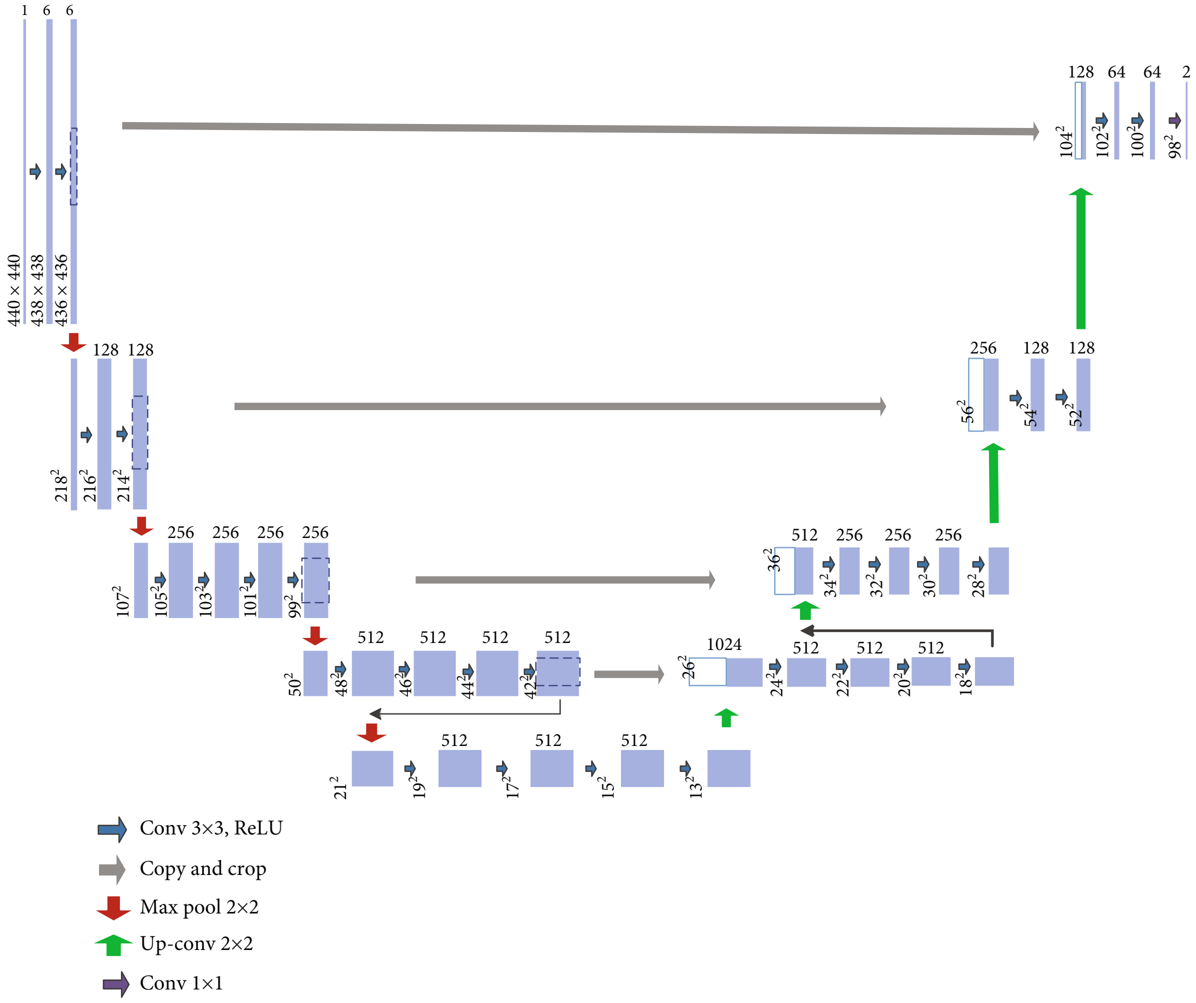


Figure 2: The framework of our segmentation model.

2.5. Evaluation of the Classification and Segmentation. The segmentation results are evaluated both visually and quantitatively, given the ground truth, our classification and segmentation results. The segmentation performance is evaluated by these evaluation metrics: accuracy, sensitivity, specificity, and dice similarity coefficient (DSC). We use TP, FP, TN, and FN to represent true positive, false positive, true negative, and false negative. And L_1 and L_2 represent the

Table 1: Classification accuracy obtained for different methods.

Category	VGG	VGG+SRCN	VGG+SPP	VGG+TL	STVGG
Tumor	0.98	0.98	0.94	0.96	0.98
Normal tissue	0.90	0.94	0.99	0.95	0.99
Polyp	0.7	0.84	0.91	0.89	0.95
Average accuracy	0.76	0.87	0.93	0.91	0.96

manual annotation and our method segmentation results, respectively, and these indexes are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (4)$$

$$\text{Sensitivity} = \frac{|L_1 \cap L_2\,|}{|L_1\,|}, \quad (5)$$

$$\text{Specificity} = \frac{|\sim (L_1 \mid L_2)|}{|\sim L_1\,|}, \quad (6)$$

$$\text{DSC} = \frac{2\,|\,L_1 \cap L_2\,|}{(|L_1|+|L_2\,|\,)}. \quad (7)$$

3. Experimental Results and Discussion

In this paper, the experimental environment was set in Python3.6.0, Tensorflow-GPU 1.7.0, Keras 2.1.3, SimpleITK 0.8.1, Nvidia Titan Pascal GPU (1080 Titan), and Cudnn V9.0. Verified by experiment, the colorectal image size of 440 was superior to 330 and 540. The average accuracy of the final test set was 0.03 and 0.01 higher than the latter two, respectively. The colorectal data with original image size range of 228 to 586 was resized to 440 × 440.

Choosing optimization functions. Experiments showed that the average classification accuracy of SGD is 0.02 and 0.05 higher than RmSprop and Adam, respectively. Therefore, SGD was finally selected as our optimization function.

Selecting transfer layers. Experiments froze the parameters of the 4, 8, 12, and 16 layers, respectively. It showed that when the parameters of the first 4 layers were frozen, the best average classification accuracy was achieved.

Cross-entropy are chosen as loss function. The relevant parameters were as follows: learning rate is 0.0001, decay = $1e-6$, and the parameter of Nesterov Momentum was set to 0.9. The batch size was 8. λ was initialized to 1.1, and the updating step was 0.05. As training began, λ became larger and the tolerance of difficult samples was greater.

3.1. Colorectal Endoscopy Image Classification. The colorectal endoscopy image classification accuracy is shown in Table 1. It can be seen from Table 1 that the polyp accuracy of VGG network was the lowest. The classification accuracies of VGG+SRCN on polyps and normal tissue are improved as it could decrease the intraclass differences. VGG+SPP also improved the classification accuracies of normal tissue and colorectal polyp but polyp accuracy was relatively low because the edges of normal tissue and polyp types are blurred. As an improvement strategy, the accuracy of VGG+TL was also improved. But compared with strategies of SRVN and SPP, the effect is not significant. In this study, STVGG was proposed and the experimental results showed that the overall accuracies were greatly improved.

The main reason is that STVGG can classify difficult-to-classify samples, for example, the small inflammatory or hyperplastic polyps which are very similar to normal colonoscopy images, and the polyps with ulcers, large areas of bleeding, and reticulated polyps, which are closer to the characteristics of tumor. The STVGG method can significantly improve the accuracy of polyp under the condition of ensuring the classification accuracy of tumor and normal ones, and the method converges in about 10 generations of training.

3.2. Polyp Image Segmentation. In the above colorectal endoscopy image classification task, a relatively good classification result was obtained by STVGG model. Therefore, polyp segmentation was designed based on classification task. Doctors usually used polyp's images to make a decision whether surgical resection is required based on pathological diagnosis. Figure 3 shows five sequences of polyp images. (Ai) is the original image, (Bi) is ground truth, (Ci) is the segmentation result of Segnet network, (Di) is the segmentation result of Unet network, (Ei) is the segmentation result of TLVGG network, and (Fi) is the segmentation result of STVGG network.

Figure 3 indicates that the results of Segnet method are greatly affected by the surrounding environment, and the segmentation result is not good. The results of Unet method are more superior than those of Segnet in big target but the effect is not obvious. Compared to those methods, the results of TLVGG method made great progress especially in surrounding and small target, but the segmentation results of large targets are not ideal. Our method shows the best results, no matter it is segmentation of large and small objects or environmental interference.

Table 2 shows that the Segnet segmentation method does not segment the polyp in its complete shape. Using Unet for segmentation of polyps is accurate, but oversegmentation is also obvious. The Unet is more sensitive to light spots in the imaging process, and it is easy to treat the reflective part as a polyp.

The segmentation performances of the TLVGG network were obviously better than Segnet and Unet. The segmentation target contour was close to the real target, but there were still missed detections. The STVGG model worked best because ImageNet network parameters were transferred to VGG network to acquire good initial network. And self-paced learning was used to optimize the network so that the classification performance of label unbalanced samples was improved.

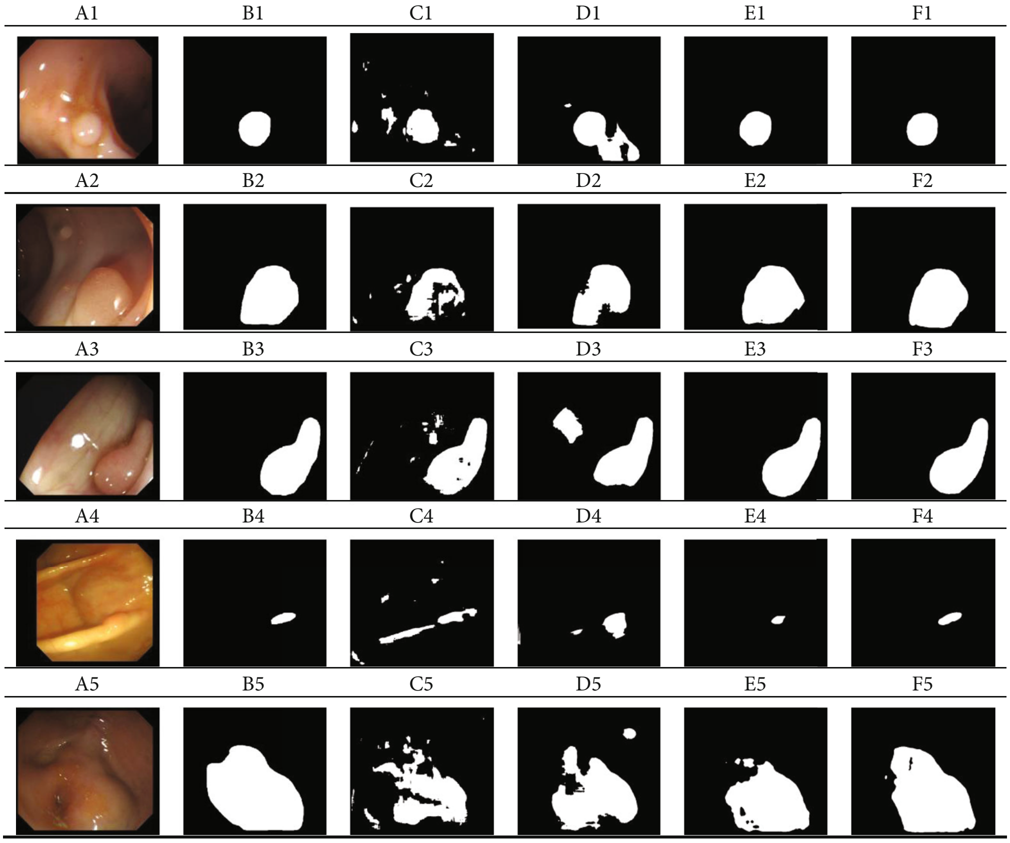


Figure 3: Segmentation in colorectal endoscopy images. (Ai) Original images, (Bi) ground truth, (Ci) Segnet, (Di) Unet, (Ei) TLVGG, and (Fi) ours.

Table 2: Segmentation indexes obtained from different methods.

	Segnet	Unet	TLVGG	STVGG
DSC	0.6210 ± 0.2370	0.6980 ± 0.3005	0.8267 ± 0.2066	0.8455 ± 0.2030
Sen	0.6916 ± 0.2677	0.7591 ± 0.3317	0.8222 ± 0.2462	0.8323 ± 0.2201
Spe	0.9766 ± 0.0180	0.9834 ± 0.0235	0.9933 ± 0.0095	0.9949 ± 0.0067

4. Conclusions

To address this issue of label unbalanced and difficult colorectal data, we presented an automatic processing pipeline for classification and segmentation based on colorectal images. STVGG network used transfer learning and self-paced learning in order to acquire good initial network and solve the problem of label unbalanced and difficult sample classification. And then STVGG network was shared as the encoding part of Unet as encoder of the segmentation task, and image segmentation task was achieved. The experimental illustrated that the proposed method obtained higher classification accuracy (96%) and segmentation performance compared with other a few methods. This proposed method may be applied to other image researches, such as stomach, ear, nose, and throat. Possible future improvements can be made in parameter adaptation.

Data Availability

The data used in the article are from the anorectal department of a hospital in Shaanxi Province, China, under ethical approval.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Yao Yao designed the algorithms and made the experiments. Shuiping Gou participated in guiding research plan. Ru Tian did the comparative experiment. Xiangrong Zhang polished the article. Shuixiang He provided the data and verified the clinical effect.

Acknowledgments

This study is supported by the Natural Science Foundation of Shaanxi Province under Grant Nos. 2019ZDLGY03-02-02 and 2017JM8026, Research Plan of Improving Public Scientific Quality in Shaanxi Province, No. E219360001, the Fundamental Research Funds for the Central Universities, No. JC2001, and Focus on developing of Shaanxi Province under Grant No. 2020GY-049. This is a research project of visiting engineers in colleges and universities of Zhejiang Province in 2020.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] A. G. Zauber, S. J. Winawer, M. J. O'Brien et al., "Colonoscopic polypectomy and long-term prevention of colorectal cancer deaths," *Obstetrical & Gynecological Survey*, vol. 67, pp. 687–696, 2012.

[3] K. Søreide, B. S. Nedrebø, A. Reite, K. Thorsen, and H. Kørner, "Endoscopy, morphology, morphometry and molecular markers: predicting cancer risk in colorectal adenoma," *Expert Review of Molecular Diagnostics*, vol. 9, no. 2, pp. 125–137, 2014.

[4] D. A. Joseph, R. G. S. Meester, and A. G. Zauber, "Colorectal cancer screening: estimated future colonoscopy need and current volume and capacity," *Cancer*, vol. 122, no. 16, pp. 2479–2486, 2016.

[5] M. J. Van den Bent, "Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective," *Acta Neuropathologica*, vol. 120, no. 3, pp. 297–304, 2010.

[6] D. G. Hewett, T. Kaltenbach, Y. Sano et al., "Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging," *Gastroenterology*, vol. 143, no. 3, pp. 599–607, 2012.

[7] M. G. Ertosun and D. L. Rubin, "Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks," in *In AMIA Annual Symposium Proceedings*, vol. 2015, pp. 1899–1908, American Medical Informatics Association, 2015.

[8] B. E. Bejnordi, M. Veta, P. J. Van Diest et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[9] A. Teramoto, T. Tsukamoto, and Y. Kiriyama, "Automated classification of lung cancer types from cytological images using deep convolutional neural networks," *Biomedical Research*, vol. 4067832, 2017.

[10] J. Wang et al., "Notice of Violation of IEEE Publication Principles: Bag-of-Features Based Medical Image Retrieval via Multiple Assignment and Visual Words Weighting," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1996–2011, 2011.

[11] A. Meier, K. Nekolla, S. Earle et al., "End-to-end learning to predict survival in patients with gastric cancer using convolutional neural networks," *Annals Oncology*, vol. 29, 2018.

[12] G. Litjens, C. I. Sánchez, N. Timofeeva et al., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.

[13] H. Y. Chang, C. K. Jung, J. I. Woo et al., "Artificial intelligence in pathology," *Journal of pathology and translational medicine*, vol. 53, no. 1, pp. 1–12, 2019.

[14] Y. Xu, Z. Jia, L. B. Wang et al., "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC bioinformatics*, vol. 18, p. 281, 2017.

[15] N. Thakur, H. Yoon, and Y. Chong, "Current trends of artificial intelligence for colorectal cancer pathology image analysis: a systematic review," *Cancers*, vol. 12, no. 7, p. 1884, 2020.

[16] B. Korbar, A. M. Olofson, A. P. Miraflor et al., "Deep learning for classification of colorectal polyps on whole-slide images," *Journal of pathology informatics*, vol. 8, 2017.

[17] P. Sena, R. Fioresi, F. Faglioni, L. Losi, G. Faglioni, and L. Roncucci, "Deep learning techniques for detecting preneoplastic and neoplastic lesions in human colorectal histological images," *Oncology Letters*, vol. 18, pp. 6101–6107, 2019.

[18] O. Iizuka, F. Kanavati, K. Kato, M. Rambeau, K. Arihiro, and M. Tsuneki, "Deep learning models for histopathological classification of gastric and colonic epithelial tumours," *Scientific Reports*, vol. 10, p. 1504, 2020.

[19] G. Urban, P. Tripathi, T. Alkayali et al., "Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy," *Gastroenterology*, vol. 155, pp. 1069–1078.e8, 2018.

[20] P. Klare, C. Sander, M. Prinzen et al., "Automated polyp detection in the colorectum: a prospective study (with videos)," *Gastrointestinal Endoscopy*, vol. 89, pp. 576–582.e1, 2019.

[21] P. Wang, T. M. Berzin, J. R. G. Brown et al., "Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study," *Gut*, vol. 68, no. 10, pp. 1813–1819, 2019.

[22] Q. Li, G. Yang, and Z. Chen, "Colorectal polyp segmentation using a fully convolutional neural network," in *2017 10th International Congress on Image and Signal Processing*, pp. 17–21, ShangHai, 2017.

[23] Y. J. Yang, B.-J. Cho, M.-J. Lee et al., "Automated classification of colorectal neoplasms in white-light colonoscopy images via deep learning," *Journal of clinical medicine*, vol. 9, 2020.

[24] R. Zhang, Y. Zheng, and T. W. C. Mak, "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain," *IEEE Journal of biomedical and health informatics*, vol. 21, no. 1, pp. 41–47, 2017.

[25] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Medical Image Analysis*, vol. 60, p. 101619, 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2014.

Hindawi
BioMed Research International
Volume 2021, Article ID 2043830, 9 pages
https://doi.org/10.1155/2021/2043830

Research Article

DVH Prediction for VMAT in NPC with GRU-RNN: An Improved Method by Considering Biological Effects

Yongdong Zhuang,[1,2] Yaoqin Xie,[2] Luhua Wang,[1,3] Shaomin Huang,[4] Li-Xin Chen,[4] and Yuenan Wang[5]

[1] *Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen 518116, China*
[2] *Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[3] *Department of Radiation Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China*
[4] *Department of Radiation Oncology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou 510060, China*
[5] *Department of Radiation Oncology, Peking University Shenzhen Hospital, Shenzhen 518036, China*

Correspondence should be addressed to Luhua Wang; wlhwq@yahoo.com, Li-Xin Chen; chenlx@sysucc.org.cn, and Yuenan Wang; yuenan.wang@gmail.com

Received 7 August 2020; Revised 28 October 2020; Accepted 4 January 2021; Published 19 January 2021

Academic Editor: Damiano Caruso

Purpose. A recurrent neural network (RNN) and its variants such as gated recurrent unit-based RNN (GRU-RNN) were found to be very suitable for dose-volume histogram (DVH) prediction in our previously published work. Using the dosimetric information generated by nonmodulated beams of different orientations, the GRU-RNN model was capable of accurate DVH prediction for nasopharyngeal carcinoma (NPC) treatment planning. On the basis of our previous work, we proposed an improved approach and aimed to further improve the DVH prediction accuracy as well as study the feasibility of applying the proposed method to relatively small-size patient data. *Methods*. Eighty NPC volumetric modulated arc therapy (VMAT) plans with local IRB's approval in recent two years were retrospectively and randomly selected in this study. All these original plans were created using the Eclipse treatment planning system (V13.5, Varian Medical Systems, USA) with ≥95% of PGTVnx receiving the prescribed doses of 70 Gy, ≥95% of PGTVnd receiving 66 Gy, and ≥95% of PTV receiving 60 Gy. Among them, fifty plans were used to train the DVH prediction model, and the remaining were used for testing. On the basis of our previously published work, we simplified the 3-layer GRU-RNN model to a single-layer model and further trained every organ at risk (OAR) separately with an OAR-specific equivalent uniform dose- (EUD-) based loss function. *Results*. The results of linear least squares regression obtained by the new proposed method showed the excellent agreements between the predictions and the original plans with the correlation coefficient $r = 0.976$ and 0.968 for EUD results and maximum dose results, respectively, and the coefficient r of our previously published method was 0.957 and 0.946, respectively. The Wilcoxon signed-rank test results between the proposed and the previous work showed that the proposed method could significantly improve the EUD prediction accuracy for the brainstem, spinal cord, and temporal lobes with a p value < 0.01. *Conclusions*. The accuracy of DVH prediction achieved in different OARs showed the great improvements compared to the previous works, and more importantly, the effectiveness and robustness showed by the simplified GRU-RNN trained from relatively small-size DVH samples, fully demonstrated the feasibility of applying the proposed method to small-size patient data. Excellent agreements in both EUD results and maximum dose results between the predictions and original plans indicated the application prospect in a physically and biologically related (or a mixture of both) model for treatment planning.

1. Introduction

Due to the complex tumor volumes in close proximity to critical structures, the nasopharyngeal carcinoma (NPC) radiation therapy (RT) plan was of great difficulty and experience-dependent [1–4]. In recent years, numbers of researches to aid in treatment planning using knowledge-based planning (KBP) techniques had improved the consistency of the plan quality and reduced the required optimization time [5–13]. The most popular tools [14–18] were developed to predict the dose-volume histogram from the organ at risk (OAR)—planning target volume (PTV) anatomy, which could assist in treatment planning by giving the appropriate OAR constraints and enabling the production of high-quality plans. The most widely used tools for quantifying the OAR-PTV anatomy, namely, the overlap volume histogram (OVH) [15, 16] and the distance-to-target histogram (DTH) [17, 18], were equivalent when the Euclidean form of the distance function was used in the DTH.

Compared to 3D-dose prediction [19–32] in the stage of academic research, DVH prediction has been clinically applied for years; for example, the commercial software named RapidPlan was developed based on the DTH approach by Varian Medical Systems (Palo Alto, California, US). However, one concern regarding the DTH and OVH was that their simplicity might lead to inaccurate presentation of the interpatient variations in anatomical features, which might have an impact on the dose deposition [12, 15, 33]. Another concern regarding the existent research was that the ignorance of the radiobiological difference in different structures or the different key features make dose distribution acceptable or unacceptable in clinic. For example, for an organ like the spinal cord, the maximum dose was considered to have the highest priority.

In our previously published works [12, 13], a multilayer gated recurrent unit-based recurrent neural network (GRU-RNN) was established to predict the DVHs for NPC treatment planning using the DVHs generated by the nonmodulated beams of different orientation. Using dosimetric information such as GRU-RNN inputs, the GRU-RNN was capable of accurate DVH prediction. Similar results were also obtained by other dosimetric information-driven researches [11, 30, 31]. RNN and its variants, such as the GRU-RNN used in this study, were particularly suitable for predicting the entire DVH. Its directionality was of great relevance for predicting the sequential data, such as DVH, a monotone decreasing sequence. And more importantly, compared to other models such as CNN, a great reduction of the parameter number in RNN and its variants indicated great potential in robust learning when applying to small-size data. The equivalent uniform dose (EUD) was the homogeneous dose inside an organ that has the same radiobiological effect as the given arbitrary dose distribution [34]. On the basis of our previous work, an EUD-based loss function was introduced in this study. By considering biological characters in different structures, the new method could pay more attention to the key dosimetric features such as maximum dose for the spinal cord and make the predicted DVH of more clinical value.

Aiming to improve the DVH prediction accuracy for NPC RT treatment planning, we proposed an improved approach in this work, which trained every OAR separately using a simplified GRU-RNN model with an equivalent uniform dose- (EUD-) based loss function, and study the feasibility of applying the proposed method to relatively small-size patient data.

2. Materials and Method

2.1. Data Acquisition. 80 NPC volumetric modulated arc therapy (VMAT) plans in recent two years with local IRB's approval were retrospectively and randomly selected for this study. Of these original plans, 50 were randomly selected for training and the remaining were used for testing. Following the ICRU-83 report, radiation oncologists delineated the gross tumor volume of the nasopharynx (GTVnx), the gross tumor volume of the metastatic lymph node (GTVnd), the clinical target volume (CTV), and the OARs in the planning CT. A margin of 3 mm was applied around the GTVnx, GTVnd, and CTV to create the planning GTVnx (PGTVnx), the planning GTVnd (PGTVnd), and the planning CTV (PTV), respectively. All the original VMAT plans were created using the Eclipse treatment planning system (V13.5, Varian Medical Systems, USA) with ≥95% of PGTVnx receiving the prescribed doses of 70 Gy, ≥95% of PGTVnd receiving 68 Gy, and ≥95% of PTV receiving 60 Gy. In this work, the DVHs were resampled by volume bin in percentage (1% in practice) rather than in absolute volume or dose values, making the DVHs of equal length. The DVHs of the nonmodulated beams were generated by a nine-field conformal plan with multileaf collimators fitting to PTV and normalizing 95% of PGTV dose to 70 Gy. An example of DVHs induced by nonmodulated beams and that of the original plan from a patient's spinal cord are shown in Figure 1.

2.2. GRU-RNN. A single-layer GRU-RNN as shown in Figure 2 was established using the PyTorch (Facebook, US) framework for DVH prediction. D_v was the dose of original plans at percent volume v as shown in Figure 1, D'_v was the predicted dose, and h_v was the hidden state at volume v. A dropout layer was inserted between GRU and FC to randomly zero the parameters of h_v with a probability of 0.5. The GRU was trained by the Adam optimizer with the goal of minimizing the loss function by a learning rate of $1e-3$.

2.3. Loss Function. The concept of equivalent uniform dose (EUD) assumes that any two dose distributions are equivalent if they cause the same radiobiological effect, which can be calculated as follows [34]:

$$\mathrm{EUD} = \left(\sum_{i=0} v \cdot D_i^a \right)^{1/a}. \qquad (1)$$

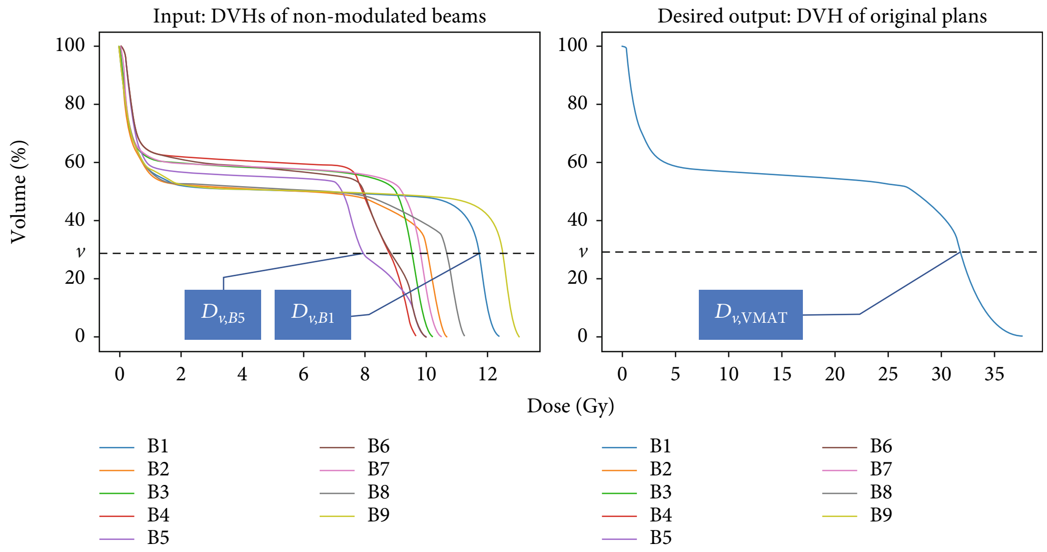


Figure 1: An example of DVHs generated by nonmodulated beams, $B1, B2, B3, \cdots, B9$, with a gantry angle of 160, 200, 240, 280, 320, 0, 40, 80, and 120 degrees and DVH of an original VMAT plan from a patient spinal cord.

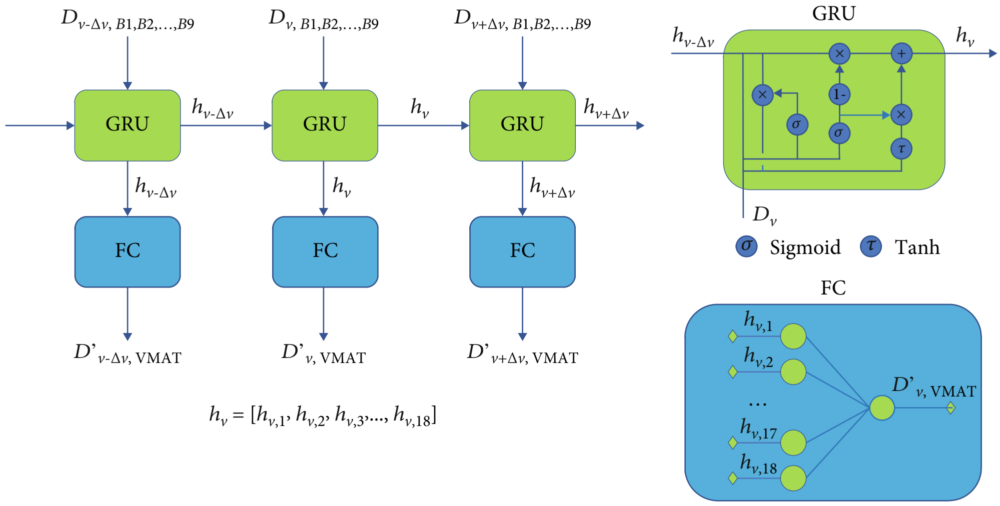


Figure 2: Architecture of the GRU-RNN model. In the practical experiments, Δv was 1% OAR volume.

Equation (2) shows that different dose values take different weights in EUD calculation when $a \neq 1$.

$$\frac{d(\mathrm{EUD})}{d(D_i)} = (\mathrm{EUD})^{1-a} * D_i^{a-1}, \qquad (2)$$

$$s(D_i, a) = \frac{D_i^{a-1}}{\sum_{i=0} D_i^{a-1}}. \qquad (3)$$

$s(a)$ represents the sensitivity of the dose value to the EUD value. The loss function to be minimized in this study is defined as equation (4) to meet the different dose requirements for different OARs. For example, for an OAR like the spinal cord, the maximum dose is considered to have the highest priority; therefore, the GRU-RNN model was individually trained with $k \gg 1$.

$$f\left(\mathrm{DVH}', \mathrm{DVH}, k\right) = \frac{1}{n}\sum_{p}^{n}\sum_{i=0} s(D_i, k)\left(D_i' - D_i\right)^2. \qquad (4)$$

Here, DVH and DVH$'$ were the DVH of the original plan and prediction. k was a positive integer and determined by trial and error with the goal of accurately predicting both EUD and maximum dose (only for serial OARs). The trial and error results were shown, when 8 was used for the brainstem, 15 for the spinal cord, 3 and 2 for the left and right optic nerves, respectively, and 1 for the chiasm, larynx, parotid glands, and temporal lobes; the most accurate EUD results (recommended by Allen Li et al. [35], $a = 8$ was used for serial

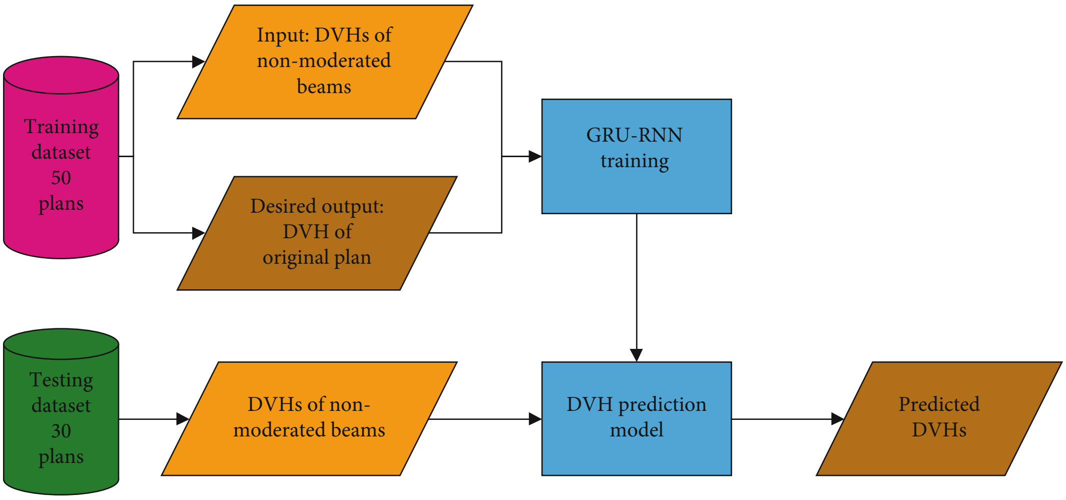

Training dataset 50 plans
Input: DVHs of non-moderated beams
Desired output: DVH of original plan
GRU-RNN training
Testing dataset 30 plans
DVHs of non-moderated beams
DVH prediction model
Predicted DVHs

Figure 3: The flowchart showed the dosimetric information of nonmodulated beam-driven DVH prediction.

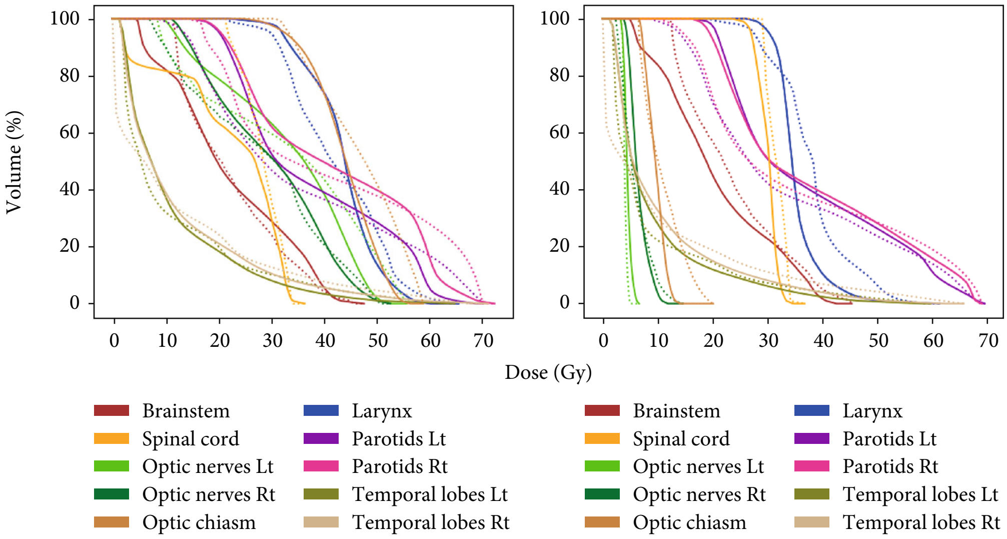


Figure 4: Comparisons between DVHs of predictions and original plans from two testing patients. Dash line: predicted DVHs; solid line: original plans' DVHs; Lt: left side; Rt: right side.

organs including the brainstem, spinal cord, optic nerves, and chiasm and $a = 1$ was used for parallel organs including the parotids, larynx, and temporal lobes in EUD calculation) were obtained. A flowchart of the dosimetric information of nonmodulated beam-driven DVH prediction is shown in Figure 3.

2.4. Model Evaluation. $\mu \pm \sigma$ was calculated to evaluate the GRU-RNN performance:

$$\delta_i = D_i' - D_i, \tag{5}$$

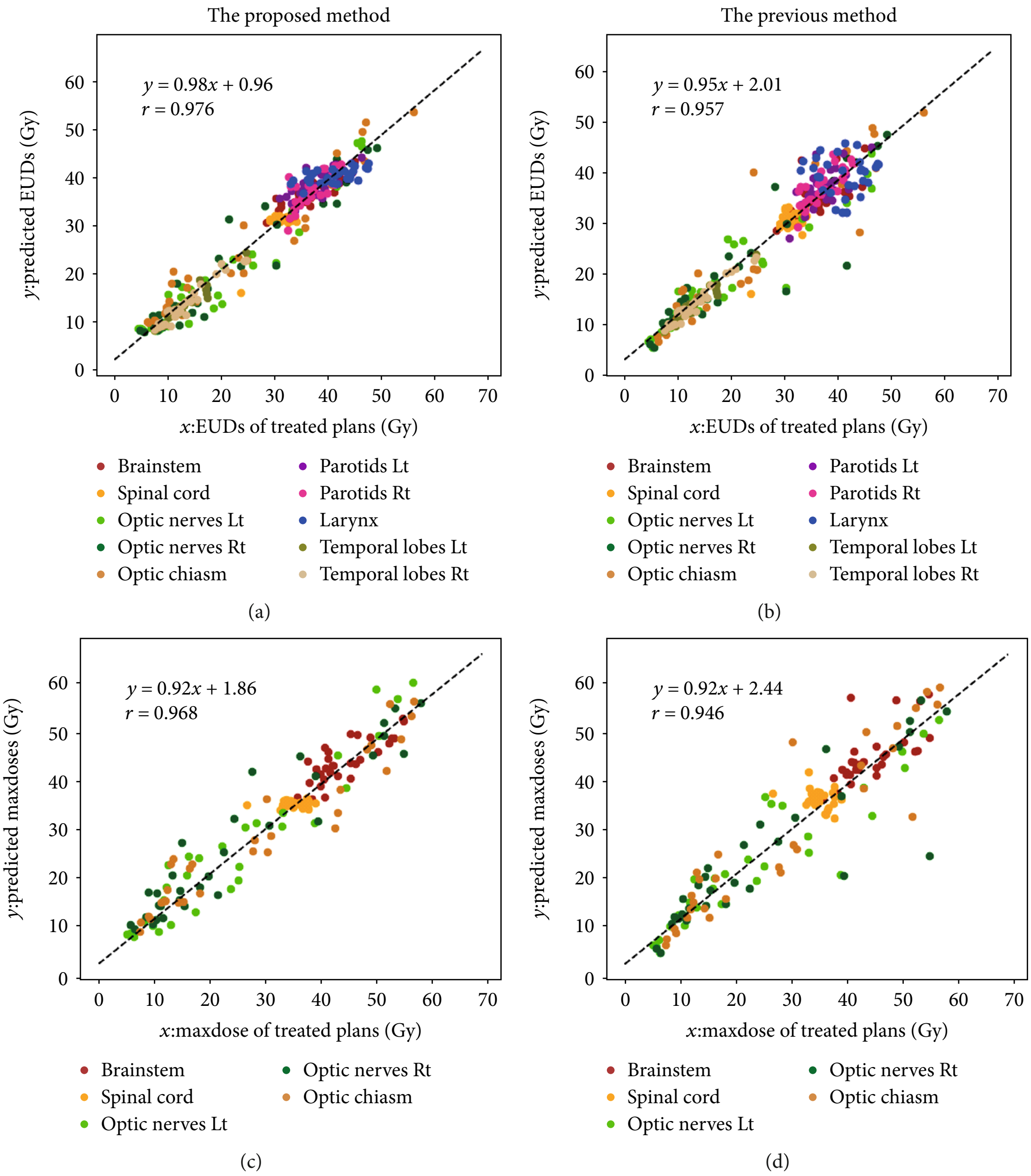
$$\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_i, \tag{6}$$

$$\sigma = \sqrt{\frac{\sum(\delta_i - \mu)^2}{n}}, \tag{7}$$

where i represents a testing patient and n is the number of testing patients, $D_i{}'$ denotes the predicted EUD or maximum dose, and D_i denotes the result of the original plan. The results were also compared to those obtained by our previous work [13] to demonstrate the improvements of the new proposed method in this study. Wilcoxon signed-rank tests were employed to compare the prediction error, $\delta_{i=1,2,\cdots,30}$, among the 30 testing patients between the proposed method and the previous one. Differences were considered statistically significant at $p < 0.05$.

3. Results

Two randomly selected testing patients' DVH prediction results are demonstrated in Figure 4. Though obvious differences could be seen in the deposited dose between two



(a) (b)

(c) (d)

Figure 5: The correlations of the EUD (a, c) and maximum dose (b, d) results between the predictions and the original plans for the 30 test patients.

patients especially at the optic nerves and chiasm, the predicted DVHs of the two patients were still very close to the original plans.

The correlations of the EUD and maximum dose results between the predictions and the original plans for the 30 test patients are plotted in Figure 5. The results of linear least squares regression showed that the predicted EUD results and maximum dose results of the proposed method in this study were both in good agreement with the original plans with correlation coefficient $r = 0.976$ (Figure 5(a)) and 0.968 (Figure 5(c)), respectively. The prediction results obtained following our previous work [13] are also demonstrated in Figures 5(b) and 5(d), and the coefficient r values were 0.957 (Figure 5(a)) and 0.946 (Figure 5(c)). The proposed method in this study has better consistency between the predicted results and those of original plans, which could be seen in the scatter plots and the coefficient r results quantitatively.

Table 1 provides a summary ($\mu \pm \sigma$) and a comparison (p value) over the prediction error ($\delta_{i=1,2,\cdots,30}$) of the EUD results and maximum dose results in the 30 testing patients. The $\mu \pm \sigma$ results showed σ decreased by the proposed method in almost all the OARs except for temporal lobes for both maximum doses and EUDs. The patient-wise Wilcoxon signed-rank tests results over δ_i between the proposed and the previous method [12] showed that the proposed

Table 1: A summary ($\mu \pm \sigma$ over δ) and a comparison (patient-wise p value over δ) of prediction accuracy in maximum doses and EUDs for the 30 test patients. Prop was the results obtained by the proposed method, and Prev was the results obtained by the previous method.

OARs	δ	$\mu \pm \sigma$ (Gy)		p value
		Prop	Prev	Prop vs. Prev
Brainstem	$D_{\max}$	0.34 ± 3.22	−0.27 ± 4.23	0.16
	EUD	0.64 ± 2.61	−1.24 ± 3.11	<0.01
Spinal cord	$D_{\max}$	0.58 ± 2.29	0.79 ± 3.23	0.29
	EUD	0.10 ± 1.96	−0.62 ± 2.20	<0.01
Optic chiasm	$D_{\max}$	−0.33 ± 6.18	−0.50 ± 6.02	0.57
	EUD	0.089 ± 4.58	0.27 ± 5.27	0.51
Optic nerves Lt	$D_{\max}$	0.68 ± 4.13	−0.52 ± 6.11	0.27
	EUD	0.50 ± 3.13	−0.32 ± 4.65	0.43
Optic nerves Rt	$D_{\max}$	1.23 ± 4.70	0.02 ± 7.46	0.37
	EUD	1.22 ± 3.71	0.26 ± 5.20	0.41
Larynx	EUD	−0.66 ± 3.46	−1.48 ± 5.59	0.22
Parotids Lt	EUD	0.64 ± 2.32	0.12 ± 2.80	0.13
Parotids Rt	EUD	0.24 ± 2.47	0.07 ± 2.14	0.37
Temporal lobes Lt	EUD	0.17 ± 0.92	0.69 ± 0.84	<0.01
Temporal lobes Rt	EUD	0.32 ± 1.07	0.73 ± 0.98	<0.01

method could significantly improve the EUD prediction accuracy of the brainstem, spinal cord, and temporal lobes with p value < 0.01.

The differences between results of the original plans and the predictions obtained by the proposed method (Prop) as well as the previous method (Prev) were expressed with a boxplot in Figure 6. The bottom and top of each box were the 25th and 75th percentiles of the differences, respectively. The distance between the bottom and top of each box was the interquartile difference range, and the lines in the middle of each box were the median differences. The whiskers were lines extending above and below each box. Whiskers went from the end of the interquartile range to the largest difference. Differences beyond the whisker length were marked as outliers, which were more than 1.5 times the interquartile range away from the bottom or top of the box. The proposed method in this study as shown in Figure 6, compared to the previous method in both maximum doses and EUDs, had the median differences closer to 0, smaller interquartile differences, and less outliers indicating better prediction accuracy and better reliability.

4. Discussion

In this study, we proposed an improved approach, which trained every OAR separately with a simplified GRU-RNN model and an equivalent uniform dose- (EUD-) based loss function. As shown in Figure 5, the new proposed method in this study improved the consistency of EUD results and maximum dose results between the predictions and original plans. For parallel OARs such as the larynx, parotid glands, and temporal lobes, the k values in equation (4) were set to 1.0, making the $s(D_i, k)$ term ineffective. The improved prediction accuracy in these OARs indicated that training different OARs separately was helpful. In our preliminary trials, we had also trained the GRU-RNN with $k = 1$ for the brainstem and spinal cord, the EUD results of $\mu \pm \sigma$ were 0.76 ± 2.94 Gy and 0.29 ± 2.03 Gy, and the maximum dose results of $\mu \pm \sigma$ were 0.58 ± 3.16 Gy and 0.69 ± 2.37 Gy. The results showed the $s(D_i, k)$ term was able to improve the prediction accuracy of EUD and maximum dose. Excellent agreements in both EUD values and maximum doses between the predictions and original plans obtained by the new proposed method indicated the application prospect in a physically and biologically related (or a mixture of both) model for treatment planning.

The GRU-RNN models were trained from only 50 DVH samples, which could reasonably be considered relatively small-size data. The excellent agreements of the results between the predictions and original plans fully demonstrated the feasibility of applying the proposed method to small-size patient data. As mentioned above, RNN and its variants, such as GRU-RNN in this study, are particularly suitable for predicting the entire DVH rather than only fixed amount of interesting points. Compared to CNN and other models, its directionality was of great relevance for predicting the sequential data, such as DVH, a monotone decreasing sequence. In this study, we trained different OARs separately and focused the training attention on the interpatient variations in deposited dose with no need of figuring out the different OARs. Decreasing the training difficulty allowed the usage of a further simplified model, a single-layer GRU-RNN in this study, which was of great significance in small-size sample training. In addition, due to the greatly reduced complexity of the modeling task, the training time is less than 100 seconds for every OAR with a computer equipped with i7-4770K CPU, Geforce GTX Titan GPU, and 16 GB memory.

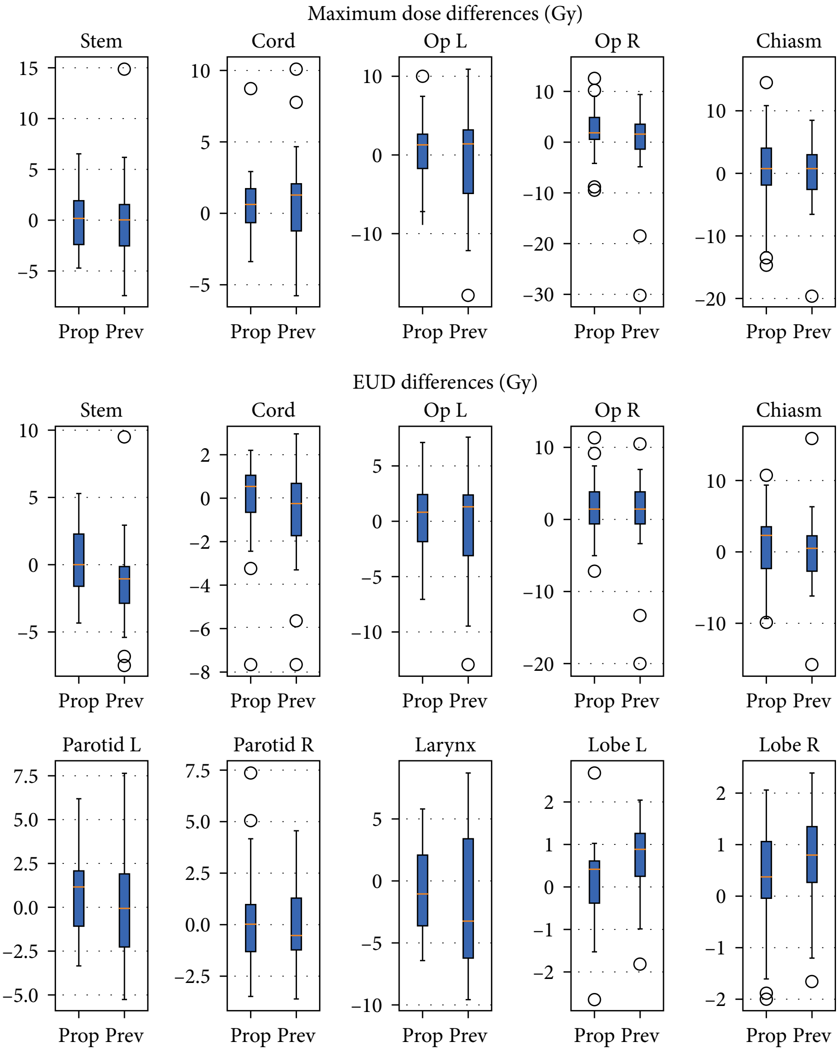


Figure 6: The differences between results of treated plans and the predicted results obtained by the proposed method (Prop) as well as the previous method (Prev) for the brainstem (stem), spinal cord (cord), left and right optic nerves (op L and op R), optic chiasm (chiasm), left and right parotid glands (parotid L and parotid R), larynx, and left and right temporal lobes (lobe L and lobe R).

Different k values of 3.0 and 1.6 for optic nerves seem unreasonable and illogical due to the similar size, the symmetrical distribution, and the same biological characteristics in left\right optic nerves. A possible reason might be that the small size of the samples was not enough to represent the broader cases. In other words, the proposed method in this study might be a possible way to backtrack the value of "a" in equation (1) with the results of the existing plan data, but the values in this study seem too data-dependent to be repeated.

5. Conclusion

The accuracy of DVH prediction achieved in different OARs showed the great improvements compared to the previous works [12, 13] and the potential of this approach being extended to other disease sites. More importantly, the effectiveness and robustness showed by the simplified and well-trained GRU-RNN models trained from relatively small-size DVH samples fully demonstrated the feasibility of applying the proposed method to small-size patient data. In addition, excellent agreements in both EUD values and maximum doses between the predictions and original plans indicated the application prospect in a physical and biologically related (or a mixture of both) model for treatment planning.

Data Availability

All datasets generated for this study are included in the article.

Ethical Approval

The studies involving human participants were reviewed and approved by IRB, CAMS Shenzhen Cancer Hospital.

Consent

Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Conflicts of Interest

The authors declare that they have no conflict of interest related to this work.

Authors' Contributions

All authors have reviewed and approved the final manuscript.

Acknowledgments

This study was sponsored by the National Natural Science Foundation of China (11375278) and Sanming Project of Medicine in Shenzhen (SZSM201812062).

References

[1] I. W.-K. Tham, S. W. Hee, R. M.-C. Yeo et al., "Treatment of nasopharyngeal carcinoma using intensity-modulated radiotherapy—the National Cancer Centre Singapore experience," *International Journal of Radiation Oncology • Biology • Physics*, vol. 75, no. 5, pp. 1481–1486, 2009.

[2] V. Batumalai, M. G. Jameson, D. F. Forstner, P. Vial, and L. C. Holloway, "How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case," *Practical Radiation Oncology*, vol. 3, no. 3, pp. e99–e106, 2013.

[3] S. L. Berry, A. Boczkowski, R. Ma, J. Mechalakos, and M. Hunt, "Interobserver variability in radiation therapy plan output: results of a single-institution study," *Practical Radiation Oncology*, vol. 6, no. 6, pp. 442–449, 2016.

[4] B. E. Nelms, G. Robinson, J. Markham et al., "Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems," *Practical Radiation Oncology*, vol. 2, no. 4, pp. 296–305, 2012.

[5] J. P. Tol, A. R. Delaney, M. Dahele, B. J. Slotman, and W. F. A. R. Verbakel, "Evaluation of a knowledge-based planning solution for head and neck cancer," *International Journal of Radiation Oncology • Biology • Physics*, vol. 91, no. 3, pp. 612–620, 2015.

[6] J. P. Tol, M. Dahele, A. R. Delaney, B. J. Slotman, and W. F. A. R. Verbakel, "Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans?," *Radiation Oncology*, vol. 10, no. 1, p. 234, 2015.

[7] A. T. Y. Chang, A. W. M. Hung, F. W. K. Cheung et al., "Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy," *International Journal of Radiation Oncology • Biology • Physics*, vol. 95, no. 3, pp. 981–990, 2016.

[8] A. Fogliata, G. Reggiori, A. Stravato et al., "RapidPlan head and neck model: the objectives and possible clinical benefit," *Radiation Oncology*, vol. 12, no. 1, p. 73, 2017.

[9] M. M. Korani, P. Dong, and L. Xing, "MO-G-201-03: deep-learning based prediction of achievable dose for personalizing inverse treatment planning," *Medical Physics*, vol. 43, no. 6-Part32, pp. 3724–3724, 2016.

[10] G. Yu, Y. Li, Z. Feng et al., "Knowledge-based IMRT planning for individual liver cancer patients using a novel specific model," *Radiation Oncology*, vol. 13, no. 52, pp. 1–8, 2018.

[11] M. Ma, N. Kovalchuk, M. K. Buyyounouski, L. Xing, and Y. Yang, "Dosimetric features-driven machine learning model for DVH prediction in VMAT treatment planning," *Medical Physics*, vol. 46, no. 2, pp. 857–867, 2019.

[12] Y. Zhuang, J. Han, L. Chen, and X. Liu, "Dose-volume histogram prediction in volumetric modulated arc therapy for nasopharyngeal carcinomas based on uniform-intensity radiation with equal angle intervals," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 23NT03, 2019.

[13] W. Cao, Y. Zhuang, L. Chen, and X. Liu, "Application of dose-volume histogram prediction in biologically related models for nasopharyngeal carcinomas treatment planning," *Radiation Oncology*, vol. 15, no. 1, p. 216, 2020.

[14] K. L. Moore, R. S. Brame, D. A. Low, and S. Mutic, "Experience-based quality control of clinical intensity-modulated radiotherapy planning," *International Journal of Radiation Oncology • Biology • Physics*, vol. 81, no. 2, pp. 545–551, 2011.

[15] B. Wu, F. Ricchetti, G. Sanguineti et al., "Patient geometry-driven information retrieval for IMRT treatment plan quality control," *Medical Physics*, vol. 36, no. 12, pp. 5497–5505, 2009.

[16] B. Wu, F. Ricchetti, G. Sanguineti et al., "Data-driven approach to generating achievable dose–volume histogram objectives in intensity-modulated radiotherapy planning," *International Journal of Radiation Oncology • Biology • Physics*, vol. 79, no. 4, pp. 1241–1247, 2011.

[17] X. Zhu, Y. Ge, T. Li, D. Thongphiew, F. F. Yin, and Q. J. Wu, "A planning quality evaluation tool for prostate adaptive IMRT based on machine learning," *Medical Physics*, vol. 38, no. 2, pp. 719–726, 2011.

[18] L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu, "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans," *Medical Physics*, vol. 39, no. 11, pp. 6868–6878, 2012.

[19] S. Shiraishi and K. L. Moore, "Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy," *Medical Physics*, vol. 43, no. 1, pp. 378–387, 2016.

[20] W. G. Campbell, M. Miften, L. Olsen et al., "Neural network dose models for knowledge-based planning in pancreatic SBRT," *Medical Physics*, vol. 44, no. 12, pp. 6148–6158, 2017.

[21] Q. Jia, T. Song, Y. Li et al., "OAR dose distribution prediction and gEUD based automatic treatment planning optimization for intensity modulated radiotherapy," *IEEE Access, Access, IEEE.*, vol. 7, pp. 141426–141437, 2019.

[22] D. Nguyen, T. Long, X. Jia et al., "Dose prediction with u-net: a feasibility study for predicting dose distributions from contours using deep learning on prostate IMRT patients," p. ar Xiv: 1709.09233, 2017.

[23] R. Mahmood, A. Babier, A. McNiven, A. Diamant, and T. C. Y. Chan, "Automated treatment planning in radiation therapy using generative adversarial networks," *Proceedings of Machine Learning Research*, vol. 85, pp. 1–15, 2018.

[24] V. Kearney, J. W. Chan, S. Haaf, M. Descovich, and T. D. Solberg, "DoseNet: a volumetric dose prediction algorithm using 3D fully-

convolutional neural networks," *Physics in Medicine and Biology*, vol. 63, no. 23, 2018.

[25] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu, "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique," *Medical Physics*, vol. 46, no. 1, pp. 370–381, 2019.

[26] X. Chen, K. Men, Y. Li, J. Yi, and J. Dai, "A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning," *Medical Physics*, vol. 46, no. 1, pp. 56–64, 2019.

[27] D. Nguyen, X. Jia, D. Sher et al., "3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture," *Physics in Medicine & Biology.*, vol. 64, no. 6, article 065020, 2019.

[28] A. Babier, R. Mahmood, A. L. McNiven, A. Diamant, and T. C. Y. Chan, "Knowledge-based automated planning with three-dimensional generative adversarial networks," *Medical Physics*, vol. 427, no. 2, pp. 297–306, 2018.

[29] A. M. Barragán-Montero, D. Nguyen, W. Lu et al., "Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations," *Medical Physics*, vol. 46, no. 8, pp. 3679–3691, 2019.

[30] Y. Xing, D. Nguyen, W. Lu, M. Yang, and S. Jiang, "Technical note: a feasibility study on deep learning-based radiotherapy dose calculation," *Medical Physics*, vol. 47, no. 2, pp. 753–758, 2019.

[31] M. Ma, N. Kovalchuk, M. K. Buyyounouski, L. Xing, and Y. Yang, "Incorporating dosimetric features into the prediction of 3D VMAT dose distributions using deep convolutional neural network," *Physics in Medicine & Biology*, vol. 64, no. 12, article 125017, 2019.

[32] P. Dong and L. Xing, "Deep DoseNet: a deep neural network for accurate dosimetric transformation between different spatial resolutions and/or different dose calculation algorithms for precision radiation therapy," *Physics in Medicine and Biology*, vol. 65, no. 3, article 035010, 2020.

[33] S. X. Jiao, L. X. Chen, J. H. Zhu, M. L. Wang, and X. W. Liu, "Prediction of dose-volume histograms in nasopharyngeal cancer IMRT using geometric and dosimetric information," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 23NT04, 2019.

[34] A. Niemierko, "A generalized concept of equivalent uniform dose (EUD)," *Medical Physics*, vol. 26, no. 6, p. 1100, 1999.

[35] X. Allen Li, M. Alber, J. O. Deasy et al., "The use and QA of biologically related models for treatment planning: short report of the TG-166 of the therapy physics committee of the AAPM," *Medical Physics*, vol. 39, no. 3, pp. 1386–1409, 2012.

Hindawi
BioMed Research International
Volume 2020, Article ID 5042356, 8 pages
https://doi.org/10.1155/2020/5042356

Research Article

Differential Diagnosis of Solitary Fibrous Tumor/Hemangiopericytoma and Angiomatous Meningioma Using Three-Dimensional Magnetic Resonance Imaging Texture Feature Model

Junyi Dong,[1] Meimei Yu,[2] Yanwei Miao,[1] Huicong Shen,[2] Yi Sui,[3] Yangyingqiu Liu,[1] Liang Han,[1] Xiaoxin Li,[1] Meiying Lin,[2] Yan Guo,[4] and Lizhi Xie[5]

[1]*Department of Radiology, The First Affiliated Hospital of Dalian Medical University, Dalian 116000, China*
[2]*Department of Radiology, Beijing Tian Tan Hospital, Capital Medical University, Beijing 100050, China*
[3]*Department of Hepatobiliary Surgery, The First Affiliated Hospital of Dalian Medical University, Dalian 116000, China*
[4]*Life Science, GE Healthcare, Shenyang 110000, China*
[5]*GE Healthcare, MR Research China, Beijing, China*

Correspondence should be addressed to Yanwei Miao; ywmiao716@163.com and Huicong Shen; shenhuicong@126.com

Received 9 August 2020; Revised 25 October 2020; Accepted 16 November 2020; Published 2 December 2020

Academic Editor: Wenjian Qin

Background. Intracranial solitary fibrous tumor(SFT)/hemangiopericytoma (HPC) is an aggressive malignant tumor originating from the intracranial vasculature. Angiomatous meningioma (AM) is a benign tumor with a good prognosis. The imaging manifestations of the two are very similar. Thus, novel noninvasive diagnostic method is urgently needed in clinical practice. Texture analysis and model building through machine learning may have good prospects. *Aim*. To evaluate whether a 3D-MRI texture feature model could be used to differentiate malignant intracranial SFT/HPC from AM. *Method*. A total of 97 patients with SFT/HPC and 95 with AM were included in this study. Patients from each group were randomly divided into the train (70%) and test (30%) sets. ROIs were drawn along the edge of the tumor on each section of T1WI, T2WI, and contrasted T1WI using ITK-SNAP software. The segmented image was imported into the AK software for texture feature extraction, and the 3D ROI signal intensity histograms of T1WI, T2WI, and contrasted T1WI were automatically obtained along with all the parameters. Modeling was performed using the language R. Confusion matrix was used to analyze the accuracy of the model. ROC curve was constructed to assess the grading ability of the logistic regression model. *Results*. After Lasso dimension reduction, 5, 9, and 7 texture features were extracted from T1WI, T2WI, and contrasted T1WI, respectively; additional 8 texture features were extracted from the combined sequence for modeling. The ROC analyses on four models resulted in an area under the curve (AUC) of 0.885 (sensitivity 76.1%, specificity 87.9%) for T1WI model, 0.918 (73.1%, 95.5%) for T2WI model, 0.815 (55.2%, 93.9%) for contrasted T1WI model, and 0.959 (92.5%, 84.8%) for the combined sequence model and were enough to correctly distinguish the two groups in 71.2%, 81.4%, 69.5%, and 83.1% of cases in test set, respectively. *Conclusions*. The radiological model based on texture features could be used to differentiate SFT/HPC from AM.

1. Introduction

Intracranial solitary fibrous tumor (SFT)/hemangiopericytoma (HPC) is a rare malignant tumor originating from the intracranial vasculature, which comprises only 1% of all primary central nervous system (CNS) tumors [1]. In the past, it was believed that intracranial SFT/HPC originates from the meninges and thus was considered as a subtype of meningioma [1]. However, with the development of molecular genetics, it was discovered that SFT/HPC originates from arachnoid cap cells [2]. SFT/HPC is an aggressive type of neoplasma, which can easily relapse and metastasize to extracranial tissues.

Angiomatous meningioma (AM) is a rare World Health Organization (WHO) grade I histological subtype of meningioma with a good prognosis, accounting for 2.1% of all meningiomas [3]. AM can be effectively cured through resection. In radiological images, SFT/HPC mimics AM that is usually benign [3]. Therefore, preoperative identification of both is essential.

Screening MRI is the primary method to identify SFT/HPC and AM; yet, considering that images of both tumors are very similar, tumor differentiation can be very challenging. Imaging omics is aimed at maximizing the potential of medical imaging in disease diagnosis through high-dimensional image texture features containing pathophysiological information [4]. Previous studies have shown that texture analysis software can be used to segment the tumor area on the image and perform texture analysis [4–6]; briefly, the characteristic parameters in the image can be extracted for differential comparison, and the tumor imaging heterogeneity can be quantitatively analyzed to provide unrecognizable images by the naked eye. Objective information does not depend on the experience and subjective judgment of the imaging physician and has excellent clinical application value. So far, texture analysis has been applied to identify intracranial tumors [5, 6], grade meningioma [7], and for assessment and survival analysis of the therapeutic response of glioma to chemotherapy [6, 8–10]. More importantly, Kanazawa et al. [11] have suggested that magnetic resonance imaging texture analysis can be useful for distinguishing SFT/HPC from meningioma, especially AM. Still, his study has certain limitations: (1) it was a relatively small sample size retrospective study; (2) this study analyzed only three texture parameters.

This study adopted three-dimensional texture (3D texture) characteristics based on the overall tumor, which can more comprehensively and objectively reflect the heterogeneity of the tumor. The purpose of this study was to further improve the diagnostic levels of these two diseases by using the texture parameters of conventional MRI sequences and to build the models through machine learning.

2. Materials and Methods

2.1. Patient Selection. The institutional review board approved the current study. The preoperative MRI was performed on 95 patients with AM (47 males and 48 females; mean age: 51.54 ± 11.54 years) and 97 with SFT/HPC (47 males and 50 females; mean age: 42.97 ± 14.35 years) at our institution from May 2012 to March 2019. All MRI results were retrospectively analyzed.

2.2. Data Acquisition. All MR images were obtained with a 3.0 T MR imager (Signa HDxt; GE Medical Systems, Milwaukee, WI) with an eight-channel head coil. The imaging protocol included unenhanced axial and sagittal T1-weighted sequences, axial and coronal T2-weighted sequences, and contrast-enhanced axial, sagittal, and coronal T1-weighted sequences. The scanning parameters were T1WI (TR/TE, 350 msec/9 msec); T2WI (TR/TE, 3,500 msec/110 msec); thickness, 6.0 mm; spacing, 1.0 mm; FOV, 220 × 220 mm; matrix, 448 × 256; sagittal and coronal slice, 8.0 mm; and layer spacing, 2.0 mm. An enhanced scan bolus Gd-DTPA (DTPA magnetic display) was given intravenously at a concentration of 0.1 mmol/kg body weight with a flow rate of 3 ml/sec.

2.3. Image Processing. First, based on image segmentation of the whole tumor, all T1WI, T2WI, and contrasted T1WI data with Digital Imaging and Communication in Medicine (DICOM) format were transferred from the picture archiving and communication system (PACS) workstation (Centricity PACS 3.1.1.4, GE Healthcare) to ITK-SNAP software. Two radiologists (residents and deputy chief physicians), who were blind to the grouping, manually selected the regions of interest (ROIs) along the edge of the tumor parenchyma on the contrasted T1WI, T1WI, and T2WI images; T2WI and contrast-enhanced T1WI were used as a reference to determine tumor areas. The ROIs were then manually drawn along the margin of the tumor parenchyma in each slice, with the intent to encompass the whole tumor volume. Consequently, the ROIs of all layers were merged into a 3D ROI (see Figure 1). Finally, the segmented image was imported into the AK (Artificial intelligence kit) software for texture feature extraction, and the 3D ROI signal intensity histograms of T1WI, T2WI, and contrasted T1WI were automatically obtained along with all the parameters (see Figure 2).

2.4. Statistical Methods and Modeling. Modeling was performed using the language R (RStudio Version 1.0.143–© 2009-2016 RStudio, Inc.). Approx. 70% of cases from each group were classified into the train set (133 cases); AM group (66 cases) and SFT/HPC group (67 cases) were used to establish the model. The remaining 30% were classified into the test set (59 cases), AM group (29 cases) and SFT/HPC group (30 cases), to verify the accuracy of the established model.

A comparison of texture features in T1WI sequences was analyzed using independent sample t-test and Kruskal-Wallis test; a P value < 0.05 was considered statistically significant. Univariate logistic regression analysis ($P < 0.05$) and Spearman's correlation analysis ($P \geq 0.05$ or $P < 0.05$, $r < 0.9$) were used to screen for the parameters with high predictive power. T2WI and contrasted T1WI sequence texture feature used the Lasso method to reduce dimensionality and selected high-performance parameters. Parameters with high predictive power in the three sequences were further eliminated using the stepwise iterative method, and the remaining high-performance parameters were fed into a multivariate logistic regression analysis to determine an optimal logistic regression model for tumor classification. The confusion matrix was used to analyze the accuracy of the model. ROC curve was constructed to assess the grading ability of the logistic regression model.

3. Results

3.1. Establishment of T1WI, T2WI, and Contrasted T1WI Texture Feature Models. After applying the dimension reduction and stepwise iterative method, the high-performance parameters of the T1WI texture feature model were kurtosis

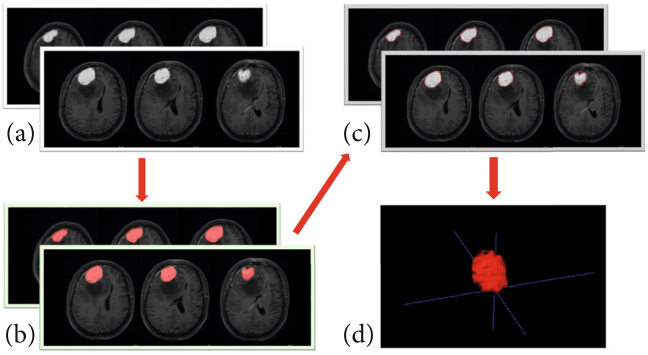
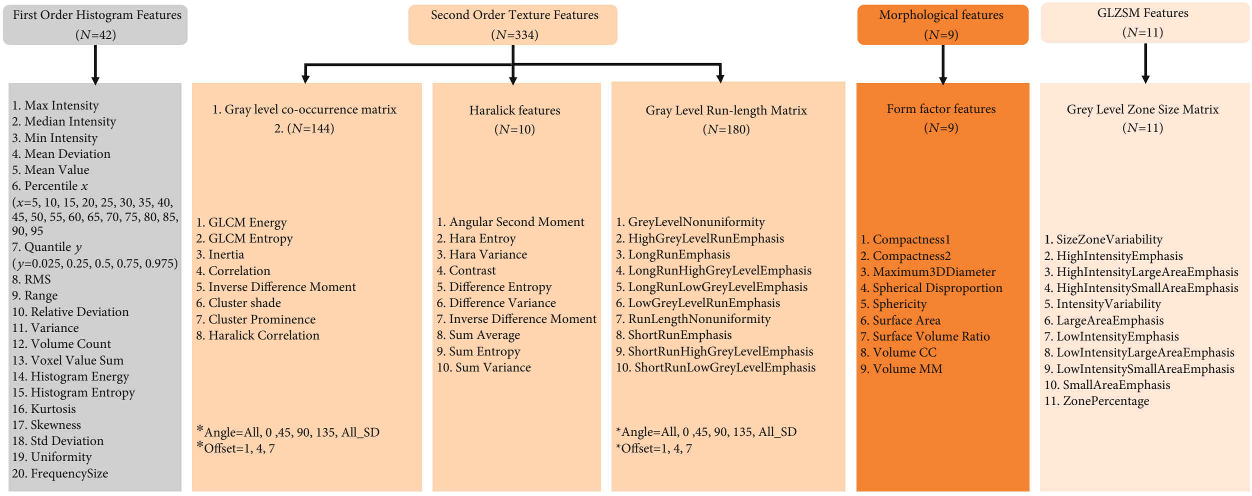


Figure 1: (a) Contrasted T1WI. (b) Contrasted T1WI image generated by ITK-SNAP software to depict the ROI of the tumor. (c) 3D ROI image of the tumor (red area) that is calculated to superimpose at all levels in the contrasted T1WI map. (d) A three-dimensional image of the tumor.



First Order Histogram Features
(N=42)

1. Max Intensity
2. Median Intensity
3. Min Intensity
4. Mean Deviation
5. Mean Value
6. Percentile x
(x=5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95
7. Quantile y
(y=0.025, 0.25, 0.5, 0.75, 0.975)
8. RMS
9. Range
10. Relative Deviation
11. Variance
12. Volume Count
13. Voxel Value Sum
14. Histogram Energy
15. Histogram Entropy
16. Kurtosis
17. Skewness
18. Std Deviation
19. Uniformity
20. FrequencySize

Second Order Texture Features
(N=334)

1. Gray level co-occurrence matrix
2. (N=144)

1. GLCM Energy
2. GLCM Entropy
3. Inertia
4. Correlation
5. Inverse Difference Moment
6. Cluster shade
7. Cluster Prominence
8. Haralick Correlation

*Angle=All, 0 ,45, 90, 135, All_SD
*Offset=1, 4, 7

Haralick features
(N=10)

1. Angular Second Moment
2. Hara Entroy
3. Hara Variance
4. Contrast
5. Difference Entropy
6. Difference Variance
7. Inverse Difference Moment
8. Sum Average
9. Sum Entropy
10. Sum Variance

Gray Level Run-length Matrix
(N=180)

1. GreyLevelNonuniformity
2. HighGreyLevelRunEmphasis
3. LongRunEmphasis
4. LongRunHighGreyLevelEmphasis
5. LongRunLowGreyLevelEmphasis
6. LowGreyLevelRunEmphasis
7. RunLengthNonuniformity
8. ShortRunEmphasis
9. ShortRunHighGreyLevelEmphasis
10. ShortRunLowGreyLevelEmphasis

*Angle=All, 0 ,45, 90, 135, All_SD
*Offset=1, 4, 7

Morphological features
(N=9)

Form factor features
(N=9)

1. Compactness1
2. Compactness2
3. Maximum3DDiameter
4. Spherical Disproportion
5. Sphericity
6. Surface Area
7. Surface Volume Ratio
8. Volume CC
9. Volume MM

GLZSM Features
(N=11)

Grey Level Zone Size Matrix
(N=11)

1. SizeZoneVariability
2. HighIntensityEmphasis
3. HighIntensityLargeAreaEmphasis
4. HighIntensitySmallAreaEmphasis
5. IntensityVariability
6. LargeAreaEmphasis
7. LowIntensityEmphasis
8. LowIntensityLargeAreaEmphasis
9. LowIntensitySmallAreaEmphasis
10. SmallAreaEmphasis
11. ZonePercentage

Figure 2: A total of 396 texture parameters extracted by AK software.

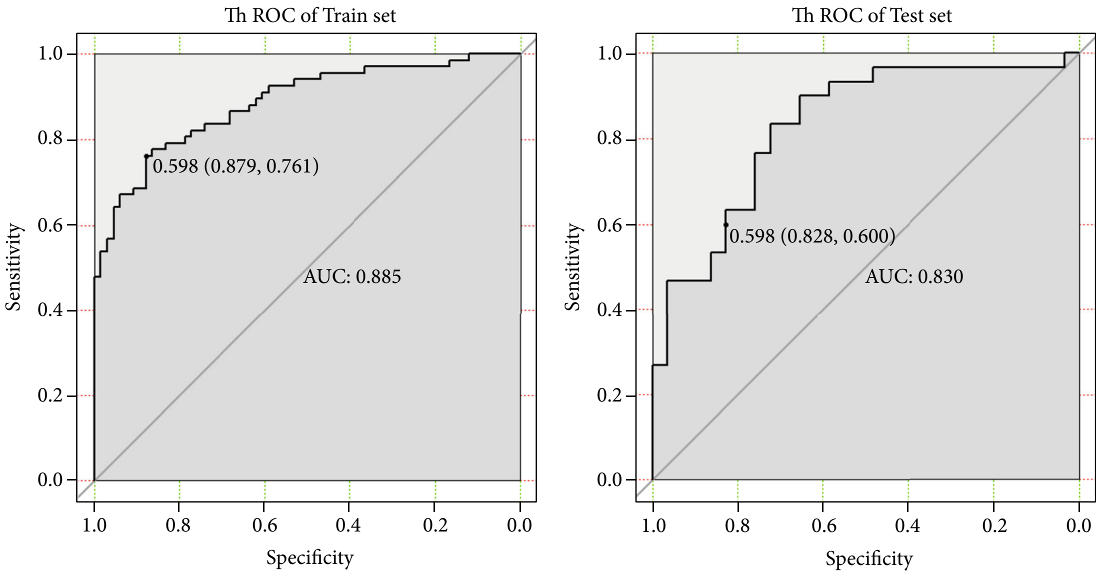


Figure 3: T1WI texture feature model for identification of HPC and AM performance.

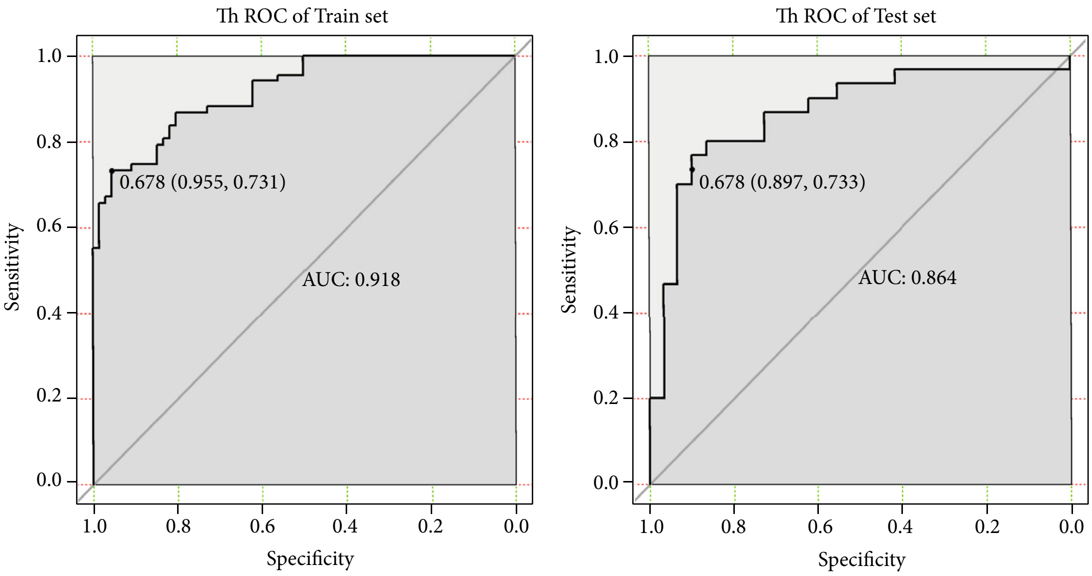


FIGURE 4: T2WI texture feature model for identification of HPC and AM performance.

(KU), skewness (SK), stdDeviation (ST), GLCMEntropy_angle90_offset1 ($GLCME_{90\text{-}1}$), and SmallAreaEmphasis (SAE). The T1WI texture feature modeling formula was the following:

$$f(\mathrm{T1WI}) = -13.4078 + 0.3066 \times \mathrm{KU} - 1.2143 \times \mathrm{SK} + 0.0351 \times \mathrm{ST} - 1.1441 \times \mathrm{GLCME}_{90\text{-}1} + 23.9030 \times \mathrm{SAE}. \tag{1}$$

In the train set, the ACC of the T1WI model in identifying SFT/HPC and AM was 0.820, and the area under the ROC curve (AUC) was 0.885, with the cutoff value of 0.598, a sensitivity of 76.1%, and a specificity of 87.9%. In the test set, ACC was 0.712 and AUC was 0.830, with a cutoff value of 0.598, a sensitivity of 60.0%, and a specificity of 82.8% (see Figure 3).

The high-performance parameters of the T2WI texture feature model were GLCMEnergy_AllDirection_offset1_SD ($GLCME_{A\text{-}1\text{-}SD}$), Inertia_angle90_offset7 ($IN_{90\text{-}7}$), InverseDifferenceMoment_AllDdirection_offset7_SD ($IDM_{A\text{-}7\text{-}SD}$), LongRunLowGreyLevelEmphasis_AllDirection_offset4_SD ($LRLGLE_{A\text{-}4\text{-}SD}$), LowGreyLevelRunEmphasis_AllDirection_offset7_SD ($LGLRE_{A\text{-}7\text{-}SD}$), ShortRunEmphasis_angle135_offset1 ($SRE_{135\text{-}1}$), ShortRunHighGreyLevelEmphasis_angle90_offset4 ($SRHGLE_{90\text{-}4}$), HighIntensitySmallAreaEmphasis (HISAE), and LowIntensityLargeAreaEmphasis (LILAE). The T2WI texture feature modeling formula was the following:

$$\begin{aligned} f(\mathrm{T2WI}) = {} & 9.78\mathrm{e}^{2} + 1.23\mathrm{e}^{8} \times \mathrm{GLCME}_{\text{A-1-SD}} - 2.81\mathrm{e}^{-3} \\ & \times \mathrm{IN}_{90\text{-}7} - 2.53\mathrm{e}^{4} \times \mathrm{IDM}_{\text{A-7-SD}} + 8.56\mathrm{e}^{-1} \\ & \times \mathrm{LRLGLE}_{\text{A-4-SD}} - 9.72\mathrm{e}^{10} \times \mathrm{LGLRE}_{\text{A-7-SD}} \\ & - 9.79\mathrm{e}^{2} \times \mathrm{SRE}_{135\text{-}1} - 1.38\mathrm{e}^{-4} \times \mathrm{SRHGLE}_{90\text{-}4} \\ & + 9.66\mathrm{e}^{-7} \times \mathrm{HISAE} + 8.04\mathrm{e}^{4} \times \mathrm{LILAE}. \end{aligned} \tag{2}$$

In the train set, the ACC of T2WI model in identifying SFT/HPC and AM was 0.842, and the area under the ROC curve (AUC) was 0.918, with a cutoff value of 0.678, a sensitivity of 73.1%, and a specificity of 95.5%. In the test set, ACC was 0.814 and AUC was 0.864, with a cutoff value of 0.678, a sensitivity of 73.3%, and a specificity of 89.7% (see Figure 4).

The high-performance parameters involved in the contrasted T1WI texture feature model were $\mathrm{Quantile}_{0.025}$, RelativeDeviation (RD), VoxelValueSum (VVS), ClusterProminence_AllDirection_offset1_SD ($CP_{A\text{-}1\text{-}SD}$), GLCMEntropy_AllDirection_offset7_SD ($GLCME_{A\text{-}1\text{-}SD}$), LongRunHighGreyLevelEmphasis_AllDirection_offset1_SD ($LRHGLE_{A\text{-}1\text{-}SD}$), and ShortRunHighGreyLevelEmphasis_AllDirection_offset4_SD ($SRHGLE_{A\text{-}4\text{-}SD}$). The contrasted T1WI texture feature model modeling formula was the following:

$$\begin{aligned} f(\mathrm{contrastedT1WI}) = {} & -6.05\mathrm{e}^{-1} + 2.84\mathrm{e}^{-3} \times \mathrm{Quantile}_{0.025} \\ & + 8.38\mathrm{e}^{-1} \times \mathrm{RD} + 3.12\mathrm{e}^{-8} \times \mathrm{VVS} \\ & - 1.68\mathrm{e}^{-14} \times \mathrm{CP}_{\text{A-1-SD}} + 2.89\mathrm{e}^{-1} \\ & \times \mathrm{GLCME}_{\text{A-1-SD}} - 1.05\mathrm{e}^{-4} \\ & \times \mathrm{LRHGLE}_{\text{A-1-SD}} - 1.07\mathrm{e}^{-5} \\ & \times \mathrm{SRHGLE}_{\text{A-4-SD}}. \end{aligned} \tag{3}$$

In the train set, the ACC of contrasted T1WI model in identifying SFT/HPC and AM was 0.744, and the area under the ROC curve (AUC) was 0.815, which had a cutoff value of 0.676, a sensitivity of 55.2%, and a specificity of 93.9%. In the test set, ACC was 0.695 and AUC was 0.772, which had a cutoff value of 0.676, a sensitivity of 60.0%, and a specificity of 79.3% (see Figure 5).

3.2. Establishment of Total Sequences Combine Texture Feature Model.

After the dimension reduction and stepwise

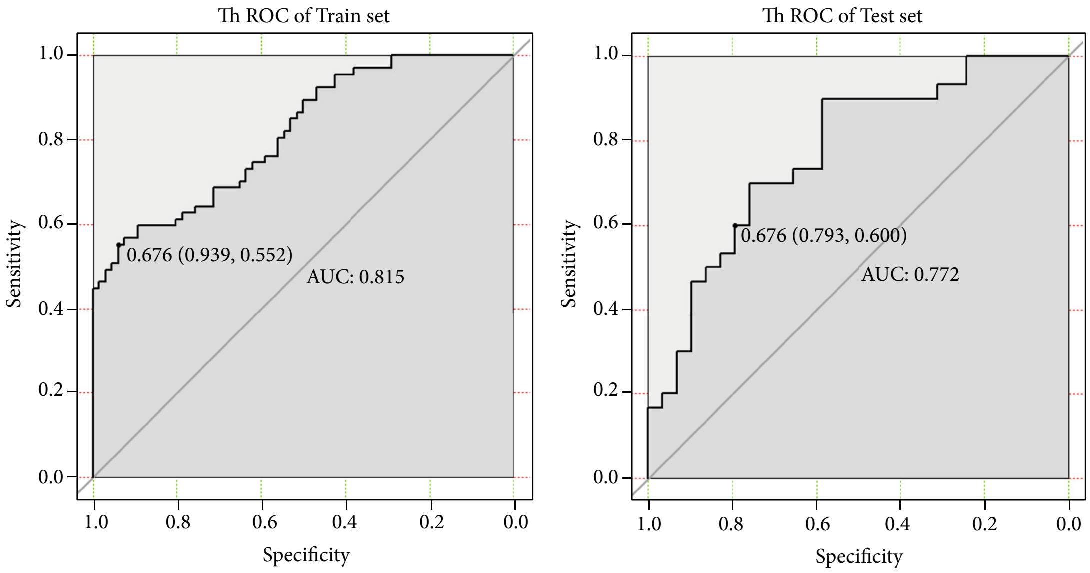


Figure 5: Contrasted T1WI texture feature model for identification of HPC and AM performance.

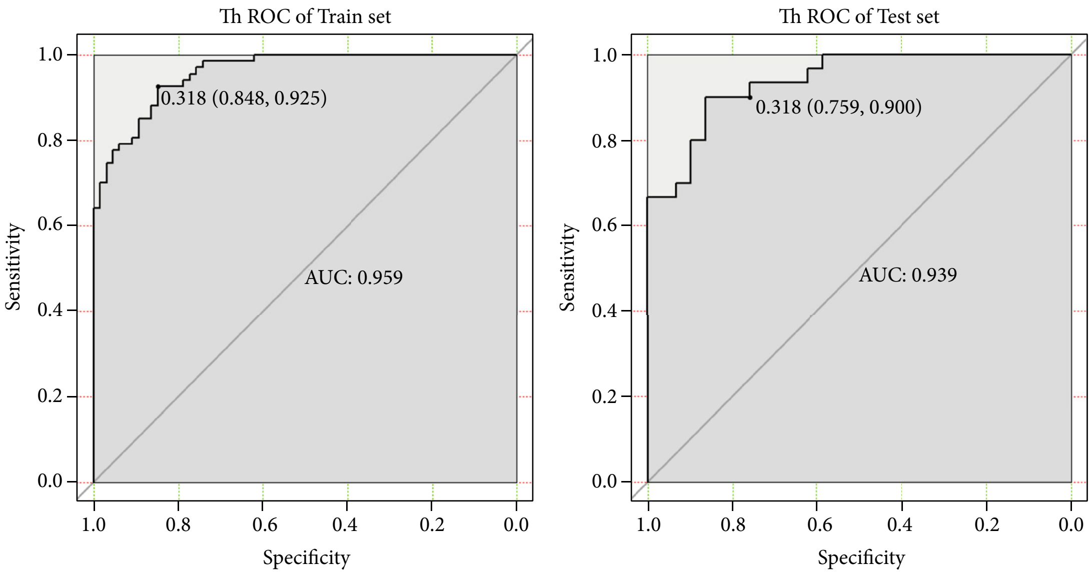


Figure 6: Total sequences combine the texture feature model for identification of HPC and AM performance.

Table 1: AUC, ACC, cut-off, sensitivity, and specificity of the four texture feature model.

Task	ACC	AUC	Cut-off	Sensitivity	Specificity
Train (T1WI)	0.820	0.885	0.598	76.1%	87.9%
Test (T1WI)	0.712	0.830	0.598	60.0%	82.8%
Train (T2WI)	0.842	0.918	0.678	73.1%	95.5%
Test (T2WI)	0.814	0.864	0.678	73.3%	89.7%
Train (contrasted-T1WI)	0.744	0.815	0.676	55.2%	93.9%
Test (contrasted-T1WI)	0.695	0.772	0.676	60.0%	79.3%
Train (combined)	0.887	0.959	0.318	92.5%	84.8%
Test (combined)	0.831	0.939	0.318	90.0%	75.9%

iterative method, the high-performance parameters of the total sequences combined texture feature model were Percentile$_{95\mathrm{T1WI}}$, KU$_{\mathrm{T1WI}}$, GLCMEntropy_angle90_offset1(GLCME$_{90\text{-}1}$)$_{\mathrm{T1WI}}$, HaraEntroy(HE)$_{\text{contrasted T1WI}}$, RunLengthNonuniformity_angle90_offset7(RLNU$_{90\text{-}1}$)$_{\text{contrasted T1WI}}$, Range(RA)$_{\mathrm{T2WI}}$, ClusterProminence_angle135_offset4(CP$_{135\text{-}4}$)$_{\mathrm{T2WI}}$, and LongRunHighGreyLevelEmphasis_angle45_offset4(LRHGLE$_{45\text{-}4}$)$_{\mathrm{T2WI}}$. The total sequences combined model modeling formula was the following:

$$\begin{aligned} f_{(\text{total sequences combine})} = {} & -8.91 + 1.93\mathrm{e}^{-2} \times \text{Percentile}_{95(\mathrm{T1WI})} \\ & + 3.09\mathrm{e}^{-1} \times \mathrm{KU}_{(\mathrm{T1WI})} - 1.22 \\ & \times \mathrm{GLCME}_{90-1(\mathrm{T1WI})} + 2.56\mathrm{e}^{1} \\ & \times \mathrm{HE}_{(\text{contrasted T1WI})} + 3.14\mathrm{e}^{-5} \\ & \times \mathrm{RLNU}_{90-1(\text{contrasted T1WI})} - 3.76\mathrm{e}^{-3} \\ & \times \mathrm{RA}_{(\mathrm{T2WI})} + 1.01\mathrm{e}^{-8} \times \mathrm{CP}_{135\text{-}4(\mathrm{T2WI})} \\ & - 1.41\mathrm{e}^{-4} \times \mathrm{LRHGLE}_{45\text{-}4(\mathrm{T2WI})}. \end{aligned} \tag{4}$$

According to ROC analysis, the combined model used to identify AM and SFT/HPC in the train set had an AUC of 0.959 (cutoff value = 0.318, specificity of 84.8%, and sensitivity of 92.5%), and the accuracy of the combined model was 0.887. In the test set, AUC was 0.939, with a cutoff value of 0.318, a sensitivity of 90.0%, and a specificity of 75.9%, and the accuracy of the combined model was 0.831 (see Figure 6).

Finally, the AUC, ACC, cut-off, sensitivity, and specificity of the four models are summarized in Table 1.

4. Discussion

Image segmentation is a critical session for the MRI images to be used in brain tumor studies. In recent years, semiautomatic and fully automatic algorithms for brain tumor segmentation have been developed rapidly. A study presented a fully automatic brain tumor detection and segmentation method using the U-Net based deep convolution network and demonstrated that this method can provide both efficient and robust segmentation compared to manual delineated ground truth [12]. Soltaninejad et al. [13] proposed a supervised learning based method for segmentation tumour in multimodal MRI brain images. Supervoxels were calculated using information fusion from multimodal MRI images, which also demonstrated promising results in the segmentation of brain tumor. Even so, there are still several opening challenges for this task mainly due to the high variation of brain tumors in size, shape, regularity, location, and their heterogeneous appearance. In addition, AM and HPC/SFP are rare diseases, and the data is relatively rare compared to common diseases. We have certain reasons to believe that segmentation based on big data may have certain errors. Considering the above reasons, the segmentation was still relied on manual delineation by human operators in this study.

In this study, radiomics method was used to construct four models to identify the 3D-texture features of SFT/HPC and AM based on conventional MRI sequence images, including the T1WI model, T2WI model, contrasted T1WI model, and a combined sequence model. Briefly, the combined sequence model showed the best performance, followed by the T2WI model. As a noninvasive predictive method, all four models can provide reference information for preoperative treatment planning and patient prognosis. Due to the relatively large number of cases, we have established a relatively accurate MRI radiological model for preoperative identification of SFT/HPC and AM. To the best of our knowledge, this is the first study that established an MRI radiological model, which can be used to differentiate SFT/HPC from AM.

Texture features are essential markers for intratumoral homogeneity. Among the twenty-three texture features that were involved in building our models, eight were histogram-based features (KU, SK, ST, RD,VVS, RA, Quantile$_{0.025}$, and Percentile$_{95}$), and twelve were matrix-based features, including five GLCM features (GLCMEnergy, GLCMEntropy, IN, IDM, and CP) and one Haralick feature (HE). Besides, there were six GLRM features (LRHGLE, LRLGLE, LGLRE, RLNU, SRE, and SRHGLE), and three GLZSM features (HISAE, LILAE, SAE). Histogram-based features are first-order statistics that primarily rely on intensity information (or brightness information) within and around the tumor. These features are used to investigate the overall distribution of intensity information within and around the tumor. For example, "kurtosis" is a measure of the "tailedness" of the median distribution of image ROI, which can be used to describe the concentration of image brightness information. Higher kurtosis means that the mass of the distribution is concentrated at the tail. "Skewness" represents the measure of "skewness" of the median distribution of the image ROI and is used to describe the degree of asymmetric distribution in the histogram. The percentile (%) of a distribution is defined as the brightness value. IDM represents the uniformity of pixel signal strength in the image, which can reflect the heterogeneity of tumor tissues.

Matrix-based features are second-order statistics that can be used to analyze the complexity within the tumor and around the tumor, changes in the hierarchy, and thickness of the texture. For example, inertia reflects the clarity of the image and texture groove depth. The contrast is proportional to the texture groove; high groove values produce more clarity, while small values lead to small contrast and fuzzy image. GLCMEntropy measures the average amount of information required to encode an image value. SRHGLE measures the joint distribution of shorter run lengths with higher grey-level values. Larger value leads to a more complex image and smaller image grey value. LRLGLE measures the joint distribution of longer-run lengths with lower grey-level values. SRE is a measure of short lengths, with larger values representing shorter lengths and finer textures. GLZSM is particularly efficient to characterize the texture homogeneity, nonperiodicity, or speckle like texture.

So far, many studies have reported the use of radiological models based on the texture features of CT and MRI images for the identification/differentiation of tumors. Chen et al. [14] found that the radiomics model based on contrast-enhanced computed tomography (CECT) could be used for predicting acute pancreatitis (AP) recurrence. As a quantitative method, radiomics exhibits promising performance in alerting relapsed patients to potential preventive measures. Kang et al. [15] tested the technical feasibility, generalizability, and diagnostic performance of a radiomics model using ADC maps for identification of atypical primary central nervous system lymphoma (PCNSL) mimicking glioblastoma. His model showed good generalizability and improved diagnostic performance than single-parameter measurements in identifying atypical PCNSL mimicking glioblastoma by providing robust high-dimensional analyses of conventional and physiological imaging features. Furthermore, Chen and colleagues [16] confirmed that an MRI-based combined radiography nomogram can effectively predict the immune score of HCC and help to make treatment decisions.

Our study showed that the combined sequence model was superior to any single sequence model in differentiating SFT/HPC from AM. T2WI sequence is the most commonly used sequence to evaluate brain pathology and the degree of tumor invasion. T1WI and contrasted T1WI sequences provide anatomical information, while tissue enhancement reflects increased blood-brain barrier permeability [17]. Considering that each sequence has different functions, combining multiple sequences may improve the accuracy in differentiating SFT/HPC from AM. Also, Tian et al. [18] verified the superiority of radiomics features extracted by multiparameter MRI in glioma grading and found that the combined application of multiparameter MRI has higher classification efficiency, which was consistent with our data. It is worth noting that in the three sequences of conventional MRI, the AUC of the radiological model based on T2WI image texture features was higher than in the other two sequences. One explanation for this may be that the T2WI sequence has a relatively long echo time and high contrast between tissues, so the image contains many differential texture features with discriminative value. Among the related studies on breast, one study suggested that T2WI images have a significant role in the differentiation of benign and malignant diseases of nonmass breast tumors [19]. Li et al. [20] confirmed that texture features of SPAIR T2W-MRI can be classified into three different types of single-liver lesions and may serve as an adjunct tool for accurate diagnosis of these diseases. Surprisingly, we also found that the contrasted T1WI model had the lowest AUC among the three conventional MRI sequence models. Furthermore, Zhang et al. [21] found that T1w+Gd had the lowest AUC in all MRI sequences when evaluating the feasibility of texture analysis on preoperative conventional MRI images in predicting early malignant transformation from low- to high-grade glioma, which was consistent with our results. Nevertheless, T1WI+C were very useful for visual evaluation of tumors.

5. Conclusions

The radiological model based on texture features could be used to differentiate SFT/HPC from AM. Besides, our texture analysis results, which extract many quantitative features from various kinds of digital images, provide the basis for further radiomics analyses and are a rapidly expanding research area [22, 23].

5.1. Limitations of the Study. Limitations of this study must be addressed. (1) This study was a retrospective study, which means that further prospective studies of a larger range of patients and multivariate analysis are necessary to verify these results. (2) In this study, only the parenchymal part of the tumor was selected for texture analysis. The peritumoral edema area of the two tumors was not analyzed, and the MRI signs were further combined with the texture parameters to improve the discrimination efficiency. (3) The ROIs were manually determined. Automatic segmentation algorithms may facilitate the procedure. (4) The correlation between the significance of various parameters of texture analysis and the biological mechanism of tumors was still insufficient; thus, further research is required to confirm our findings.

Data Availability

The data of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declared that they have no conflicts of interest to this work.

Authors' Contributions

Junyi Dong and Meimei Yu contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 81671646).

References

[1] D. Choi, D. Na, H. Byun et al., "Salivary gland tumors: evaluation with two-phase helical CT," *Radiology*, vol. 214, no. 1, pp. 231–236, 2000.

[2] Y. Tsushima, M. Matsumoto, and K. Endo, "Parotid and parapharyngeal tumours: tissue characterization with dynamic magnetic resonance imaging," *The British Journal of Radiology*, vol. 67, no. 796, pp. 342–345, 1994.

[3] M. Hasselblatt, K. Nolte, and W. Paulus, "Angiomatous meningioma: a clinicopathologic study of 38 cases," *The American Journal of Surgical Pathology*, vol. 28, no. 3, pp. 390–393, 2004.

[4] P. Lambin, E. Rios-Velazquez, R. Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[5] G. Yang, T. L. Jones, T. R. Barrick, and F. A. Howe, "Discrimination between glioblastoma multiforme and solitary metastasis using morphological features derived from thep:qtensor decomposition of diffusion tensor imaging," *NMR in Biomedicine*, vol. 27, no. 9, pp. 1103–1111, 2014.

[6] G. Yang, T. L. Jones, F. A. Howe, and T. R. Barrick, "Morphometric model for discrimination between glioblastoma multiforme and solitary metastasis using three-dimensional shape analysis," *Magnetic Resonance in Medicine*, vol. 75, no. 6, pp. 2505–2516, 2016.

[7] X. Li, Y. Miao, L. Han et al., "Meningioma grading using conventional MRI histogram analysis based on 3D tumor measurement," *European Journal of Radiology*, vol. 110, pp. 45–53, 2019.

[8] R. Rahman, A. Hamdan, R. Zweifler et al., "Histogram analysis of apparent diffusion coefficient within enhancing and nonenhancing tumor volumes in recurrent glioblastoma patients treated with bevacizumab," *Journal of Neuro-Oncology*, vol. 119, no. 1, pp. 149–158, 2014.

[9] Y. Choi, S. S. Ahn, D. W. Kim et al., "Incremental prognostic value of ADC histogram analysis over MGMT promoter methylation status in patients with glioblastoma," *Radiology*, vol. 281, no. 1, pp. 175–184, 2016.

[10] A. Zolal, T. Juratli, J. Linn et al., "Enhancing tumor apparent diffusion coefficient histogram skewness stratifies the postoperative survival in recurrent glioblastoma multiforme patients undergoing salvage surgery," *Journal of Neuro-Oncology*, vol. 127, no. 3, pp. 551–557, 2016.

[11] T. Kanazawa, Y. Minami, M. Jinzaki, M. Toda, K. Yoshida, and H. Sasaki, "Preoperative prediction of solitary fibrous tumor/-hemangiopericytoma and angiomatous meningioma using magnetic resonance imaging texture analysis," *World Neurosurgery*, vol. 120, pp. e1208–e1216, 2018.

[12] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-net based fully convolutional networks," in *Medical Image Understanding and Analysis. MIUA 2017*, Communications in Computer and Information Science, M. Valdés Hernández and V. González-Castro, Eds., pp. 506–517, Springer, Cham, 2017.

[13] M. Soltaninejad, G. Yang, T. Lambrou et al., "Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 69–84, 2018.

[14] Y. Chen, T.-w. Chen, C.-q. Wu et al., "Radiomics model of contrast-enhanced computed tomography for predicting the recurrence of acute pancreatitis," *European Radiology*, vol. 29, no. 8, pp. 4408–4417, 2019.

[15] D. Kang, J. E. Park, Y.-H. Kim et al., "Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation," *Neuro-Oncology*, vol. 20, no. 9, pp. 1251–1261, 2018.

[16] S. Chen, S. Feng, J. Wei et al., "Pretreatment prediction of immunoscore in hepatocellular cancer: a radiomics-based clinical model based on Gd-EOB-DTPA-enhanced MRI imaging," *European Radiology*, vol. 29, no. 8, pp. 4177–4187, 2019.

[17] K. I. Ly and E. R. Gerstner, "The role of advanced brain tumor imaging in the care of patients with central nervous system malignancies," *Current Treatment Options in Oncology*, vol. 19, no. 8, p. 40, 2018.

[18] Q. Tian, L.-F. Yan, X. Zhang et al., "Radiomics strategy for glioma grading using texture features from multiparametric MRI," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 6, pp. 1518–1528, 2018.

[19] P. A. T. Baltzer, M. Dietzel, and W. A. Kaiser, "Nonmass lesions in magnetic resonance imaging of the breast: additional T2-weighted images improve diagnostic accuracy," *Journal of Computer Assisted Tomography*, vol. 35, no. 3, pp. 361–366, 2011.

[20] Z. Li, Y. Mao, W. Huang et al., "Texture-based classification of different single liver lesion based on SPAIR T2W MRI images," *BMC Medical Imaging*, vol. 17, no. 1, p. 42, 2017.

[21] S. Zhang, G. C.-Y. Chiang, R. S. Magge et al., "Texture analysis on conventional MRI images accurately predicts early malignant transformation of low-grade gliomas," *European Radiology*, vol. 29, no. 6, pp. 2751–2759, 2019.

[22] I. Hassan, A. Kotrotsou, A. S. Bakhtiari et al., "Radiomic texture analysis mapping predicts areas of true functional MRI activity," *Scientific reports*, vol. 6, no. 1, article 25295, 2016.

[23] H. J. W. L. Aerts, "The potential of radiomic-based phenotyping in precision medicine: a review," *JAMA Oncology*, vol. 2, no. 12, pp. 1636–1642, 2016.

Hindawi
BioMed Research International
Volume 2020, Article ID 8864756, 8 pages
https://doi.org/10.1155/2020/8864756

Research Article

Separability of Acute Cerebral Infarction Lesions in CT Based Radiomics: Toward Artificial Intelligence-Assisted Diagnosis

Yun Guan,[1,2] Peng Wang,[1,3] Qi Wang,[1,2] Peihao Li,[4] Jianchao Zeng,[1,2] Pinle Qin,[1,2] and Yanfeng Meng[1,3]

[1]*North University of China-Taiyuan Central Hospital Joint Innovation Institute, 3 Xueyuan Road, Taiyuan, Shanxi 030051, China*
[2]*College of Big Data, North University of China, 3 Xueyuan Road, Taiyuan, Shanxi 030051, China*
[3]*Taiyuan Central Hospital of Shanxi Medical University, 5 Dong San Dao Lane, Jiefang Street, Taiyuan, Shanxi 030009, China*
[4]*School of Information and Communication Engineering, North University of China, 3 Xueyuan Road, Taiyuan, Shanxi 030051, China*

Correspondence should be addressed to Yanfeng Meng; yanfeng.m@163.com

Received 16 September 2020; Revised 2 November 2020; Accepted 3 November 2020; Published 16 November 2020

Academic Editor: Zhiguo Zhou

This study aims at analyzing the separability of acute cerebral infarction lesions which were invisible in CT. 38 patients, who were diagnosed with acute cerebral infarction and performed both CT and MRI, and 18 patients, who had no positive finding in either CT or MRI, were enrolled. Comparative studies were performed on lesion and symmetrical regions, normal brain and symmetrical regions, lesion, and normal brain regions. MRI was reconstructed and affine transformed to obtain accurate lesion position of CT. Radiomic features and information gain were introduced to capture efficient features. Finally, 10 classifiers were established with selected features to evaluate the effectiveness of analysis. 1301 radiomic features were extracted from candidate regions after registration. For lesion and their symmetrical regions, there were 280 features with information gain greater than 0.1 and 2 features with information gain greater than 0.3. The average classification accuracy was 0.6467, and the best classification accuracy was 0.7748. For normal brain and their symmetrical regions, there were 176 features with information gain greater than 0.1, 1 feature with information gain greater than 0.2. The average classification accuracy was 0.5414, and the best classification accuracy was 0.6782. For normal brain and lesions, there were 501 features with information gain greater than 0.1 and 1 feature with information gain greater than 0.5. The average classification accuracy was 0.7480, and the best classification accuracy was 0.8694. In conclusion, the study captured significant features correlated with acute cerebral infarction and confirmed the separability of acute lesions in CT, which established foundation for further artificial intelligence-assisted CT diagnosis.

1. Introduction

Globally, stroke is still the leading cause of mortality and disability, and there are substantial economic costs for post-stroke care [1–4]. In practice, CT is the preferred radiologic modality for patients with stroke-like clinical manifestation, since it is immediately available, cost effective, and capable of differentiating brain disorders [5]. CT is very sensitive in detecting intracranial hemorrhagic stroke and chronic ischemic stroke. CT detects acute cerebral infarction (ACI) in terms of decrease of CT attenuation, loss of gray-white matter differentiation, sulcal effacement, and other indirect signs [6]. In practice, radiologists often encounter poor accuracy in diagnosing acute infarct by CT, with accuracy rate ≤67% within 3 hours [5].

Patient, who has stroke-like symptom but CT showed negative findings, needs MRI [7]. MR diffusion-weighted imaging (DWI) can detect ischemic lesions within minutes

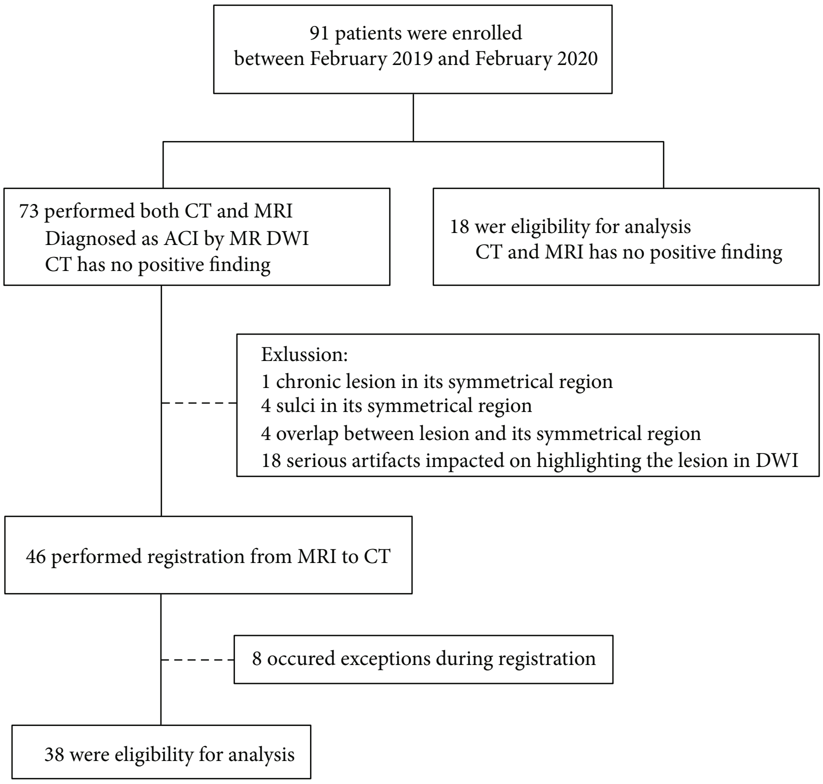

91 patients were enrolled
between February 2019 and February 2020
73 performed both CT and MRI
Diagnosed as ACI by MR DWI
CT has no positive finding
18 wer eligibility for analysis
CT and MRI has no positive finding
Exlussion:
1 chronic lesion in its symmetrical region
4 sulci in its symmetrical region
4 overlap between lesion and its symmetrical region
18 serious artifacts impacted on highlighting the lesion in DWI
46 performed registration from MRI to CT
8 occured exceptions during registration
38 were eligibility for analysis

Figure 1: Flowchart of the recruitment produce for this study. ACI denotes acute cerebral infarction.

of symptoms, which is an extremely sensitive technology and used for estimating whether thrombolysis is appropriate. However, MRI is still not the primary modality, because it is time-consuming, which may lead to miss time window of thrombolysis, costing expensive, and various contradictions [8, 9]. Studies demonstrated that better clinical outcomes correlated with earlier diagnosis of ischemic stroke [10]. Therefore, it is essential to improve the accuracy rate of early recognizing ACI by CT within the time window.

In this study, we hypothesize that ACI lesions in CT are separable by combining image registration, accurate lesion location, radiomic feature extraction, and information gain calculation, so as to establish a foundation for artificial intelligence-assisted CT diagnosis.

2. Materials and Methods

2.1. Patients. This retrospective study protocol was reviewed and approved by the institutional review board of our hospital. Written informed consent was waived.

Between February 2019 and February 2020, we retrospectively studied 38 patients, who have performed both CT and MRI and diagnosed as ACI by DWI; meanwhile, CT has no positive finding by radiologist. Another 18 patients with no positive finding in either CT or MRI were also enrolled as normal control (Figure 1).

2.2. CT and MRI. CT images were acquired on a 320-MDCT scanner (Aquilion ONE, Toshiba Medical System, Otawara, Japan) with the following parameters: 120 kV and 300 mA, 5-mm slice thickness, 512 × 512 matrix, and 0.6 mm collimation.

MRI was acquired on 1.5 T MR scanner (Sonata, Siemens Healthcare, Erlangen, Germany) and 3.0 T MR scanner (MAGNETOM Skyra, Siemens Healthcare, Erlangen, Germany). The parameters of 1.5 T DWI were as follows: TR/TE == 3800 ms/84 ms, slice thickness = 5 mm, matrix = 128 × 128, FOV = 200 × 220, b value = 1000. The parameters of 3.0 T DWI were as follows: TR/TE == 4950 ms/64 ms, slice thickness = 5 mm, matrix = 164 × 164, FOV = 220 × 220, b value = 1000.

2.3. Registration and Candidate Region Acquisition. The pipeline of our methodology was shown in Figure 2. Since the position and angle of CT were different from MRI for one patient, the DWI had to be registered to CT images (Figure 3). Herein, the CT images were not adjusted to avoid loss of intact information. The DWI were multiple planners reconstructed to get a consistent angle with CT and achieve coarse registration. Then, a series of affine transformations were performed to get a consistent position and achieve fine registration, including translation, rotation, and scaling transformation.

Early cerebral infarction was obvious on DWI which was sensitive to the restricted Brownian movement of water molecules in brain tissue [5]. Immediately after registration, we highlighted the lesion regions in DWI through adjusting the window width and window level. Because of CT and DWI were matched, the salient lesion position of DWI was also used as the lesion label for CT.

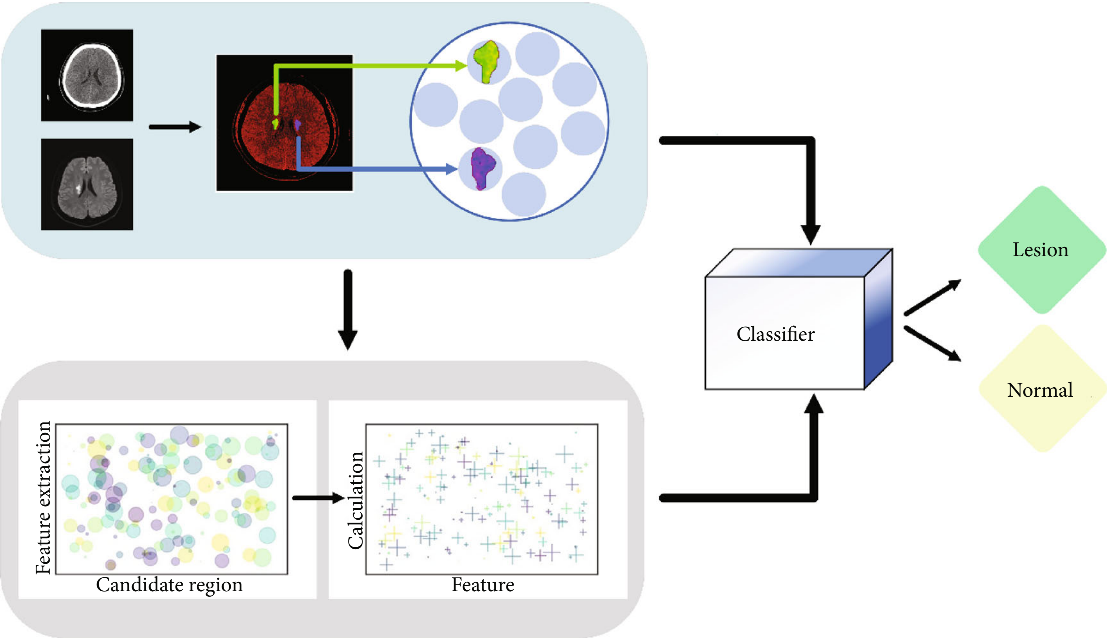


Figure 2: The pipeline of our methodology included three steps: registration and candidate region acquisition, feature extraction and analysis, and classifiers establishment. Firstly, CT and MRI were input to obtain lesion regions and their symmetrical regions as candidate regions through registration. Then, features were extracted and calculated from candidate regions to capture useful features for auxiliary separating acute cerebral infarction. Finally, the classifiers were introduced to separate candidate region with selected features.

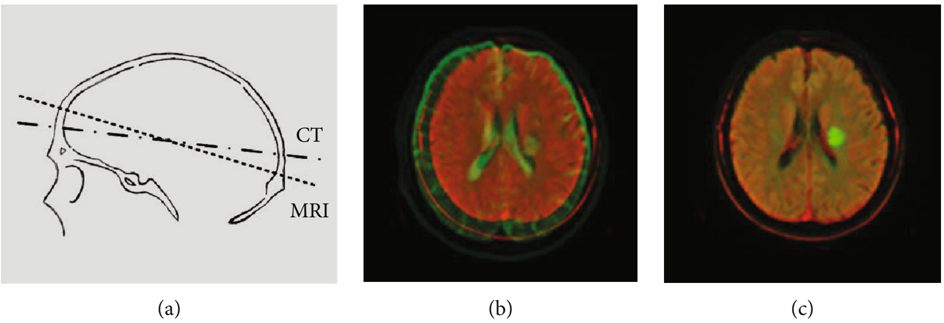


Figure 3: Image registration. (a) DWI was adjusted by multiple planner reconstruction to obtain a consistent angle with CT. Dotted line denoted MRI and point solid line denoted CT. (b) CT and DWI were put together to achieve coarse registration. (c) Fine registration was performed by a series of affine transformation including translation, rotation, and scaling.

Besides delineating exact lesion regions in CT projected from DWI, we took the midline of the brain as the axis of symmetry and depicted the profile of symmetrical regions (Figure 4). Instead of simply comparing the left and right sides of the brain [11–13], the lesion regions and their symmetrical regions were served as candidate regions to reduce redundant information and achieve accurate comparative analysis.

2.4. Feature Extraction and Analysis. Unlike Lo et al. [13] who improved a texture feature of radiomics, our scheme was to extract features from the image firstly, and then used machine learning techniques to learn these features, so that the computer can mine the information of cerebral infarction in CT according to the acquired characteristics and then identify. Radiomic feature extraction and statistical analysis were performed to complete the plentiful features extraction

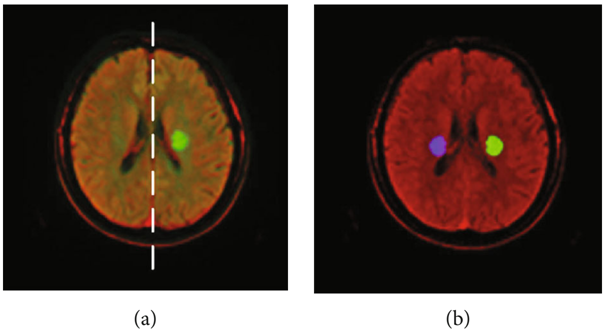


Figure 4: Candidate region acquisition. (a) The midline of the brain was the axis of symmetry for projecting symmetric position. (b) Depict the profile of symmetrical regions for achieving comparative analysis. The lesion regions and their symmetrical regions were served as candidate regions.

Table 1: Demographic characteristic and multivariate logistic regression results.

Characteristic	Total ($n = 56$)	Patients with ACI ($n = 38$)	Patients with no ACI ($n = 18$)	OR$^{\alpha\#}$ (OR 95% CI)
Age*		64.71 ± 12.92	34.17 ± 6.52	1.458 (1.086~1.957)
Sex(y)†				
Woman	24	17 (44.74)	7 (38.89)	1.000
Man	32	21 (55.26)	11 (61.11)	2.748 (0.108~69.973)
Predict value				
Negative	10	4 (10.53)	6 (33.33)	1.000
Positive	46	34 (89.47)	12 (66.67)	43.530 (0.640~2960.497)
MRI		Diagnosed as ACI	No positive finding	
CT		No positive finding	No positive finding	

[#]The value of OR was obtained from binary logistic regression by adjusting $\alpha_{\text{in}} = 0.1$, and $\alpha_{\text{out}} = 0.15$. Dependent is the true value, and covariates are sex, age, and the predicted value. All the covariates were calculated by the enter method. *Data are mean ± standard deviation. †Data in parentheses are percentages. ACI denotes acute cerebral infarction.

and choose features which are contributing to classification, respectively.

High-throughput feature extraction was applied to search abundant information in CT images. In this study, we followed the radiomic method by Lambin et al. [14] for the extraction process, which were divided into two steps: (1) image transformation and (2) feature calculation[15]. For image transformation, feature map was constructed nonlinearly from the original images, including wavelet, square root, and logarithm. For feature calculation, the feature was calculated on the original and transformed images, including first-order statistics and gray level cooccurrence matrix.

The information gain was further introduced as a statistical standard to measure the correlation between radiomic features and ACI, which was devoted to select significative information from a great deal features of above [16–18]. Each feature was calculated out a value in terms of information gain for dichotomy, lesion, or normal regions, by the equation below:

$$H(\mathrm{X}_i) = \sum_{x \in \mathrm{X}_i} - p(x) \log p(x), \tag{1}$$

$$H(Y \mid \mathrm{X}_i) = \sum_{x \in \mathrm{X}_i} p(x) H(Y \mid \mathrm{X}_i = x), \tag{2}$$

$$IG(\mathrm{X}_i, Y) = H(Y) - H(Y \mid \mathrm{X}_i), \tag{3}$$

where X_i represents the random variable of ith feature value, $x \in \mathrm{X}_i$ denotes the possible value of random variable X_i, $p(x)$ represents the probability when the random variable X_i takes the value x, and Y denotes the random variable of whether or not a cerebral infarction. $IG(X_i, Y)$ represents the information gain which is used to measure the reduction of uncertainty of event Y after X is known.

Theoretically, the feature was effective when the value was greater than zero, but we chose 0.1 as the minimal threshold to prevent calculating error and sampling error [18]. That is, features below the threshold were considered to be insignificant. The higher the information gain value of features, the greater contribution to remove noise and retain significant feature information.

2.5. Classifier Establishment. The classifier was established to demonstrate the separability of ACI. Given a candidate region, classifier automatically distinguished lesion or normal region in terms of the selected features. The classifiers probably make mistakes, so a classification accuracy score was calculated when all candidate regions were performed, which represents the separability of ACI.

In the experiment, we obtained different classification scores with different features, respectively, which was to confirm the effectiveness of features under different thresholds. We chose 10 common classifiers to obtain a reliable result, calculated the average classification accuracy, and selected the best classification accuracy as the final result. Each classifier experiment was repeated 100 times for average, and 4-fold cross validation was operated to get stable result.

The separation analysis was operated on ACI regions and their symmetrical regions in CT images. In addition, to exclude the separability of the left and right sides of the normal brain, and to explain the separability of the lesion and normal brain at same region, we performed the same experiments on the normal brain and their symmetrical normal regions, as well as normal brain and lesion regions.

3. Results

Demographic characteristics of all the patients in this study were shown in Table 1. For each of the 38 patients, one slice from CT was selected, which had a prominent lesion in MRI correspondingly. A total of 38 slices, which are 38 ACI regions and 38 symmetric noninfarct regions, were obtained. We extracted 1301 radiomic features from the candidate regions; meanwhile, the information gain was calculated to extract key information from abundant features. As shown in Table 2, there were 280 features with information gain greater than 0.1, which were considered to be contributory to classify candidate regions. There were 23 features with information gain greater than 0.2, and 2 features with information gain greater than 0.3, which showed potential

Table 2: Feature number under different thresholds of information gain on candidate region.

Candidate region	Threshold					
	0.0	0.1	0.2	0.3	0.4	0.5
Feature number of lesions and their symmetrical regions	1292	280	23	2	0	0
Feature number of normal and their symmetrical regions	1279	176	1	0	0	0
Feature number of lesions and normal regions	1295	501	126	51	18	1

Table 3: The classification accuracy result with selected features under different thresholds of information gain on candidate region.

Classifier	Lesions and their symmetrical regions threshold				Normal and their symmetrical regions threshold			Lesions and normal regions threshold					
	0.0	0.1	0.2	0.3	0.0	0.1	0.2	0.0	0.1	0.2	0.3	0.4	0.5
Multilayer perceptron	0.4902	0.4974	0.5694	0.7269	0.5061	0.4983	0.4434	0.5476	0.5242	0.5291	0.5292	0.5125	0.7185
Decision tree	0.5980	0.6138	0.6603	0.6655	0.5941	0.6333	0.5841	0.7155	0.7423	0.7782	0.7742	0.7982	0.8230
Random forest	0.5897	0.6452	0.7036	0.7206	0.6055	0.6782	0.5775	0.7700	0.7932	0.8401	0.8260	0.8162	0.8360
Adaboost	0.5850	0.6263	0.6811	0.6818	0.5946	0.6651	0.6001	0.7071	0.7276	0.7671	0.7862	0.7748	0.8291
Gradient boosting	0.5977	0.6346	0.6838	0.7463	0.5931	0.6505	0.5950	0.7517	0.7564	0.7903	0.7978	0.8017	0.8694
Bagging	0.6217	0.6567	0.6973	0.7249	0.6065	0.6529	0.5745	0.7530	0.7767	0.8169	0.8282	0.8144	0.8307
Bernoulli naive Bayes	0.5100	0.6164	0.6724	0.7105	0.4413	0.5175	0.4318	0.6557	0.6748	0.7253	0.7594	0.7566	0.6785
Gaussian naive Bayes	0.4743	0.6203	0.6661	0.6984	0.4801	0.4574	0.4439	0.3737	0.3935	0.8098	0.7842	0.8323	0.8605
Support vector machine	0.4184	0.4223	0.6903	0.4211	0.4382	0.4299	0.4326	0.6785	0.6785	0.6650	0.8123	0.7942	0.6785
K-nearest neighbor	0.2686	0.4563	0.7188	0.7748	0.2690	0.3492	0.6137	0.5812	0.5585	0.6437	0.8153	0.8010	0.8673
Average		0.5789	0.6743	0.6870		0.5532	0.5296		0.6625	0.7365	0.7712	0.7701	0.7991
			0.6467			0.5414					0.7480		

capability for separating lesion regions and their symmetric noninfarct regions. The related features were used to build up 10 classifiers to verify the feature effectiveness in candidate regions. The average classification accuracy was 0.6467, and the best classification accuracy was 0.7748 (Table 3).

For each of the 18 patients with no positive finding in either CT or MRI, three slices from CT were selected to augment data, which depicted by projecting the lesion labels obtained from 38 aforementioned MRI. A total of 54 slices, which are 54 normal brain tissue regions and 54 symmetrical regions, were obtained. As shown in Table 2, there were only 176 features with information gain greater than 0.1, 1 feature with information gain greater than 0.2, and no feature with information gain greater than 0.3. Although the best classification accuracy was 0.6782, from the overall classification results, the classification results were generally low and the average classification accuracy was only 0.5414 (Table 3).

For each of the 56 aforementioned patients, one slice from CT were selected. A total of 56 slices, which are 38 lesion regions and 18 normal brain tissue regions, were obtained. As shown in Table 2, there were 501 features with information gain greater than 0.1, 126 features with information gain greater than 0.2, 51 features with information gain greater than 0.3, 18 features with information gain greater than 0.4, and 1 feature with information gain greater than 0.5. The average classification accuracy was 0.7480, and the best classification accuracy was as high as 0.8694 (Table 3).

Besides, feature map that features reflected on candidate regions were shown to illustrate the effectiveness of feature analysis (Figure 5). We visualized one of the first three features ordered by information gain on the candidate region. Among them, it is a clear distinction on lesion and its symmetrical region, which explains the separability of ACI. The left and right sides of the normal region showed no obvious difference, which confirmed the inseparability of the left and right sides of the normal brain. The difference between lesion and normal region was also prominent, which indicated separability of lesion and normal region.

4. Discussion

Sensitively recognizing acute cerebral infarction is a valuable research for clinical treatment, within effective thrombolytic time. To the best of our knowledge, the finding of analyzing the separability of acute cerebral infarction lesions in CT based on image registration, precision positioning, radiomic feature extraction, and information gain calculation has not previously been well established in the literature. The overall concept of the algorithm was to extract and analyze the feature of regions where there is cerebral infarction, and more importantly, to separate lesions from normal regions.

Accurately recognizing acute ischemic stroke by CT remains challenging, due to the low accuracy of radiologist diagnosis. A lot of studies focused on prior knowledge, including decrease of CT attenuation, loss of gray-white matter differentiation, and sulcal effacement. However, 1/3 cases were missed since the sensitivity and specificity were low [5].

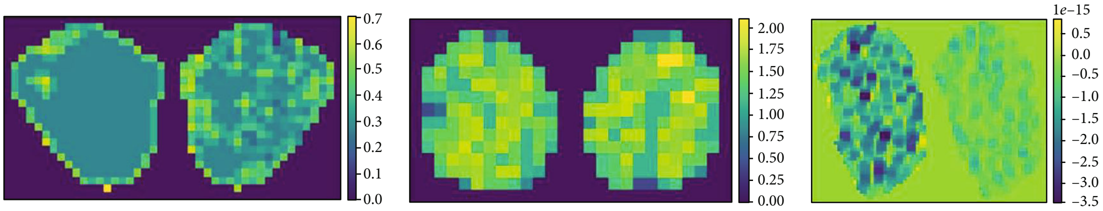


Figure 5: Feature map of (a) lesion region (left) and its symmetric region (right) showed separable by calculating short-run low gray-level emphasis on the square transformed images, (b) normal region (left) and its symmetric region (right) showed inseparable by calculating run entropy on the wavelet transformed images, and (c) lesion region (left) and same position of normal region (right) showed separable by calculating 10th percentile on the wavelet transformed images.

On the other hand, acute ischemic stroke is inconspicuous, more complex features including texture, need to be introduced [19, 20]. Later, Rajini [11] proposed a method to separate ischemic stroke regions from normal tissues in CT, which used segmentation, midline offset, and image features. Nevertheless, most of these studies involved acute lesions that were already visible. Instead, Chawla [12] investigated a two-stage classification system by comparing the image intensity differences between the two hemispheres, which can detect hemorrhagic and ischemic stroke. Recently, a predictive model based on Ranklet features to distinguish strokes and normal tissues was proposed, which achieved significantly high accuracy 81% [13]. Compared with MRI, it is still a gap which is not sufficient for practice, and artificial intelligence-assisted CT diagnosis needs more robust features. As Petrou [21] suggested, a few features, which are not sensitive to human vision and tend to be ignored, are needed to be excavated. Therefore, it is potential for improving the accuracy of detection ACI by exploring other features besides texture.

Lesion delineation is the primary premise in medical image analysis. However, defining the entire lesion boundaries in CT might be complicated because of the invisible of lesion. The next practical way is that ischemic tissues can be highlighted by comparing the left and right sides, since inherent anatomical structures in the human brain are symmetric [22, 23]. Nevertheless, it is unreliable by simply comparing both sides of the brain especially the lesion size was small because of normal brain tissue overwhelming the characteristic features from small lesion.

In order to solve the conundrum of candidate region acquisition and feature quantity insufficiency, we matched exact lesion regions from DWI to CT images. The exact lesion regions and their symmetrical regions served as candidate regions for imaging feature extraction. Inspired by Lambin et al. [14] who extracted a large number of radiomic features from medical images and used statistical analysis to identify features that could characterize disease, we extracted 1301 radiomic features through image transformation and feature extraction. Note that we do not claim any novelty in the extraction design. Instead, our contribution lies in the essence of that constructing more complex feature is necessary for selecting certain features which contribute to classification in the next step. Since not all features were effective, the information gain was further introduced as a standard to measure the correlation between features and ACI, which is according to the principle of feature distinction and independence in mathematical description [16–18]. Furthermore, machine learning is often used as a means to evaluate radiomic analysis [24–26]. Hence, 10 classifiers were established with selected features to verify the effectiveness.

The sufficient experimental data showed differences between the cerebral infarction and their symmetrical noninfarct regions, since the effective features extracted had great potential in classify lesions and their symmetrical regions. Simultaneously, to rule out these separable differences probably coming from the inherently separable between the left and right sides of the normal brain, we operated on normal brain tissue and their symmetrical regions. The results confirmed that the left and right sides of the normal brain tissues were inseparable. According to effective features achieved astounding performance in classify lesion regions and same position of another normal brain tissue, the lesions which were separable from normal tissue in CT were further confirmed. More importantly, the classification results proved the necessity and effectiveness of feature extraction and screening.

Some limitations are noteworthy in the current study. We only included 18 patients performed head CT with no positive finding. They were younger compared to 38 patients performed with both CT and MRI and diagnosed as ACI by DWI, since it is difficult to select normal brain tissue in the elderly. Besides, the size of our population might be considered small; further studies that include a larger population are needed to strengthen the statistical power of these investigations.

5. Conclusion

This study analyzed the separability of acute cerebral infarction lesions in CT, which facilitates CT diagnosis directly. Furthermore, the results of the study established a theoretical foundation for artificial intelligence-assisted CT diagnosis, which will bring potential benefits for acute cerebral infarction patients: shortening the waiting time of thrombolysis, saving the cost of examination, and improving the prognosis.

Abbreviations

ACI: Acute cerebral infarction
DWI: Diffusion-weighted imaging.

Data Availability

The data used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors report no conflicts of interest in this work.

Acknowledgments

We thank Dr. Yanpeng Cheng, PhD, from the Chinese Center for Disease Control and Prevention, for his help in Statistics. This work was supported by the National Natural Science Foundation of China (Grant No.: 81671789), the project of Health Commission of Shanxi Province (Grant No.: 2018109), the Construction project of Engineering Technology Research Center of Shanxi Province (Grant No.: 201805D121008), the Graduate Research and Innovation Projects of Shanxi Province (Grant No.: 2020SY381), and the Science and Technology Foundation of North University of China (Grant No.: 20191634).

References

[1] C. O. Johnson, M. Nguyen, G. A. Roth et al., "Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurology*, vol. 18, no. 5, pp. 439–458, 2019.

[2] S. Rajsic, H. Gothe, H. H. Borba et al., "Economic burden of stroke: a systematic review on post-stroke care," *The European Journal of Health Economics*, vol. 20, no. 1, pp. 107–134, 2019.

[3] J. Kim, T. Thayabaranathan, G. A. Donnan et al., "Global stroke statistics 2019," *International Journal of Stroke*, vol. 15, no. 8, pp. 819–838, 2020.

[4] Y. J. Wang, Z. X. Li, H. Q. Gu et al., "China Stroke Statistics 2019: A Report from the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations," *Stroke and Vascular Neurology*, vol. 5, no. 3, pp. 211–239, 2020.

[5] W. J. Powers, A. A. Rabinstein, T. Ackerson et al., "Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic Stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 50, no. 12, pp. E344–E418, 2019.

[6] C. L. Truwit, A. J. Barkovich, A. Gean-Marton, N. Hibri, and D. Norman, "Loss of the insular ribbon: another early CT sign of acute middle cerebral artery infarction," *Radiology*, vol. 176, no. 3, pp. 801–806, 1990.

[7] M. Mitomi, K. Kimura, J. Aoki, and Y. Iguchi, "Comparison of CT and DWI findings in ischemic stroke patients within 3 hours of onset," *Journal of Stroke and Cerebrovascular Diseases*, vol. 23, no. 1, pp. 37–42, 2014.

[8] J. Emberson, K. R. Lees, P. D. Lyden et al., "Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials," *The Lancet*, vol. 384, no. 9958, pp. 1929–1935, 2014.

[9] A. E. Arch, D. C. Weisman, S. G. Coca, K. Nystrom, C. R. Wira, and J. Schindler, "Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services," *Stroke*, vol. 47, no. 3, pp. 668–673, 2016.

[10] J. L. Saver, M. Goyal, A. van der Lugt et al., "Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis," *JAMA*, vol. 316, no. 12, pp. 1279–1288, 2016.

[11] N. Hema Rajini and R. Bhavani, "Computer aided detection of ischemic stroke using segmentation and texture features," *Measurement*, vol. 46, no. 6, pp. 1865–1874, 2013.

[12] M. Chawla, S. Sharma, J. Sivaswamy, and L. T. Kishore, "A method for automatic detection and classification of stroke from brain CT images, international conference of the ieee engineering in medicine and biology society," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3581–3584, Minneapolis, MN, USA, 2009.

[13] C. M. Lo, P. Hung, and K. L. Hsieh, "Computer-Aided Detection of Hyperacute Stroke Based on Relative Radiomic Patterns in Computed Tomography," *Applied Sciences*, vol. 9, no. 8, p. 1668, 2019.

[14] P. Lambin, E. Rios-Velazquez, R. T. H. Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[15] J. J. M. van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[16] N. Hoque, D. K. Bhattacharyya, and J. Kalita, "MIFS-ND: a mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[17] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, Opatija, Croatia, 2015.

[18] W. Kim, J. Park, H. Sheen et al., "Development of Deep Learning Model for Prediction of Chemotherapy Response Using PET Images and Radiomics Features," in *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pp. 1–3, Sydney, Australia, 2018.

[19] A. Ušinskas, R. A. Dobrovolskis, and B. Tomandl, "Ischemic stroke segmentation on CT images using joint features," *Informatica*, vol. 15, no. 2, pp. 283–290, 2004.

[20] I. Rekik, S. Allassonniere, T. K. Carpenter, and J. M. Wardlaw, "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal," *NeuroImage: Clinical*, vol. 1, no. 1, pp. 164–178, 2012.

[21] M. Petrou and P. García Sevilla, *Image Processing: Dealing with Texture*, John Wiley & Sons, Ltd, 2006.

[22] Y. Shieh, C. Chang, M. Shieh et al., "Computer-aided diagnosis of hyperacute stroke with thrombolysis decision support using a contralateral comparative method of CT image analysis," *Journal of Digital Imaging*, vol. 27, no. 3, pp. 392–406, 2014.

[23] J. Minnerup, G. Broocks, J. Kalkoffen et al., "Computed tomography–based quantification of lesion water uptake identifies patients within 4.5 hours of stroke onset: a multicenter observational study," *Annals of Neurology*, vol. 80, no. 6, pp. 924–934, 2016.

[24] J. Ma, Q. Wang, Y. Ren, H. Hu, and J. Zhao, "Automatic lung nodule classification with radiomics approach," in *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*, vol. 978906, San Diego, California, United States, 2016.

[25] B. Zhang, X. He, F. Ouyang et al., "Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma," *Cancer Letters*, vol. 403, pp. 21–27, 2017.

[26] R. Shiradkar, S. Ghose, I. Jambor et al., "Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: preliminary findings," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 6, pp. 1626–1636, 2018.

Hindawi
BioMed Research International
Volume 2020, Article ID 5193707, 9 pages
https://doi.org/10.1155/2020/5193707

Research Article

Comparison of Supervised and Unsupervised Deep Learning Methods for Medical Image Synthesis between Computed Tomography and Magnetic Resonance Images

Yafen Li,[1,2] Wen Li,[1,2] Jing Xiong,[1] Jun Xia,[3] and Yaoqin Xie[1]

[1] *Institute of Biomedical and Health Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen 518055, China*
[2] *Shenzhen College of Advanced Technology, University of Chinese Academy of Science, Shenzhen 518055, China*
[3] *Department of Radiology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen 518035, China*

Correspondence should be addressed to Yaoqin Xie; yq.xie@siat.ac.cn

Received 7 August 2020; Revised 8 September 2020; Accepted 23 September 2020; Published 5 November 2020

Academic Editor: Zhiguo Zhou

Cross-modality medical image synthesis between magnetic resonance (MR) images and computed tomography (CT) images has attracted increasing attention in many medical imaging area. Many deep learning methods have been used to generate pseudo-MR/CT images from counterpart modality images. In this study, we used U-Net and Cycle-Consistent Adversarial Networks (CycleGAN), which were typical networks of supervised and unsupervised deep learning methods, respectively, to transform MR/CT images to their counterpart modality. Experimental results show that synthetic images predicted by the proposed U-Net method got lower mean absolute error (MAE), higher structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) in both directions of CT/MR synthesis, especially in synthetic CT image generation. Though synthetic images by the U-Net method has less contrast information than those by the CycleGAN method, the pixel value profile tendency of the synthetic images by the U-Net method is closer to the ground truth images. This work demonstrated that supervised deep learning method outperforms unsupervised deep learning method in accuracy for medical tasks of MR/CT synthesis.

1. Introduction

Cross-modality medical image synthesis between magnetic resonance (MR) images and computed tomography (CT) images could benefit medical procedures in many ways. As a multiparameter imaging modality, magnetic resonance imaging (MRI) provides a wide range of image contrast mechanisms without ionizing radiation exposure, while CT images outperform MR images in acquisition time and resolution of bone structure. CT is also related with electron density which is critical for PET-CT attenuation correction and radiotherapy treatment planning [1]. Generating synthetic CT (sCT) images from MR images makes it possible to do MR-based attenuation correction in PET-MR system [2–6] and radiation dose calculation in MRI-guided radiotherapy planning [7–9]. Synthesizing MR images from CT images can enlarge the datasets for MR segmentation task and thus improve the accuracy of segmentation [10].

In recent years, there have been many efforts to work on medical image synthesis between MR and CT images. Among all these methods, deep learning method exhibited superior ability of learning a nonlinear mapping from one image domain to another image domain. It can be classified into two categories: supervised and unsupervised deep learning methods. Supervised deep learning methods required paired images for model training. In the MR/CT synthesis task, MR and CT images have to be well-registered at first and then used as inputs and corresponding labels for the neural network model to learn an end-to-end mapping. Nie et al. [11] used three-dimensional paired MR/CT image patches to train a three-layer fully convolutional network for estimating CT images from MR images.

Other researchers [4, 5, 12–15] have trained deeper network for MR-based CT image prediction. However, as for medical image dataset, it is not that easy to get paired MR and CT images. It may take a long time span to collect patients who are scanned by both MR and CT scanners. Registration of certain accuracy between MR and CT images are also necessary to make paired MR-CT dataset.

Unsupervised deep learning methods enabled the possibility of using unpaired images for image-to-image translation [16–20]. It was first proposed for natural image synthesis and now has been implemented by many researchers for medical image synthesis [10, 21–24]. Chartsias et al. [10] demonstrate the application of CycleGAN in synthesizing cardiac MR images from CT images, using MR and CT images of different patients. Nie et al. [21] synthesized MR images from CT images with a deep convolutional adversarial network. Since there are plenty of unpaired medical images, the available datasets could be easily enlarged.

Unlike natural images, accuracy is highly emphasized in medical images. In this paper, we aim to compare the accuracy of supervised and unsupervised learning-based image synthesis methods for pseudo-MR/CT generation tasks. Two typical networks of U-Net [25] and CycleGAN [17] were introduced as representatives of supervised and unsupervised learning methods, respectively. Mean absolute error (MAE), structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) of the synthetic results were calculated to evaluate their performance quantitatively. More detailed comparisons and discussions about the advantage and disadvantage of these methods are included in Results and Discussion.

2. Materials and Methods

2.1. Neural Network Models. In our experiments of pseudo-MR/CT generation tasks, U-Net and CycleGAN were used as the typical representative network of supervised and unsupervised deep learning methods, respectively.

U-Net has made a great achievement in segmentation tasks [25–29]. The advantage of U-Net is that it could use very few images to make a good performance. In this study, we adapted U-Net to an end-to-end image synthesis task.

The basic architecture of U-Net consists of a contracting part to capture features and a symmetric expanding part to enable precise localization. As shown in Figure 1, we added LeakyReLU [30, 31] as activation operation before convolution operation in the contracting part of the network. Activation operation of LeakyReLU was replaced with ReLU [32] in the expanding part. Batch normalization [33] was introduced to U-Net to enable faster and more stable training. In Figure 1, the number of channels is denoted on top of each of the convolution operation, and the size of feature maps is signed in the parentheses.

In the medical image synthesis task, input image and its corresponding label were fed to the proposed U-Net to train and learn an end-to-end nonlinear mapping between them. Figure 1 illustrated the MR-to-CT synthesis using U-Net architecture, which takes MR images as input and CT images as label to train a synthetic CT generating model. On the contrary, when we use CT images as input and MR images as labels, U-Net could be trained as a synthetic MR-predicting model. The loss function used in the proposed U-Net is

$$L_{\text{U-Net}} = \mathrm{E}_{\mathrm{x},\,\mathrm{y}\sim\hat{P}_{\text{data}}}\left[\|f(x) - y\|_1\right] \tag{1}$$

CycleGAN [17] which is proposed by Zhu et al. could be seen as an updated version of generative adversarial networks (GAN) [16]. GAN methods can learn a nonlinear mapping from input image domain to target image domain by adversarial training. CycleGAN introduced the idea of cycle consistency to general GAN methods. Cycle consistency adds restriction that the generated pseudoimage in target domain should be able to be transformed back to the original input image.

We used the CycleGAN architecture from Zhu et al. [17] for our medical image synthesis task. It takes unpaired MR and CT images as inputs to learn nonlinear mappings between these two image modalities. As illustrated in Figure 2, the CycleGAN architecture has two cycles, forward cycle and backward cycle. The forward cycle consists of three networks: two generative networks of G and F and one discriminator of D_{CT}. The backward cycle uses the same generative networks of F and G and a counterpart discriminator of D_{MR}.

In the forward cycle, network G was used to generate synthetic CT (sCT) from input MR images, while network F generated synthetic MR (sMR) from network G-generated sCT images. Network D_{CT} discriminates whether the generated sCT image is real CT or fake. The backward cycle works just the opposite way. Network F took CT images as input images and generated sMR; then, network G synthesized sCT from the F-generated sMR images. Network D_{MR} was used to distinguish whether the sMR image is real MR or fake.

The adversarial losses of CycleGAN are as follows:

$$\begin{aligned}
L_{\text{GAN\_}G\text{\_MR}_{\text{to}}\text{CT}} &= \mathrm{E}_{\text{CT}\sim P_{\text{data}}(\text{CT})}\left[\|\log\left(D_{\text{CT}}(\text{CT})\right)\|_1\right] \\
&\quad + \mathrm{E}_{\text{MR}\sim P_{\text{data}}(\text{MR})}\left[\|\log\left[1-\left(D_{\text{CT}}(G(\text{MR}))\right)\right]\|_1\right], \\
L_{\text{GAN\_}F\text{\_CT}_{\text{to}}\text{MR}} &= \mathrm{E}_{\text{MR}\sim P_{\text{data}}(\text{MR})}\left[\|\log\left(D_{\text{MR}}(\text{MR})\right)\|_1\right] \\
&\quad + \mathrm{E}_{\text{CT}\sim P_{\text{data}}(\text{CT})}\left[\|\log\left[1-\left(D_{\text{MR}}(F(\text{CT}))\right)\right]\|_1\right].
\end{aligned} \tag{2}$$

The cycle-consistency loss consists of forward cycle loss $L_{\text{forward\_cyc}}$ and the backward cycle loss $L_{\text{backward\_cyc}}$. It is represented as follows:

$$\begin{aligned}
L_{\text{forward\_cyc}} &= \mathrm{E}_{\text{MR}\sim P_{\text{data}}(\text{MR})}\left[\|(F(G(\text{MR})) - \text{MR})\|_1\right], \\
L_{\text{backward\_cyc}} &= \mathrm{E}_{\text{CT}\sim P_{\text{data}}(\text{CT})}\left[\|(G(F(\text{CT})) - \text{CT})\|_1\right], \\
L_{\text{Cycle}-\text{consistency}} &= L_{\text{forward\_cyc}} + L_{\text{backward\_cyc}}.
\end{aligned} \tag{3}$$

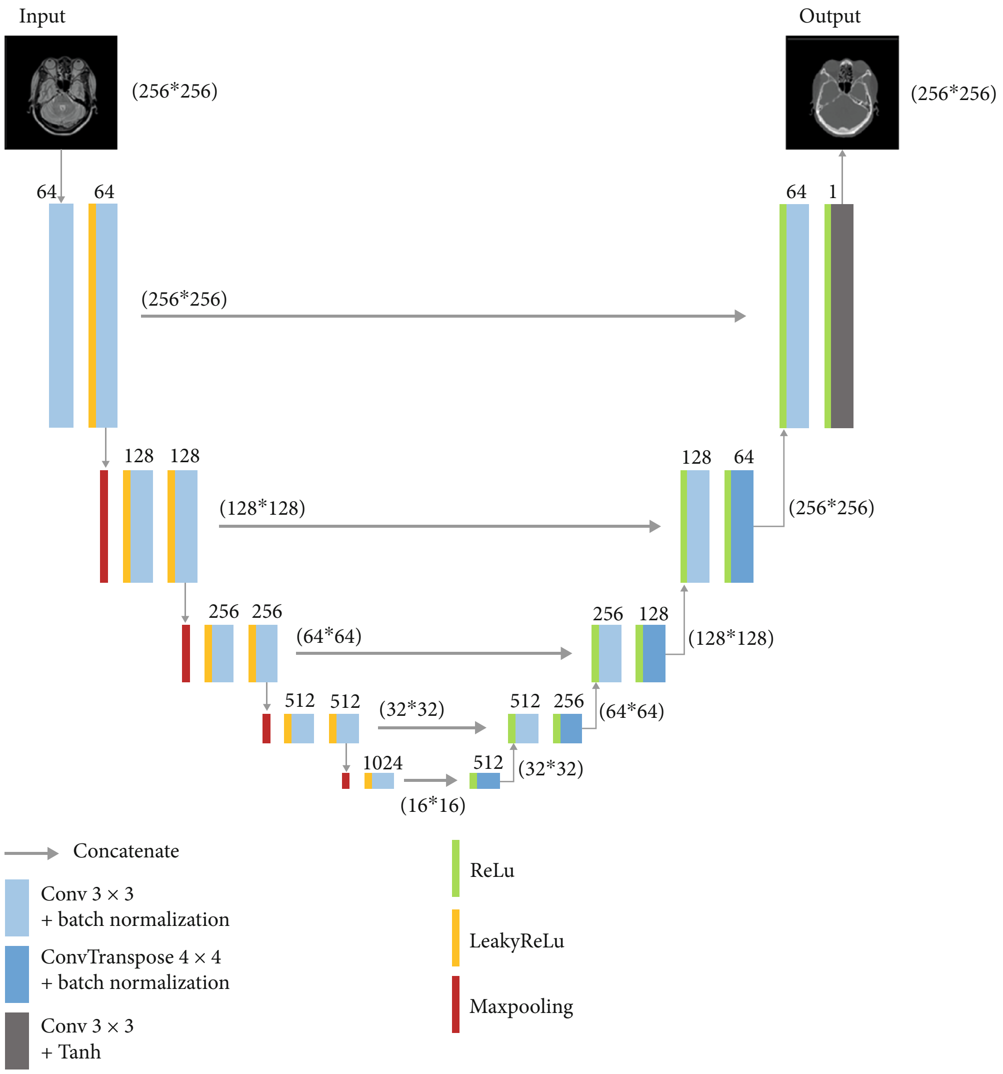


Figure 1: Architecture of proposed U-Net for image synthesis.

Then, we have the full objective as the below equation:

$$L_{\text{CycleGAN}} = L_{\text{GAN\_}G\text{\_MR}_{\text{to}}\text{CT}} + L_{\text{GAN\_}F\text{\_CT}_{\text{to}}\text{MR}} + \lambda L_{\text{Cycle-consistency}}, \quad (4)$$

where λ is the weight of the objectives of cycle consistency.

2.2. Cross-Modality MR/CT Image Synthesis and Evaluation. We used PyTorch to implement the proposed U-Net and CycleGAN. Both the networks were trained for bidirectional image synthesis, which includes learning a MR-to-CT model for generating synthetic CT images from MR images and a CT-to-MR model for generating synthetic MR images from CT images.

U-Net and CycleGAN used similar parameters for training nonlinear mapping models between MRI/CT images. Adam optimizer was adopted for both the networks. The batch size was set to 1. Both networks were trained for 200 epochs, with fixed learning rate for the first 100 epochs. The learning rate decreased linearly to 0 for the following 100 epochs.

Whole 2D slices of axial medical images with size of 256 ∗ 256 pixels were used as inputs. During the training process, the images would be padded to 286 ∗ 286 pixels and then random cropped to 256 ∗ 256 for data augmentation. While U-Net should utilize paired MR and CT datasets for training nonlinear mapping, CycleGAN can take use of unpaired MR and CT images as inputs for both the forward and backward cycles in training procedure. As for the CycleGAN method, we randomly shuffled the MR image input sequences and CT image input sequences in the paired datasets to make the input MR and CT slices unpaired. The MRI input sequence in unpaired datasets were not the same as that in paired datasets.

Three metrics were used to quantitatively characterize the accuracy of the prediction of synthetic images compared with the ground truth images. The mean absolute error (MAE) measures the discrepancies by voxels. Structural similarity index (SSIM) [34] quantifies the similarities in a whole

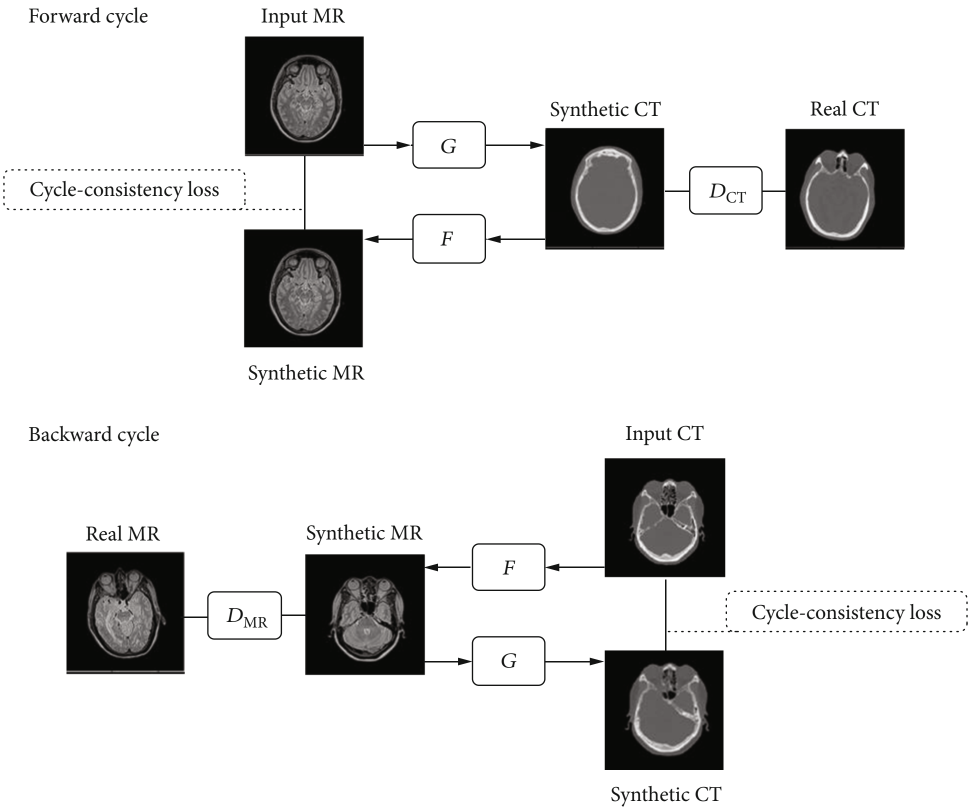


Figure 2: CycleGAN architecture for bidirection synthesis of MR and CT images. The forward cycle generated synthetic CT from input MR by G while F translate the synthetic CT back to the MR image domain. D_{CT} discriminate whether the generated images is real or fake CT. The backward cycle generated synthetic MR from input CT by F while G translate the synthetic MR back to the CT image domain. D_{MR} discriminate whether the generated images is real or fake MR. Two cycle-consistency loss was introduced to capture the intuition that the synthetic image should be translated back to the original image modality.

image scale. Peak signal-noise-ratio (PSNR) assesses the quality of prediction.

These evaluation metrics are expressed as follows:

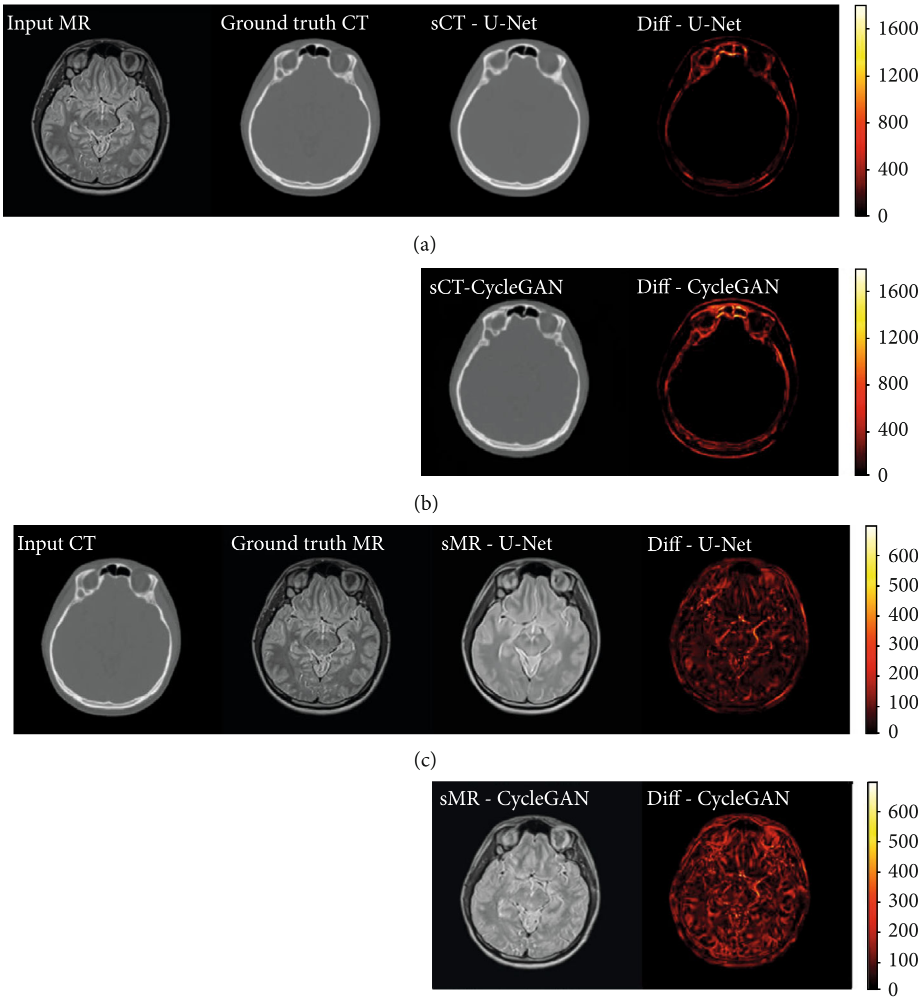
$$\begin{aligned}
\mathrm{MAE} &= \frac{1}{H * W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|X(i,j) - Y(i,j)\right|,\\
\mathrm{SSIM} &= \left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)/\left(\mu_x^2 + \mu_y^2 + c_1\right)\\
&\quad\cdot\left(\sigma_x^2 + \sigma_y^2 + c_2\right),\\
c_1 &= \left(K_1L\right)^2, c_2 = \left(K_2L\right)^2,\\
\mathrm{PSNR} &= 10\ \log_{10}(L/\mathrm{MSE}), \mathrm{MSE} = \frac{1}{H * W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(X(i,j) - Y(i,j)\right)^2,
\end{aligned} \tag{5}$$

where H and W are the height and width of the images, respectively. X is the ground truth images, and Y is the predicted synthetic images. μ_x and μ_y are the average values of ground truth images and synthetic images, respectively. σ_x^2 and σ_y^2 are the variance of ground truth images and synthetic images, respectively. σ_{xy} represents the covariance of ground truth images and synthetic images. L denotes the dynamic range of the voxel values. c_1 and c_2 are two variables to stabilize the division with a weak denominator. Here, we take $k_1 = 0.01$ and $k_2 = 0.03$ by default.

2.3. Dataset Preparing. The datasets contain 34 patients. Each patient has both T2-weighted MR images and CT images of the head region. We acquired T2-weighted MR images (TR: 2500 ms, TE: 123 ms, $1 * 1 * 1\,\mathrm{mm}^3$, $256 * 256$) on a 1.5 T Avanto scanner (Siemens). The CT images (120 kV, 330 mA, exposure time: 500 ms, $0.5 * 0.5 * 1$ mm3, $512 * 512$) were acquired on SOMATOM Definition Flash (Siemens).

In this experiment, CT images were resampled to a size of $256 * 256$ ($1 * 1\,\mathrm{mm}^2$) by bicubic interpolation [35] to match the voxel size of MR images. Binary head masks were generated by the Otsu threshold method [36] for MR and CT images to remove unnecessary background information around the head region.

Since the head region is mainly a rigid construction of bone structure, we applied rigid registration to the MR and CT images to make paired MR/CT images for the proposed U-Net. CT images were set as a fixed volume. MR images were set as a moving volume to register with CT images by Elastix toolbox [37]. The paired datasets were randomly shuffled to make an unpaired dataset for CycleGAN.



(a)

(b)

(c)

(d)

Figure 3: (a–d) From left to right: the 4 columns are input images, ground truth images, synthetic images, and the difference maps. sCT results generated by U-Net (a) and CycleGAN (b), respectively; sMR results by U-Net (c) and CycleGAN (d), respectively.

In our medical image synthesis task, 28 patients with 4063 image pairs were randomly selected for model training. The remaining 6 patients with 846 image pairs were used for evaluation procedure.

3. Results and Discussion

The results of synthetic MR and synthetic CT images generated by U-Net and CycleGAN and their ground truth are showed in Figure 3. The first column is the input images, and the second column is ground truth images. The third column showed the generated synthetic images predicted from input images by the two networks. The difference map between synthetic images and ground truth images was calculated and showed in the fourth column.

The first two rows in Figure 3 are sCT images synthesized by U-Net and CycleGAN, respectively. For the task of synthesizing CT images from MR images, the soft tissue area is translated from high contrast to low contrast. It could be seen from the difference map images that the soft tissue area of synthetic CT images by both networks is well-translated with little error. The translation error mainly occurred in the bone area. Their difference map demonstrates that the sCT by CycleGAN synthesized more error than sCT by U-Net in the bone areas.

The third and fourth rows in Figure 3 are sMR images generated by U-Net and CycleGAN, respectively. It could be seen that sMR by CycleGAN seems more realistic for it has more complex contrast information than sMR by U-Net. However, the difference map images illustrated that the CycleGAN method generated much more error than U-Net does. The abundant image contrast information in sMR by CycleGAN may be false and unnecessary.

In synthesizing CT tasks, the difference between synthetic images and ground truth mainly occurs in the bone area. But in synthesizing MR tasks, the error is evenly distributed in the whole head region. It means synthesizing high contrast images of MR from low contrast image domain of CT is tougher than its reverse synthesizing direction.

To compare the image details, 1D profiles of pixel intensity were also plotted. Figure 4 shows the 1D profiles passing through the short red lines and long blue lines as indicated in

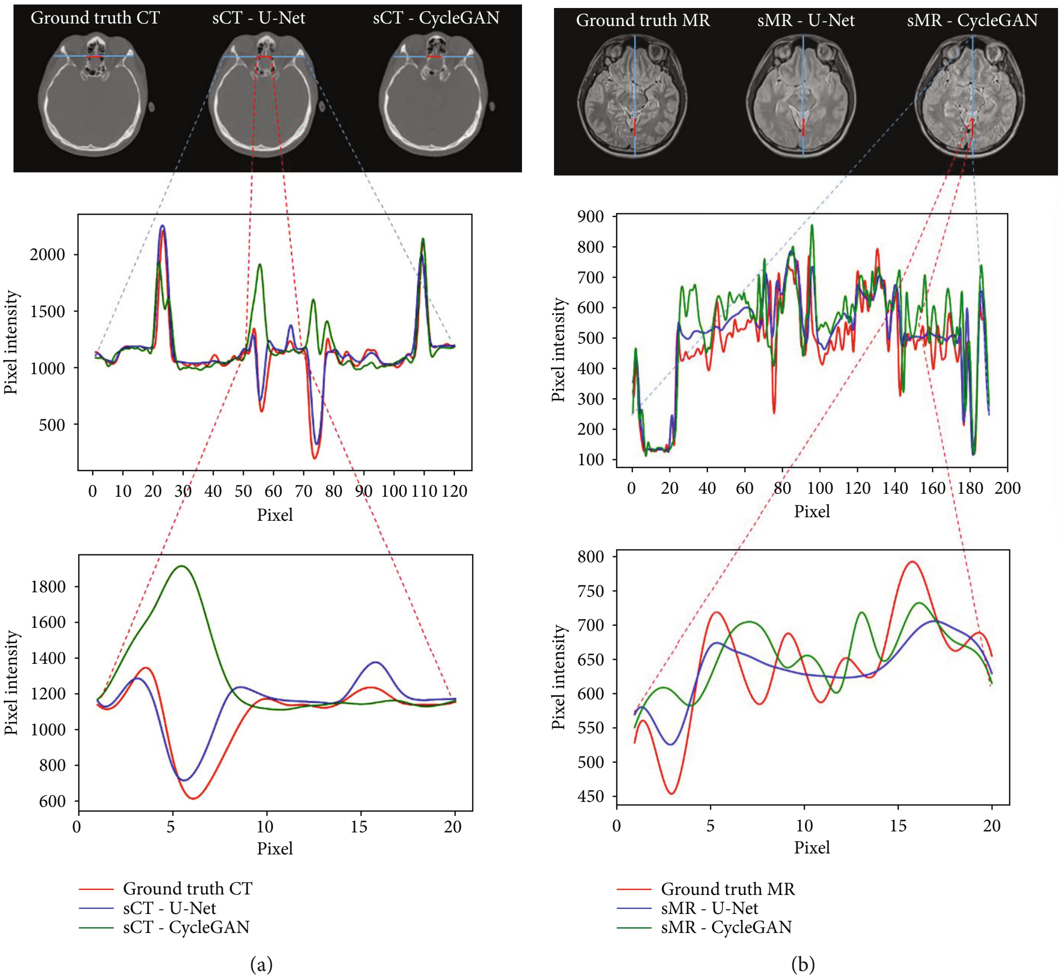


Figure 4: Comparison of 1D profiles of pixel intensity passing through the short red lines and long blue lines as indicated in the images: (a) 1D profile and its close-up marked by the horizontal lines in ground truth CT, U-Net sCT, and CycleGAN sCT images; (b) 1D profile and its close-up marked by the horizontal lines in ground truth MR, U-Net sMR, and CycleGAN sMR images.

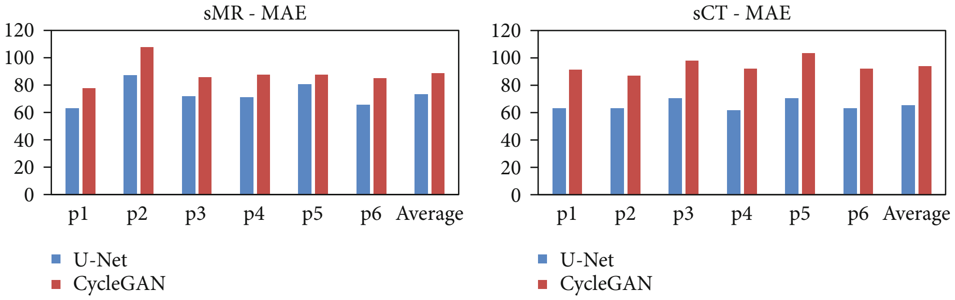
the corresponding images in the first row. The red line is overlapped with the blue lines. The 1D profile in the second row of Figure 4 demonstrates pixel intensities of the long blue lines. The 1D profiles in the third row are the pixel intensities of the short red lines of 20 pixels, which shows close-ups of part of the long blue lines' 1D profile.

In the profiles, the red curve indicates pixel intensities of ground truth CT or MR. The blue curve represented for U-Net and the green curve for CycleGAN. It could be clearly seen in Figure 4(a) that the blue curve is close to the red curve, while some of the peaks of the green curve deviated from the red curve to an opposite direction. It means that the tendency of 1D profiles in sCT by U-Net was closer to the ground truth CT, while the CycleGAN method tends to generate fake contrast information in sCT images.

The profile in Figure 4(b) shows that the blue curve vibrated less from the red curve. Some peaks of the green curve deviated more from the red curve. It could be seen in the close-up 1D profile that some peaks of the green curve are biased to the opposite from the red curve, while the tendency of the blue cure seems like a smoothened or flattened red curve. It means that the pixel value of sMR by U-Net was closer to the ground truth but may lack contrast details. The pixel value of sMR by CycleGAN exhibits more deviation from the ground truth along the profile whereas the tendency may be false or exaggerated.

The quantitative metrics have been calculated for comparison. Figure 5 shows the MAE of sCT and sMR for each of the 6 patients in the evaluation datasets and the average result. It is obvious that the U-Net method generated lower MAE either in sCT image generation or sMR image generation for all the patients. This also demonstrates the robust performance of the U-Net method in bidirection MR/CT image translation tasks.

Figures 5(a) and 5(b) show that the deviations of the MAE between the U-Net and CycleGAN method for sMR images of all the 6 patients are not as significant as those for sCT images. In Figure 3, the difference map of sMR indicated that the main predicted errors are evenly distributed in the whole head region, while the main error of sCT mainly occurs mainly in the bone structure. This could be interpreted that generating MR images of high soft tissue contrast



(a)

(b)

Figure 5: (a) MAE of sMR images for all the 6 patients in test set and their average value. (b) MAE of sCT images generated for all the 6 patients in test set and their average value. Both the blue columns denoted the U-Net method, and the red columns represented the CycleGAN method.

Table 1: Quantitative evaluation results between ground truth CT images and sCT images: MAE, SSIM, and PSNR.

Model	MAE ± SD (HU)	SSIM ± SD	PSNR ± SD (dB)
U-Net	65.36 ± 4.08	0.972 ± 0.004	28.84 ± 0.57
CycleGAN	93.95 ± 5.89	0.955 ± 0.007	26.32 ± 0.55

Table 2: Quantitative evaluation results between ground truth MR images and sMR images: MAE, SSIM, and PSNR.

Model	MAE ± SD (HU)	SSIM ± SD	PSNR ± SD (dB)
U-Net	73.43 ± 9.16	0.946 ± 0.004	32.35 ± 0.78
CycleGAN	88.71 ± 10.04	0.924 ± 0.003	30.79 ± 0.73

from CT images of low soft tissue contrast is much complex than the inverse direction synthesis of generating CT from MR images.

Table 1 shows the overall statistics of three quantitative metrics for sCT by both the U-Net and CycleGAN methods. The SSIM values indicate that the sCT images by both methods have fairly high similarity with the ground truth CT images. The U-Net method outperformed the CycleGAN method with a much lower MAE of 65.36 HU, a higher SSIM of 0.972, and a higher PSNR of 28.84 dB. The average sCT MAE deviation between the two methods is nearly 30 HU.

Table 2 shows the overall statistics of three quantitative metrics for sMR images by the U-Net method and CycleGAN method. The U-Net method outperformed the CycleGAN method with a lower MAE of 73.43 HU, a higher SSIM of 0.946, and a higher PSNR of 32.35 dB.

The qualitative and quantitative results demonstrate that the proposed U-Net, a typical supervised learning method, outperforms CycleGAN, a representative advanced unsupervised learning method, in synthesis accuracy of medical image translation task. Since medical images highly value accuracy for the purpose of disease diagnosing, clinical treatment, and therapeutic effect evaluation, the supervised learning method is more recommended in medical practice.

Nevertheless, the success of supervised learning cannot do without well-registered image pairs. The performance of the trained model also depends on the registration accuracy of the paired images. Unlike natural images, paired medical images are not that easy to get. It would take a long time span to collect enough patients who need to be scanned for both MR and CT images at the same time. It is well-known that big amount of datasets could greatly improve the performance of the deep learning method. Though it outperforms the unsupervised learning method, the limit of dataset volume may constrain the further improvement of the supervised learning method in medical image synthesis tasks.

From the experiments discussed above, the image synthesis by using unsupervised learning methods still has a long way to go for practical application in clinic due to their relatively low accuracy. But still, the unsupervised learning method could benefit when there is lack of paired medical image datasets. The good news is that there are abundant easy-to-obtain retrospective unpaired MR and CT images for the unsupervised learning method to take advantage of. No registration is needed.

Our experiments show that when the same datasets were taken as inputs, the unsupervised learning method got inferior quality in the synthesis accuracy for medical image translation. But nonetheless, if the dataset is large enough, it could be expected that the performance of the unsupervised learning method would be improved to a certain acceptable extent in clinical practice.

4. Conclusions

Cross-modality medical image synthesis between MR and CT images could benefit a lot from the fast growing of deep learning methods. In this paper, we compared different deep learning-based image synthesis methods for pseudo-MR/CT generation, including the unsupervised learning method of CycleGAN and supervised learning methods of the proposed U-Net. Synthetic images produced by the CycleGAN method contain more but fake contrast information in the whole image scale. Though the proposed U-Net method blurred the generated pseudoimages, its pixel value profile tendency is basically close to the ground truth images. The quantitative results also indicate that the U-Net method outperformed the CycleGAN method, especially in synthesizing CT image task.

As accuracy is highly demanded in medical procedures, we recommend the supervised method such as the proposed U-Net in cross-modality medical image synthesis at present clinical practice.

Data Availability

The datasets of MR and CT images used to support the findings in this study are restricted by the Medical Ethics Committee of Shenzhen Second People's Hospital in order to protect patient privacy.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work is supported in part by grants from the National Key Research and Development Program of China (2016YFC0105102), the Leading Talent of Special Support Project in Guangdong (2016TX03R139), the Fundamental Research Program of Shenzhen (JCYJ20170413162458312), the Natural Science Foundation of Guangdong Province (2017B020229002, 2015B020233011, and 2014A030312006), the Beijing Center for Mathematics and Information Interdisciplinary Sciences, and the National Natural Science Foundation of China (61871374).

References

[1] P. Dirix, K. Haustermans, and V. Vandecaveye, "The value of magnetic resonance imaging for radiotherapy planning," *Seminars in Radiation Oncology*, vol. 24, no. 3, pp. 151–159, 2014.

[2] N. Burgos, M. J. Cardoso, K. Thielemans et al., "Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies," *Ieee Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2332–2341, 2014.

[3] A. Mehranian, H. Arabi, and H. Zaidi, "Vision 20/20: magnetic resonance imaging-guided attenuation correction in PET/MRI: challenges, solutions, and opportunities," *Medical Physics*, vol. 43, no. 3, pp. 1130–1155, 2016.

[4] F. Liu, H. Jang, R. Kijowski, T. Bradshaw, and A. B. McMillan, "Deep learning MR imaging-based attenuation correction for PET/MR imaging," *Radiology*, vol. 286, no. 2, pp. 676–684, 2018.

[5] A. P. Leynes, J. Yang, F. Wiesinger et al., "Zero-echo-time and Dixon deep pseudo-CT (ZeDD CT): direct generation of pseudo-CT images for pelvic PET/MRI attenuation correction using deep convolutional neural networks with multiparametric MRI," *Journal of Nuclear Medicine*, vol. 59, no. 5, pp. 852–858, 2018.

[6] Y. Wu, W. Yang, L. Lu et al., "Prediction of CT substitutes from MR images based on local diffeomorphic mapping for brain PET attenuation correction," *Journal of Nuclear Medicine*, vol. 57, no. 10, pp. 1635–1641, 2016.

[7] M. A. Schmidt and G. S. Payne, "Radiotherapy planning using MRI," *Physics in Medicine and Biology*, vol. 60, no. 22, pp. R323–R361, 2015.

[8] J. A. Dowling, J. Sun, P. Pichler et al., "Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences," *International Journal of Radiation Oncology • Biology • Physics*, vol. 93, no. 5, pp. 1144–1153, 2015.

[9] A. M. Dinkla, J. M. Wolterink, M. Maspero et al., "MR-only brain radiation therapy: dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network," *International Journal of Radiation Oncology • Biology • Physics*, vol. 102, no. 4, pp. 801–812, 2018.

[10] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," Springer International Publishing, Cham, 2017.

[11] D. Nie, X. H. Cao, Y. Z. Gao, L. Wang, and D. G. Shen, "Estimating CT image from MRI data using 3D fully convolutional networks," in *Deep Learning and Data Labeling for Medical Applications*, vol. 10008, pp. 170–178, 2016.

[12] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Medical Physics*, vol. 44, no. 4, pp. 1408–1419, 2017.

[13] F. Liu, P. Yadav, A. M. Baschnagel, and A. B. McMillan, "MR-based treatment planning in radiation therapy using a deep learning approach," *Journal of Applied Clinical Medical Physics*, vol. 20, no. 3, pp. 105–114, 2019.

[14] J. Fu, Y. Yang, K. Singhrao et al., "Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging," *Medical Physics*, vol. 46, no. 9, pp. 3788–3798, 2019.

[15] K. Gong, J. Yang, K. Kim, G. El Fakhri, Y. Seo, and Q. Li, "Attenuation correction for brain PET imaging using deep neural network based on Dixon and ZTE MR images," *Phys Med Biol*, vol. 63, no. 12, p. 125011, 2018.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in Neural Information Processing Systems 27 (Nips 2014)*, vol. 27, 2014.

[17] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 Ieee International Conference on Computer Vision (Iccv)*, pp. 2242–2251, Venice, Italy, 2017.

[18] Z. L. Yi, H. Zhang, P. Tan, and M. L. Gong, "DualGAN: unsupervised dual learning for image-to-image translation," in *2017 Ieee International Conference on Computer Vision (Iccv)*, pp. 2868–2876, Venice, Italy, 2017.

[19] P. Isola, J. Y. Zhu, T. H. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 5967–5976, Hawaii, America, 2017.

[20] J. X. Lin, Y. C. Xia, T. Qin, Z. B. Chen, and T. Y. Liu, "Conditional image-to-image translation," in *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 5524–5532, Salt Lake City, America, 2018.

[21] D. Nie, R. Trullo, J. Lian et al., "Medical image synthesis with context-aware generative adversarial networks," *Medical Image Computing and Computer-Assisted Intervention*, pp. 417–425, 2017.

[22] M. Maspero, M. H. F. Savenije, A. M. Dinkla et al., "Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy," *Phys Med Biol*, vol. 63, no. 18, p. 185001, 2018.

[23] H. Emami, M. Dong, S. P. Nejad-Davarani, and C. K. Glide-Hurst, "Generating synthetic CTs from magnetic resonance

images using generative adversarial networks," *Medical Physics*, vol. 45, no. 8, pp. 3627–3636, 2018.

[24] P. Costa, A. Galdran, M. I. Meyer et al., "End-to-end adversarial retinal image synthesis," *Ieee Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2018.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Munich, Germany, 2015.

[26] Y. Yang, C. Feng, and R. Wang, "Automatic segmentation model combining U-Net and level set method for medical images," *Expert Systems with Applications*, vol. 153, p. 113419, 2020.

[27] S. Chen, H. Yang, J. Fu et al., "U-Net plus: deep semantic segmentation for esophagus and esophageal cancer in computed tomography images," *Ieee Access*, vol. 7, pp. 82867–82877, 2019.

[28] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U -Net for lesion segmentation," in *2019 Ieee 16th International Symposium on Biomedical Imaging (Isbi 2019)*, pp. 683–687, Venice, Italy, 2019.

[29] Y. Y. Zeng, X. Y. Chen, Y. Zhang, L. F. Bai, and J. Han, "Dense-U-Net: densely connected convolutional network for semantic segmentation with a small number of samples," in *Tenth International Conference on Graphics and Image Processing (Icgip 2018)*, vol. 11069, Chengdu, China, 2019.

[30] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network. ICML deep learning workshop," 2015, http://arxiv.org/abs/1505.00853.

[31] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, Haifa, Israel, 2010.

[33] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, Lille, France, 2015.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Ieee Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[35] H. Prashanth, H. Shashidhara, and K. N. Balasubramanya Murthy, "Image scaling comparison using universal image quality index," in *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference*, pp. 859–863, Trivandrum, Kerala, India, December 2009.

[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[37] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *Ieee Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

Hindawi
BioMed Research International
Volume 2020, Article ID 5491963, 13 pages
https://doi.org/10.1155/2020/5491963

Research Article

Application of BERT to Enable Gene Classification Based on Clinical Evidence

Yuhan Su,[1] **Hongxin Xiang**,[1] **Haotian Xie**,[2] **Yong Yu**,[1] **Shiyan Dong**,[3] **Zhaogang Yang**,[3] **and Na Zhao**[1]

[1]*National Pilot School of Software, Yunnan University, Kunming, 650091, China*
[2]*Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA*
[3]*Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA*

Correspondence should be addressed to Zhaogang Yang; zhaogang.yang@utsouthwestern.edu and Na Zhao; zhaonayx@126.com

Received 31 July 2020; Revised 31 August 2020; Accepted 7 September 2020; Published 7 October 2020

Academic Editor: Zhiguo Zhou

The identification of profiled cancer-related genes plays an essential role in cancer diagnosis and treatment. Based on literature research, the classification of genetic mutations continues to be done manually nowadays. Manual classification of genetic mutations is pathologist-dependent, subjective, and time-consuming. To improve the accuracy of clinical interpretation, scientists have proposed computational-based approaches for automatic analysis of mutations with the advent of next-generation sequencing technologies. Nevertheless, some challenges, such as multiple classifications, the complexity of texts, redundant descriptions, and inconsistent interpretation, have limited the development of algorithms. To overcome these difficulties, we have adapted a deep learning method named Bidirectional Encoder Representations from Transformers (BERT) to classify genetic mutations based on text evidence from an annotated database. During the training, three challenging features such as the extreme length of texts, biased data presentation, and high repeatability were addressed. Finally, the BERT+abstract demonstrates satisfactory results with 0.80 logarithmic loss, 0.6837 recall, and 0.705 F-measure. It is feasible for BERT to classify the genomic mutation text within literature-based datasets. Consequently, BERT is a practical tool for facilitating and significantly speeding up cancer research towards tumor progression, diagnosis, and the design of more precise and effective treatments.

1. Introduction

Nowadays, genomic, transcriptomic, and epigenomic studies have been benefited from the development of inexpensive next-generation sequencing technologies, which play essential roles in exploring tumor biology [1–3]. Tumors usually possess heterogeneities, and the genomic profiling of tumors normally contains various types of genetic mutations [4–7]. However, only a small proportion of mutation genes are involved in boosting tumor growth, whereas most of them are neutral and irrelevant to tumor progression [8, 9]. Characterization and identification of cancer driver genes are important a in clinical trials to reveal tumor pathogenesis and facilitate diagnosis, prognosis, and personalized therapy [10–13]. Despite the importance of gene classification, the following analysis is challenging due to the significant amount of manual work for interpreting genomics, which is time-consuming, laborious, and subjective. With the increasing availability of electronic unstructured and semistructured data sources, automatically categorizing documents has emerged as a potential tool for information organization. Machine learning (ML), as a promising optimization tool, has been widely used in credit scoring, fraud detection, retailers, market segmentation, manufacturing, education, and healthcare [14–18]. Hence, using ML to analyze clinical contextual data automatically is favorable [19–21]. For example, in 1986, Swanson first discovered the undiscovered links in a large number of scientific literature [22]. Also, Marcotte et al. used Naive Bayesian classification to classify the literature focusing on protein-protein interaction [23].

Despite the achievements traditional ML methods have made, potential drawbacks such as low accuracy exist when they are applied on clinical text classification. In 2018, Google proposed that the BERT method achieved state-of-the-art results in 11 projects, including text classification [24]. Descriptions about clinical research acadamic papers show high similarities , which blurs the classification boundary, increases the inconsistancy, and lows the accuracy. Consequently, the advanced ML methods, such as Light Gradient Boosting Machine (LightGBM), has been proposed to enable gene multiclassification based on complex literature [25]. Nevertheless, these methods are limited by complex calculations when applied to large-scale datasets, particularly for genomic-related literature datasets that contain millions, or billions, of annotated training examples [26, 27]. In addition, the performances of ML are dependent on feature extraction that requires professional knowledge and long-term processing [28–31].

To overcome these difficulties, deep learning (DL) has emerged to handle large-scale and complex datasets since its performance increases with the enlargement of datasets [32–34]. For example, the convolutional neural networks (CNN) [35], recurrent neural networks (RNN) [36], and their combination [37] have been applied to the sentence classification successfully. Also, In 2018, Google proposed that the BERT method achieved state-of-the-art results in 11 projects, including text classification.

Hence, we fine-tune the BERT model to classify mutation effects (9 classes) using an expert-annotated oncology knowledge base. Our BERT method is developed based on the original BERT model and is capable of obtaining different syntactic and semantic information. Three main characters of training datasets including extreme length of text entry, data imbalance, and repetitive description are engineered during training challenges. We propose three truncation methods including abstract+head, head only, and head+tail to deal with extreme length of text entry and repetitive description. Besides, data imbalance is relieved by negative sampling. Overall, we improve the BERT method to classify complex clinical texts, and obtain 0.8074 logarithmic loss, 0.6837 recall, and 0.705 F-measure scores.

2. Problem Statement

The treatment of cancer is closely related to the identification of mutant genes [38]. At present, clinicians need review and classify each mutant gene manually according to the evidence in text-based clinical literature, which is a complicated, time-consuming, and error-prone method [39–42]. To solve this problem, Memorial Sloan Kettering Cancer Center (MSKCC) has provided an expert-annotated precision oncology knowledge base with thousands of mutations manually annotated by world-class researchers and oncologists for studying gene classification using computer-based method [43]. On top of that, we design an artificial intelligence algorithm to automatically and accurately classify mutations for avoiding mistakes caused by manual classification, and provide further help for cancer treatments.

In recent years, with the rise of artificial intelligence, natural language processing, which uses linguistics, computers, mathematics, and other scientific methods to communicate between human beings and computers, has developed rapidly [44–46]. Among them, text classification is one of the most basic and critical tasks in natural language processing [47]. Text classification is the process of associating a given text within one or more categories according to characteristics of texts (content or attributes) under a predefined classification system [48–50]. The process of text classification mainly includes three steps. Firstly, the text is preprocessed, then the vector representation of the text is extracted. Finally, the classifier is trained to classify the text [48]. Text classification can be divided into single-label text classification and multilabel text classification according to the number of labels to which the text belongs. The single-label text refers to each text belonging to only one category, while multilabel text refers to each text belonging to one or more categories [51–53]. The calculation formula for text classification can be defined as follows:

$$F(D, C) = \{\text{True}, \text{False}\}. \tag{1}$$

In the formula, the collection $D = \{d_1, d_2, \cdots d_n\}$ refers to the set of texts classified, where the ith classified text is represented by $\mathrm{d_i}$, and n is the number of classified texts. The collection $C = \{c_1, c_2, \cdots, c_m\}$ is a collection of predefined classification categories, where the jth category is represented by c_j, and m is the number of predefined categories. F is a function representing a mapping relationship.

Currently, the most common methods for text classification are statistical ML and DL-based methods. Statistical ML methods usually preprocess texts in the first place, then manually extract high-dimensional sparse features. Consequently, they use statistical ML algorithms to obtain classification results. In 1998, Joachims first employed support for vector machine (SVM) in text classification and achieved favorable results [54]. In the following research, many methods based on statistical ML are used in text classification, including Naïve Bayes classifier [55], K-nearest Neighbor method (KNN) [56], decision tree [57], boosting [58], and LightGBM [59]. Among them, LightGBM is widely used in classification problems due to its fast speed, low memory consumption, and relatively high accuracy [60]. Although LightGBM gets good classification results in some scenes, research related to this approach runs basically into bottleneck due to its strong dependence on the effectiveness of features. Also, it is time-consuming and labor-intensive during feature extraction process.

Although the traditional statistical ML models can classify texts faster than the manual method, they require manual feature extraction, which leads to a large amount of labor cost and is difficult to obtain effective features [61–63]. On the other hand, the DL methods are superior to traditional statistical ML methods in terms of text feature expression and automatic acquisition of feature expression capabilities, thus eliminating complex manual feature engineering processes and reducing possible application costs [64]. As we all know, large-scale pretraining language models have become a new driving force for various natural language processing tasks [65]. For example, BERT models can significantly improve model performance by fine-tuning downstream tasks. Google first proposed the BERT model, and it completely subverted the logic of training word vectors before training specific tasks in natural language processing [24]. Methods of fine-tuning the BERT model, such as extended text preprocessing and layer adjustment, have been proved to improve the results substantially [66]. Wu et al. proposed a conditional BERT method, which can enhance the text classification ability of original BERT method by predicting the conditions of masked words [67]. To sum up, it is feasible to employ the fine-tuned model based on the original BERT to classify genetic mutations.

Hence, we propose an improved BERT model with high classification accuracy after analyzing the MSKCC mutation gene interpretation database thoroughly. We believe this method can be successfully applied to genetic mutation classification. The main contributions of our work are summarized as follows:

(1) The text description of the individual sample shows considered lengths. There are differences in text lengths between different categories of samples. Some categories contain shorter words, while others contain miscellaneous descriptions. Generally, texts in a dataset range from hundreds to thousands of words in length. However, the lengths of the gene mutation in this paper are much longer than usual. We use the BERT method to truncate texts and extract valuable information in the texts using different methods, thus avoiding adverse impacts of excessive differences in text lengths on the results.

(2) There is a deviation of total gene number in all categories. Individual genes are unevenly distributed in different categories. Some genes belong to five or more groups, while others only present in two categories. To solve the vast differences in the number of samples between different categories in the dataset, we choose an undersampled data processing method to balance the data deviations between different categories.

(3) The whole dataset has a high repeated description. Different examples belong to different categories share the same text entry. Some categories show a high correlation, which may lead to low accuracy. To solve this problem, we improve the BERT model and splice the last three layers of the initial model, which increases the accuracy of the model and reduces the running time.

(4) To a certain extent, we illustrate the effectiveness of using DL in the classification of genetic clinical texts. As the data set increases, the DL model represented by BERT will learn the characteristics of the sample better to achieve exceptional results. In the future, DL models will have better performances on similar tasks.

3. Materials and Methods

3.1. Description of Datasets. MSKCC sponsored the training and test datasets in this study for method development and validation. For the past several years, world-class experts have created a clinical evidence annotated precision oncology knowledge database. The annotations contain information about which genes are oncology clinically actionable. We sum up three characteristics of the MSKCC datasets mentioned below:

(i) Textual descriptions of individual samples exhibit considerable lengths. The text lengths among different classes show variabilities. Some of the classes contain shorter words while other classes contain redundant descriptions.

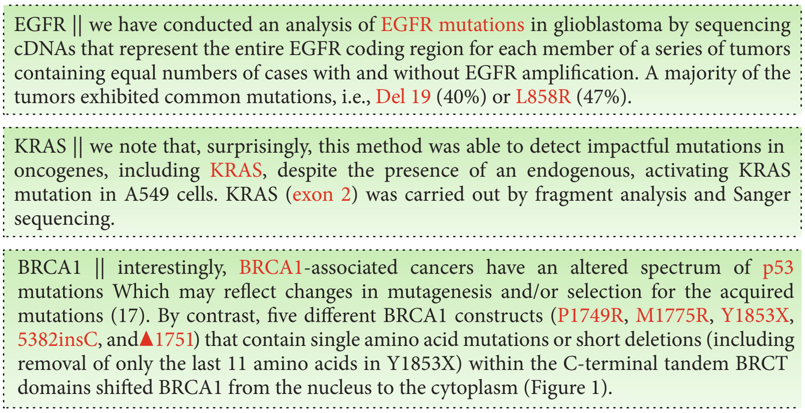
(ii) The overall gene numbers presented among the whole classes show biases. The distribution of individual genes in different classes is unequal. Some genes belong to five classes or more, and some of the genes only fit in two classes.

(iii) High repetitive descriptions exist in the whole datasets. Different samples belong to different classes that share the same text entry. Classes demonstrate high correlations.

3.1.1. Length of Entry Text. It is reasonable to analyze the length of the entry text as a prior task for textual-based classification. We find that extremely long descriptions with massive irrelevant information are correlated with samples (Figure 1). We plot the distribution of text lengths (Figure 2), and our datasets contain more counted words than the normal classification datasets in reviews [68]. Consequently, we examine the distribution of text lengths among different target classes to better understand the uniformity of datasets. Variabilities are demonstrated among different classes (Figure 3). Comparing the density of the length distributions, we divide the classes into three groups. Classes 3, 5, and 6 contain the shortest counted words; classes 1, 2, 4, and 7 exhibite medium counted words; and classes 8 and 9 show the most counts. Overall, two features that increase the task difficulty are attracted: considerable lengths of words and the unequal text length distribution among different classes.

3.1.2. Analysis of the Data Distribution. Analyzing the composition of datasets can help us construct algorithms at an early stage. We sum up the frequency of genes among 9 classes (Figure 4). The 9 classes correspond to mutation effects but are annotated using numbers instead of real textural information to avoid artificial labeling, thus improving the reliability of our algorithms during the training. The true information of these labels is listed in Table 1. The distribution of genes among 9 classes exhibited bias. Genes in class 7 are significantly higher than genes in classes 3, 8, and 9.

We also examin the interactions among different features within target classes. To reduce calculations, we select the top 20 gene types to illustrate the interrelations instead of the whole gene types (Table 2). Selected genes are sorted by



EGFR || we have conducted an analysis of EGFR mutations in glioblastoma by sequencing cDNAs that represent the entire EGFR coding region for each member of a series of tumors containing equal numbers of cases with and without EGFR amplification. A majority of the tumors exhibited common mutations, i.e., Del 19 (40%) or L858R (47%).

KRAS || we note that, surprisingly, this method was able to detect impactful mutations in oncogenes, including KRAS, despite the presence of an endogenous, activating KRAS mutation in A549 cells. KRAS (exon 2) was carried out by fragment analysis and Sanger sequencing.

BRCA1 || interestingly, BRCA1-associated cancers have an altered spectrum of p53 mutations Which may reflect changes in mutagenesis and/or selection for the acquired mutations (17). By contrast, five different BRCA1 constructs (P1749R, M1775R, Y1853X, 5382insC, and▲1751) that contain single amino acid mutations or short deletions (including removal of only the last 11 amino acids in Y1853X) within the C-terminal tandem BRCT domains shifted BRCA1 from the nucleus to the cytoplasm (Figure 1).

Figure 1: The cut-off document views of the datasets.

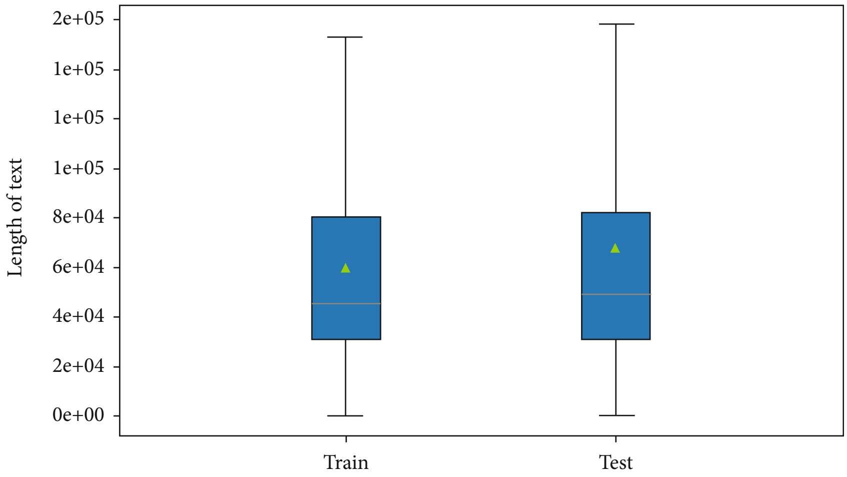


Figure 2: Distribution of the text entry lengths.

classes (Figure 5). The distribution of genes demonstrate huge variabilities among different classes. We find that classes 8 and 9 contain almost none of these genes, and class 3 contain a few of these genes. These distribution biases are in accordance with our previous gene frequency summary based on the whole gene types. Similarly, the trends in classes 1, 2, 4, and 7 correspond to our previous results. These comparable results indicate that the whole datasets are highly associated with selected genes. Consequently, discriminatory differences among classes can impede the feature learning performances of our algorithms and low the accuracy of the text classification.

We further explore the distribution of individual genes within classes, which demonstrates inequitable distributions. For instance, genes such as CDKN2A, PTEN, and TSC2 only present in a limited number of classes (lesser than three). In contrast, BRCA1, ERBB2, FGFR2, and RET are possessed in the majority of classes. Compared with genes only present in a few groups, genes that spread among classes are generally difficult to classify because elaborate texture descriptions can blur the classification standard. Hence, the accuracy of classifications is dependent on the gene compositions. Commonly, genes distributed in lesser classes can show more satisfactory results.

3.1.3. Characteristics of the Datasets. Using typical genes as samples, we find that these typical genes presented in classes demonstrated variabilities. To better recognize these biases and complete potential influences behind them, we conduct a statistical analysis of the whole datasets from the text entry aspects. We find that different samples share the same text entries after extracting common words. The highly repetitive descriptions increase the difficulties of classification, especially when samples in different classes share the same sketches. The worst scenario is the fact that samples belong to different classes that have the same name, but other clue information is missing. For example, five possible mutations of gene BRCA1, the mutation P1749R, M1775R, Y1853X, 5382insC, and $\Delta 1751$, may belong to different classes, but their descriptions are close, even in the same sentence. Similarly, two mutations of EGFR, such as Del 19 and L858R, also show in pairs (Figure 1). Hence, we can assume that it is tough to categorize the samples into correct classes by relying on the name of mutations with limited or without other valuable information.

Also, class-dependent word similarities are evaluated using full word lists (Figure 6). Correlation coefficients exhibited high connections (higher than 60%) between classes. Among them, classes 2 and 7 and classes 1 and 4 demonstrate

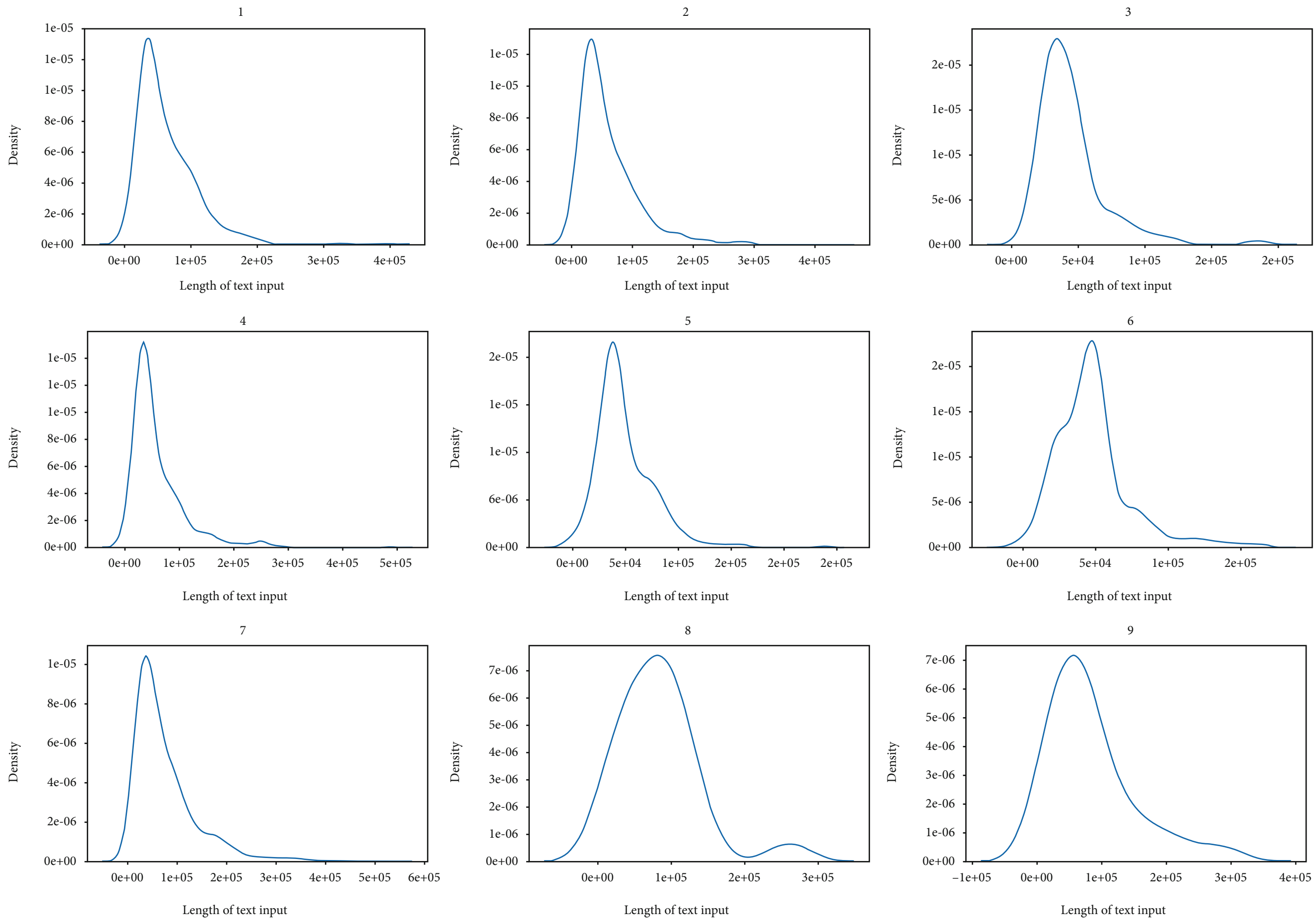


Figure 3: Distribution of the text entry lengths among different classes.

extremely high correlations with 97% and 93% coefficients, respectively. Therefore, we think substantial work needs to be done to clarify samples that share similar descriptions in high correlative classes. Besides, we can not expect high accuracy when classifying samples with these properties.

3.2. BERT. Compared with traditional ML methods, DL demonstrates better performances in text feature expression and automatically obtains feature expression capabilities, thus removing the complicated manual feature engineering process and decreasing its application cost. BERT is a new language representation model based on DL, which was released by the AI team of Google company in October 2018. The BERT model is divided into two parts: pretraining and fine-tuning.

3.2.1. Pretraining of Modified BERT Model. In the pretraining process, a large-scale unlabeled text corpus is used to complete the deep vector representation of text content in the deep bidirectional neural network through an unsupervised training method, thus forming the corresponding text pretraining model. Google has trained two pretrained models. One is the BERT-base model, which includes 12 transformers, 12 self-attention heads, and 768 hidden sizes. The other is the BERT-large model, which contains 24 transformers, 16 self-attention heads, and 1024 hidden sizes. Parameters of BERT-base methods are loaded into the downstream BERT classification model so that our model parameters can be fine-tuned based on these pretrained models, which significantly reduces the convergence time of the model and increases the accuracy of the model. During the pretraining process, BERT randomly masks out, replaces some words, and predicts these missing or replaced words through the remaining ones. The transformer must maintain a distributed representation of each input token. The transformer is likely to remember the word masked without this masking and predicting procedure.

3.2.2. Fine-Tuning of Modified BERT Model. Since the generalization ability of the pretrained model is powerful, the BERT pretrained model can be applied to various downstream tasks after fine-tuning the parameters of the pretrained model. For example, it is possible to meet the needs of a text classification task by adding pooling, full connect, and Softmax function to the output layer sequence of fine-tuned BERT model. The fine-tuning process requires much lesser training resources compared to the pretraining process. The method of fine-tuning BERT model, such as truncation and layer adjustment, has been proved to be capable of improving the result [18]. It implements the process of unsupervised learning through the mask, thereby predicting the vocabulary that will appear in the sentence and

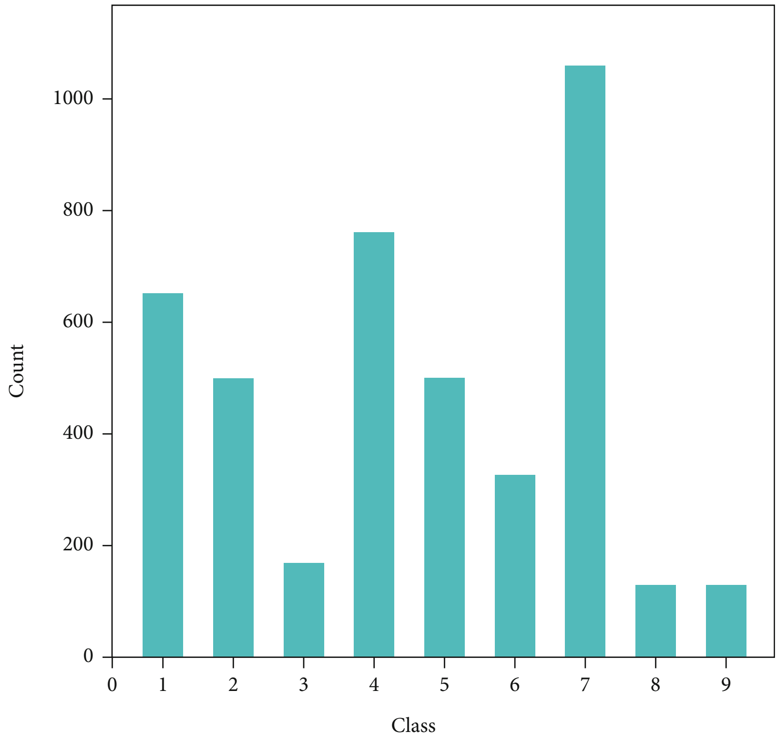


Figure 4: Distribution of the number of genes among 9 classes.

Table 1: Class information corresponds to the annotated number.

Annotated number	Class information
1	Likely loss of function
2	Likely gain of function
3	Neutral
4	Loss of function
5	Likely neutral
6	Inconclusive
7	Gain of function
8	Likely switch of function
9	Switch of function

Table 2: List of top 20 genes in the datasets.

Rank	Gene name	Rank	Gene name
1	EGFR	11	FLT3
2	TP53	12	MTOR
3	CDKN2A	13	MAP2K1
4	ERBB2	14	PTEN
5	PDGFRA	15	BRCA1
6	TSC2	16	BRAF
7	PIK3CA	17	BRCA2
8	FGFR2	18	KIT
9	ALK	19	KRAS
10	VHL	20	RET

understanding the specific meaning of the sentence according to the context.

3.3. Evaluation Equation. This paper evaluates the performances of the model using several evaluation indicators: Logloss, recall (REC), precision (PRE), F1 score, receiver operating characteristic (ROC) curve, and confusion matrix. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) can be used to calculate some of the indicators mentioned above. TP is the number of categories that are correctly predicted. TN is the number of categories that are correctly predicted as another class. FP is the number of categories that are wrongly predicted. FN is the number of categories that are wrongly predicted as another class.

In multiclassification tasks, Logloss is one of the most common loss functions, where the predicted input is a probability value distribution between 0 and 1, and it can be defined as follows:

$$\text{Logloss} = -\frac{1}{S_n} \sum_{m=1}^{S_n} \sum_{n=1}^{N} y_{mn} \log \left(p(y_{mn})\right), \tag{2}$$

where M is the number of samples and N is the number of classifications. y_{mn} is the predicted result of classification, such as 0 and 1. $p\,(y_{mn})$ is the predicted probability of y_{mn}.

PRE defines the proportion of genes identified correctly belonging to this type of mutation:

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

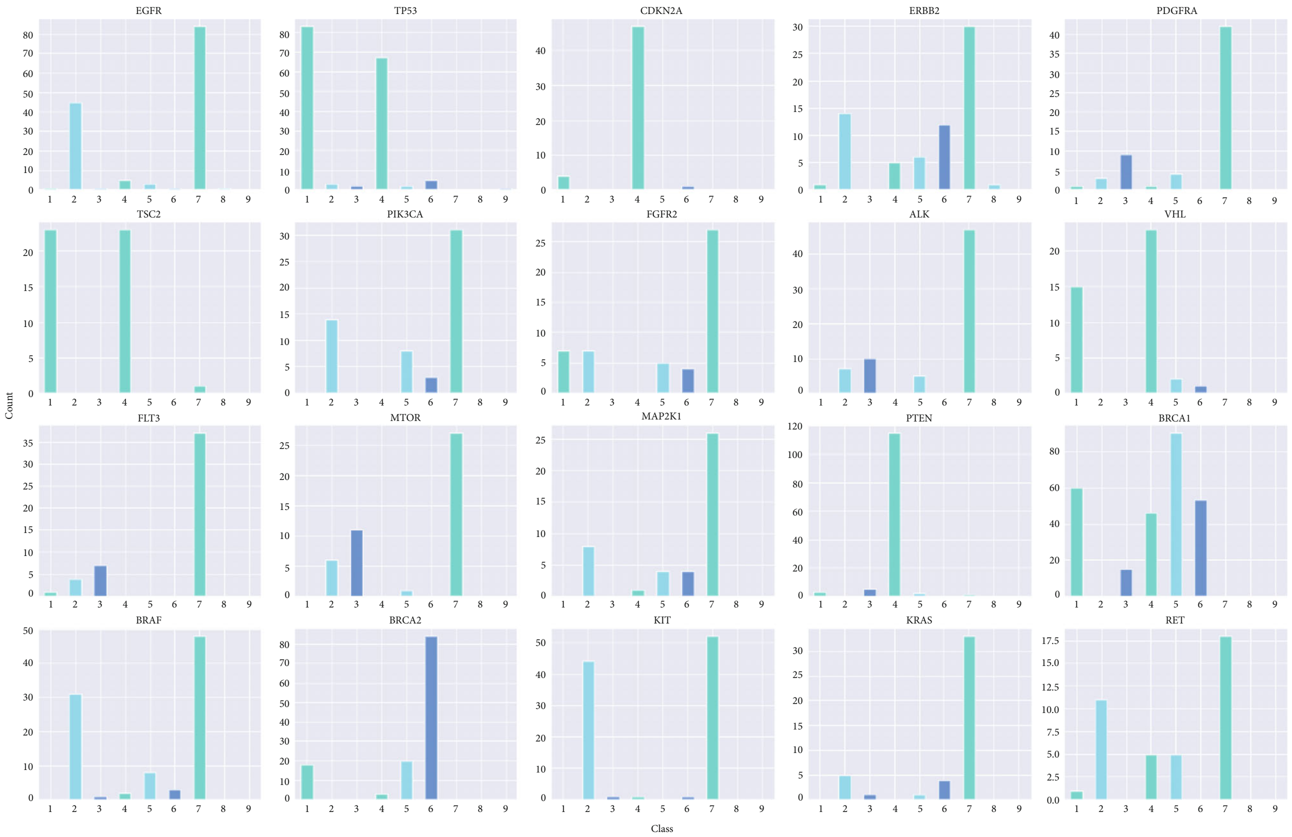


Figure 5: Distribution of genes among classes.

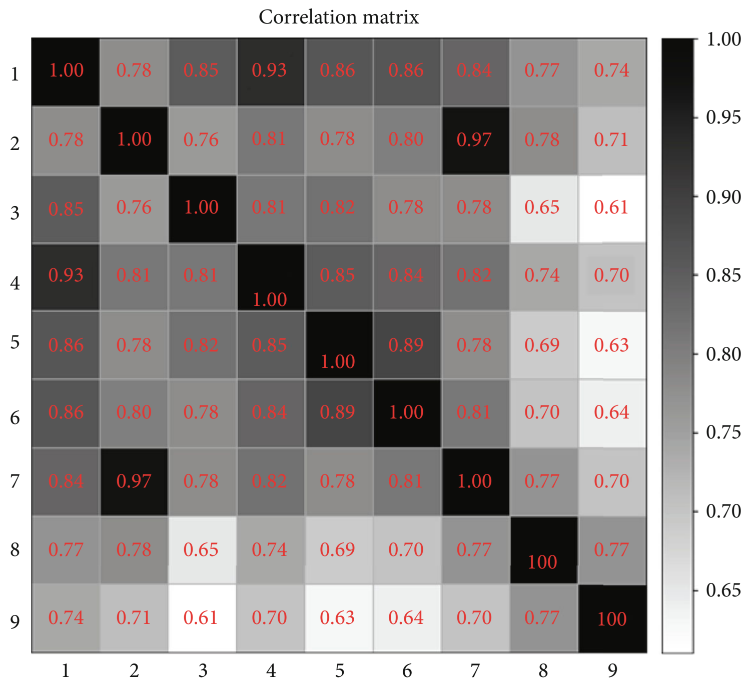


Figure 6: Confusion matrix analysis of the similarity of the texts in different classes.

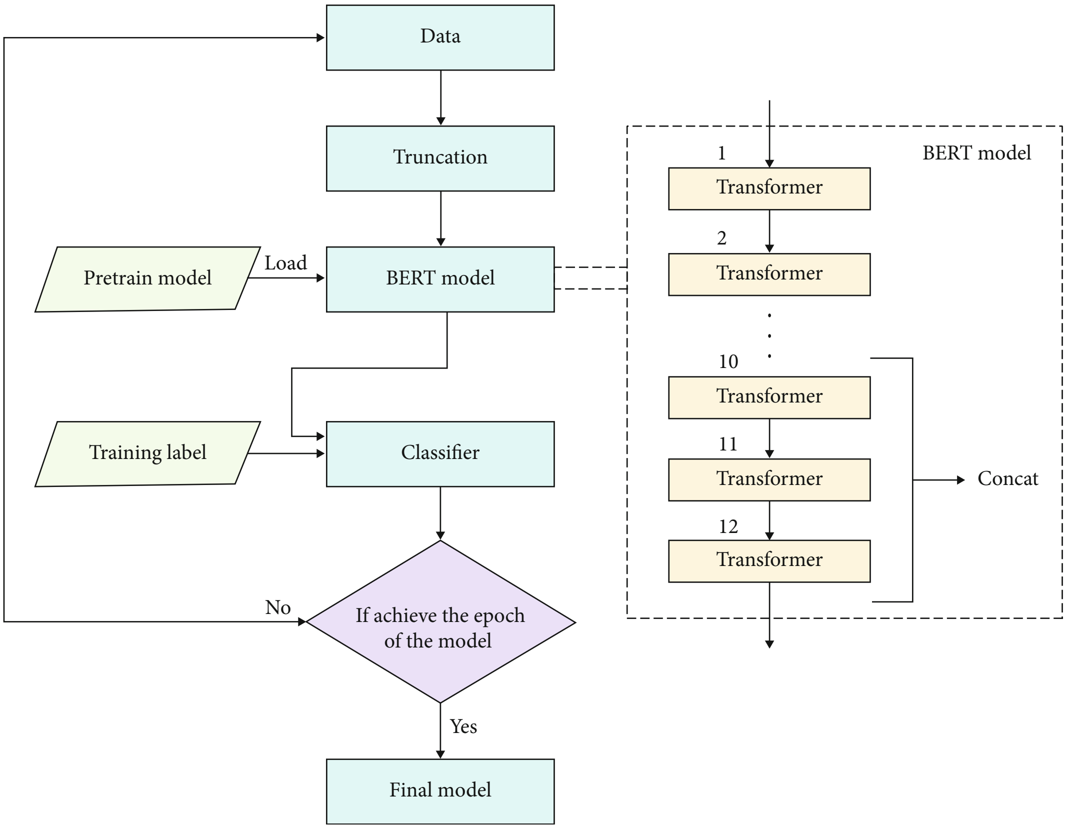


Figure 7: Scheme of the training.

REC calculates the proportion of genes identified correctly belonging to this type of mutation in all this type of gene:

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (4)$$

F1 score takes into account the factors of PRE and REC. F1 is the standard metric for this task. It combines precision and recall. Macro-F1 is a parameter index that can best reflect the effectiveness and stability of the model:

$$\mathrm{F1} = \frac{2\mathrm{PRE} * \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}. \quad (5)$$

The ROC curve is created by plotting the TP against the FP at various threshold settings.

The confusion matrix is a specific table capable of visualizing the performance of an algorithm. Individual rows of the matrix represent the predicted gene classses, while each column represents the genes in the actual classes.

4. Experiments

For easier comparison with other methods, our training process uses the GPU of the server in the lab for training. There are 3136 training sets and 553 verification sets in total. The Python language is selected as the programming language in this experiment. The experiment is completed on Tensorflow's open-source framework and BERT-base. We use the parameters on BERT-base trained by Google through a large number of corpus on Wikipedia as pretraining parameters to accelerate the convergence speed and reduce the convergence difficulty. Our experimental parameters are batch size 128, learning rate $3e-5$, and warmup period 0.06; the whole experiment runs for 30 cycles; the maximum sequence length of BERT input is 512; and the optimizer is Adam optimizer, while other model parameters remain unchanged.

4.1. Experiment Procedure. The BERT model can automatically complete the process of converting each word in the text into a one-dimensional vector by querying the word vector table and inputting it in the model. The input of the model contains three sections: the token embeddings, the segmentation embeddings, and position embeddings.

Because BERT is a pretraining model with high generalization ability, the output layer of BERT can be externally connected with corresponding layers to complete downstream tasks. For example, in this experiment, the processed data is substituted into the BERT model for training, and the output layer will connect Softmax function for classification tasks (Figure 7).

BERT is an unsupervised model that uses whether the sentences are related to each other as labels and masks some words to make the masked words as labels, thus avoiding the tedious process of manually labeling data. Generally, the data in the dataset are not balanced. Take the samples in the 7th and 8th categories of the dataset as an example. The

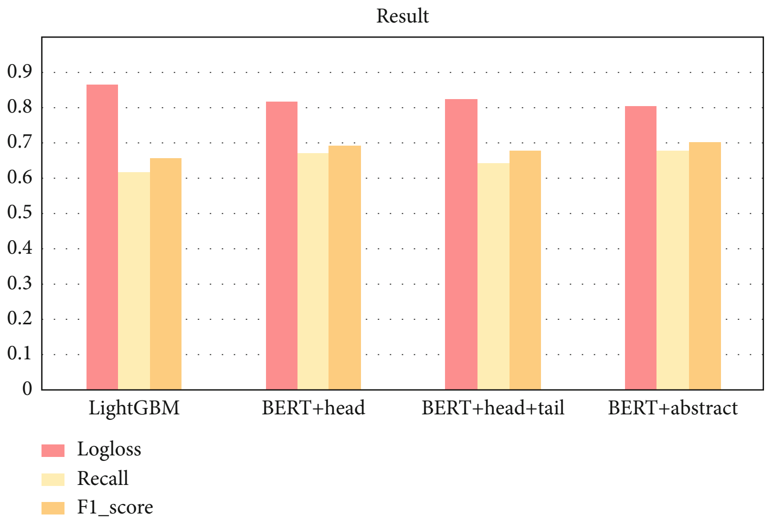


Figure 8: Evaluation of four methods.

difference between their numbers even reaches more than 10 times. In this case, the default classification method makes classifiers pay too much attention to the category with a larger number of samples, thus making the generalization ability of the model weak and unable to obtain satisfactory results. Therefore, we use random sampling to eliminate the imbalance between data and extract only a part of samples from the category within a larger number of samples to balance the sample number differences between classes.

Simultaneously, because the length of the gene text in the dataset is greater than 512 tokens, which is the longest length that can be retained by BERT, we need to use the truncation method to intercept part of the information in the text. We take three ways to solve this problem. The head only truncation method intercepts the first 512 tokens (at most) as input, the head+tail method intercepts part of the head and part of the tail to form 512 tokens (at most) as input, and the abstract +head method sorts the gene text according to importance, then select the most important 512 tokens (at most) as input.

Finally, the processed data are substituted into the BERT model for training. Numerous previous works have shown that fine-tuning a pretrained model which has been trained with a large amount of corpus can significantly improve the classification result. As BERT can learn different contents in different layers, stitching some of the layers together can make the model get richer information, thereby improving the accuracy of the model, so the last three layers in the BERT model are concatenated. Max pooling, fully connected, and Softmax function are added after the concatenated output layer to realize the classification of gene text to improve the classification accuracy of the model.

4.2. Experiment Results and Discussions. It can be seen from the figure that compared with the LightGBM method, the BERT methods using three types of truncation have higher ACC, REC, and F1 score. The confusion matrix shows our classification situation in a visual way (Figure 6). The red numbers are nonzero values. It can be observed that type 1 is easy to be confused with types 4 and 5. There are more machine judgment errors of texts between type 7 and type 2. Overall, the classification of data-lacking types 8 and 9 is more complicated than other types, possibly because there are fewer samples of types 8 and 9, and these two types have fewer intersections with other types of mutation. The lack of intersection leads to difficulties in distinguishing types 8 and 9 from different types of mutations. The ROC curve can evaluate the accuracy of the model prediction.

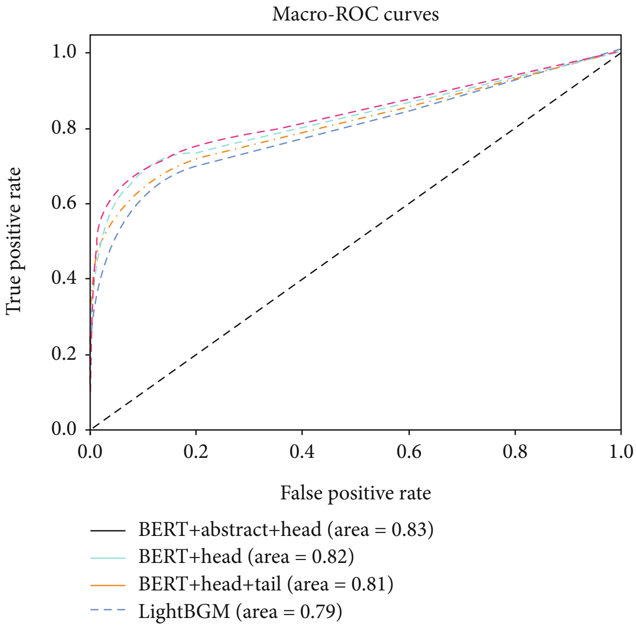


Figure 9: ROC curves of the proposed methods.

The performance and ranking of the entries for the proposed four methods are shown in Figure 8. All methods share the same setting of hyperparameters for an unbiased comparison. Overall, deep learning-based algorithms (BERT) perform slightly better than machine learning-based methods (LightGBM). Among the three models using BERT, the BERT+abstract truncation method has the best performance

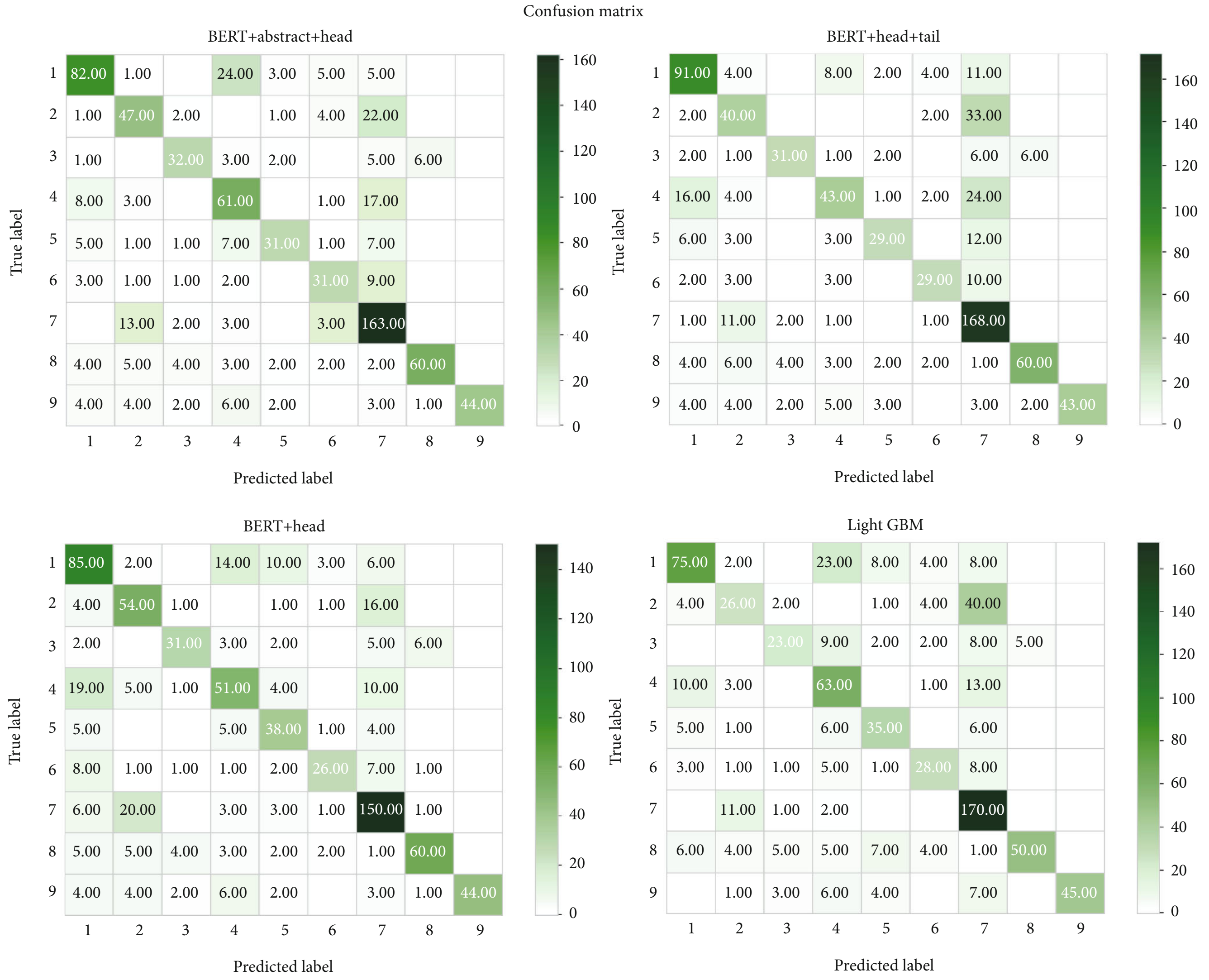


Figure 10: Confusion matrix tables of proposed four methods.

as a single model with 0.8074 logarithmic loss and 0.6837 recall. The 0.705 *F*-measure score is limited by the extreme shortage of training data. Better performance should be obtained when it is applied to large-scale datasets.

Besides, the ROC curves of the other three methods are below the ROC curve of the BERT+abstract (Figure 9). The ROC curves for the BERT+abstract, the BERT+head, the BERT+head+tail, and the LightGBM with the highest and lowest AUCs are also shown in Figure 9. Compared results indicate better performance of the BERT+abstract since the AUC assesses the algorithm's inherent validity using an effective and combined measure of sensitivity and specificity. The accuracy of predicted results is highly dependent on the datasets. The performances of our model are limited by the size of available datasets in our case. However, the capabilities of deep networks can be improved using expanded data. Our proposed model is a proof-of-concept, and we believe it is applicable when applied on large-scale datasets.

Moreover, we compare confusion matrix tables using predict classes versus true classes among different methods (Figure 10). The confusion matrix table is an error matrix which can be used to evaluate the performance of the algorithm. In summary, individual classes of genes are predicted precisely using the BERT+abstract method, corresponding with results of Logloss and F1 measurements.

It can be observed that class 1 is easy to be confused with classes 4 and 5. Furthermore, there are more machine judgment errors of texts between type 7 and type 2. These phenomena can be easily attributed to the similarity of texts among these classes as we previously described. Also, it is apparent that classifying classes 8 and 9 is complicated. The computer may misjudge mutation texts with real labels of 8 or 9 as other types but hardly underestimate other types of mutation texts as type 8 or 9 since there are fewer samples in classes 8 and 9. The shortage of samples in classes 8 and 9 also fails to provide sufficient data to distinguish themselves from other classes since there are no intersections. Contrastingly, the classification of class 7 is easier due to the abundant samples. Therefore, the abundance of data plays essential roles in improving the efficiency of classification.

5. Conclusion

In this study, we propose a deep learning algorithm to identify genomic information within texture-based literature abstracts. Aiming to address the classification problem in an extremely long, imbalanced, and repetitive dataset, we test four methods, including LightGBM and three different truncation BERT methods. By analyzing their Logloss, recall, F1 score, ROC curve, and AUC scores, we notice that the abstract+head truncation BERT method has superior results than other algorithms in all indicators.

In this study, our BERT method is limited due to the shortage of datasets, and its performance can be improved dramatically with the size of datasets increasing. Moreover, our approach will be potentially applied on diagnosing and treating more than 120,000 patients every year around the world based on the announcement of the MSKCC, which will provide our opportunity to enhance our methods further when large-scale datasets are available. We believe BERT is a promising tool for accelerating tumor genomic-related research and facilitating tumor diagnosis and treatments. Besides, this text-based classifier algorithm demonstrated high universality, and it is applicable not only in tumor-specific research but also in other types of diseases and in other nonacademic areas.

Data Availability

All data generated or analyzed during this study are included in this published article.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

N Zhao and ZG Yang conceived and designed the experiments. YH Su, HX Xiang, HT Xie, and Y Yu analyzed and extracted data. YH Su and HT Xie constructed the figures. YH Su, HT Xie, and ZG Yang participated in table construction. All authors participated in the writing, reading, and revising of the manuscript and approved the final version of the manuscript.

Acknowledgments

This work has been supported by the National Key Research and Development Program No.2018YFB2100100, Data-Driven Software Engineering innovation team of Yunnan province of China No.2017HC012, Postdoctoral Science Foundation of China No.2020M673312, Innovation and Entrepreneurship training projects for College Students of Yunnan University No.20201067307, Postdoctoral Science Foundation of Yunnan Province, Project of the Yunnan Provincial Department of Education scientific research fund No. 2019J0010, and DongLu Young and Middle-aged backbone Teachers Project of Yunnan University.

References

[1] M. L. Metzker, "Sequencing technologies — the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[2] Z. Yang, J. Shi, J. Xie et al., "Large-scale generation of functional mRNA-encapsulating exosomes via cellular nanoporation," *Nature Biomedical Engineering*, vol. 4, no. 1, pp. 69–83, 2020.

[3] M. Masseroli, A. Canakoglu, P. Pinoli et al., "Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data," *Bioinformatics*, vol. 35, no. 5, pp. 729–736, 2019.

[4] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[5] P. J. Stephens, The Oslo Breast Cancer Consortium (OSBREAC), P. S. Tarpey et al., "The landscape of cancer genes and mutational processes in breast cancer," *Nature*, vol. 486, no. 7403, pp. 400–404, 2012.

[6] C. Ma, F. Jiang, Y. Ma, J. Wang, H. Li, and J. Zhang, "Isolation and detection technologies of extracellular vesicles and application on cancer diagnostic," *Dose-response : a publication of International Hormesis Society*, vol. 17, no. 4, pp. 155932581989100–1559325819891004, 2019.

[7] N. Walters, L. T. H. Nguyen, J. Zhang, A. Shankaran, and E. Reátegui, "Extracellular vesicles as mediators ofin vitroneutrophil swarming on a large-scale microparticle array," *Lab on a Chip*, vol. 19, no. 17, pp. 2874–2884, 2019.

[8] M. H. Bailey, C. Tokheim, E. Porta-Pardo et al., "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, pp. 371–385.e18, 2018.

[9] J. R. Pon and M. A. Marra, "Driver and passenger mutations in cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 10, no. 1, pp. 25–50, 2015.

[10] A. Youn and R. Simon, "Identifying cancer driver genes in tumor genome sequencing studies," *Bioinformatics*, vol. 27, no. 2, pp. 175–181, 2011.

[11] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas et al., "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Scientific Reports*, vol. 3, no. 1, p. 2650, 2013.

[12] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[13] Y. Liu, Y. Ma, J. Zhang, Y. Yuan, and J. Wang, "Exosomes: a novel therapeutic agent for cartilage and bone tissue regeneration," *Dose Response*, vol. 17, no. 4, article 1559325819892702, 2019.

[14] C.-F. Tsai and J.-W. Wu, "Using neural network ensembles for bankruptcy prediction and credit scoring," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2639–2649, 2008.

[15] B. Norgeot, B. S. Glicksberg, and A. J. Butte, "A call for deep-learning healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 14-15, 2019.

[16] Z.-L. Sun, T.-M. Choi, K.-F. Au, and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing," *Decision Support Systems*, vol. 46, no. 1, pp. 411–419, 2008.

[17] C. Shen, D. Nguyen, Z. Zhou, S. B. Jiang, B. Dong, and X. Jia, "An introduction to deep learning in medical physics: advantages, potential, and challenges," *Physics in Medicine & Biology*, vol. 65, no. 5, p. 05TR01, 2020.

[18] R. Wang, Y. Weng, Z. Zhou, L. Chen, H. Hao, and J. Wang, "Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy," *Physics in Medicine & Biology*, vol. 64, no. 24, p. 245005, 2019.

[19] H. Wu, G. Toti, K. I. Morley et al., "SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research," *Journal of the American Medical Informatics Association*, vol. 25, no. 5, pp. 530–537, 2018.

[20] C. Gulden, M. Kirchner, C. Schüttler et al., "Extractive summarization of clinical trial descriptions," *International Journal of Medical Informatics*, vol. 129, pp. 114–121, 2019.

[21] S. Li, P. Xu, B. Li et al., "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features," *Physics in Medicine & Biology*, vol. 64, no. 17, p. 175012, 2019.

[22] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986.

[23] E. M. Marcotte, I. Xenarios, and D. Eisenberg, "Mining literature for protein-protein interactions," *Bioinformatics*, vol. 17, no. 4, pp. 359–363, 2001.

[24] M.-W. C. Jacob Devlin, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2019, https://arxiv.org/abs/1810.04805.

[25] D. Wang, Y. Zhang, and Y. Zhao, "Association for Computing Machinery," in *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, pp. 7–11, Newark, NJ, USA, 2017.

[26] Y. Oytam, F. Sobhanmanesh, K. Duesing, J. C. Bowden, M. Osmond-McLeod, and J. Ross, "Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets," *BMC Bioinformatics*, vol. 17, no. 1, p. 332, 2016.

[27] A. E. Woerner, M. P. Cox, and M. F. Hammer, "Recombination-filtered genomic datasets by information maximization," *Bioinformatics*, vol. 23, no. 14, pp. 1851–1853, 2007.

[28] S. Raschka and V. Mirjalili, *Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*, Packt Publishing Ltd, 2019.

[29] L. Li, H. Ruan, C. Liu et al., "Machine-learning reprogrammable metasurface imager," *Nature Communications*, vol. 10, pp. 1–8, 2019.

[30] X.-D. Zhang, *A Matrix Algebra Approach to Artificial Intelligence*, Springer, 2020.

[31] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[32] A. Messac and X. Chen, "EMBC 2019 Speakers," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2182–2185, Berlin, Germany, 2019.

[33] Z. Tian, A. Yen, Z. Zhou, C. Shen, K. Albuquerque, and B. Hrycushko, "A machine-learning–based prediction model of fistula formation after interstitial brachytherapy for locally advanced gynecological malignancies," *Brachytherapy*, vol. 18, no. 4, pp. 530–538, 2019.

[34] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh, and B. Gupta, "Deep learning for Arabic NLP: a survey," *Journal of Computational Science*, vol. 26, pp. 522–531, 2018.

[35] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.

[36] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 515–520, San Diego, California, 2016.

[37] C. M. Shiou Tian Hsu, P. Jones, and N. Samatova, "A hybrid CNN-RNN alignment model for phrase-aware sentence classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 443–449, Valencia, Spain, 2017.

[38] P. D. Stenson, M. Mort, E. V. Ball et al., "The human gene mutation database: 2008 update," *Genome Medicine*, vol. 1, pp. 1–6, 2009.

[39] P. D. Stenson, E. V. Ball, M. Mort et al., "Human gene mutation database (HGMD®): 2003 update," *Human Mutation*, vol. 21, no. 6, pp. 577–581, 2003.

[40] P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. D. Phillips, and D. N. Cooper, "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine," *Human Genetics*, vol. 133, no. 1, pp. 1–9, 2014.

[41] G. J. G. Prelich, "Gene overexpression: uses, mechanisms, and interpretation," *Genetics*, vol. 190, pp. 841–854, 2012.

[42] A. Gertych, Z. Swiderska-Chadaj, Z. Ma et al., "Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides," *Scientific Reports*, vol. 9, article 1483, 2019.

[43] (MSKCC), "M. S. K. C. C," http://www.mskcc.org/.

[44] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, CRC Press, 2010.

[45] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc., 2009.

[46] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, MIT press, 1999.

[47] A. Kao and S. R. Poteet, *Natural Language Processing and Text Mining*, Springer Science & Business Media, 2007.

[48] A. Moschitti and R. Basili, *European Conference on Information Retrieval*, Springer, 2007.

[49] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing Automated Text Classification Methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, 2019.

[50] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: a review," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3797–3816, 2019.

[51] H. Amazal, M. Ramdani, and M. Kissi, *International Conference on Smart Applications and Data Analysis*, Springer, 2020.

[52] W. Chen, X. Liu, D. Guo, and M. Lu, *International Conference on Data Mining and Big Data*, Springer, 2018.

[53] O. Einea, A. Elnagar, and R. Al Debsi, "Sanad: single-label Arabic news articles dataset for automatic text categorization," *Data in Brief*, vol. 25, article 104076, 2019.

[54] T. Joachims, *European Conference on Machine Learning*, Springer, 2005.

[55] J. Chen, H. Huang, S. Tian, and Y. Qu, “Feature selection for text classification with naïve Bayes,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[56] N. Suguna and K. Thanushkodi, “An improved k-nearest neighbor classification using genetic algorithm,” *International Journal of Computer Science Issues*, vol. 7, pp. 18–21, 2010.

[57] T. Pranckevičius and V. Marcinkevičius, “Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification,” *Baltic Journal of Modern Computing*, vol. 5, p. 221, 2017.

[58] S. Bloehdorn and A. Hotho, *International Workshop on Knowledge Discovery on the Web*, Springer, 2004.

[59] X. S. Zhang, D. Chen, Y. Zhu et al., “A multi-view ensemble classification model for clinically actionable genetic mutations,” 2018, https://arxiv.org/abs/1806.09737.

[60] E. Al Daoud, “Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset,” *International Journal of Computer and Information Engineering*, vol. 13, pp. 6–10, 2019.

[61] B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, “Machine learning for detecting gene-gene interactions,” *Applied Bioinformatics*, vol. 5, pp. 77–88, 2006.

[62] T. J. Cleophas, A. H. Zwinderman, and H. I. Cleophas-Allers, *Machine learning in medicine*, Springer, 2013.

[63] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: concerns and ways forward,” *PLoS One*, vol. 13, article e0194889, 2018.

[64] J. Schmidhuber, “Deep learning in neural networks: an overview,” *Neural Networks*, vol. 61, pp. 85–117, 2015.

[65] D. W. Otter, J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2020.

[66] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, *International Conference on Computational Science*, pp. 84–95, Springer.

[67] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?,” in *China National Conference on Chinese Computational Linguistics*, Springer, 2019.

[68] G. Giannakopoulos, P. Mavridi, G. Paliouras, G. Papadakis, and K. Tserpes, “Article 13 (Association for Computing Machinery),” in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, Craiova, Romania, 2012.

Hindawi
BioMed Research International
Volume 2020, Article ID 6287545, 10 pages
https://doi.org/10.1155/2020/6287545

Research Article

Machine Learning-Based Differentiation of Nontuberculous Mycobacteria Lung Disease and Pulmonary Tuberculosis Using CT Images

Zhiheng Xing,[1,2] Wenlong Ding,[2] Shuo Zhang,[2] Lingshan Zhong,[2] Li Wang,[2] Jigang Wang,[2] Kai Wang,[2] Yi Xie,[2] Xinqian Zhao,[2] Nan Li,[2] and Zhaoxiang Ye[1]

[1] *Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin, Tianjin's Clinical Research Center for Cancer, Tianjin, China*
[2] *Haihe Hospital, Tianjin University, Tianjin Institute of Respiratory Diseases, Tianjin, China*

Correspondence should be addressed to Zhaoxiang Ye; yezhaoxiang@163.com

Received 6 August 2020; Revised 11 September 2020; Accepted 22 September 2020; Published 30 September 2020

Academic Editor: Zhiguo Zhou

An increasing number of patients infected with nontuberculous mycobacteria (NTM) are observed worldwide. However, it is challenging to identify NTM lung diseases from pulmonary tuberculosis (PTB) due to considerable overlap in classic manifestations and clinical and radiographic characteristics. This study quantifies both cavitary and bronchiectasis regions in CT images and explores a machine learning approach for the differentiation of NTM lung diseases and PTB. It involves 116 patients and 103 quantitative features. After the selection of informative features, a linear support vector machine performs disease classification, and simultaneously, discriminative features are recognized. Experimental results indicate that bronchiectasis is relatively more informative, and two features are figured out due to promising prediction performance (area under the curve, 0.84 ± 0.06; accuracy, 0.85 ± 0.06; sensitivity, 0.88 ± 0.07; and specificity, 0.80 ± 0.12). This study provides insight into machine learning-based identification of NTM lung diseases from PTB, and more importantly, it makes early and quick diagnosis of NTM lung diseases possible that can facilitate lung disease management and treatment planning.

1. Introduction

Nontuberculous mycobacteria (NTM) is a major cause of morbidity and mortality in progressive lung diseases; unfortunately, an increasing number of patients with NTM lung disease (NTM-LD) are witnessed worldwide [1, 2]. As the etiologic agents, NTM have been found in a variety of environmental sources, and the clinical relevance of NTM-LD indicates the geographical heterogeneity in distribution and pathogenicity [3, 4]. Due to similar manifestations, it is difficult to recognize the lung infection caused by NTM or by pulmonary tuberculosis (PTB) for early diagnosis [5–9]. In clinic, as the first choice, microscopic examination of sputum smear for acid-fast bacillus (AFB) is used to screen mycobacterial lung infections; however, the presence of pulmonary mycobacterial infection could also be traced by AFB-positive [10–13]. Besides elaborate safety precautions, a definite diagnosis of NTM based on bacterial culture and strain identification lasts for about two months each time [6, 14]. Once being suspected of PTB with positive sputum AFB, a patient will take empirical anti-TB medicine for treatment when the test is ongoing to identify the bacteria. That means a part of patients receive potentially unnecessary treatment. It might cause the patients the risk of drug adverse reaction and thus nonessential healthcare cost [14]. Therefore, early diagnosis of NTM-LD can improve patients' life quality and facilitate disease treatment, and in particular, it benefits developing countries with resource-poor healthcare systems [1–3].

One challenging task is to differentiate NTM-LD from PTB lung disease (PTB-LD). Clinical manifestations are first considered, such as chronic cough, sputum production, and appetite loss. Moreover, clinical and radiographic characteristics are investigated, such as age, history of smoking, and previous TB treatment, since these characteristics are more frequently found in patients with NTM-LD than those with PTB-LD. However, considerable overlaps exist in classic manifestations, clinical characteristics, and radiographic features, making the diagnosis subjective and unstable [7–10, 14–19]. According to the radiographic features of cavities and bronchiectasis, NTM-LD can be generally classified into two distinct subtypes. One is characterized by cavities with areas of increased opacity and usually located in the upper lobes, and the other is by bronchiectasis and bronchiolar nodules which are predominant in the middle lobe and/or lingual. In comparison to PTB-LD patients with cavities or bronchiectasis, CT findings indicate that radiographic changes of NTM-LD could lead to subtle differences, such as thin-walled cavities and less bronchogenic but more contiguous spread of disease [14, 16, 17]. However, these observed differences are qualitative or subtle, which are not sufficient or discriminative to differ the NTM-LD from PTB-LD patients.

Some studies have explored machine learning methods for PTB screening. An artificial neural network (ANN) was used for the prediction of PTB infection [20]. The study examined blood samples of 115 PTB-LD patients and 60 normal subjects. Based on 39 features, the accuracy of two-hidden-layered ANN was up to 93.93%. An approach incorporating a fuzzy logic controller and an artificial immune recognition system was proposed [21] which utilized 20 features to represent each of 175 data samples and resulted in high accuracy, sensitivity, and specificity. A convolutional neural network (CNN) was designed for PTB examination [22]. The network enabled an end-to-end training from images to labels and required no objective-specific manual feature engineering. Its classification performance was larger than 0.85 (AUC (area under the curve)) on three real data sets [22]. Transferred learning, deep network, data augmentation, and radiologist involvement were considered, and high performance of PTB diagnosis was achieved [23]. These machine learning approaches are advancing the techniques for PTB-LD diagnosis [24].

The present study explores to build a machine learning model for the differentiation of NTM-LD and PTB-LD by using CT images. To the best of our knowledge, there are no machine learning models available to this challenging task. The contribution of this study is manifold. First, a machine learning approach is designed. It involves 116 patients, and to each patient case, 103 quantitative features are analyzed. Second, the effectiveness of different regions (cavities, bronchiectasis, and their combination) is investigated. Third, experimental results indicate that bronchiectasis is more informative, and two discriminative features are figured out. In addition, a simple and interpretable machine learning model is built which achieves promising classification performance. This study provides insight into machine learning-based differentiation of NTM-LD and PTB-LD patients, and most importantly, it provides some feasible clues on the early and quick diagnosis of lung diseases, benefiting disease management and treatment planning.

2. Material and Methods

2.1. Data Collection. From January 2019 to January 2020, a total of 1291 AFB smear-positive sputum specimens of previously untreated cases were retrospectively retrieved in Tianjin Haihe Hospital, Tianjin University, China. The sputum test is required to be conducted at least twice to show varying degrees of AFB smear positive. After being cultured and strain-identified, the smear-positive sputum was tested. The test result verified that 287 specimens were NTM, and 1004 were PTB. Details of PTB and NTM diagnosis are as follows. In order to find the mycobacteria in a tissue section, an AFB stain is done for all sputum samples. Based on PCR assays, a TB polymerase chain reaction (PCR) was performed with in-house IS6110. Mycobacterium culture was carried out using Löwenstein-Jensen Medium. Specifically, PTB diagnosis was in accordance with mycobacteria culture results and guidelines from the Chinese Medical Association, and NTM was based on mycobacterial culture results and guidelines of the American Thoracic Society (ATS) [25].

The chosen patients were with reliable CT imaging data, and CT scan images were reviewed independently by three experienced radiologists (XZH, WL, and ZS) who were blind to patients' microbiology results. With regard to the chest CT findings, the final decisions were determined by consensus. As shown in Figure 1, after an independent review of CT images, 116 cases (57 M. tuberculosis and 59 NTM) with lung cavities and/or with bronchiectasis were identified for retrospective analysis.

In addition, clinical characteristics of patients in both groups are shown in Table 1. It indicates that most patients show similar symptoms, including cough, sputum production, and fever. It is also found that some patients are smokers and some are with diabetes mellitus. Most importantly, no significant difference in symptoms is found between the two groups of patients.

2.2. CT Image Acquisition. All chest CT examinations were performed within 3 months of the AFB smear test by using a helical CT scanner (Aquilion Prime 128, Canon Medical Systems, Otawara, Japan). Patients were scanned from the lung apices to the adrenal glands during full inspiration, and the procedure was repeated during full expiration. The CT scanning parameters were as follows: 64 × 0.5 mm collimation, 120 kV automatic tube current modulation, and 0.5 s gantry rotation time. Contiguous inspiratory CT images were obtained with a thickness of 5.0 mm, at 5.0 mm intervals. Images were exported in DICOM format and forwarded to observers. In addition, CT scans were interpreted at window settings that were optimal for lung parenchyma (reconstruction kernel, FC 52; window level, -600 HU; window width, 1500 HU) and soft tissue (reconstruction kernel, FC 30; window level, 400 HU; window width, 40 HU).

2.3. Label Annotation. Both cavitary and bronchiectasis are labeled by using the software 3D Slicer (version 3.10.2, http://www.slicer.org/). Seven radiologists participated in this task. To ensure the accuracy, six radiologists (1 to 3 years' experience) were trained in a trial-and-error manner.

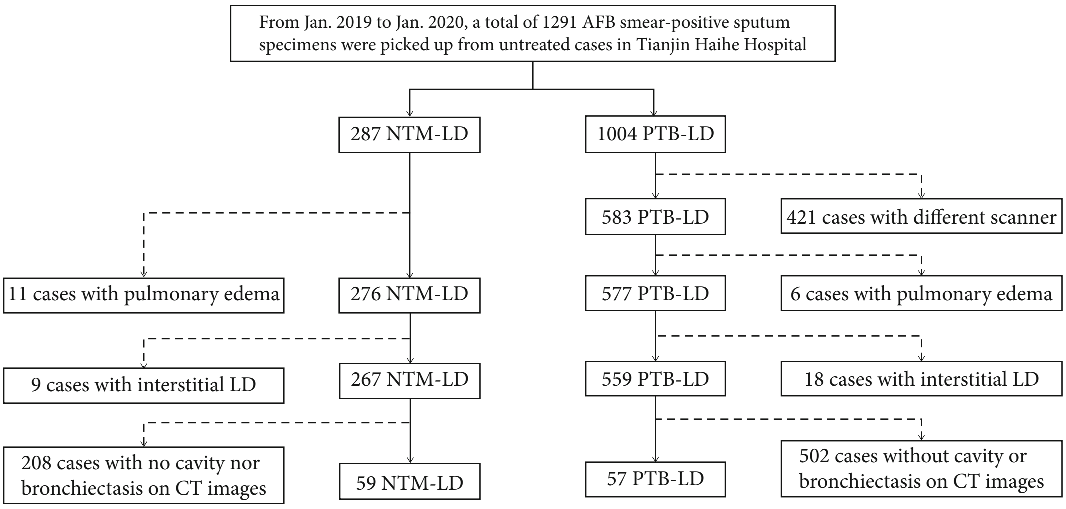

From Jan. 2019 to Jan. 2020, a total of 1291 AFB smear-positive sputum specimens were picked up from untreated cases in Tianjin Haihe Hospital
287 NTM-LD
1004 PTB-LD
583 PTB-LD
421 cases with different scanner
11 cases with pulmonary edema
276 NTM-LD
577 PTB-LD
6 cases with pulmonary edema
9 cases with interstitial LD
267 NTM-LD
559 PTB-LD
18 cases with interstitial LD
208 cases with no cavity nor bronchiectasis on CT images
59 NTM-LD
57 PTB-LD
502 cases without cavity or bronchiectasis on CT images

Figure 1: The procedure of data collection. After review of CT images, 116 cases remain for analysis.

Table 1: Clinical characteristics of patients.

	NTM-LD ($n = 59$)	PTB-LD ($n = 57$)	Chi-squared test	p value
Cough	27 (45.76%)	36 (63.16%)	3.535	0.060
Sputum production	25 (42.37%)	31 (54.39%)	1.676	0.196
Fever	17 (28.81%)	20 (35.09%)	0.525	0.469
Chest pain	3 (5.08%)	8 (14.04%)	2.706	0.100
Hemoptysis	7 (11.86%)	7 (12.28%)	0.005	0.945
Fatigue	4 (6.78%)	1 (1.75%)	0.766	0.382
Emaciation	4 (6.78%)	2 (3.51%)	0.141	0.707
Shortness of breath	1 (1.69%)	4 (7.02%)	0.910	0.340
Smoker	15 (25.42%)	14 (24.56%)	0.011	0.915
Diabetes mellitus	9 (15.25%)	8 (14.04%)	0.034	0.853
COPD	5 (8.47%)	5 (8.77%)	0.000	1.000

COPD stands for chronic obstructive pulmonary disease; $p < 0.05$ indicates significant difference.

Furthermore, to ensure the consistency, after training and case annotation, a senior radiologist with 10 years' experience performed the label verification without clinical information. Meanwhile, the senior radiologist performed as a supervisor and summarized the errors and cautions in label annotation and further gave the junior radiologists a second chance to rectify their errors. As shown in Figure 2, the whole procedure involves 2-round training, 2-round case labeling, 2-round modification, 2-round summarization, and 3-round verification until the labels can be used for the follow-up analysis.

Figure 3 shows representative examples of cavity (red) and bronchiectasis (yellow) from NTM-LD and PTB-LD patients. In CT images, both cavity and bronchiectasis are well-defined [26]. A cavity is a gas-filled space which is seen as a lucency or low-attenuation area, within pulmonary consolidation, a mass, or a nodule, and notably, no content is in a cavity. A thin-walled purification cavity is with a basically uniform wall thickness less than 3 mm and a thick-walled purification cavity is with a substantially uniform wall thickness greater than or equal to 3 mm, while a wall-less cavity is a gas density stove with no walls and smooth inner edges and located in the consolidated lung tissue. In addition, cavitary is a cavity that can be clearly imaged on the basis of consolidation. Whether a thick or thin wall, it is always marked as a cavity, and the outer wall of the lesion edge is the boundary mark. Morphological criteria of bronchiectasis consider bronchial dilatation with respect to accompanying pulmonary artery (signet ring sign), lack of tapering of bronchi, and identification of bronchi within 1 cm of the pleural surface. There are three types of labeling for bronchiectasis: (1) saccular: the inner diameter of the bronchus greater than 1.5 times the diameter of the accompanying artery. (2) Columnar: dilated bronchi with the same proximal and distal ends of the bronchi, longer than 2 cm. (3) Varicose veins: dilated bronchus with an uneven wall and tortuous course. The inner wall was marked as the boundary.

2.4. Feature Extraction. The open-source package Pyradiomics (https://pyradiomics.readthedocs.io) was used in this study, and 103 features were extracted regarding annotated bronchiectasis and cavity in original-resolution CT images.

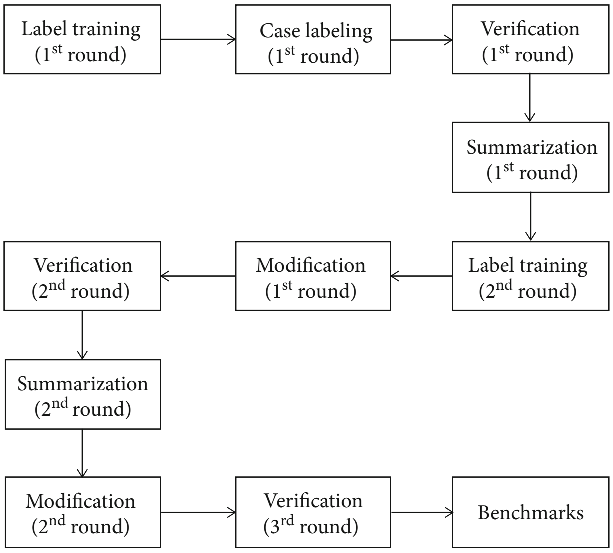


Figure 2: The procedure of cavitary and bronchiectasis annotation. Seven radiologists participated in this task. Six radiologists were trained in a trial-and-error manner (training, labeling, and modification), and one senior radiologist helped the verification, summarization, and training of the six radiologists.

The features consist of 14 shape features, 21 first-order features, 22 Gray-Level Cooccurrence Matrix (GLCM) features, 16 Gray-Level Run Length Matrix (GLRLM) features, 16 Gray-Level Size Zone Matrix (GLSZM) features, and 14 Gray-Level Differential Matrix (GLDM) features. These features have been widely used for data representation and disease diagnosis [27, 28].

2.5. A Machine Learning Approach. A simple and interpretable machine learning approach is desirable. Given the data, to simplify the retrieval of informative features, Gini importance is used to measure the feature importance, since it defines dependence and independence of variables [29]. Further, to reduce the computation burden, several important features are considered in the follow-up analysis. Due to limited patient cases, to retrieve a few discriminative features is reasonable. At last, for good interpretability, linear SVM [30] performs the differentiation of the NTM-LD and the PTB-LD patients.

Figure 4 shows the flow chart which attempts to build a machine learning approach for interpretable diagnosis. The dashed lines indicate offline feature ranking. Features are sorted in terms of Gini importance. Assuming k features are extracted from each data sample, a resultant vector $<f_1, f_2, \cdots, f_k>$ stands for the indexes of the most to the least important features (1). Then, i top most important features are kept (2), and all combinations of feature subsets using 2 or 3 features are provided (3).

Potential feature subsets are prepared, and the optimal one is selected by comparing classification performance as shown in solid lines in Figure 4. For instance, if a subset of features is selected, the patient cases were randomly grouped into the training and the testing set (4). Using the training set, the parameters of the linear SVM classifier are experimentally determined (5). Once the model is trained, the testing set is fed into the model (6), and the performance is evaluated with classification metrics (7).

2.6. Experiment Design. Four experiments are conducted, and three are shown in Table 2. For each experiment, the number of patient cases, sex, and ages are reported. The first (TA), the second (TB), and the third (TC), respectively, use the cavity, the bronchiectasis, and both for retrieving the most discriminative features in an automated fashion. It should be noted that the fourth experiment is used to verify the effectiveness of the combination of retrieved features from TA and TB for disease classification.

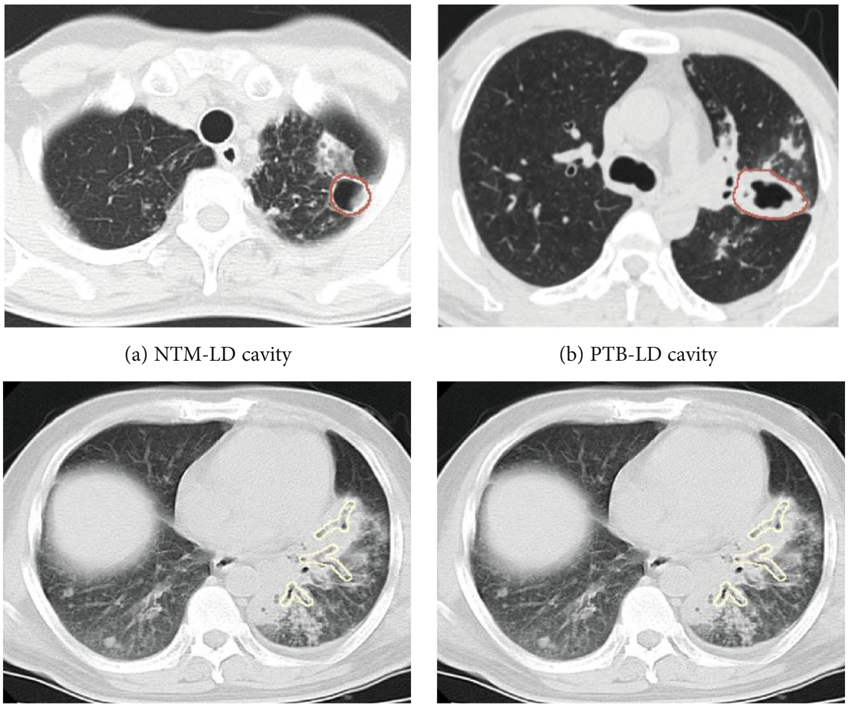
With regard to each experiment, a total of 100 times of data splitting are conducted at random, and nearly 80% of cases are portioned into the training set and the rest into the testing set. After each time of data splitting, all feature subsets are used one by one for machine learning-based disease classification.

2.7. Performance Evaluation and Statistical Analysis. Four metrics are used to evaluate the classification performance, and they are the area under the curve (AUC), accuracy (ACC), sensitivity (SEN), and specificity (SPE). To figure out the best performance, i.e., the subset with the most discriminative features, statistical analyses were conducted using SPSS 17.0 software for Windows (SPSS Inc., Chicago, IL, USA), and performance metrics were compared by a paired t-test.

3. Results

3.1. Gini Importance-Based Feature Importance Ranking. Table 3 lists the top 10 most important features with regard to different forms used for lung disease analysis. The indexes of features that are derived from intensity statistics, shape representation, and texture analysis are, respectively, highlighted in italic, bold, and underline. Analysis of the cavitary form identifies 6 intensity statistics features and 4 texture analysis features, and analysis of the bronchiectatic form figures out 4 shape representation features and 6 texture analysis features, while analysis of the combined form indicates that all features are from the bronchiectatic form (feature indexes larger than 103), including one intensity statistics feature, three shape representation features, and six texture analysis features.

3.2. Cavity-Based Lung Disease Differentiation. Based on the cavity analysis and automated retrieval of discriminative features, three subsets achieving superior performance are listed in Table 4. It shows that the subset using the 22nd and the 99th features (in bold) obtains the best or competitive result in terms of four metrics, while no significant difference is found (p value > 0.23). The 30th feature is also recognized as important; however, no improvement is observed in disease classification. As to the discriminative features, one (the 22nd) quantifies the intensity distribution, and the other (the 99th) shows the texture analysis of the cavity.



(a) NTM-LD cavity

(b) PTB-LD cavity

(c) NTM-LD bronchiectasis

(d) PTB-LD bronchiectasis

Figure 3: Representative examples of annotated cavity and bronchiectasis. Thick-walled, thin-walled, and wall-less cavities are marked as a cavity, and the outer wall of the lesion edge is the boundary mark, while bronchiectasis annotation should concern bronchial dilatation with respect to different factors.

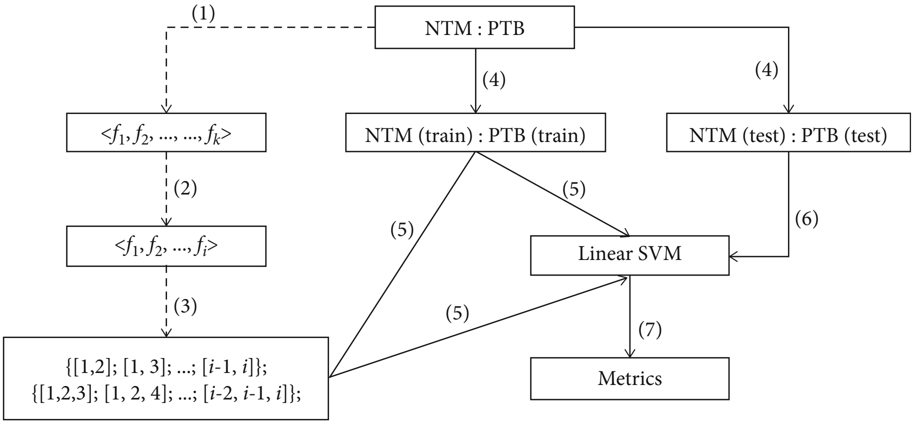

(1)
NTM : PTB
(4)
(4)
<$f_1, f_2, ..., ..., f_k$>
NTM (train) : PTB (train)
NTM (test) : PTB (test)
(2)
(5)
(5)
(6)
<$f_1, f_2, ..., f_i$>
Linear SVM
(3)
(5)
(7)
{[1,2]; [1, 3]; ...; [i-1, i]};
{[1,2,3]; [1, 2, 4]; ...; [i-2, i-1, i]};
Metrics

Figure 4: The framework for machine learning-based differentiation of NTM-LD and PTB-LD patients. The dashed lines indicate offline processing, and the solid ones stand for the retrieval of discriminative features for accurate disease diagnosis.

Table 2: The number of patient cases, sex, and age in experiment design.

	NTM (male/female/age)	PTB (male/female/age)
TA	44 (28/16/60 ± 15)	54 (40/14/48 ± 18)
TB	45 (28/17/62 ± 15)	54 (41/13/49 ± 17)
TC (TA∩TB)	32 (21/11/64 ± 12)	46 (34/12/49 ± 18)

3.3. Bronchiectasis-Based Lung Disease Differentiation. Table 5 shows three subsets of features that lead to superior performance with regard to analyzing bronchiectasis. It suggests that the subset consisting of the 13th and the 87th features results in the best performance in terms of AUC and SPE, and the competitive performance in terms of ACC and SEN. It is worth noting that there is no significant difference of each performance metric between any two feature subsets (p value > 0.37). Moreover, the 48th and the 6th features are identified for their importance in disease differentiation, and adding one of them causes no enhancement. In the subset of discriminative features, one (the 13th) aims for shape representation, and the other (the 87th) analyzes tissue textures.

Table 3: Ten most important features via Gini importance-based feature ranking.

Form	Ranked index of features from the most to less important ones									
Cavitary form	*2*	*23*	80	*35*	95	60	99	*22*	*30*	*25*
Bronchiectatic form	**13**	49	58	94	87	**7**	48	**11**	67	**6**
Combined form	*123*	190	**116**	152	**109**	161	197	170	**114**	151

Table 4: Cavity-based LD differentiation.

Feature subsets	AUC	ACC	SEN	SPE
[99, 22]	0.70 ± 0.07	0.71 ± 0.06	0.72 ± 0.09	0.68 ± 0.14
[99, 30]	0.70 ± 0.08	0.70 ± 0.08	0.70 ± 0.10	0.66 ± 0.15
[22, 99, 30]	0.69 ± 0.07	0.70 ± 0.07	0.72 ± 0.09	0.68 ± 0.11

#The 22nd feature, original_firstorder_interquartilerange; the 30th feature, original_firstorder_robustmeanabsolutedeviation; the 99th feature, original_gldm_largedependencelowgraylevelemphasis.

Table 5: Bronchiectatic form-based differentiation of lung diseases.

Feature subsets	AUC	ACC	SEN	SPE
[13, 87]	0.84 ± 0.06	0.85 ± 0.06	0.88 ± 0.07	0.80 ± 0.12
[13, 87, 48]	0.82 ± 0.07	0.84 ± 0.07	0.89 ± 0.09	0.74 ± 0.13
[13, 87, 6]	0.83 ± 0.07	0.85 ± 0.07	0.89 ± 0.09	0.76 ± 0.10

#The 6th feature, original_shape_leastaxislength; the 13th feature, original_shape_minoraxislength; the 48th feature, original_glcm_Imc1; the 87th feature, original_glszm_zoneentropy.

Table 6: Disease differentiation using both the cavity and the bronchiectasis.

Feature subsets	AUC	ACC	SEN	SPE
[190, 152]	0.82 ± 0.08	0.78 ± 0.08	0.76 ± 0.11	0.88 ± 0.13
[190, 116, 152]	0.81 ± 0.10	0.75 ± 0.09	0.75 ± 0.06	0.89 ± 0.16
[190, 116, 151]	0.82 ± 0.10	0.77 ± 0.06	0.75 ± 0.06	0.86 ± 0.15

#The 116th feature, original_shape_minoraxislength; the 151st feature, original_glcm_Imc1; the 152nd feature, original_glcm_Imc2; the 190th feature, original_glszm_zoneentropy.

3.4. Combined Form for Lung Disease Differentiation. Based on both the cavity and the bronchiectasis, the subsets of features with good performance are presented in Table 6. The subset including the 190th and the 152nd features leads to the overall best performance in terms of three metrics (AUC, ACC, and SEN), and no significant difference is observed between the performance derived from each of the three subsets (p value > 0.52). Moreover, the 151st feature is figured out for its importance in disease classification, while again, no improvement is found. In addition, both discriminative features are from texture analysis.

3.5. Performance Comparison. Table 7 shows the performance of lung disease differentiation with regard to different regions (TA: cavity; TB: bronchiectasis; TC: combined analysis by using automated feature selection; TD: combined analysis by using retrieved features from TA and TB). It demonstrates that the subset of retrieved features from the bronchiectasis (TB) is the most discriminative in comparison to each of the other retrieved features. It also indicates that combining feature subsets (TD) does not improve the differentiation performance, and on the contrary, a slight decrease is observed from each metric. In particular, it is found that the subset of features retrieved from the cavity results in inferior performance with AUC 0.70 on average.

Error-bar plots in Figure 5 show the performance of lung disease differentiation by analyzing different regions. In general, using bronchiectasis (TB) achieves the highest AUC, ACC, and SEN and the second best SPE; using combined subsets of features (TD) obtains comparative performance, while using the cavity (TA) produces the worst performance in lung disease differentiation.

ROC curves are shown in Figure 6. Different colors correspond to different methods. The bronchiectasis (TB, red) results in the best performance (AUC 0.86), followed by both regions with combined features (TD, green) with AUC 0.82 and both regions using automated feature selection (TC, blue) with AUC 0.81, and the worst is the cavitary form (TA, pink) with AUC 0.73.

4. Discussion

The increasing prevalence of NTM-LD is observed worldwide. Bacterial culture and strain identification remain the unique way to identify NTM, while the procedure takes a long time. Early and quick diagnosis of NTM-LD is urgently important yet challenging. Massive studies investigate the manifestations, clinical characteristics, radiographic findings, and clinical relevance. However, due to considerable overlap of symptoms and subtle difference in CT images, these findings are not sufficient to differentiate NTM-LD from PTB-LD patient cases. This study is the first work that explores machine learning to identify the NTM-LD patients from the PTB-LD ones, and in CT images, both the cavity and the bronchiectasis regions are delineated for quantitative analysis. Experimental results suggest that the proposed machine learning model achieves promising performance when two features are used to represent the bronchiectasis.

Quantified bronchiectasis plays an important role in the machine learning model for the differentiation between NTM-LD and PTB-LD cases. It enables high performance (AUC, 0.84 ± 0.06; ACC, 0.85 ± 0.06; SEN, 0.88 ± 0.07; and SPE, 0.80 ± 0.12) which is obviously higher than those corresponding metrics from the quantified cavity (AUC, 0.70 ± 0.07; ACC, 0.71 ± 0.06; SEN, 0.72 ± 0.09; and SPE, 0.68 ± 0.14). Its performance is slightly superior or competitive to that using both cavity and nodular bronchiectasis. Predominance of cavities and bronchiectasis is observed in

Table 7: LD differentiation using selected features with regard to different regions.

	Retrieved features	AUC	ACC	SEN	SPE
TA	[99, 22]	0.70 ± 0.07	0.71 ± 0.06	0.72 ± 0.09	0.68 ± 0.14
TB	**[13, 87]**	0.84 ± 0.06	0.85 ± 0.06	0.88 ± 0.07	0.80 ± 0.12
TC	[190, 152]	0.82 ± 0.08	0.78 ± 0.08	0.76 ± 0.11	0.88 ± 0.13
TD	[99, 22]+[13, 87]	0.81 ± 0.09	0.83 ± 0.07	0.85 ± 0.08	0.78 ± 0.18

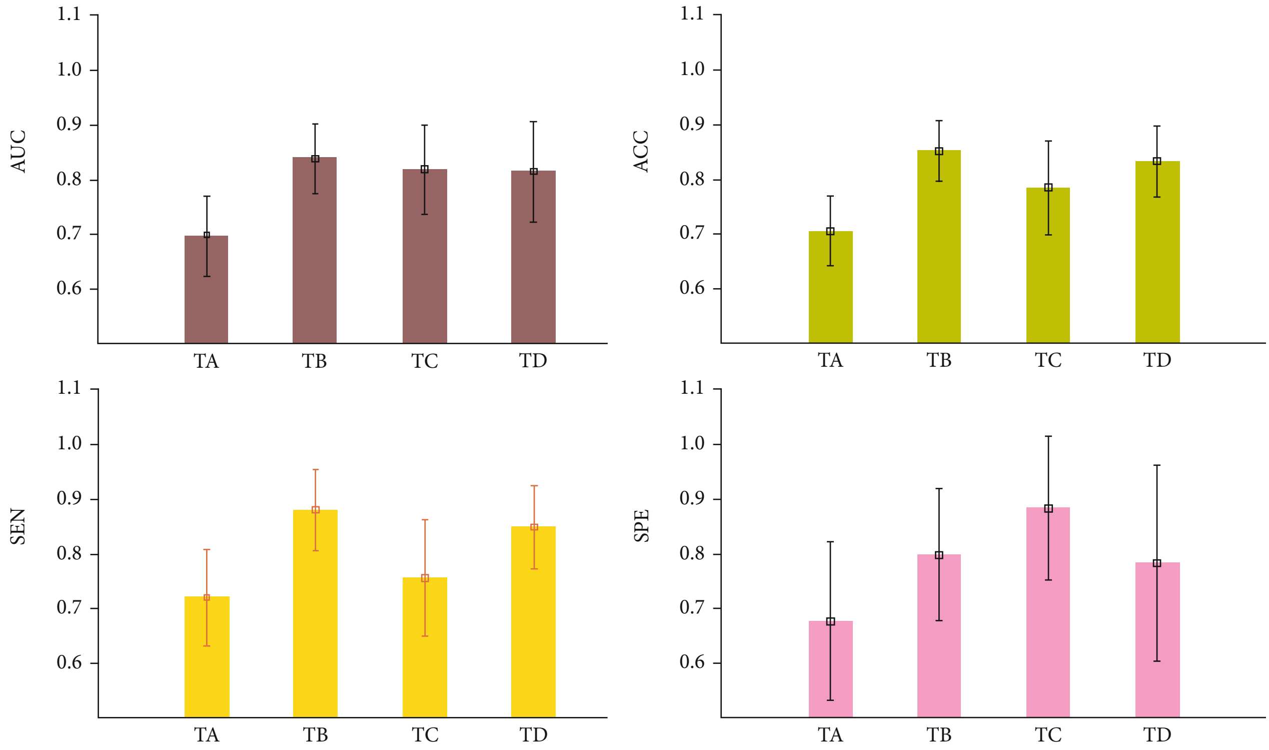


Figure 5: The performance of disease differentiation via analyzing different regions (TA, cavity; TB, bronchiectasis; TC, combined analysis using automated feature selection; TD, combined analysis using retrieved features from TA and TB). It shows that using bronchiectasis (TB) achieves overall best performance.

radiographic findings of NTM-LD cases. One study indicated that of the 19 patients evaluated, 84.2% cases were with bronchiectasis, and 73.7% were with cavities [31]. One study with 34 patients figured out that nodular lesions (100%) and bronchiectasis (85.29%) were the most frequent CT features of Mycobacterium simiae pulmonary infection [32]. A meta-analysis study reported that 9.3% of NTM-LD patients were with bronchiectasis [33]. A comparison of CT findings between NTM-LD and PTB-LD has also been considered. A study analyzed 95 CT scans from 159 patients with AFB smear-positive sputum (75 scans from PTB-LD patients and 20 scans from NTM-LD patients) and claimed that the presence of bronchiectasis changes in CT scans was strongly associated with patients with NTM-LD [16]. A study investigated a total of 4167 untreated cases with AFB smear-positive sputum (124 cases were with NTM-LD, and 210 cases with PTB-LD were randomly selected from the remaining cases), and bronchiectasis and thin-walled cavity were identified independent predictors for NTM-LD diagnosis via multivariate analysis [14]. A cavity analysis study (128 NTM-LD and 128 PTB-LD patients with matched age and gender) discovered that the major cavities in NTM disease generally have thinner and more even walls than those in PTB cases [17]. Thus, to investigate cavity and bronchiectasis in CT images for lung disease differentiation is reasonable. Most importantly, the current study points out that the quantified bronchiectasis seems more informative than the cavity in differing the NTM-LD from PTB-LD cases.

The machine learning model is well built, and it is simple and interpretable. It makes use of two quantitative features for the representation of bronchiectasis in CT images. In the original images, one feature describes the minor (second-largest) axis length of shape, and the other is the zone entropy of GLSZM texture which describes the randomness in the distribution of zone sizes and gray levels. Interestingly, both features have been reported in related clinical studies. For instance, the minor axis length of shape is important in the detection of clinically significant prostate cancer in multiparametric MR images [34], and the zone entropy of GLSZM reflects the areas with different gray intensities within the

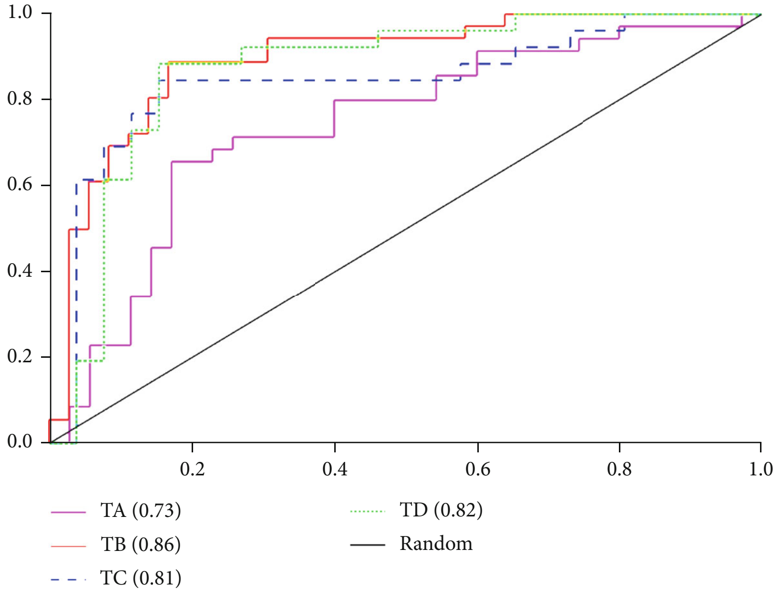


Figure 6: ROC curves of disease differentiation via the analysis of different regions.

nodules for lung cancer detection [35]. However, it should be noted that both features cannot be perceived directly, and thus, accurate segmentation of the bronchiectasis regions becomes indispensable. Moreover, the model utilizes an interpretable classifier of linear SVM, which is widely used in knowledge discovery. It is worth noting that SVM with a nonlinear kernel could map data samples into high-dimension space, and the classification performance might be further improved. In addition, this simple model supports good generalization and evolving, and it can avoid the curse of dimensionality in high-throughput feature analysis.

There are several limitations to the current study. First, the number of patient cases should be increased, and a multi-institution study would be better, as it can make the results more convincing, generalizable, and applicable. Therefore, our future work will focus on data collection and multicenter collaboration. Second, advanced techniques [23, 24, 27, 28] could be used to improve the diagnosis performance, and the hybrid techniques [36–38] that integrate manifestations and clinical and radiographic features are feasible. Third, automated annotation and quantification of bronchiectasis and cavity are also appealing. For instance, the thickness of cavity walls is helpful, since cavity walls of NTM-LD patients are found significantly thinner and more even than those of PTB-LD [17]. However, it requires advanced algorithms for accurate and objective quantification. In the end, this study involves a single hospital and a limited number of cases. For further verification of our findings, a large-scale experiment should be conducted.

5. Conclusion

The increasing incidence and prevalence of NTM-LD have become a major public health problem. This study explores a machine learning approach, and both bronchiectasis and cavity are delineated for differing NTM-LD patients from PTB-LD patients. Bronchiectasis is found more informative, and two quantitative features are identified discriminative for disease differentiation. The built machine learning model makes early and quick diagnosis of NTM-LD possible, and it could further facilitate disease management and treatment planning and improve patients' life quality.

Data Availability

The CT images supporting the findings of this study are restricted by the Medical Ethics Committee of Haihe Hospital in order to protect patient privacy. If interested, requests for access to the extracted features can be made to the corresponding author Zhaoxiang Ye (yezhaoxiang@163.com).

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work is funded by the Tianjin Municipal Health Science and Technology Project From Tianjin Municipal Health and Health Committee (2020xkm03).

References

[1] Y. Tan, B. Su, W. Shu et al., "Epidemiology of pulmonary disease due to nontuberculous mycobacteria in Southern China, 2013-2016," *BMC pulmonary medicine*, vol. 18, no. 1, pp. 1–7, 2018.

[2] A. K. Maurya, V. L. Nag, S. Kant et al., "Prevalence of nontuberculous mycobacteria among extrapulmonary tuberculosis cases in tertiary care centers in Northern India," *BioMed Research International*, vol. 2015, 5 pages, 2015.

[3] C. Okoi, S. T. Anderson, M. Antonio, S. N. Mulwa, F. Gehre, and I. M. Adetifa, "Non-tuberculous mycobacteria isolated

from pulmonary samples in sub-Saharan Africa-a systematic review and meta analyses," *Scientific Reports*, vol. 7, no. 1, pp. 1-2, 2017.

[4] H. F. Schiff, S. Jones, A. Achaiah, A. Pereira, G. Stait, and B. Green, "Clinical relevance of non-tuberculous mycobacteria isolated from respiratory specimens: seven year experience in a UK hospital," *Scientific Reports*, vol. 9, no. 1, pp. 1–6, 2019.

[5] H. J. Jun, K. Jeon, S. W. Um, O. J. Kwon, N. Y. Lee, and W. J. Koh, "Nontuberculous mycobacteria isolated during the treatment of pulmonary tuberculosis," *Respiratory Medicine*, vol. 103, no. 12, pp. 1936–1940, 2009.

[6] N. W. Schluger, "Tuberculosis and nontuberculous mycobacterial infections in older adults," *Clinics in chest medicine*, vol. 28, no. 4, pp. 773–781, 2007.

[7] R. Gopalaswamy, S. Shanmugam, R. Mondal, and S. Subbian, "Of tuberculosis and non-tuberculous mycobacterial infections–a comparative analysis of epidemiology, diagnosis and treatment," *Journal of Biomedical Science*, vol. 27, no. 1, p. 74, 2020.

[8] B. A. Kendall, C. D. Varley, D. Choi et al., "Distinguishing tuberculosis from nontuberculous mycobacteria lung disease, Oregon, USA," *Emerging infectious diseases*, vol. 17, no. 3, pp. 506–509, 2011.

[9] Y. K. Kim, S. Hahn, Y. Uh et al., "Comparable characteristics of tuberculous and non-tuberculous mycobacterial cavitary lung diseases," *The International journal of tuberculosis and lung disease*, vol. 18, no. 6, pp. 725–729, 2014.

[10] I. Abubakar, R. K. Gupta, M. X. Rangaka, and M. Lipman, "Update in tuberculosis and nontuberculous mycobacteria 2017," *American Journal of Respiratory and Critical Care Medicine*, vol. 197, no. 10, pp. 1248–1253, 2018.

[11] P. J. McShane and J. Glassroth, "Pulmonary disease due to nontuberculous mycobacteria: current state and new insights," *Chest*, vol. 148, no. 6, pp. 1517–1527, 2015.

[12] J. E. Stout, W. J. Koh, and W. W. Yew, "Update on pulmonary disease due to non-tuberculous mycobacteria," *International Journal of Infectious Diseases*, vol. 45, pp. 123–134, 2016.

[13] N. Wassilew, H. Hoffmann, C. Andrejak, and C. Lange, "Pulmonary disease caused by non-tuberculous mycobacteria," *Respiration*, vol. 91, no. 5, pp. 386–402, 2016.

[14] H. Q. Chu, B. Li, L. Zhao et al., "Chest imaging comparison between non-tuberculous and tuberculosis mycobacteria in sputum acid fast bacilli smear-positive patients," *European Review for Medical and Pharmacological Sciences*, vol. 19, no. 13, pp. 2429–2439, 2015.

[15] T. R. Aksamit, J. V. Philley, and D. E. Griffith, "Nontuberculous mycobacterial (NTM) lung disease: the top ten essentials," *Respiratory Medicine*, vol. 108, no. 3, pp. 417–425, 2014.

[16] M. K. Yuan, C. Y. Chang, P. H. Tsai, Y. M. Lee, J. W. Huang, and S. C. Chang, "Comparative chest computed tomography findings of non-tuberculous mycobacterial lung diseases and pulmonary tuberculosis in patients with acid fast bacilli smear-positive sputum," *BMC pulmonary medicine*, vol. 14, no. 1, 2014.

[17] C. Kim, S. H. Park, S. Y. Oh et al., "Comparison of chest CT findings in nontuberculous mycobacterial diseases vs. Mycobacterium tuberculosis lung disease in HIV-negative patients with cavities," *PLoS One*, vol. 12, no. 3, article e0174240, 2017.

[18] Y. S. Kwon and W. J. Koh, "Diagnosis of pulmonary tuberculosis and nontuberculous mycobacterial lung disease in Korea," *Tuberculosis and respiratory diseases*, vol. 77, no. 1, pp. 1–5, 2014.

[19] M. J. Nasiri, H. Dabiri, A. A. Fooladi, S. Amini, G. Hamzehloo, and M. M. Feizabadi, "High rates of nontuberculous mycobacteria isolation from patients with presumptive tuberculosis in Iran," *New microbes and new infections*, vol. 21, pp. 12–17, 2018.

[20] O. Er, F. Temurtas, and A. Ç. Tanrıkulu, "Tuberculosis disease diagnosis using artificial neural networks," *Journal of medical systems*, vol. 34, no. 3, pp. 299–302, 2010.

[21] S. Shamshirband, S. Hessam, H. Javidnia et al., "Tuberculosis disease diagnosis using artificial immune recognition system," *International journal of medical sciences*, vol. 11, no. 5, pp. 508–514, 2014.

[22] S. Hwang, H. E. Kim, J. Jeong, and H. J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," *Medical imaging 2016: computer-aided diagnosis*, vol. 9785, p. 97852W, 2016.

[23] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[24] P. Dande and P. Samant, "Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review," *Tuberculosis*, vol. 108, pp. 1–9, 2018.

[25] C. L. Daley, J. M. Iaccarino, C. Lange et al., "Treatment of nontuberculous mycobacterial pulmonary disease: an official ATS/ERS/ESCMID/IDSA clinical practice guideline," *European Respiratory Journal*, vol. 56, no. 1, p. 2000535, 2020.

[26] D. M. Hansell, A. A. Bankier, H. Mac Mahon, T. C. McLoud, N. L. Muller, and J. Remy, "Fleischner Society: glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[27] J. J. Van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[28] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[29] S. Zhang, X. Dang, D. Nguyen, D. Wilkins, and Y. Chen, "Estimating feature-label dependence using Gini distance statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, 2019.

[30] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[31] R. Mogami, T. Goldenberg, P. G. Marca, F. C. Mello, and A. J. Lopes, "Pulmonary infection caused by Mycobacterium kansasii: findings on computed tomography of the chest," *Radiologia brasileira*, vol. 49, no. 4, pp. 209–213, 2016.

[32] A. Baghizadeh, P. Mehrian, and P. Farnia, "Computed tomography findings of PulmonaryMycobacterium simiaeInfection," *Canadian respiratory journal*, vol. 2017, 5 pages, 2017.

[33] H. Chu, L. Zhao, H. Xiao et al., "Prevalence of nontuberculous mycobacteria in patients with bronchiectasis: a meta-analysis," *Archives of medical science: AMS*, vol. 10, no. 4, pp. 661–668, 2014.

[34] R. Cuocolo, A. Stanzione, A. Ponsiglione et al., "Clinically significant prostate cancer detection on MRI: a radiomic shape features study," *European journal of radiology*, vol. 116, pp. 144–149, 2019.

[35] S. Chauvie, A. De Maggi, I. Baralis et al., “Artificial intelligence and radiomics enhance the positive predictive value of digital chest tomosynthesis for lung cancer detection within SOS clinical trial,” *European Radiology*, vol. 30, no. 7, pp. 4134–4140, 2020.

[36] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, pp. 1–15, Springer, Berlin, Heidelberg, 2000.

[37] Z. Zhou, M. Folkert, P. Iyengar et al., “Multi-objective radiomics model for predicting distant failure in lung SBRT,” *Physics in Medicine & Biology*, vol. 62, no. 11, pp. 4460–4478, 2017.

[38] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” *Advances in neural information processing systems*, pp. 2546–2554, 2011.

Hindawi
BioMed Research International
Volume 2020, Article ID 9258649, 9 pages
https://doi.org/10.1155/2020/9258649

Research Article

Automatic Prediction of MGMT Status in Glioblastoma via Deep Learning-Based MR Image Analysis

Xin Chen,[1] Min Zeng,[1] Yichen Tong,[2] Tianjing Zhang,[3] Yan Fu,[4] Haixia Li,[2] Zhongping Zhang,[3] Zixuan Cheng,[1] Xiangdong Xu,[1] Ruimeng Yang,[1] Zaiyi Liu,[5] Xinhua Wei,[1] and Xinqing Jiang[1]

[1] *Department of Radiology, Guangzhou First People's Hospital, Guangzhou Medical University, Guangzhou 510180, China*
[2] *Sun Yat-sen University, Guangzhou, China*
[3] *Philips Healthcare, Guangzhou, China*
[4] *EPFL, Lausanne, Switzerland*
[5] *Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China*

Correspondence should be addressed to Zaiyi Liu; zyliu@163.com, Xinhua Wei; weixinhua@aliyun.com, and Xinqing Jiang; gzcmmcjxq@163.com

Received 16 June 2020; Revised 27 August 2020; Accepted 3 September 2020; Published 23 September 2020

Academic Editor: Zhiguo Zhou

Methylation of the O^6-methylguanine methyltransferase (MGMT) gene promoter is correlated with the effectiveness of the current standard of care in glioblastoma patients. In this study, a deep learning pipeline is designed for automatic prediction of MGMT status in 87 glioblastoma patients with contrast-enhanced T1W images and 66 with fluid-attenuated inversion recovery(FLAIR) images. The end-to-end pipeline completes both tumor segmentation and status classification. The better tumor segmentation performance comes from FLAIR images (Dice score, 0.897 ± 0.007) compared to contrast-enhanced T1WI (Dice score, 0.828 ± 0.108), and the better status prediction is also from the FLAIR images (accuracy, 0.827 ± 0.056; recall, 0.852 ± 0.080; precision, 0.821 ± 0.022; and F_1 score, 0.836 ± 0.072). This proposed pipeline not only saves the time in tumor annotation and avoids interrater variability in glioma segmentation but also achieves good prediction of MGMT methylation status. It would help find molecular biomarkers from routine medical images and further facilitate treatment planning.

1. Introduction

Glioblastoma multiforme (GBM) is the most common and aggressive type of primary brain tumor in adults. It accounts for 45% of primary central nervous system tumors, and the 5-year survival rate is around 5.1% [1, 2]. The standard treatment for GBM is surgical resection followed by radiation therapy and temozolomide (TMZ) chemotherapy, which improves median survival by 3 months compared to radiotherapy alone [3]. Several studies indicated that O^6-methylguanine-DNA methyltransferase (MGMT) gene promoter methylation reported in 30-60% of glioblastomas [4] can enhance the response to TMZ, which has been proven to be a prognostic biomarker in GBM patients [3, 5]. Thus, determination of MGMT promoter methylation status is important to medical decision-making.

Genetic analysis based on surgical specimens is the reference standard to assess the MGMT methylation status, while a large tissue sample is required for testing MGMT methylation status using methylation-specific polymerase chain reaction [6]. In particular, the major limitations are the possibility of incomplete biopsy samples due to tumor spatial heterogeneity and high cost [7]. Besides, it cannot be used for real-time monitoring of the methylation status.

Magnetic resonance imaging (MRI) is a standard conventional examination in diagnosis, preoperative planning, and therapy evaluation of GBM [8, 9]. Recently, radiomics, extracting massive quantitative features from medical

Table 1: Dataset distribution of each experiment.

	Phase	Cases (methylation/unmethylation)	CE-T1WI slices (methylation/unmethylation)	FLAIR slices (methylation/unmethylation)
FLAIR	Training	51 (25/26)	676 (288/388)	—
	Testing	15 (7/8)	167 (62/105)	
CE-T1WI	Training	70 (36/34)	—	1208 (609/599)
	Testing	17 (10/7)		220 (109/111)

Note: FLAIR: fluid-attenuated inversion recovery; CE-T1WI; contrast-enhanced T1-weighted imaging.

images, has been proposed to explore the correlation between image features and underlying genetic traits [10–12]. There is growing evidence that radiomics can be used in predicting the status of MGMT promoter methylation [13–15]. However, most previous works utilized handcrafted features. This procedure includes tumor segmentation, feature extraction, and informatics analysis [16–19]. In particular, tumor segmentation is a challenging and important step because most works depend on manual delineation. This step is burdensome and time consuming, and inter- or intraobserver disagreement is unavoidable. Deep learning which can extract features automatically has been emerging as an innovative technology in many fields [20]. The convolutional neural network (CNN) is proven to be effective in image segmentation, disease diagnosis, and other medical image analysis tasks [21–25]. Compared to traditional methods with handcrafted features, deep learning shows several advantages of being robust to distortions such as changes in shape and lower computational cost. A few studies have shown that deep learning can be used to segment tumors and predict MGMT methylation status for glioma [26]. However, to the best of our knowledge, there is no previous report regarding building a pipeline for both glioma tumor segmentation and MGMT methylation status prediction in an end-to-end manner. Therefore, we investigate the feasibility of integrating the tumor segmentation and status prediction of GBM patients into a deep learning pipeline in this study.

2. Methods

2.1. Data Collection. A total of 106 GBM patients were analyzed in our study. MR images, including presurgical axial contrast-enhanced T1-weighted images (CE-T1WI) and T2-weighted fluid-attenuated inversion recovery (FLAIR) images, were collected from The Cancer Imaging Archive (http://www.cancerimagingarchive.net). The images were originated from four centers (Henry Ford Hospital, University of California San Francisco, Anderson Cancer Center, and Emory University). Clinical and molecular data were also obtained from the open-access data tier of the TCGA website.

Genomic data were from the TCGA data portal. MGMT methylation status analysis was performed on Illumina HumanMethylation27 and HumanMethylation450 BeadChip platforms. A median cutoff using the level 3 beta-value present in the TCGA was utilized for categorizing methylation status. Illumina Human Methylation probes (cg12434587 and cg12981137) were selected in this study [27].

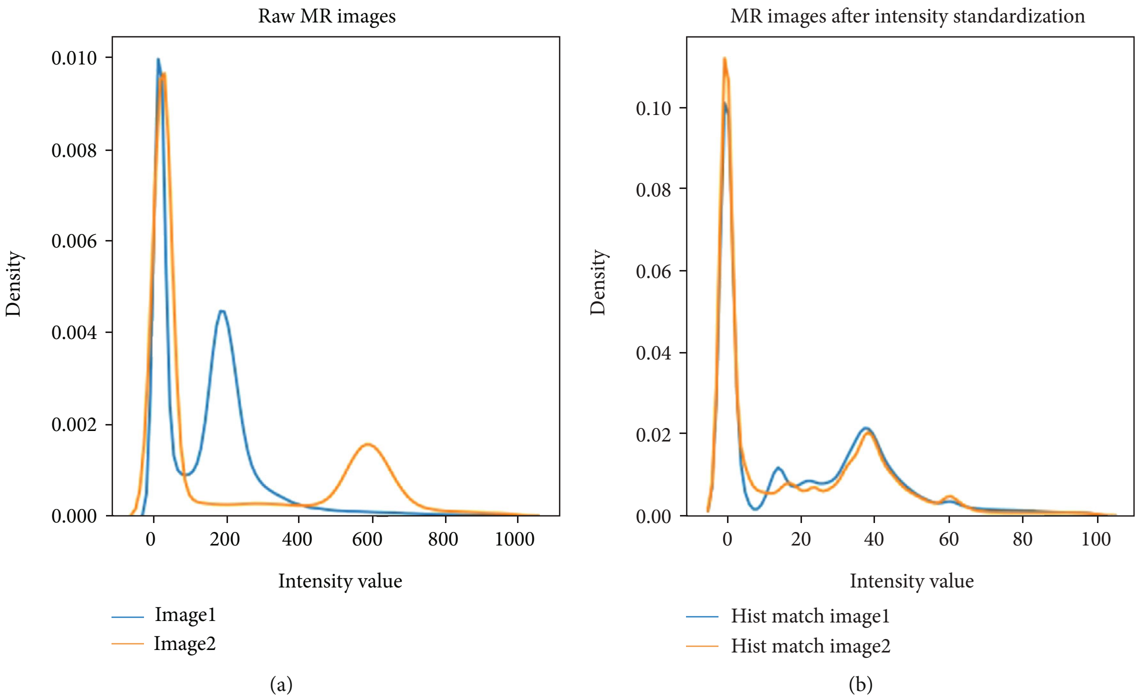
Of 106 GBM cases, 87 cases were with CE-T1W images, and 66 cases with FLAIR images. We randomly split the cases into training and testing sets with the ratio of 8 : 2 and applied 10-fold cross-validation to the training set with scikit-learn library (https://scikit-learn.org/stable/). The dataset distribution is listed in Table 1.

2.2. Image Preprocessing. For general images, the pixel values contain reliable image information. However, MR images do not have a standard intensity scale. In Figure 1(a), we show the density plot of two raw MR images. In each plot, there are two peaks, the peak around 0 refers to background pixels, and the other peak refers to white matter. The white matter peaks of the two images are far away. Thus, MR images normalization is needed to guarantee that the grey values of the same tissue among different MR images are close to each other [28].

The piece-wise linear histogram matching was used to normalize the intensity distribution of MR images [29]. Firstly, we studied standard histogram distribution via averaging the 1st to 99th percentile of all images. Then, we linearly mapped the intensities of each image to this standard histogram. In Figure 1(b), we can see that the white matter peaks of two images coincide with each other after normalization. Secondly, the images were normalized to zero mean and unit standard deviation only on valued voxels. At last, data augmentation was used to increase the dataset size to avoid overfitting. We rotated images for every 5 degrees from -20 to +20 degrees, resulting in a 9-fold increment in the number of MRI scans.

2.3. Segmentation. As for tumor segmentation, one state-of-the-art model [30] in BraTS 2018 challenge (Multimodal Brain Tumor Segmentation 2018 Challenge http://braintumorsegmentation.org/) was adapted. The whole network architecture is shown in Figure 2.

In short, the deep learning model added a variational autoencoder (VAE) branch to a fully convolutional network model. The decoder part was shared for both segmentation and VAE tasks. The prior distribution taken for the KL divergence in the VAE part is $N(0, 1)$. ResNet blocks used in the architecture [31] included two 3×3 convolutions with normalization and ReLU as well as skip connections. In the encoder part, the image dimension was downsampled using stride convolution by 2 and increased channel size by 2. For the decoder part, the structure was similar to that of the



(a) (b)

Figure 1: Density plot of two different MR images (a) before and (b) after piece-wise linear histogram matching.

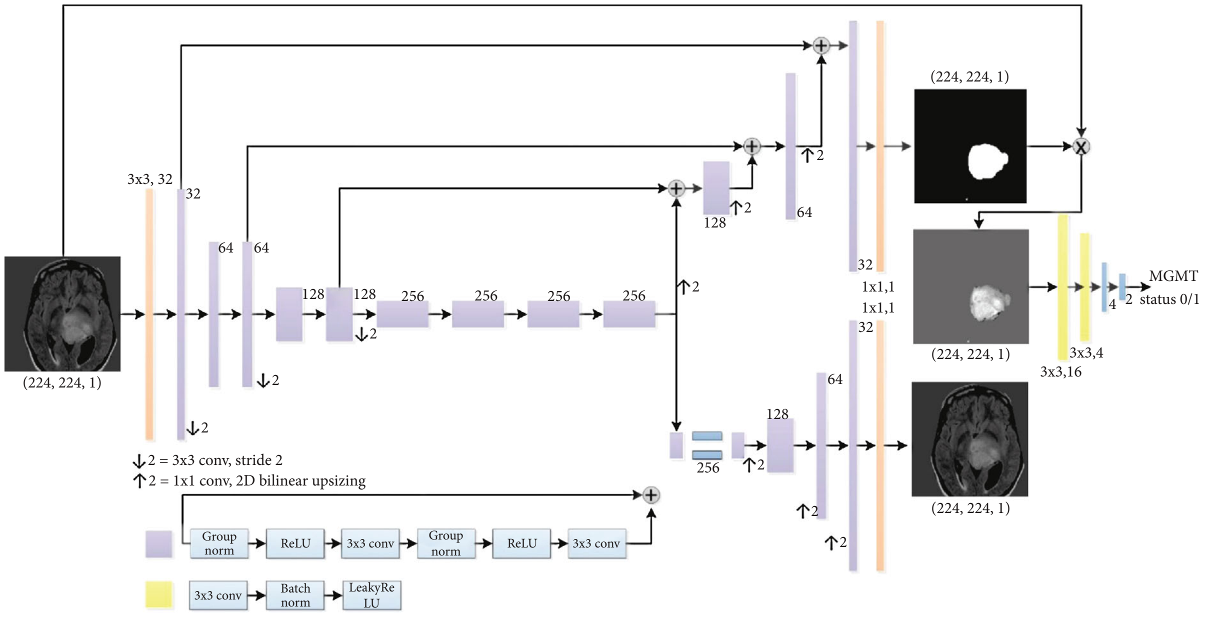


Figure 2: An end-to-end deep learning pipeline for both tumor segmentation and status classification.

encoder part but using upsampled. The decoder endpoint had the same size as the input image followed by sigmoid activation, and its output was for tumor segmentation. As for the VAE part, the encoder output was reduced to 256, and the input image was reconstructed by using a similar structure as the decoder without skip connection. The segmentation part output the tumor segmentation and the VAE branch attempted to reconstruct the input image. Except for the input and output layers, all blocks in Figure 2 utilized the ResNet block with different channel numbers (depicted aside each layer). For the input layer, a 3×3 convolution was with 3 channels; and for both output layers, a 3×3 convolution with a dropout rate of 0.2 and L_2 regularization with weight $1e-3$ were used to avoid overfitting. The loss function consists of 3 terms as shown in

$$L = L_{\text{Dice}} + 0.1 \times L_{L2} + 0.1 \times L_{KL}, \quad (1)$$

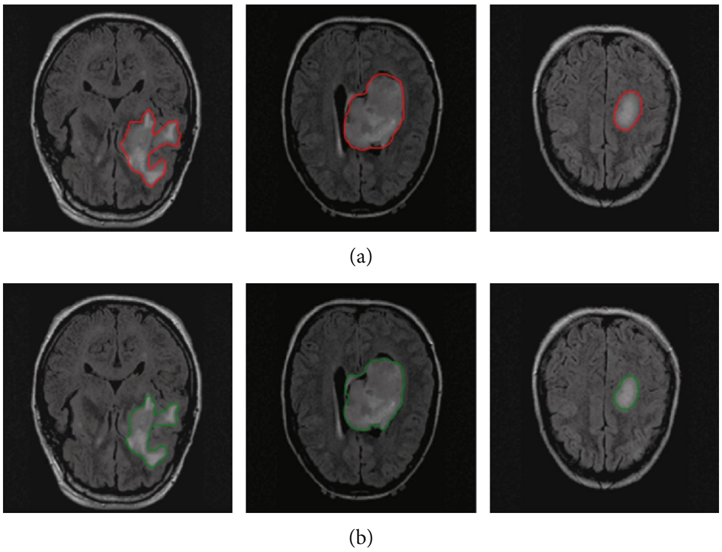


Figure 3: Automatic segmentation results of brain tumors with FLAIR images. (a) The ground truth of tumor boundaries in FLAIR images and (b) automatic segmentation results using the proposed network with FLAIR images.

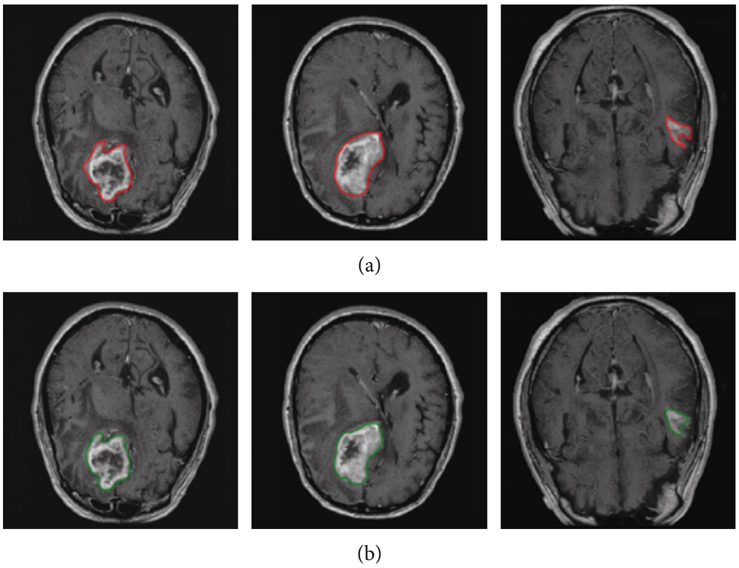


Figure 4: Three representative cases of brain tumor manual annotation and automatic segmentation with CE-T1WI images. (a) The manual annotation and (b) the automatic segmentation results with our proposed network.

where L_{Dice} is the soft Dice loss between the predicted segmentation and the ground truth labels. The ground truth labels were manually annotated with ImageJ (https://imagej.nih.gov) by one neuroradiologist with 10 years' experience specialized in brain disease diagnosis. L_{L2} is the L_2 loss on the VAE branch output image and the input image, and L_{KL} is the standard VAE penalty term [32, 33]. Then, the Dice coefficient as defined in function (2) was calculated to assess the performance of segmentation:

$$\text{Dice} = \sum_i \frac{2 \cdot p_i \cdot \widehat{p_i}}{\|p_i\|^2 + \|p_i \wedge\|^2 + \text{epsilon}}, \tag{2}$$

where p_i is the ground truth, $\widehat{p}_i$ is the prediction for pixel i, and epsilon = $1e - 8$.

2.4. Status Classification. Meanwhile, for the classification of MGMT methylation status, a 4-layer CNN was designed. Further, the classification model was cascaded with the tumor segmentation model. At the stage of the tumor segmentation model design, the classification network was tried with different numbers of convolutional layers [2–5], and we found that 2 convolutional layers with 2 fully connected (FC) layers performed the best for this task. The first convolutional layer had 16 filters, and the second one had 4 filters. All the convolutional layers had a kernel size of 3×3 and stride of

Table 2: Dice scores of the deep network on tumor segmentation using MR images.

Modality	Training	Validation	Testing
CE-T1WI	0.832 ± 0.009	0.831 ± 0.012	0.828 ± 0.108
FLAIR	0.893 ± 0.004	0.892 ± 0.008	0.897 ± 0.007

Note: the number in the table referred to the mean ± standard deviation values of 10 cross-validation experiments. CE-T1WI: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery.

Table 3: Inference time (seconds) of one MR slice for glioma segmentation.

Modality	Manual annotation	Deep model
CE-T1WI	50 s	0.11 s
FLAIR	60 s	0.07 s

Note: CE-T1WI: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery.

1 followed by LeakyReLU, batch normalization, and max pooling. LeakyReLU was an advanced ReLU activation that avoids dead neurons by setting a negative half-axis slope 0.3 instead of 0. Its advantages include good performance in eliminating gradient saturation, low computational cost, and faster convergence. Batch normalization was used to normalize features by the mean and variance within a small batch. It helped to solve the covariance shift issue and ease optimization. Max pooling with a 4 × 4 filter was used to downsample image features extracted through convolutional layers and then fed into 2 FC layers. ReLU and softmax were adapted as activation functions for the first and second FC layers, respectively. The weight initialization of all layers was done by He-normal [34].

2.5. Parameter Settings and Software. All experiments were conducted under the open-source framework Keras (https://keras.io/) on one GeForce RTX 2080Ti GPU. The numbers of parameters of the segmentation and classification model are, respectively, 6,014,721 and 3,498. In tumor segmentation, Adam optimizer was adapted with a self-designed learning rate scheduler which was initialized with a learning rate $1e-4$; then, the learning rate was divided by 2 when the validation loss did not reduce in the past 5 epochs. The epoch was set at 50 and batch size at 8. Every epoch took around 50 seconds. In tumor classification, 4-CNN was trained for 50 epochs which utilized Adam with learning rate $2e-4$, and the batch size was 32. If the validation accuracy was observed stable for over 10 epochs, the training process would be ended. The averaged elapsed time for each epoch was 5 seconds.

2.6. Statistical Analysis. The Dice coefficient was calculated for evaluating the performance of tumor segmentation. For the MGMT methylation status classification, the accuracy rate, recall, precision, and F_1 score were calculated according to equations listed below. In addition, the receiver operating characteristic (ROC) curve was plotted, and the area under the ROC curve (AUC) was reported to measure the classification accuracy. All the parameters were calculated in PyCharm with the programming language of Python (version 3.6.8; Wilmington, DE, USA; http://www.python.org/):

$$\begin{aligned}\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},\\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}},\\ F_1 &= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}},\end{aligned} \tag{3}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.

3. Results

3.1. Tumor Segmentation

3.1.1. Qualitative Observation. Tumors could be accurately delineated by the proposed pipeline. Figure 3 shows the annotated ground truth (the first row) and corresponding segmentation results (the second row) of GBM in FLAIR images. It is observed that tumor boundaries could be accurately localized by using the deep learning network, and the major hyperintense regions are delineated. The three cases show that automatic segmentation is quite close to the ground truth.

Figure 4 shows the GBM in CE-T1WI images, and the ground truth (the first row) and the segmentation results (the second row) are presented. Tumor boundaries are localized, and it seems that there is no obvious difference between the manual annotation and its corresponding segmentation results obtained from our proposed network, and the suspicious regions are mainly contoured. The three cases show that segmentation results from the deep network approximate the manual delineation.

3.1.2. Quantitative Evaluation. The quantitative performance of automatic tumor segmentation is summarized in Table 2. The deep network obtained good testing performance on tumor segmentation using CE-T1WI (Dice score, 0.828 ± 0.108) and FLAIR (Dice score, 0.897 ± 0.007). And the Dice scores from FLAIR were slightly higher than those from CE-T1WI across training, validation, and testing sets. The maximum difference of the Dice score between average Dice scores from CE-T1W images in training and validation sets was 0.026, indicating that the model was not overfitting.

3.1.3. Computational Performance. Time consumption between manual annotation and automatic prediction per MR slice is compared as shown in Table 3. For the evaluation of time consumption, we recorded the total time and divided it by the number of slices. So, the time listed in Table 3 was the average segmentation time per slice. It was observed that the deep network was more efficient, and it took less than 0.2 seconds to complete the segmentation of an MR slice, while manual annotation required more than 30 seconds.

Table 4: Results of MGMT methylation status classification.

Modality	Phase	Classification			
		Accuracy	Recall	Precision	F_1 score
CE-T1WI	Training	0.894 ± 0.012	0.906 ± 0.007	0.886 ± 0.018	0.896 ± 0.010
	Validation	0.839 ± 0.046	0.866 ± 0.044	0.823 ± 0.051	0.845 ± 0.045
	Testing	0.804 ± 0.011	0.818 ± 0.033	0.798 ± 0.014	0.808 ± 0.015
FLAIR	Training	0.941 ± 0.056	0.943 ± 0.104	0.947 ± 0.026	0.945 ± 0.081
	Validation	0.885 ± 0.090	0.941 ± 0.105	0.857 ± 0.028	0.889 ± 0.101
	Testing	0.827 ± 0.056	0.852 ± 0.080	0.821 ± 0.022	0.836 ± 0.072

Note: the number in the table referred to the mean ± standard deviation values of 10 cross-validation experiments. CE-T1WI: contrast-enhanced T1-weighted imaging; FLAIR: fluid-attenuated inversion recovery.

3.2. Classification of MGMT Promoter Methylation Status. Table 4 shows the prediction performance of MGMT promoter methylation status which is evaluated from four classification metrics (accuracy, recall, precision, and F_1 score) on three stages (training, validation, and testing) when using different MR images (CE-T1WI, FLAIR). In general, the model trained with FLAIR achieves better results for all metrics across three stages, followed by the model trained with CE-T1WI images. Specifically, the accuracy, recall, precision, and F_1 score of the deep model trained with FLAIR images reach 0.827, 0.852, 0.821, and 0.836 in the testing stage, respectively.

ROC curves of the prediction results are demonstrated in Figures 5 and 6. Figure 5 shows the best status classification when using FLAIR images for a deep model, which achieves an AUC of 0.985 (yellow curve), 0.968 (green curve), and 0.905 (red curve) on the training, validation, and testing datasets, respectively.

The best status classification when using CE-T1WI images for deep model training is shown in Figure 6. The well-trained deep model obtains AUC up to 0.973 (yellow curve), 0.942 (green curve), and 0.887 (red curve) on the training, validation, and testing datasets, respectively.

4. Discussion

This study presents an MR-based deep learning pipeline for automatic tumor segmentation and MGMT methylation status classification in an end-to-end manner for GBM patients. Experimental results demonstrate promising performance on accurate glioma delineation (Dice score, 0.897) and MGMT status prediction (accuracy, 0.827; recall, 0.852; precision, 0.821; and F_1 score, 0.836) coming from the model trained with FLAIR images. In addition, the proposed pipeline dramatically shortens the inference time on glioma segmentation.

For glioma segmentation, one state-of-the-art deep model is utilized and obtains impressive performance on the involved MGMT dataset for GBM segmentation. Its performance is close to these deep network-based tumor segmentation studies. Hussain et al. [35] reported a CNN approach for glioma MRI segmentation, and the model achieved a Dice score of 0.87 on the BRATS 2013 and 2015 datasets. Cui et al. [36] proposed an automatic semantic segmentation model on the BRATS 2013 dataset, and the Dice score was near 0.80 on the combined high- and low-grade glioma datasets. Kaldera et al. [37] proposed a faster RCNN method and achieved a Dice score of 0.91 on 233 patients' data. These studies suggest that deep networks are full of potential for accurate tumor segmentation in MR images.

Several deep models have been designed for the classification of MGMT methylation status in GBM patients. Chang et al. [38] proposed a deep neural network which achieved a classification accuracy of 83% for 259 gliomas patients with T1W, T2W, and FLAIR images. Korfiatis et al. [26] compared different sizes of the ResNet baseline model and reached the highest accuracy of 94.9% in 155 GBM patients with T2W images. Han et al. [39] proposed a bidirectional convolutional recurrent neural network architecture for MGMT methylation classification, while the accuracy was around 62% for 262 GBM patients with T1W, T2W, and FLAIR images. In this study, a shallow CNN is used, and the classification performance is promising. The best performance comes from the model trained with FLAIR images, and we achieved a satisfactory result with the highest accuracy of 0.827 and recall of 0.852 in consideration of the relatively small dataset.

In the previous studies, Drabycz et al. [40] analyzed handcrafted features to distinguish methylated from unmethylated GBM and figured out that texture features from T2-weighted images were important for the prediction of MGMT methylation status. Han et al. [41] found that MGMT promoter-methylated GBM was prone to more tumor necrosis, while T2-weighted FLAIR sequence may be more sensitive to necrosis than T1-weighted images. Interestingly, we also find that better performances of both GBM segmentation and molecular classification are achieved on FLAIR images in our study although the images of CE-T1W and FLAIR did not come from the same patients.

The strengths of this study lie in the fully automatic glioma segmentation and predicting the MGMT methylation status based on a small dataset. Generally, it takes a radiologist about one minute per slice in tumor annotation, while the inference time of the deep learning model is about 0.1 seconds which is around 1/600 times used in manual annotation. Additionally, manual annotation is burdensome and prone to introduce inter- and intraobserver variability. While once well trained, a deep learning model can continuously

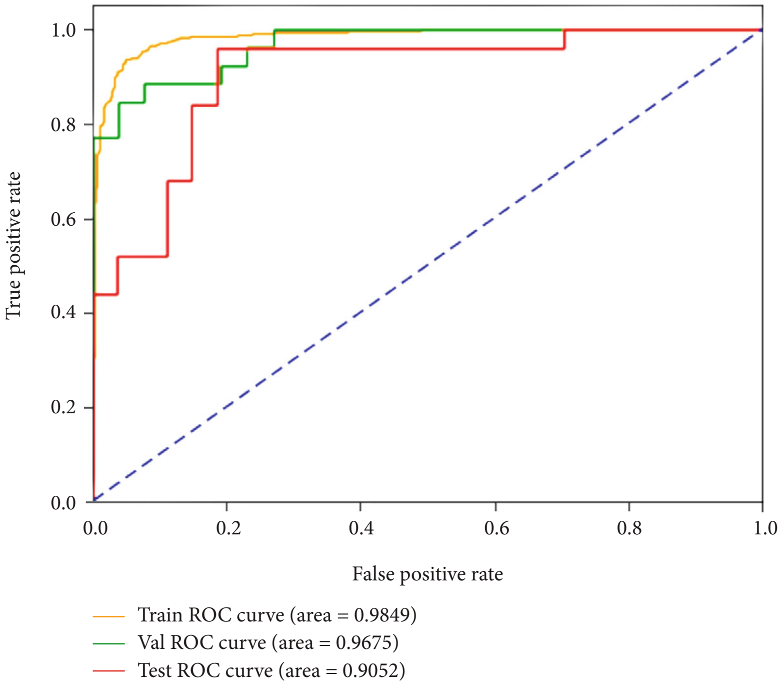


Figure 5: ROC curves of the best result on the FLAIR images for MGMT promoter methylation status classification on the training, validation, and testing datasets.

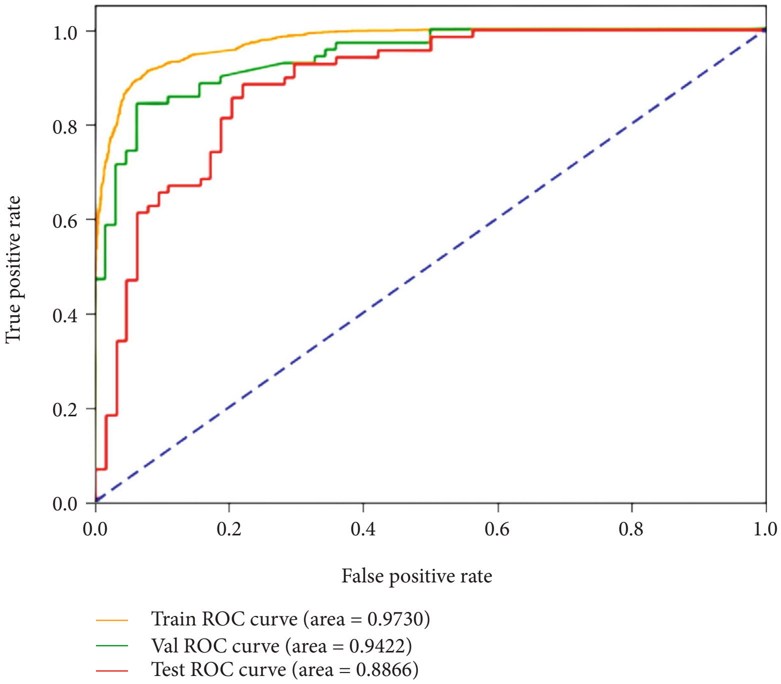


Figure 6: ROC curves of the best result on the CE-T1W images for MGMT promoter methylation status classification in the training, validation, and testing datasets.

and repeatedly perform tumor segmentation regardless of the observers. On the other hand, the training strategy in this study is beneficial for small dataset analysis. In general, a deep model requires a large number of training instances. However, it is challenging or impossible to provide massive high-quality images in medical imaging. Finally, although several studies tried to use deep networks for automatic glioma segmentation [35, 36, 42] or molecular classification [26, 38, 39], the proposed network in this study could integrate both glioma segmentation and classification in a seamless connection pipeline. And the performance is competitive to the state-of-the-art studies in tumor segmentation and classification.

There are several limitations to our study. First, the sample size is small in the study; we will further confirm the findings in a study with larger samples. Second, a multicenter research trial is helpful to validate the capability of the proposed pipeline, while the variations of MR imaging sequences, equipment venders, and other factors could impose difficulties on model building. Third, we failed to investigate the value of combined CE-T1WI and FLAIR in tumor segmentation and classification considering the fewer samples. In the future, we will explore multiple MR sequences for MGMT methylation status prediction, such as amide-proton-transfer-weighted imaging and diffusion-weighted imaging. These may have great potential to improve the performance of MGMT methylation status prediction.

5. Conclusion

An MRI-based end-to-end deep learning pipeline is designed for tumor segmentation and MGMT methylation status prediction in GBM patients. It can save time and avoid interobserver variability in tumor segmentation and help discover molecular biomarkers from routine medical images to aid in diagnosis and treatment decision-making.

Data Availability

All MRI data are available in the cancer imaging archive (https://www.cancerimagingarchive.net/), and clinical and molecular data are obtained from the open-access data tier of the TCGA website.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Authors' Contributions

Xin Chen, Min Zeng, and Yichen Tong contributed equally.

Acknowledgments

This study was supported by the Key R&D Program of Guangdong Province (2018B030339001), the National Science Fund for Distinguished Young Scholars (No. 81925023), the National Natural Scientific Foundation of China (Nos. 81601469, 81771912, 81871846, 81971574, and 81802227), the Guangzhou Science and Technology Project of Health (No. 20191A011002), the Natural Science Foundation of Guangdong Province (2018A030313282), the Fundamental Research Funds for the Central Universities, SCUT (2018MS23), and the Guangzhou Science and Technology Project (202002030268 and 201804010032).

References

[1] L. L. Morgan, "The epidemiology of glioma in adults: a "state of the science" review," *Neuro-Oncology*, vol. 17, no. 4, pp. 623-624, 2015.

[2] Q. T. Ostrom, H. Gittleman, G. Truitt, A. Boscia, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2011-2015," *Neuro-oncology*, vol. 20, supplement_4, pp. iv1–iv86, 2018.

[3] M. E. Hegi, A. C. Diserens, T. Gorlia et al., "*MGMT* Gene Silencing and Benefit from Temozolomide in Glioblastoma," *The New England Journal of Medicine*, vol. 352, no. 10, pp. 997–1003, 2005.

[4] M. Weller, R. Stupp, G. Reifenberger et al., "*MGMT* promoter methylation in malignant gliomas: ready for personalized medicine?," *Nature Reviews Neurology*, vol. 6, no. 1, pp. 39–51, 2010.

[5] A. A. Brandes, E. Franceschi, A. Tosoni et al., "*MGMT* Promoter Methylation Status Can Predict theIncidence and Outcome of Pseudoprogression After Concomitant Radiochemotherapy in Newly Diagnosed Glioblastoma Patients," *Journal of Clinical Oncology*, vol. 26, no. 13, pp. 2192–2197, 2008.

[6] L. Wang, Z. Li, C. Liu et al., "Comparative assessment of three methods to analyze MGMT methylation status in a series of 350 gliomas and gangliogliomas," *Pathology-Research and Practice*, vol. 213, no. 12, pp. 1489–1493, 2017.

[7] N. R. Parker, A. L. Hudson, P. Khong et al., "Intratumoral heterogeneity identified at the epigenetic, genetic and transcriptional level in glioblastoma," *Scientific Reports*, vol. 6, no. 1, article 22477, 2016.

[8] P. Y. Wen, D. R. Macdonald, D. A. Reardon et al., "Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group," *Journal of clinical oncology*, vol. 28, no. 11, pp. 1963–1972, 2010.

[9] M. J. van den Bent, J. S. Wefel, D. Schiff et al., "Response assessment in neuro-oncology (a report of the RANO group): assessment of outcome in trials of diffuse low-grade gliomas," *The lancet oncology*, vol. 12, no. 6, pp. 583–593, 2011.

[10] L. Macyszyn, H. Akbari, J. M. Pisapia et al., "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques," *Neuro-Oncology*, vol. 18, no. 3, pp. 417–425, 2015.

[11] L. S. Hu, S. Ning, J. M. Eschbacher et al., "Radiogenomics to characterize regional genetic heterogeneity in glioblastoma," *Neuro-Oncology*, vol. 19, no. 1, pp. 128–137, 2017.

[12] E. K. Hong, S. H. Choi, D. J. Shin et al., "Radiogenomics correlation between MR imaging features and major genetic profiles in glioblastoma," *European radiology*, vol. 28, no. 10, pp. 4350–4361, 2018.

[13] V. G. Kanas, E. I. Zacharaki, G. A. Thomas, P. O. Zinn, V. Megalooikonomou, and R. R. Colen, "Learning MRI-based classification models for MGMT methylation status prediction in glioblastoma," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 249–257, 2017.

[14] Y. B. Xi, F. Guo, Z. L. Xu et al., "Radiomics signature: a potential biomarker for the prediction of MGMT promoter methylation in glioblastoma," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 5, pp. 1380–1387, 2018.

[15] F. Tixier, H. Um, D. Bermudez et al., "Preoperative MRI-radiomics features improve prediction of survival in

glioblastoma patients over MGMT methylation status alone," *Oncotarget*, vol. 10, no. 6, pp. 660–672, 2019.

[16] P. Kickingereder, U. Neuberger, D. Bonekamp et al., "Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma," *Neuro-Oncology*, vol. 20, no. 6, pp. 848–857, 2018.

[17] Z. Liu, X. Y. Zhang, Y. J. Shi et al., "Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer," *Clinical Cancer Research*, vol. 23, no. 23, pp. 7253–7262, 2017.

[18] P. Lambin, E. Rios-Velazquez, R. Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[19] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, 2014.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] P. Rajpurkar, J. Irvin, R. L. Ball et al., "Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, article e1002686, 2018.

[22] Y. Yang, L. F. Yan, X. Zhang et al., "Glioma grading on conventional MR images: a deep learning study with transfer learning," *Frontiers in Neuroscience*, vol. 12, 2018.

[23] A. Akay and H. Hess, "Deep learning: current and emerging applications in medicine and technology," *EEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 906–920, 2019.

[24] G. S. Tandel, M. Biswas, O. G. Kakde et al., "A review on a deep learning perspective in brain cancer classification," *Cancers*, vol. 11, no. 1, 2019.

[25] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, "A technical review of convolutional neural network-based mammographic breast cancer diagnosis," *Computational and mathematical methods in medicine*, vol. 2019, Article ID 6509357, 16 pages, 2019.

[26] P. Korfiatis, T. L. Kline, D. H. Lachance, I. F. Parney, J. C. Buckner, and B. J. Erickson, "Residual deep convolutional neural network predicts MGMT methylation status," *Journal of Digital Imaging*, vol. 30, no. 5, pp. 622–628, 2017.

[27] P. Bady, D. Sciuscio, A.-C. Diserens et al., "*MGMT* methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status," *Acta Neuropathologica*, vol. 124, no. 4, pp. 547–560, 2012.

[28] J. C. Reinhold, B. E. Dewey, A. Carass, and J. L. Prince, "Evaluating the impact of intensity normalization on MR image synthesis," in *SPIE Medical Imaging*, vol. 10949, San Diego, California, United States, 2019.

[29] M. Shah, Y. Xiao, N. Subbanna et al., "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis," *Medical Image Analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[30] A. Myronenko, *3D MRI Brain Tumor Segmentation Using Autoencoder Regularization*, International MICCAI Brainlesion Workshop: Springer, 2019.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[32] C. Doersch, "Tutorial on variational autoencoders," 2016, https://arxiv.org/abs/1606.05908.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, https://arxiv.org/abs/1312.6114.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, Santiago, Chile, 2015.

[35] S. Hussain, S. M. Anwar, and M. Majid, "Segmentation of glioma tumors in brain using deep convolutional neural network," *Neurocomputing*, vol. 282, pp. 248–261, 2018.

[36] S. Cui, L. Mao, J. Jiang, C. Liu, and S. Xiong, "Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network," *Journal of healthcare engineering*, vol. 2018, Article ID 4940593, 14 pages, 2018.

[37] H. Kaldera, S. Gunasekara, and M. B. Dissanayake, "MRI based glioma segmentation using deep learning algorithms," in *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 51–56, Colombo, Sri Lanka, 2019.

[38] P. Chang, J. Grinband, B. D. Weinberg et al., "Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas," *AJNR. American Journal of Neuroradiology*, vol. 39, no. 7, pp. 1201–1207, 2018.

[39] L. Han and M. R. Kamdar, "MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks," *Biocomputing*, vol. 23, pp. 331–342, 2018.

[40] S. Drabycza, G. Roldánbcde, P. Roblesbde et al., "An analysis of image texture, tumor location, and *MGMT* promoter methylation in glioblastoma using magnetic resonance imaging," *NeuroImage*, vol. 49, no. 2, pp. 1398–1405, 2010.

[41] Y. Han, L. F. Yan, X. B. Wang et al., "Structural and advanced imaging in predicting MGMT promoter methylation of primary glioblastoma: a region of interest based analysis," *BMC Cancer*, vol. 18, no. 1, 2018.

[42] S. Wu, H. Li, D. Quang, and Y. Guan, "Three-plane-assembled deep learning segmentation of gliomas," *Radiology: Artificial Intelligence*, vol. 2, no. 1, article e190011, 2020.

Hindawi
BioMed Research International
Volume 2020, Article ID 6863231, 12 pages
https://doi.org/10.1155/2020/6863231

Research Article

MAGE-Targeted Gold Nanoparticles for Ultrasound Imaging-Guided Phototherapy in Melanoma

Xuelin Li,[1] Shigen Zhong,[2] Cuncheng Zhang,[2] Pan Li,[3] Haitao Ran,[3] and Zhigang Wang[3]

[1] *Department of Geriatric, Chongqing General Hospital (University of Chinese Academy of Sciences Chongqing Hospital), Chongqing 400014, China*
[2] *Department of Ultrasound, Chongqing General Hospital (University of Chinese Academy of Sciences Chongqing Hospital), Chongqing 400014, China*
[3] *Institute of Ultrasound Imaging of Chongqing Medical University, Chongqing Key Laboratory of Ultrasound Molecular Imaging, Chongqing 400010, China*

Correspondence should be addressed to Zhigang Wang; wzg62942443@163.com

Received 27 June 2020; Revised 22 August 2020; Accepted 3 September 2020; Published 18 September 2020

Academic Editor: Erlei Zhang

Gold nanorods exhibit a wide variety of applications such as tumor molecular imaging and photothermal therapy (PTT) due to their tunable optical properties. Several studies have demonstrated that the combination of other therapeutic strategies may improve PTT efficiency. A method called optical droplet vaporization (ODV) was considered as another noninvasive imaging and therapy strategy. Via the ODV method, superheated perfluorocarbon droplets can be vaporized to a gas phase for enhancing ultrasound imaging; meanwhile, this violent process can cause damage to cells and tissue. In addition, active targeting through the functionalization with targeting ligands can effectively increase nanoprobe accumulation in the tumor area, improving the sensitivity and specificity of imaging and therapy. Our study prepared a nanoparticle loaded with gold nanorods and perfluorinated hexane and conjugated to a monoclonal antibody (MAGE-1 antibody) to melanoma-associated antigens (MAGE) targeting melanoma, investigated the synergistic effect of PTT/ODV therapy, and monitored the therapeutic effect using ultrasound. The prepared MAGE-Au-PFH-NPs achieved complete eradication of tumors. Meanwhile, the MAGE-Au-PFH-NPs also possess significant ultrasound imaging signal enhancement, which shows the potential for imaging-guided tumor therapy in the future.

1. Introduction

Melanoma is a malignant neoplasm sourced from melanocytes skin cells—with poor prognosis at advanced stages. Standard cancer treatments can be highly toxic to healthy tissues without differentiating malignant from normal cells, causing significant adverse effects in patients. Nanoparticle-based photothermal ablation therapy assisted by near-infrared (NIR) laser holds a promise to eliminate tumors noninvasively, reduce tumor resistance, and prevent recurrence [1–3]. In addition, active targeting ability through functionalization with specific ligands can effectively enable nanoprobes to accumulate in the tumor area [2]. Melanoma-associated antigens (MAGE) are a specific and highly expressed family of antigens in malignant melanoma [4–6]. Therefore, MAGE proteins could also be used to functionalize nanoprobes for molecular imaging and accurate treatment of melanoma.

Among the available nanoparticle systems, gold nanorods (GNRs/Au-NRs) have attracted particular attention in cancer imaging and photothermal therapy [7–9] due to several advantages: biocompatibility, high photothermal conversion efficiency, well-established methods for synthesis in a wide range of sizes, and ease of biomodification [10]. However, complete tumor eradication is so far difficult to achieve with the introduction of these photothermal nanomaterials. In particular, their low tissue bioabsorption and utilization result in limited curative effect in deeply located tumors [11]. Recently, it has been reported that PTT efficiency may be improved with the combination of other therapeutic strategies [12–14]. Phase-changeable liquid fluorocarbon emulsions can be vaporized and transformed from droplets to microbubbles under optical irradiation (ODV) [15–18] which has become an attractive noninvasive theranostic protocol based on ultrasound imaging and physical therapy [19]. The emulsion significantly increased acoustic impedance between the tumor and surrounding tissues [20] and causes physical damage to tumor cells [21].

Poly(D,L-lactide-co-glycolide) (PLGA), as a type of biodegradable nanomaterial with an excellent safety profile in humans [22] and good film-forming ability, has been approved by FDA for vaccine and drug delivery as well as tissue engineering [23–25]. Thus, we prepared a MAGE-targeted PLGA nanomolecular probe encapsulating liquid perfluorohexane (PFH) and Au-NRs (MAGE-Au-PFH-NPs) in this work, specifically targeting melanoma cells, which have been confirmed to possess the potential to enhance photoacoustic (PA) and ultrasonic imaging in melanoma in our previous study. The results from our previous study showed that MAGE-Au-PFH-NPs could accumulate and retain in the tumor area, allowing for therapeutic guidance and monitoring [26]. This study went further in exploring the role of the MAGE-Au-PFH-NPs in the treatment of melanoma based on our previous study. Benefiting from the high photothermal-conversion efficiency and ODV effect in MAGE-Au-PFH-NPs, the efficient tumor ablation was achieved in melanoma-bearing mice, which provides a promising alternative strategy for imaging-guided phototherapy of cancer.

2. Materials and Methods

2.1. Materials. Gold nanorods (Au-NRs, 780 nm) were purchased from NanoSeedz Ltd. (Hong Kong SAR), perfluorohexane (PFH) was from Ji'nan Daigang Biological Engineering Co. Ltd. (Jinan, China), and the B16 mouse melanoma cell line was purchased from the Punuosai Company (Wuhan, China). MAGE-1 antibody was from the Bioye Company (Shanghai, China), and propidium iodide (PI) was purchased from Sigma-Aldrich (St. Louis, MO, USA). Calcein acetoxymethyl ester (Calcein-AM) was purchased from Santa Cruz Biotechnology (TX, USA). Anti-HSP70 Rabbit pAb was from Servicebio (Wuhan, China). Cy3-conjugated goat anti-rabbit IgG was from Servicebio (Wuhan, China).

2.2. Cell Culture and Animal Experiment. B16 mouse melanoma cells were cultured in T75 flasks containing Roswell Park Memorial Institute (RPMI) 1640 medium supplemented with 10% foetal bovine serum and 1% penicillin and streptomycin (antibiotics) and incubated at 37°C under 5% CO_2, with medium changes every 2-3 days. For all the experiments, the cells were harvested using 0.25% trypsin solution and were then resuspended in fresh medium before plating.

All the animals (male BALB/c nude mice: ~20 g, 4–6 weeks) were purchased from the Experimental Animal Center of Chongqing Medical University and bred at constant temperature and humidity, with food and water provided *ad libitum*. The animals were maintained in accordance with the National Guidelines for Experimental Animal Welfare (MOST, China, 2006) at the Centre for Animal Experiments, and all the experiments and procedures were approved by the Institutional Animal Care and Use Committee of Chongqing Medical University. B16 cells were suspended in PBS (1×10^6 B16 cells in 100 μL of PBS per mouse) and then injected subcutaneously into the buttock of the BALB/c nude mice to establish tumor-bearing mice.

2.3. Characterization of the NPs. By reference to our previous study [26], MAGE-Au-PFH-NPs were prepared by the double emulsion method. The carbodiimide method was employed to modify the Au-PFH-NPs with MAGE antibody to prepare the targeted nanoparticles (MAGE-Au-PFH-NPs).

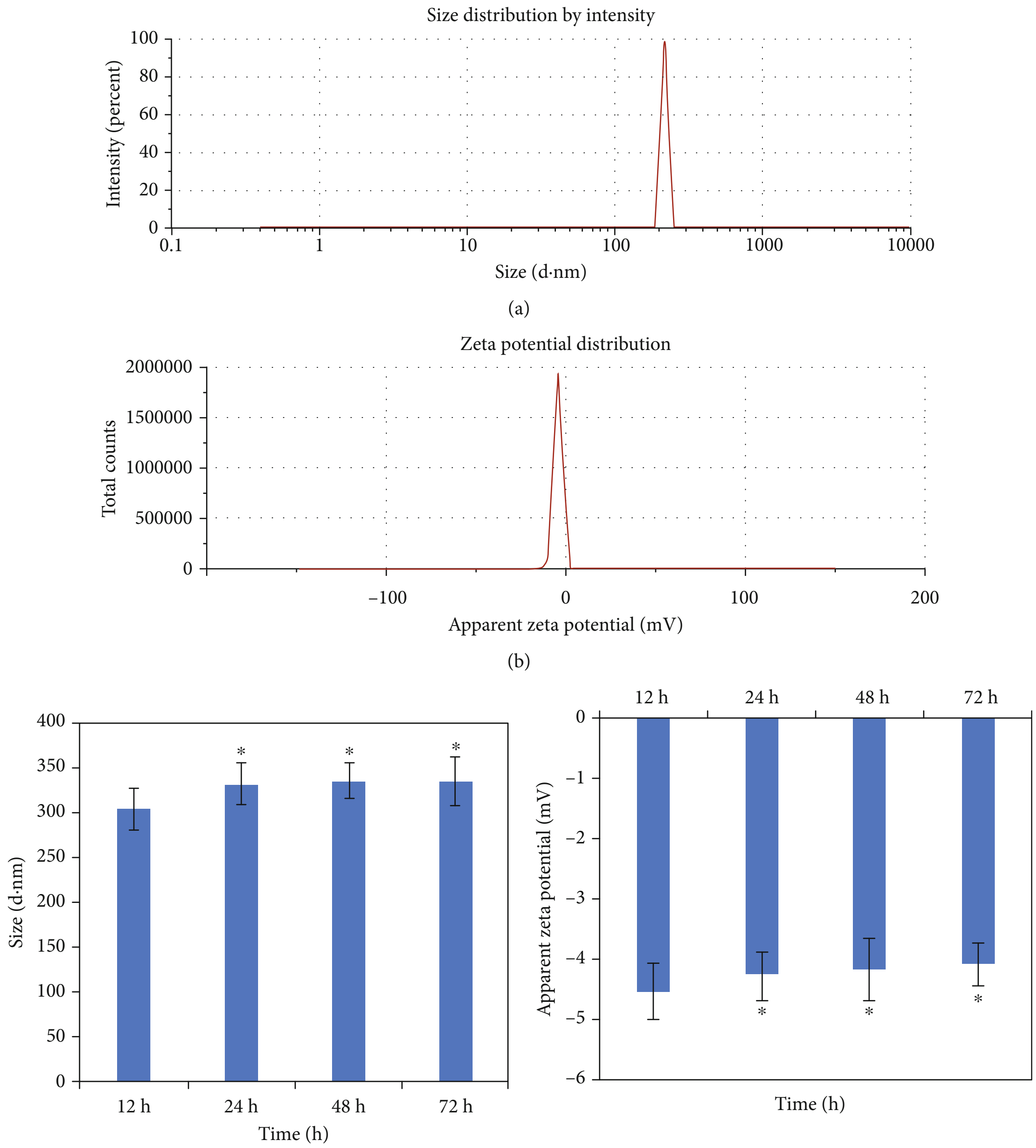
The prepared NPs were observed under a confocal laser scanning microscope (Nikon) and transmission electron microscope (TEM) (H-7500; Hitachi, ×1.5 k Zoom-1 HC-1 80.0 kv).

The size and zeta potential of MAGE-Au-PFH-NPs were measured by a Malvern laser particle size analyzer (Malvern, England). Furthermore, to assess the stability of MAGE-Au-PFH-NPs, the NP size and zeta potential changes were tested for 72 hours while incubating in plasma at 37°C. Meanwhile, the NPs underwent PAI scanning at different wavelengths ranging from 680 nm to 970 nm (interval = 5 nm) to determine the maximum absorbance for optimized PAI by a Vevo LAZR Photoacoustic Imaging System (Vevo LAZR, Toronto, Canada).

2.4. In Vitro Photothermal Properties of NPs. In our previous study, MAGE-Au-PFH-NP temperature was increased to 70°C through photothermal conversion after absorbing near infrared light [26]. In this study, we evaluated photothermal capability of NPs in different concentrations at different power densities by laser irradiation using an 808 nm laser for 5 minutes *in vitro*. The infrared radiation (IR) thermal images and temperature changes were recorded by an infrared thermal-imaging camera.

2.5. In Vitro Photothermal Ablation against B16 Cells. To test the photothermal ablation effects of MAGE-Au-PFH-NPs *in vitro*, B16 cancer cells were divided into 4 groups: control (normal saline (NS) only), NPs only, laser only, and NPs +laser, and were seeded onto four confocal-specific cell-culture dishes (1×10^5 cells per dish) overnight. After cell adhesion, NP suspension was added into two dishes (two groups: NPs only and NPs+laser), and equal volume serum-free RPMI 1640 medium was added into the other two dishes and incubated for another 12 h; cells of two groups (laser only and NPs+laser) were exposed to laser (1.00 W/cm^2) for 10 min. Then, the medium was removed, and the cells of each dish were washed three times with PBS. The cells of four groups were scanned by confocal microscopy after costaining with Calcein-AM and propidium iodide.

2.6. Photothermal Conversion Evaluation and Heat Shock Protein (HSP) Evaluation In Vivo. To investigate the *in vivo*



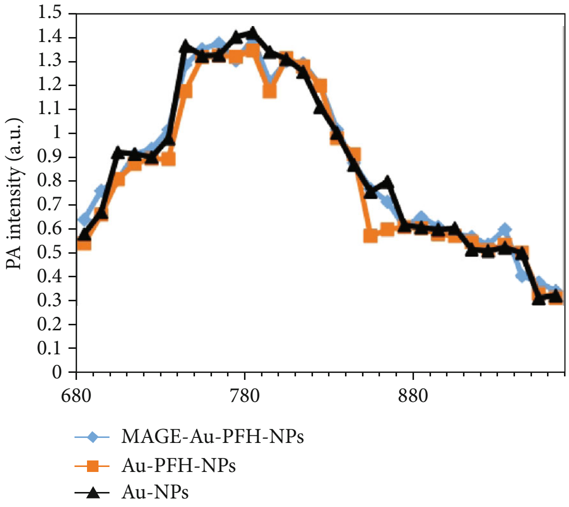
(a)

(b)

(c)

(d)

Figure 1: Continued.



(e)

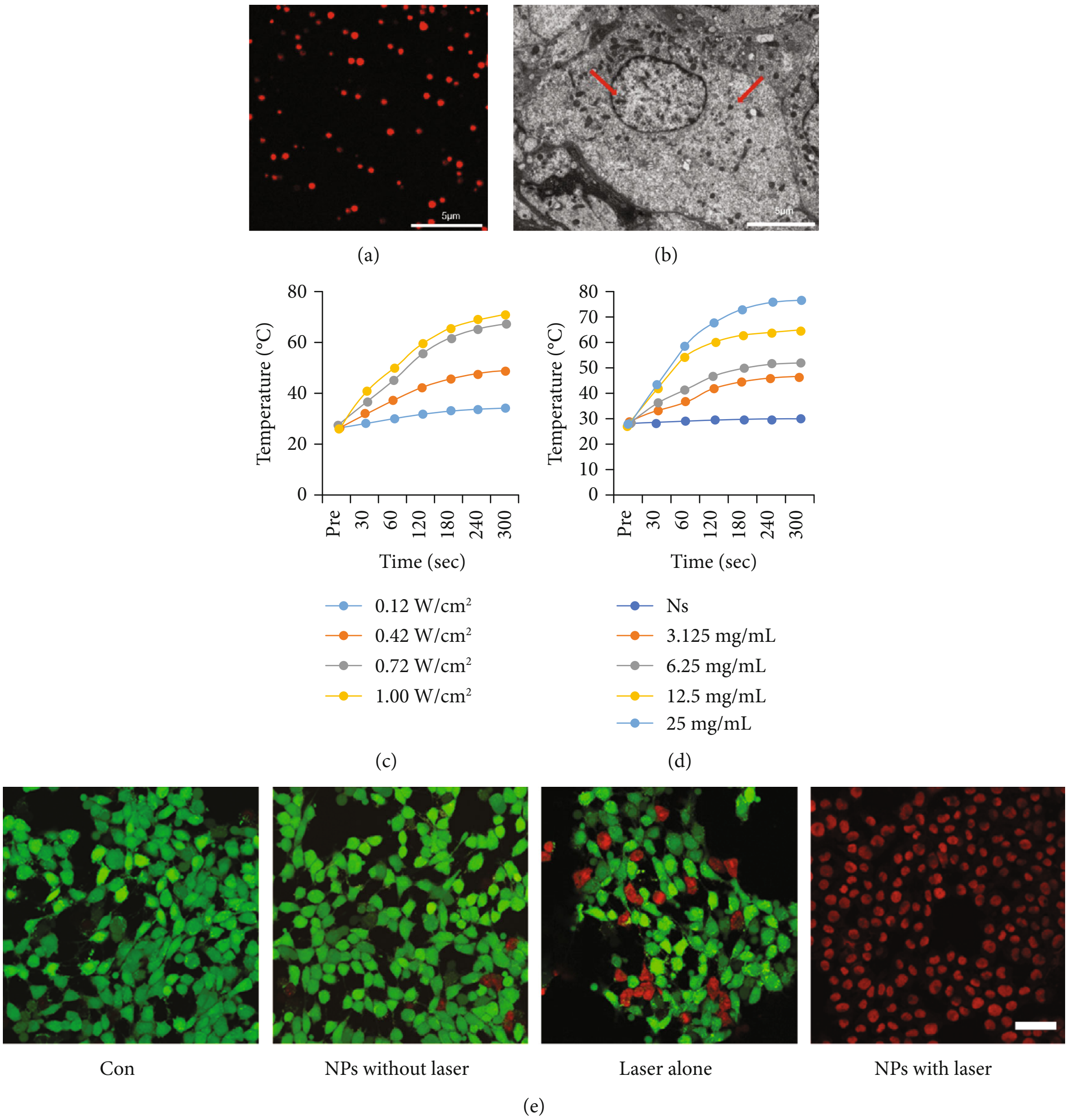
Figure 1: Physicochemical characterization of MAGE-Au-PFH-NPs. (a) Size distribution of MAGE-Au-PFH-NPs. (b) Zeta potential of MAGE-Au-PFH-NPs. (c) Size distribution of MAGE-Au-PFH-NPs incubating in plasma at 37°C after 12, 48, and 72 h ($^*P > 0.05$). (d) Zeta potential of MAGE-Au-PFH-NPs incubating in plasma at 37°C after 12, 48, and 72 h ($^*P > 0.05$). (e) PA spectrum of MAGE-Au-PFH-NPs, Au-PFH-NPs, and Au-NPs (25 mg/mL) from 680 nm to 960 nm.

photothermal efficiency of MAGE-Au-PFH-NPs, twenty B16 tumor-bearing mice were randomly divided into four groups (5 mice per group) when the tumor volumes reached about 100 mm^3, including MAGE-Au-PFH-NP, Au-PFH-NP, Au-NP, and NS groups. The mice were intravenously injected with 100 μL MAGE-Au-PFH-NPs, Au-PFH-NPs, and Au-NPs at a concentration of 25 mg/mL, respectively. The same volume of normal saline was injected into the mice in the control NS group. The tumors were exposed to the 808 nm laser (1.00 W/cm^2) for 10 min after injection of NPs and NS 2 minutes later. The temperature changes in tumors and infrared radiation (IR) thermal images were recorded by an infrared thermal-imaging camera. To further analyze the effect of photothermal therapy on local hyperthermia, we detected the HSP70 expression within tumors by immunofluorescent staining. All the dose and laser irradiation conditions were adjusted to the same level as mentioned above. At the second day after treatment, mice were sacrificed to collect tumors for HSP70 immunofluorescent staining. Anti-HSP70 rabbit pAb as primary antibodies was added in the section and incubated overnight at -4°C; Cy3 conjugated goat anti-rabbit IgG as secondary antibodies was added in the section away from light for 50 min at room temperature, followed by DAPI hyperchromatic nucleus and sealing piece. Eventually, the sections were observed and imaged under a fluorescence microscope.

2.7. Targeting Ability In Vitro and In Vivo. To assess the targeting ability *in vitro*, immunofluorescence imaging observed under confocal microscopy has been performed. The melanoma-associated antigens were combined to the targeted nanoparticles to obtain the blocking group (MAGE-R-Au-PFH-NPs). MAGE-Au-PFH-NPs and the blocking group (MAGE-R-Au-PFH-NPs) were all treated with DiI fluorescent dye in the first step of synthesis before the sonication. B16 cells were seeded in confocal laser dishes at 1×10^5 and coincubated with dyed NPs (MAGE-Au-PFH-NPs, MAGE-R-Au-PFH-NPs) at 37°C for 30 min and then observed under a laser scanning confocal microscope after being fixed with 4% paraformaldehyde and stained with 20 μL of 4′,6-diamidino-2-phenylindole (DAPI).

To detect whether MAGE-Au-PFH-NPs have a long circulation time compared to that of the Au-PFH-NPs, 6 tumor-bearing mice were randomly divided into two groups (MAGE-Au-PFH-NPs and Au-PFH-NPs); the two groups of mice were intravenously injected with 100 μL of MAGE-Au-PFH-NPs and Au-PFH-NPs (25 mg/mL). The dynamic distribution of nanoparticles within the whole body was measured using fluorescence spectrum.

2.8. Photothermal/ODV Efficiency and Detection In Vivo. To evaluate the *in vivo* photothermal and ODV efficiency of MAGE-Au-PFH-NPs, twenty B16 tumor-bearing mice were used when the tumor volumes reached about 100 mm^3. The mice were divided into four groups (5 mice per group) randomly including the laser only, Au-NP+laser, Au-PFH-NP +laser, and MAGE-Au-PFH-NP+laser groups, which were intravenously injected with normal saline solution (100 μL), Au-NP (25 mg/mL, 100 μL), Au-PFH-NP (25 mg/mL, 100 μL), and MAGE-Au-PFH-NP suspension (25 mg/mL, 100 μL), respectively. Then, each mouse was exposed to the 808 nm laser for 10 min at a power density of 1.00 W/cm^2. The treatment was performed every other day, and one of the mice in each group was sacrificed on the third day for pathological section and staining. All the tumor tissues were collected and fixed in a 4% paraformaldehyde solution, stained with H&E, TdT-mediated dUTP Nick-End Labeling (TUNEL), and Proliferating Cell Nuclear Antigen (PCNA) for histopathology analysis.



(a) (b) (c) (d) (e)

FIGURE 2: Morphological distribution of MAGE-Au-PFH-NPs under the microscope and photothermal properties of NPs *in vitro*. (a) CLSM image of DiI-stained MAGE-Au-PFH-NPs (scale bar: 5 μm). (b) TEM image of MAGE-Au-PFH-NP distributions in cells (scale bar: 5 μm). Red arrow showed the NP distribution in B16 cells. (c) Plot of temperature change of MAGE-Au-PFH-NP suspension at different power densities of an 808 nm laser (0.12, 0.42, 0.72, and 1.00 W/cm^2) as a function of irradiation duration (MAGE-Au-PFH-NP concentration: 25 mg/mL and 100 μL). (d) Plot of temperature change of NS and MAGE-Au-PFH-NP suspension at different levels of concentration (MAGE-Au-PFH-NP concentrations: 3.125, 6.25, and 13.5, 25 mg/mL and 100 μL) exposure to an 808 nm laser (1.00 W/cm^2) as a function of irradiation duration. (e) Confocal fluorescence imaging of Calcein-AM and PI costained B16 cells after coincubation with NPs for 12 h followed by different treatments (scale bar = 100 μm).

The remaining four mice were maintained for 15 days, and the mouse weight and tumor volume were measured every other day after PTT. The tumor-volume changes were normalized using the relative tumor volumes V/V_0 (V_0: the initial tumor volume before the treatment).

In our previous study, CEUS imaging was significantly enhanced at the tumor site in the MAGE-Au-PFH-NP group after laser irradiation [26]. Thus, contrast-enhanced ultrasonography (Esaote MyLab 90, Genoa, Italy) was performed in tumor-bearing mice after the treatment to evaluate the therapeutic effect of MAGE-Au-PFH-NPs *in vivo* in this study. CEUS-mode images of the tumors were recorded after exposure to the 808 nm laser (1.00 W/cm^2, 10 mins). Echo intensity was acquired and analyzed using a DFY-ultrasonic image quantitative analyzer (Institute of Ultrasound Imaging of Chongqing Medical University, Chongqing, China).

2.9. Toxicity Test In Vitro and Biocompatibility In Vivo. To assess the toxicity of MAGE-Au-PFH-NPs, the typical CCK-8 assay was used to evaluate the cell viability in endothelial cells and hepatic cells. The test of liver functional markers (AST, ALT), kidney functional markers (CR,

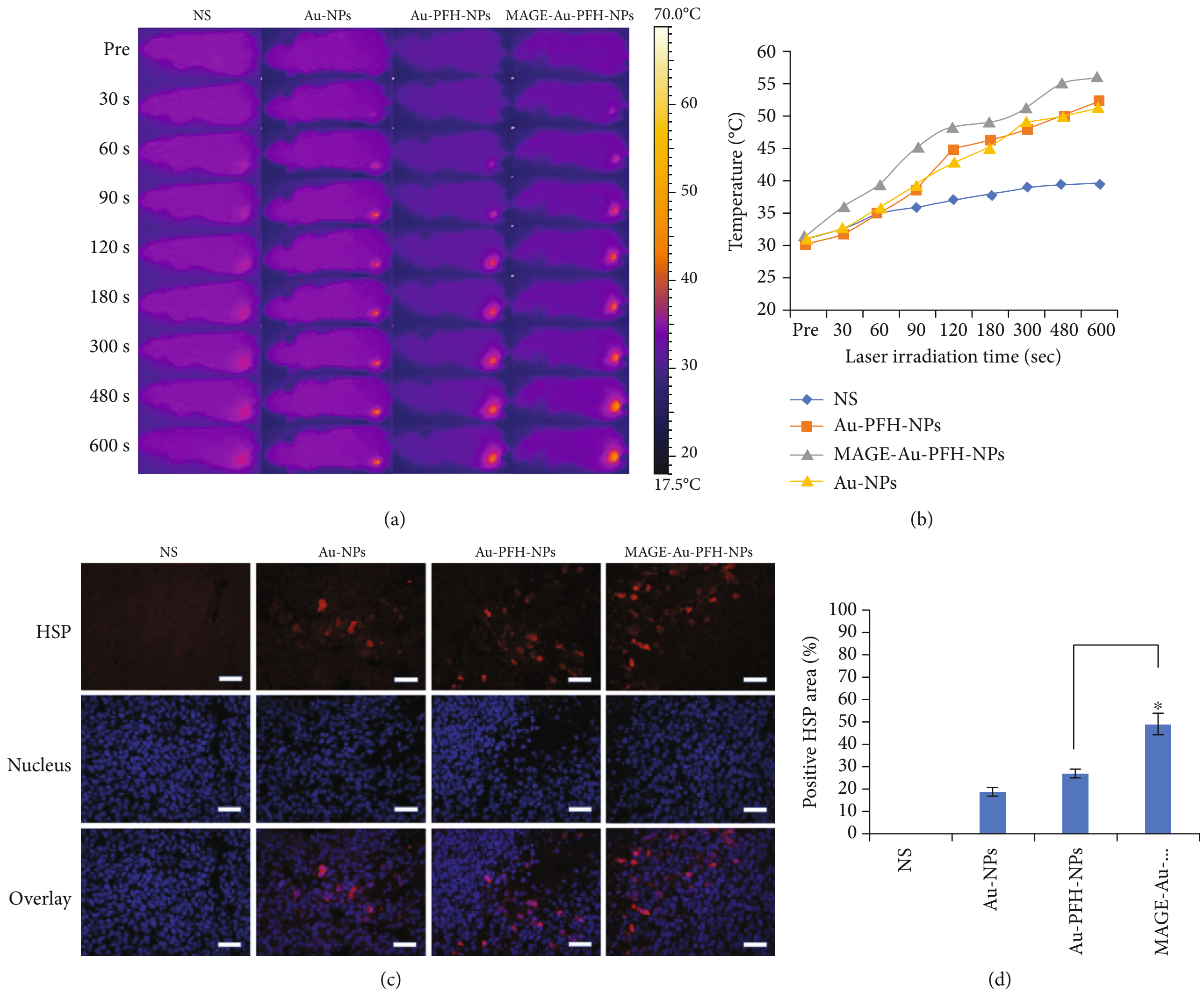


Figure 3: Photothermal conversion efficiency and heat shock protein (HSP) evaluation *in vivo*. (a) IR thermal images of B16 tumor-bearing mice of the four groups (NS+laser, Au-NP+laser, Au-PFH-NP+laser, and MAGE-Au-PFH-NP+laser groups) taken at different times. (b) Plot of temperature change of four groups (NS+laser, Au-NP+laser, Au-PFH-NP+laser, and MAGE-Au-PFH-NP+laser groups) *in vivo*. (c) Immunofluorescent staining of HSP70 of tumor tissues dissected from different groups on the 1st day post treatments. The scale bar is 50 μm. (d) Quantitative analysis of HSP expression from different groups on the 1st day post treatments. (The data is shown as mean ± SD, $n = 5$ per group; $^*P < 0.05$.).

BUN), and H&E staining of the major organs (heart, liver, and kidney) after intravenous injection and laser irradiation have been performed to evaluate the biocompatibility *in vivo*.

3. Statistical Analysis

Data were presented as the mean ± standard deviation. Statistical analyses were done using the SPSS Ver. 19.0. $P < 0.05$ was considered statistically significant according to one-way ANOVA and Student's t-test.

4. Results

4.1. Characterization of NPs and Photothermal Conversion Efficiency of NPs In Vitro. The size and zeta potential of MAGE-Au-PFH-NPs were 324.54 ± 21.76 nm and −4.76 ± 3.7 mV, respectively, measured by a Malvern laser particle size analyzer. Furthermore, results showed that the size and zeta potential changes measured at 48 h and 72 h had no statistic difference compared with 12 h that could confirm the stability of MAGE-Au-PFH-NPs (Figures 1(a)–1(d)). The result of the PAI scanning showed that the maximum absorbance of three groups (Au-NPs, Au-PFH-NPs, and MAGE-Au-PFH-NPs) had been detected at 780 nm which conformed to the spectral properties of gold nanorods (GNRs/Au-NRs) (Figure 1(e)).

The NPs showed a good dispersity under a confocal laser scanning microscope and were mainly distributed in the cytoplasm after uptake by cells under a transmission electron microscope (Figures 2(a) and 2(b)).

The temperature change of NPs after laser irradiation was recorded by an infrared thermal-imaging camera. The

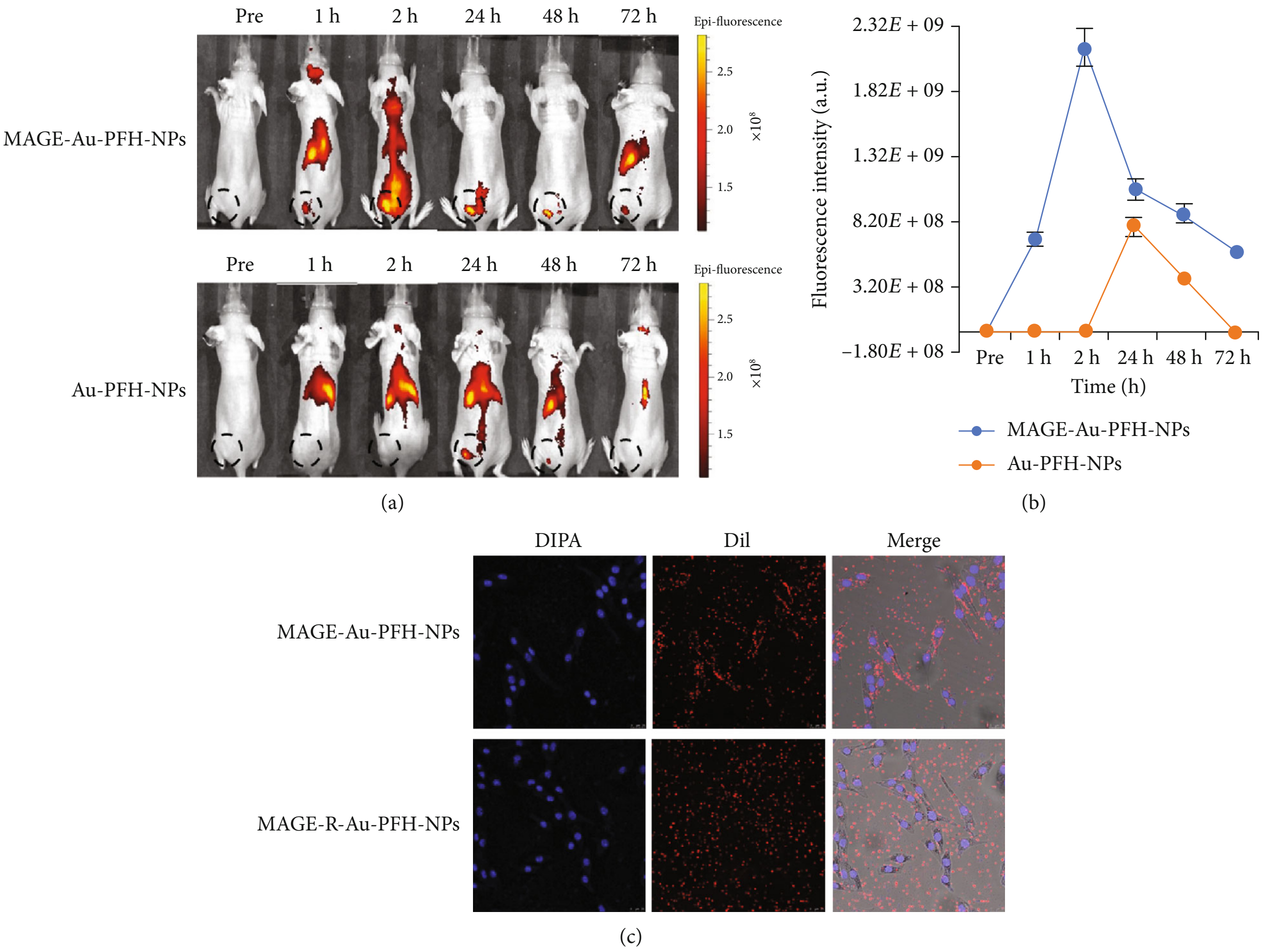


Figure 4: Fluorescence imaging *in vivo*. (a) Fluorescence images of B16 tumor-bearing mice after intravenous injection of MAGE-Au-PFH-NPs and Au-PFH-NPs at 0, 1, 2, 24, 48, and 72 h ($n = 3$ per group). (b) Plot of fluorescence images of B16 tumor-bearing mice after intravenous injection of MAGE-Au-PFH-NPs and Au-PFH-NPs at 0, 1, 2, 24, 48, and 72 h. The dotted circle instruction for the tumor mass. (c) Targeting ability of MAGE-Au-PFH-NPs and MAGE-R-Au-PFH-NPs to B16 cells observed under CLSM; the left line was cell nuclei stained by DAPI, the middle line was NP stained by DiI, and the right line was the merged result of the two fluorescence images.

temperature in a MAGE-Au-PFH-NP suspension increased significantly and reached as high as 71°C under NIR irradiation within 5 min at a concentration of 25 mg/mL as shown in Figures 2(c) and 2(d), while no temperature change was found in the normal saline (NS) group. Moreover, the temperature elevated significantly with the increase of NIR irradiation power, which demonstrated that the photothermal conversion efficiency of MAGE-Au-PFH-NPs depended on both concentration and laser power density.

4.2. In Vitro Photothermal Ablation against B16 Cells. A large number of B16 cells died and showed a strong red fluorescence (Figure 2(e)) in the MAGE-Au-PFH-NP+laser group in the experiment *in vitro*, suggesting photothermal effect induced by MAGE-Au-PFH-NPs under external NIR irradiation. In contrast, dead cells were rarely found in the control group, which displayed green fluorescence of Calcein-AM staining (Figure 2(e)). Only a number of dead cells were shown in the laser- and MAGE-Au-PFH-NP only groups, as confirmed by the strong green fluorescence and very weak red fluorescence from PI staining.

4.3. Photothermal Conversion Efficiency and HSP Evaluation In Vivo. As shown in Figures 3(a) and 3(b), the surface temperature of tumors in the MAGE-Au-PFH-NP+laser group increased from 31.6 ± 1.09°C to 56.1 ± 2.6°C under irradiation for 10 min. The Au-PFH-NP+laser group increased from 30.1 ± 1.3°C to 52.0 ± 2.1°C under irradiation for 10 min. The temperature in the Au-NP+laser group increased from 31.2 ± 0.9°C to 51.4 ± 1.7°C under irradiation for 10 min. Comparatively, only a slight temperature increase was found in the tumor region in the laser only group.

The HSP70 expression level was analyzed and is shown in Figures 3(c) and 3(d). The mice that received MAGE-Au-PFH-NPs plus laser irradiation showed a remarkably higher HSP70 expression compared to those treated with Au-PFH-NPs and Au-NPs followed by laser irradiation. In the control group (NS with laser irradiation), no evident expression of HSP70 was found.

4.4. Targeting Ability In Vitro and In Vivo. The biodistribution of nanoparticles in mice was investigated by *in vivo* fluorescence imaging. At 1 hour post injection, prominent uptake

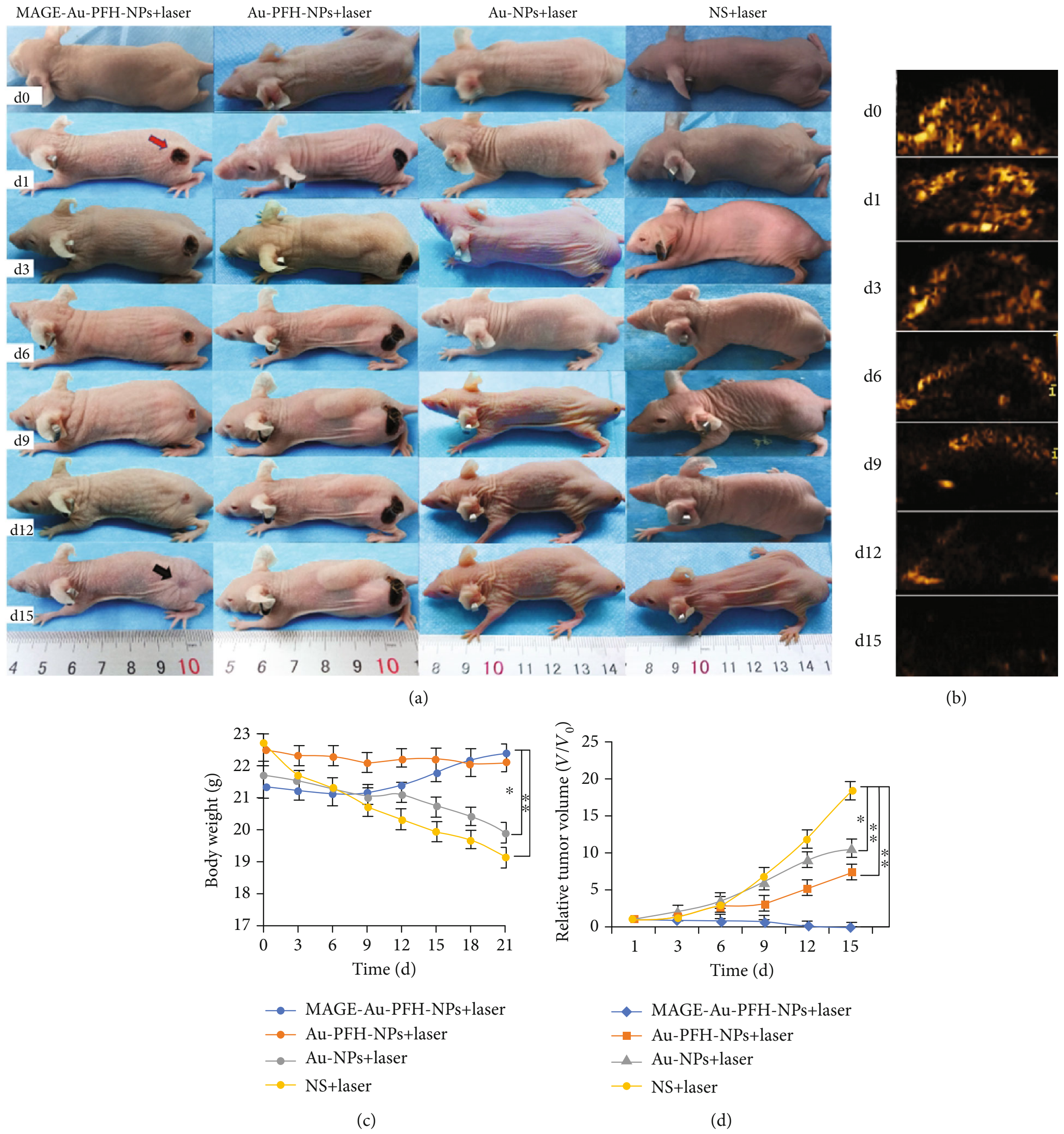
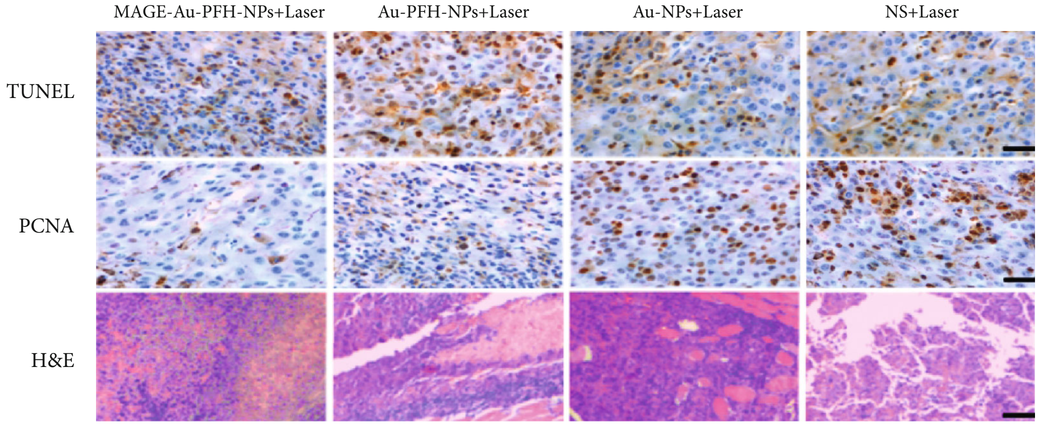


Figure 5: Continued.



(e)

Figure 5: Detection of photothermal/ODV efficiency *in vivo*. (a) Photographs of B16 tumor-bearing mice in the four groups taken during a 15-day period after the various treatments. (b) Ultrasound imaging from the region of interest in B16 tumor-bearing nude mice in the MAGE-Au-PFH-NP group using CEUS after laser irradiation during a 15-day period. (c) Body weight curves ($n = 5$, mean ± SD) of the four groups after different treatments. (d) Tumor growth curves ($n = 5$, mean ± SD) of the four groups after various treatments. (e) H&E, TUNEL, and PCNA staining on tumor sections after various treatments. All the scale bars are 50 μm. (The data is shown as mean ± SD, $n = 5$ per group; $^*P < 0.05$, $^{**}P < 0.01$.).

of nanoparticles in the tumor was observed in the MAGE-Au-PFH-NP group; the signals reached a peak 2 hours later and lasted for 72 h. While in the Au-PFH-NP group, fluorescent signals were found at 24 h and disappeared after 48 h (Figures 4(a) and 4(b)). These results confirmed the accumulation and long retention time of MAGE-Au-PFH-NPs within the tumor area. Fluorescence imaging observed under confocal microscopy demonstrated a large number of MAGE-Au-PFH-NPs concentrated in B16 cells, showing clear red fluorescence, while the blocking group (MAGE-R-Au-PFH-NPs) presented sparse distribution around the cells (Figure 4(c)). These results confirmed that the MAGE-Au-PFH-NPs could specifically target B16 cells.

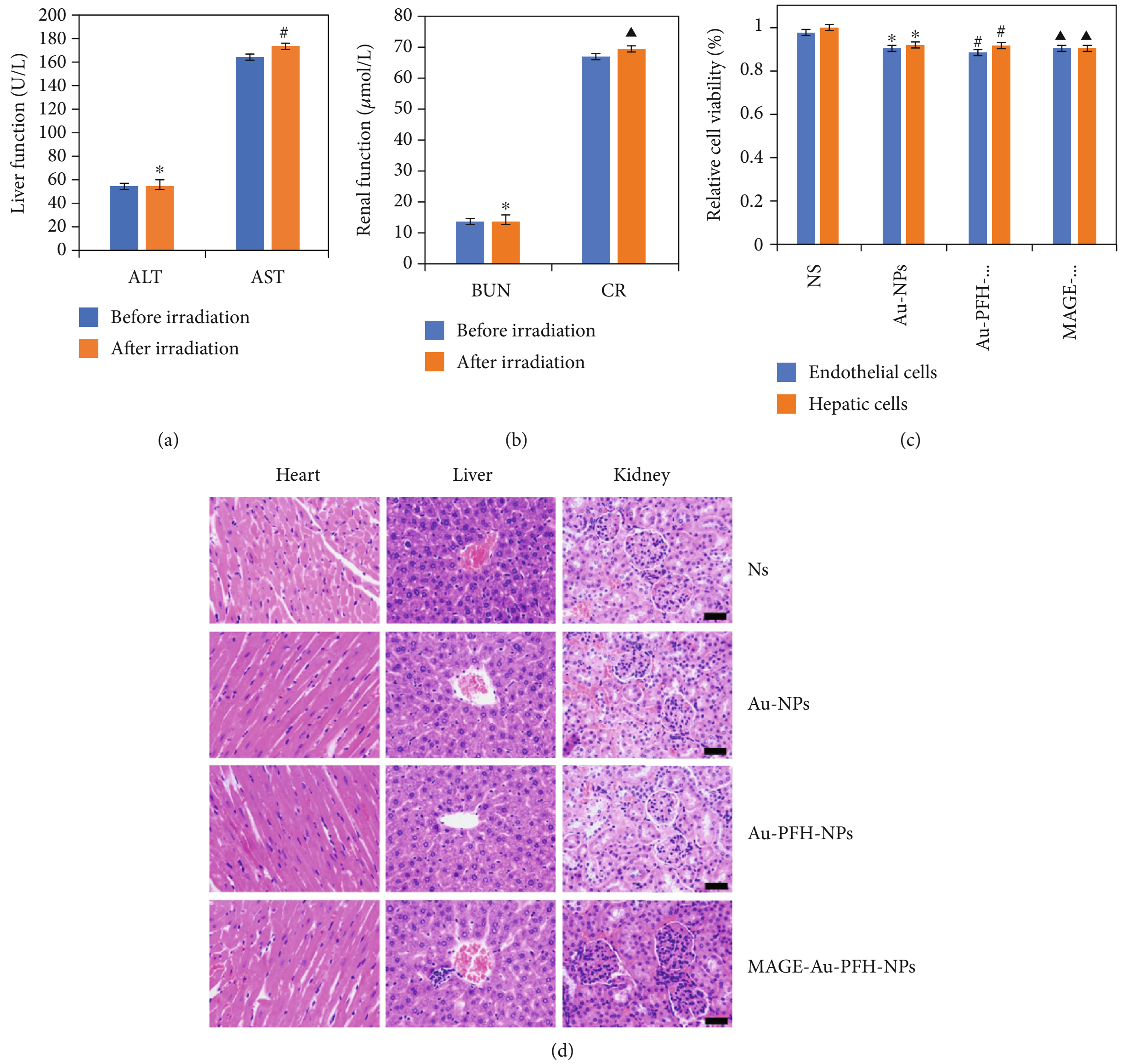
4.5. Photothermal/ODV Efficiency and Detection In Vivo. Based on the *in vivo* photoacoustic imaging experiment, MAGE-Au-PFH-NPs actively accumulated in the tumor region, and the signal reached a peak at 2 h post injection [26]. Therefore, the mice were treated for 10 min after 2 h post injection until the 5th day. Tumor tissue necrosis was found in mice in the MAGE-Au-PFH-NP+laser group the second day after treatment, leaving black scars in the initial tumor regions (Figure 5(a), red arrow). Then, they disappeared 15 days later, leading to complete tumor eradication (Figure 5(a), black arrow), while the tumors in the remaining three groups grew rapidly. The ultrasound imaging was significantly enhanced at the tumor site in the MAGE-Au-PFH-NP group under the laser irradiation. Then, the enhanced ultrasound signals gradually decreased with the increase of treatment times and disappeared by day 15 (Figure 5(b)). The weight and tumor volume of each mouse were recorded every other day (Figure 5(c)). Then, the tumor-volume change was normalized as V/V_0 (Figure 5(d)).

H&E and TUNEL staining results further confirmed tumor necrosis in mice in the MAGE-Au-PFH-NP+laser group, which was more serious compared to those in the remaining three groups (Figure 5(e)). From the result of the PCNA assay, the MAGE-Au-PFH-NP+laser group exhibited a significant suppression effect on tumor cell proliferation. In contrast, no evident proliferative inhibition was observed in the remaining three groups.

4.6. Toxicity Test In Vitro and Biocompatibility In Vivo. The cytotoxicity of MAGE-Au-PFH-NPs was investigated by the CCK-8 assay. After 24 h incubation, inconspicuous cytotoxicity of the NPs on endothelial cells and hepatic cells was observed since cell viability remained above 80% at NP concentration of 25 mg/mL. And no significant difference was found among the groups (NS, Au-NPs, Au-PFH-NPS, and MAGE-Au-PFH-NPs) (Figure 6(c)). The test of liver functional markers (AST and ALT) and kidney functional markers (CR and BUN) and H&E staining of the major organs (heart, liver, and kidney) after intravenous injection and laser irradiation showed no significant physiological toxicity (Figures 6(a), 6(b), and 6(d)).

5. Discussion

Photothermal therapy (PTT) is a minimally invasive technique for cancer treatment which uses laser-activated photoabsorbers to convert photon energy into heat sufficient to induce cell destruction via apoptosis, necroptosis, and/or necrosis. From the current studies, photothermal therapy cannot ablate the tumor thoroughly using photothermal materials due to nonuniform tumor internal heat distribution. The combination of PTT and other methods may overcome this disadvantage. Gold NPs (Au-NPs) designed to act as photothermal sensitizing agents are widely used in cancer therapy due to their high optical absorption coefficients. In our previous study, the Au-NRs and PFH were encapsulated in a PLGA shell through the double emulsion method with a high encapsulation efficiency and conjugated with the



(a)

(b)

(c)

(d)

Figure 6: Toxicity test *in vitro* and biocompatibility *in vivo*. (a, b) Haematological assay of B16 tumor-bearing nude mice. (c) Cell viability assay of toxicity of MAGE-Au-PFH-NPs to normal cell ($^{*}P > 0.05$, $^{\#}P > 0.05$, and $^{\blacktriangle}P > 0.05$). (d) H&E staining of the major organs (heart, liver, and kidney) of B16 tumor-bearing nude mice after administration of MAGE-Au-PFH-NPs.

MAGE-1 antibody for targeting melanoma cells [26]. With the application of MAGE-Au-PFH-NPs, the temperature can be raised up to 71°C under laser irradiation, showing their excellent photothermal effect [27]. These targeted nanoparticles actively accumulated in the tumor area and enhanced the PTT effect. In addition, the PFH can be vaporized by laser irradiation [28] with the application of MAGE-Au-PFH-NPs [16], which had also been confirmed in our *in vitro* study [26]. Meanwhile, from the *in vivo* results, the temperature in the tumor region increased up to 56.1 ± 2.6° C (Figure 3(b)), which was sufficient to ablate the tumor tissue [29]; meanwhile, it could convert MAGE-Au-PFH-NPs into bubbles. Then, the physical and mechanical damage induced by the phase change process can directly kill tumor cells [21].

Heat shock protein is produced under the induction of stress agents such as high temperature to induce thermoresistance for cells [30]. In our study, the results showed that HSP70 expression in the MAGE-Au-PFH-NP group was significantly higher compared to that in the groups of Au-PFH-NPs and Au-NPs with laser irradiation. Only little expression of HSP70 was found in the NS plus laser group. The results of HSP70 immunofluorescent staining corresponded well with those of photothermal conversion effect.

In our study, complete tumor eradication was observed in the MAGE-Au-PFH-NP+laser group. In contrast, the tumors of the other three groups grew rapidly. The results showed that the combination of targeted photothermal therapy and ODV phase transition physical therapy could achieve better tumor ablation effect. *In vivo* fluorescence imaging confirmed that targeted nanoparticles (MAGE-Au-PFH-NPs) had a longer circulation time compared to the nontargeted Au-PFH-NPs.

Tumor recurrence was found in the Au-PFH-NP and Au-NP plus laser therapy groups indicating limited PTT effect leading to residues in tumor tissue. By combining the PTT effect and ODV physical damage from MAGE-Au-PFH-NPs, the tumor ablation was greatly improved. Meanwhile, damage to normal tissues caused by long time laser irradiation can be avoided.

Besides, when the phase-changeable nanoparticles were transformed from droplets to microbubbles, the acoustic impedance of the tumor and surrounding tissues was increased, which thereby enhanced ultrasound imaging. In this study, the CEUS imaging was significantly enhanced at the tumor site in the MAGE-Au-PFH-NP group under laser irradiation and decreased with the increase of treatment times, providing a method for monitoring tumor ablation effect by CEUS.

6. Conclusion

We successfully constructed MAGE-targeted phase-changeable gold nanoparticles which could accumulate at tumor sites. With the combination of PTT and ODV effects, complete tumor eradiation was achieved and could be monitored by contrast-enhanced ultrasonography. Several advantages such as noninvasiveness, short recovery time, low complication rate, and monitorable treatment process were included with this protocol. These novel targeted nanoparticles could be used as a multifunctional theranostic agent for imaging-guided tumor ablation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgments

This work was supported by grants from the High-level Medical Reserved Personnel Training Project of Chongqing, China (No: 2017HBRC010) and the National Natural Science Foundation of China (No: 31630026).

References

[1] J. Sheng, B. Ma, Q. Yang, C. Zhang, Z. Jiang, and E. Borrathybay, “Tailor-made PEG-DA-CuS nanoparticles enriched in tumor with the aid of retro Diels–Alder reaction triggered by their intrinsic photothermal property,” *International Journal of Nanomedicine*, vol. Volume 13, pp. 4291–4302, 2018.

[2] M. A. Shevtsov, L. Y. Yakovleva, B. P. Nikolaev et al., “Tumor targeting using magnetic nanoparticle Hsp70 conjugate in a model of C6 glioma,” *Neuro Oncology*, vol. 16, no. 1, pp. 38–49, 2014.

[3] Z. Xiao, X. Jiang, B. Li et al., “Hydrous RuO_2 nanoparticles as an efficient NIR-light induced photothermal agent for ablation of cancer cells in vitro and in vivo,” *Nanoscale*, vol. 7, no. 28, pp. 11962–11970, 2015.

[4] F. Brasseur, D. Rimoldi, D. Liénard et al., “Expression of MAGE genes in primary and metastatic cutaneous melanoma,” *International Journal of Cancer*, vol. 63, no. 3, pp. 375–380, 1995.

[5] C. Barrow, J. Browning, D. MacGregor et al., “Tumor antigen expression in melanoma varies according to antigen and stage,” *Clinical Cancer Research*, vol. 12, no. 3, pp. 764–771, 2006.

[6] O. L. Caballero, Q. Zhao, D. Rimoldi et al., “Frequent MAGE mutations in human melanoma,” *PLoS One*, vol. 5, no. 9, article e12773, 2010.

[7] B. Van de Broek, N. Devoogdt, A. D’Hollander et al., “Specific cell targeting with nanobody conjugated branched gold nanoparticles for photothermal therapy,” *ACS Nano*, vol. 5, no. 6, pp. 4319–4328, 2011.

[8] L. Xing, B. Chen, D. Li, W. Wu, and Z. Ying, “Gold nanospheres enhanced photothermal therapy in a rat model,” *Lasers in Surgery and Medicine*, vol. 50, no. 6, pp. 669–679, 2018.

[9] P. Singh, S. Pandit, V. Mokkapati, A. Garg, V. Ravikumar, and I. Mijakovic, “Gold nanoparticles in diagnostics and therapeutics for human cancer,” *Int J Mol Sci*, vol. 19, no. 7, p. 1979, 2018.

[10] N. J. Durr, T. Larson, D. K. Smith, B. A. Korgel, K. Sokolov, and A. Ben-Yakar, “Two-photon luminescence imaging of cancer cells using molecularly targeted gold nanorods,” *Nano Lett*, vol. 7, no. 4, pp. 941–945, 2007.

[11] B. Liu, C. Li, G. Chen et al., “Synthesis and optimization of $MoS_2@Fe_3O_4$-ICG/Pt(IV) nanoflowers for MR/IR/PA bioimaging and combined PTT/PDT/chemotherapy triggered by 808 nm laser,” *Advanced Science*, vol. 4, no. 8, p. 1600540, 2017.

[12] W. Zhang, Z. Guo, D. Huang, Z. Liu, X. Guo, and H. Zhong, “Synergistic effect of chemo-photothermal therapy using PEGylated graphene oxide,” *Biomaterials*, vol. 32, no. 33, pp. 8555–8561, 2011.

[13] Y. Zhao, W. Song, D. Wang et al., “Phase-shifted PFH@PLGA/Fe_3O_4 nanocapsules for MRI/US imaging and photothermal therapy with near-infrared irradiation,” *ACS Applied Materials & Interfaces*, vol. 7, no. 26, pp. 14231–14242, 2015.

[14] C. Xu, F. Gao, J. Wu et al., “Biodegradable nanotheranostics with hyperthermia-induced bubble ability for ultrasound imaging-guided chemo-photothermal therapy,” *International Journal of Nanomedicine*, vol. Volume 14, pp. 7141–7153, 2019.

[15] A. S. Hannah, G. P. Luke, and S. Y. Emelianov, “Blinking phase-change nanocapsules enable background-free ultrasound imaging,” *Theranostics*, vol. 6, no. 11, pp. 1866–1876, 2016.

[16] G. P. Luke, A. S. Hannah, and S. Y. Emelianov, “Super-resolution ultrasound imaging in vivo with transient laser-activated nanodroplets,” *Nano Letters*, vol. 16, no. 4, pp. 2556–2559, 2016.

[17] J. D. Dove, P. A. Mountford, T. W. Murray, and M. A. Borden, “Engineering optically triggered droplets for photoacoustic

imaging and therapy," *Biomedical Optics Express*, vol. 5, no. 12, pp. 4417–4427, 2014.

[18] Q. Chen, J. Yu, and K. Kim, "Review: optically-triggered phase-transition droplets for photoacoustic imaging," *Biomedical Engineering Letters*, vol. 8, no. 2, pp. 223–229, 2018.

[19] H. Zhao, M. Wu, L. Zhu et al., "Cell-penetrating peptide-modified targeted drug-loaded phase-transformation lipid nanoparticles combined with low-intensity focused ultrasound for precision theranostics against hepatocellular carcinoma," *Theranostics*, vol. 8, no. 7, pp. 1892–1910, 2018.

[20] P. S. Sheeran, J. D. Rojas, C. Puett, J. Hjelmquist, C. B. Arena, and P. A. Dayton, "Contrast-enhanced ultrasound imaging and in vivo circulatory kinetics with low-boiling-point nanoscale phase-change perfluorocarbon agents," *Ultrasound in Medicine & Biology*, vol. 41, no. 3, pp. 814–831, 2015.

[21] Y. Sun, Y. Wang, C. Niu et al., "Laser-activatible PLGA microparticles for image-guided cancer therapy in vivo," *Advanced Functional Materials*, vol. 24, no. 48, pp. 7674–7680, 2014.

[22] S. Jain, D. T. O'Hagan, and M. Singh, "The long-term potential of biodegradable poly(lactide-co-glycolide) microparticles as the next-generation vaccine adjuvant," *Expert Review of Vaccines*, vol. 10, pp. 1731–1742, 2014.

[23] I. Amjadi, M. Rabiee, and M. S. Hosseini, "Anticancer activity of nanoparticles based on PLGA and its co-polymer: in-vitro evaluation," *Iranian Journal of Pharmaceutical Research*, vol. 12, no. 4, pp. 623–634, 2013.

[24] M. Mir, N. Ahmed, and A. U. Rehman, "Recent applications of PLGA based nanostructures in drug delivery," *Colloids and Surfaces. B, Biointerfaces*, vol. 159, pp. 217–231, 2017.

[25] D. N. Kapoor, A. Bhatia, R. Kaur, R. Sharma, G. Kaur, and S. Dhawan, "PLGA: a unique polymer for drug delivery," *Therapeutic Delivery*, vol. 6, no. 1, pp. 41–58, 2015.

[26] X. Li, D. Wang, H. Ran et al., "A preliminary study of photoacoustic/ultrasound dual-mode imaging in melanoma using MAGE-targeted gold nanoparticles," *Biochemical and Biophysical Research Communications*, vol. 502, no. 2, pp. 255–261, 2018.

[27] S. Kang, S. H. Bhang, S. Hwang et al., "Mesenchymal stem cells aggregate and deliver gold nanoparticles to tumors for photothermal therapy," *ACS Nano*, vol. 9, no. 10, pp. 9678–9690, 2015.

[28] L. Cheng, C. Wang, L. Feng, K. Yang, and Z. Liu, "Functional nanomaterials for phototherapies of cancer," *Chemical Reviews*, vol. 114, no. 21, pp. 10869–10939, 2014.

[29] W. Guo, C. Guo, N. Zheng, T. Sun, and S. Liu, "CsxWO3Nanorods coated with polyelectrolyte multilayers as a multifunctional nanomaterial for bimodal imaging-guided photothermal/photodynamic cancer treatment," *Advanced Materials*, vol. 29, no. 4, 2017.

[30] D. Liu, L. Ma, L. Liu et al., "Polydopamine-encapsulated Fe_3O_4 with an adsorbed HSP70 inhibitor for improved photothermal inactivation of bacteria," *ACS Applied Materials & Interfaces*, vol. 8, no. 37, pp. 24455–24462, 2016.

Hindawi
BioMed Research International
Volume 2020, Article ID 7103647, 10 pages
https://doi.org/10.1155/2020/7103647

Research Article

A CT-Based Radiomics Approach for the Differential Diagnosis of Sarcomatoid and Clear Cell Renal Cell Carcinoma

Xiaoli Meng,[1] **Jun Shu**,[2] **Yuwei Xia**,[3] **and Ruwu Yang**[1]

[1]*Department of Radiology, Xi'an XD Group Hospital, Shaanxi University of Chinese Medicine, Feng Deng Road No. 97 Xi'an City 710077, China*
[2]*Department of Radiology, Xijing Hospital, Fourth Military Medical University, Changle West Road No. 127 Xi'an City 710032, China*
[3]*Huiying Medical Technology Co., Ltd., Room C103, B2, Dongsheng Science and Technology Park, Haidian District, Beijing City 100192, China*

Correspondence should be addressed to Ruwu Yang; yangruwu@126.com

Received 26 April 2020; Accepted 16 June 2020; Published 25 July 2020

Guest Editor: Zhiguo Zhou

This study was aimed at building a computed tomography- (CT-) based radiomics approach for the differentiation of sarcomatoid renal cell carcinoma (SRCC) and clear cell renal cell carcinoma (CCRCC). It involved 29 SRCC and 99 CCRCC patient cases, and to each case, 1029 features were collected from each of the corticomedullary phase (CMP) and nephrographic phase (NP) image. Then, features were selected by using the least absolute shrinkage and selection operator regression method and the selected features of the two phases were explored to build three radiomics approaches for SRCC and CCRCC classification. Meanwhile, subjective CT findings were filtered by univariate analysis to construct a radiomics model and further selected by Akaike information criterion for integrating with the selected image features to build the fifth model. Finally, the radiomics models utilized the multivariate logistic regression method for classification and the performance was assessed with receiver operating characteristic curve (ROC) and DeLong test. The radiomics models based on the CMP, the NP, the CMP and NP, the subjective findings, and the combined features achieved the AUC (area under the curve) value of 0.772, 0.938, 0.966, 0.792, and 0.974, respectively. Significant difference was found in AUC values between each of the CMP radiomics model ($0.0001 \le p \le 0.0051$) and the subjective findings model ($0.0006 \le p \le 0.0079$) and each of the NP radiomics model, the CMP and NP radiomics model, and the combined model. Sarcomatoid change is a common pathway of dedifferentiation likely occurring in all subtypes of renal cell carcinoma, and the CT-based radiomics approaches in this study show the potential for SRCC from CCRCC differentiation.

1. Introduction

Sarcomatoid renal cell carcinoma (SRCC) is a special subtype of renal cell carcinoma (RCC). Rather than an independent one, it is dedifferentiated from other histological subtypes of RCC both in epithelial and mesenchymal tissues [1]. SRCC is uncommon but highly aggressive, accounting for approximately 1/6 cases of advanced kidney cancers. In particular, it results in more dismal prognosis than the common subtype of clear cell renal cell carcinoma (CCRCC) [2]. According to the newly International Society of Urological Pathology (ISUP) grading system, RCC will be classified to grade IV when a sarcomatoid component was identified [3, 4].

Previous studies report that 45%-84% of SRCC have synchronous distant metastases at the time of diagnosis [5–7]. However, most systemic therapies developed for metastatic RCC are less effective in SRCC [8]. CCRCC can benefit from surgical resection even in the setting of metastasis, while for SRCC patients, surgical resection prior to systemic targeted therapies may worsen the outcomes because it might delay the administration of systemic therapy [2, 9]. Although ablative technique is an option for small renal masses, there is no

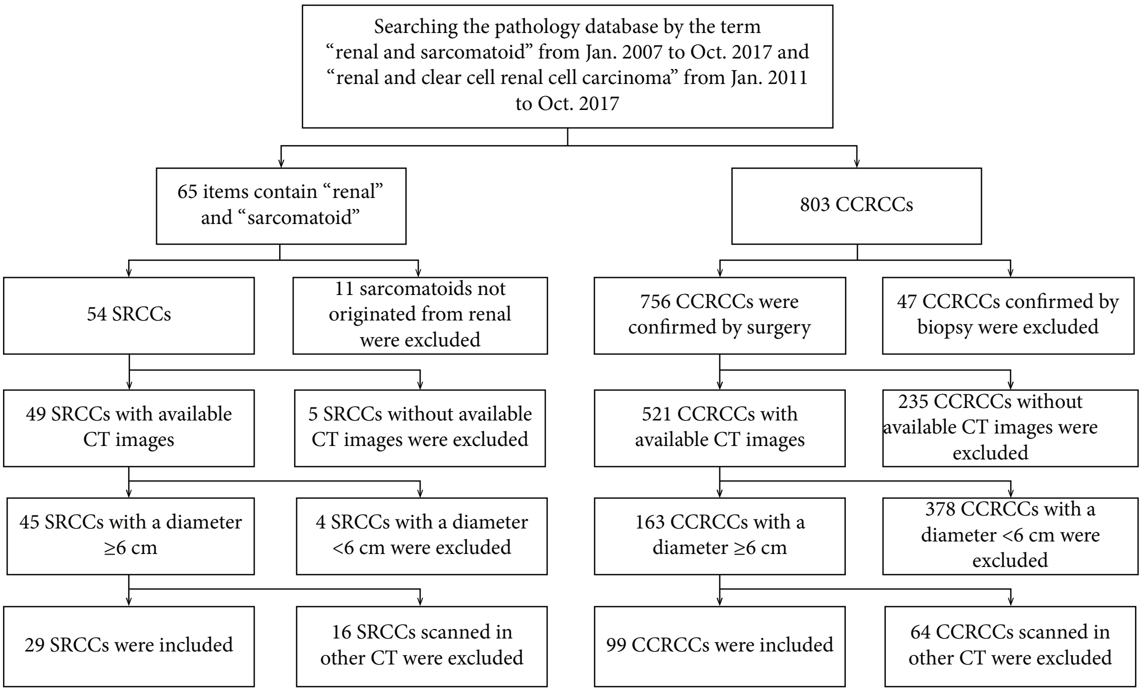

Searching the pathology database by the term "renal and sarcomatoid" from Jan. 2007 to Oct. 2017 and "renal and clear cell renal cell carcinoma" from Jan. 2011 to Oct. 2017
65 items contain "renal" and "sarcomatoid"
803 CCRCCs
54 SRCCs
11 sarcomatoids not originated from renal were excluded
756 CCRCCs were confirmed by surgery
47 CCRCCs confirmed by biopsy were excluded
49 SRCCs with available CT images
5 SRCCs without available CT images were excluded
521 CCRCCs with available CT images
235 CCRCCs without available CT images were excluded
45 SRCCs with a diameter ≥6 cm
4 SRCCs with a diameter <6 cm were excluded
163 CCRCCs with a diameter ≥6 cm
378 CCRCCs with a diameter <6 cm were excluded
29 SRCCs were included
16 SRCCs scanned in other CT were excluded
99 CCRCCs were included
64 CCRCCs scanned in other CT were excluded

Figure 1: Recruitment pathway for patients in this study.

enough support for using this technique in the small SRCC, and moreover, the infiltrative nature of SRCC tumors makes the determination of negative margin more difficult [2].

Preoperative diagnosis of SRCC is a challenging task. Since recognizable sarcomatoid elements just comprise a variable amount of the whole tumor, the use of biopsy is limited to confirm this entity [10, 11]. Despite some studies that reported that preoperative imaging could be used for predictive diagnosis of SRCC, the small sample sizes of such studies resulted in limited unconvincing consequences [12–15]. In this study, we explore to use of radiomics for the extraction and analysis of high-throughput features and both corticomedullary phase (CMP) and nephrographic phase (NP) images during CT imaging are concerned. Incorporated with clinical information, radiomics could further improve computer-aided diagnosis, prognosis, and predictive accuracy [16–18]. Hence, the purpose of this study is to build a CT-based radiomics approach that uses quantitative features and subjective CT findings for the differentiation of SRCC and CCRCC tumors in a relatively larger sample size.

2. Materials and Methods

2.1. Data Collection. The study was a retrospective study, and the informed patient consent was waived. Given the predominant number of CCRCC patients and a small number of SRCC patients in our hospital, the SRCC cases were collected from January 2007 to October 2017, and the CCRCC cases were collected from January 2011 to October 2017. To develop a study group with appropriate cases for building the radiomics models, the following inclusion criteria were used: (1) tumors originated from renal; (2) CT with contrast-enhanced CMP and NP images; and (3) tumor diameter ≥ 6 cm. The exclusion criteria were (1) CT images without sufficient quality due to motion artifacts or poor contrast injection; (2) the pathology confirmed as CCRCC only by biopsy; and (3) CT images not acquired in the specified scanner. Figure 1 shows the recruitment pathway for patient cases in this study.

2.2. Clinical Assessment of SRCC and Fuhrman Grades of CCRCC. The determination of SRCC and Fuhrman grade of CCRCC was gathered from the pathology reports, and one pathologist with 8 years of experience specializing in renal pathology reexamined all of the specimens. In this study, one RCC was considered to be SRCC when it resembles any form of sarcoma with or without atypical spindle cells, and a minimum proportion of sarcomatoid tumor was not required to make a diagnosis of sarcomatoid carcinoma. The criterion was in accordance with the ISUP 2012 Consensus Conference [3].

2.3. CT Imaging Protocol. The CT images were obtained by the scanner GE Light Speed VCT 64. The scanning parameters were as follows: tube voltage, 120 kVp; the tube current, 250-400 mA using automatic modulation; section thickness, 5 mm; and reconstruction interval, 5 mm. The patients were injected with 1.0 mL/kg of nonionic contrast material (iopromide, Ultravist 370; Bayer, Germany) at rate of 3.5 mL/s via the antecubital vein through a power injector. The CMP and the NP began 25 and 70 seconds after contrast injection, respectively.

2.4. Subjective CT Findings. Subjective CT findings for each patient were independently accessed and recorded in a blinded manner by two readers with 6 and 10 years of experience in abdominal imaging, and interreader variability was evaluated by using Kappa statistics. The solution to the

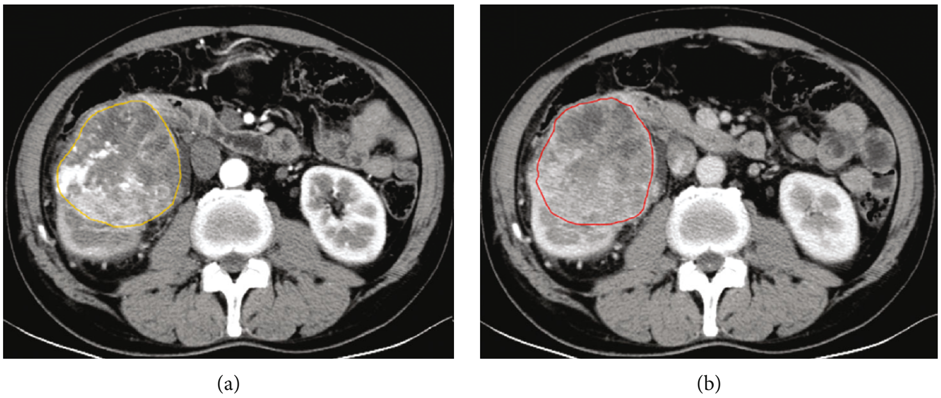


Figure 2: Manual delineation of a SRCC tumor of the same patient at different phases: (a) corticomedullary phase and (b) nephrographic phase.

divergences for the same case was to ask the readers jointly reviewing it to reach a consensus for further analysis.

Each reader evaluated the tumors for spread pattern, presence or absence of venous thrombus, intratumoral neovascularity, peritumoral neovascularity, calcification, and diameter. (a) Spread pattern was categorized into infiltrative or noninfiltrative. An infiltrative spread pattern was defined as invasion into the collecting system or neighboring organ, or interdigitation into adjacent renal parenchyma with the loss of a clear radiological capsule separating the lesion from adjacent parenchyma. (b) Intratumoral or peritumoral neovascularity means visible vascularity in the tumor parenchyma or perinephric fat adjacent to the mass. (c) Diameter was the largest transverse diameter measured at the maximum axial slice.

The difference in the subjective CT findings between the two patient groups was analyzed by using a chi-squared test or independent sample t-test, if appropriate. The findings without significant difference would not be integrated for the model building. All statistical analysis was completed by using SPSS (version 21.0).

2.5. Image Segmentation. To obtain the regions of interest (ROIs), the entire tumor of all contiguous slices was outlined except for the first and the last one which aimed to minimize the partial volume effects. Contouring was drawn slightly within the borders of the tumor masses. It included necrotic, cystic change, and hemorrhagic areas, while normal renal tissue and perinephric or sinus fat were excluded. The ROIs were drawn by the two readers both of whom were blinded to the clinical and pathological information. Figure 2 shows a representative example of manually outlined patient cases.

2.6. Radiomics Feature Extraction. To each phase image per patient, 1029 radiomics features were extracted through Radcloud platform (Huiying Medical Technology, http://radcloud.cn/). The radiomics features were divided into first-order features, shape features, and texture features. Shape features were calculated on the original ROI image, while first-order features and texture features were computed on the original ROI image and other derived images obtained by applying several filters, including exponential filter, square filter, square root filter, logarithm filter, and wavelet decomposition [19, 20]. Furthermore, texture features were derived from gray-level cooccurrence matrix (GLCM), gray-level run length matrix (GLRLM), and gray-level size zone matrix (GLSZM). As for the full details of radiomics data, please refer to supplemental S01.

2.7. Assessment of Delineation Consistency and Radiomics Feature Stability. To estimate the consistency of delineating CMP and NP images by the two readers, interclass correlation coefficient (ICC) values among 1029 features of each patient and each phase were calculated. If the ICC value of one patient in a phase was greater than 0.75, the manual delineation was considered in good agreement [21, 22] and the delineated image of the first reader would be used in follow-up model construction. Otherwise, the delineation would be repeated by the readers until the ICC met the requirement.

To ensure the stability and reproducibility of the radiomics features, the ICC was also calculated in each radiomics feature between two readers in the CMP and NP images. Features with an ICC greater than 0.75 were regarded as being in good agreement and retained for further radiomics analysis, and others were trimmed off.

2.8. Radiomics Feature Selection. Although some of the radiomics features with an ICC lower than 0.75 were removed, there still remained a great quantity of features. In order to decrease the high degree of redundancy and irrelevance, feature selection was conducted using the least absolute shrinkage and selection operator (LASSO) regression method in Anaconda3 platform (https://www.anaconda.com) with scikit-learn (https://scikit-learn.org/) and matplotlib packages (https://matplotlib.org/).

The LASSO regression method has been proved to be efficient and effective in the high-dimensional data analysis [23, 24]. It is aimed at minimizing the cost function and at keeping the features with nonzero coefficients. In this study, features passing the ICC screening were normalized by Z -score transform. Then, a 10-fold crossvalidation was carried out to choose the optimal parameters via the minimum of average mean square error. At last, the radiomics features with nonzero coefficients were used for further analysis.

2.9. Development, Diagnostic Performance, and Comparison of Classification. In order to evaluate the potential of CT-based radiomics and subjective CT findings for the differentiation of SRCC and CCRCC tumors, 5 models were built with the logistic regression method and fivefold crossvalidation strategy. These models differ from each other, since the selected features are the CMP, the NP, the CMP and NP, the subjective CT findings, and the combined features and subjective CT findings.

K-fold crossvalidation is a common model validation technique and widely used in machine learning studies. It randomly partitions the whole set into K subsets with equal or close size of data samples. Among the subsets, one is set as the validation set and the others as the training set. The experiment repeats K times to ensure that each of the subsets will be used exactly once as the validation set.

Specifically, to build the model with the subjective CT findings, the findings with statistically significant difference were concerned. In this study, infiltrative spread pattern, presence of venous thrombus, neovascularity, and calcification were set as 1, and noninfiltrative spread pattern and absence were set as 0. In the combined model, the CT findings were further selected by Akaike information criterion (AIC) and in the end integrated into a combined model with these selected CMP and NP features.

The quantitative indices used to assess the performance of these classification models were the receiver operating curve (ROC) and the area under the ROC curve (AUC), accuracy, sensitivity, and specificity. The confidence interval of AUC was computed by the exact binomial method. The ROC values of every two models were compared by using the DeLong test [25]. All the model construction, statistical computation, and figures were conducted in the Anaconda3 platform with scikit-learn and matplotlib.

3. Results

3.1. Clinical Characteristics. This study involved 128 patients (89 males and 39 females; mean age, 57.11 ± 10.52 years; range, 24-80 years). There were 29 (22.66%) SRCC and 99 (77.34%) CCRCC patients. No significant difference was found in gender ($p = 0.593$) or age ($p = 0.297$) between the patient groups, while it showed significant difference in tumor size ($p < 0.001$) and T stage ($p < 0.001$). Patient characteristics are shown in Table 1.

Specifically, among the SRCC patient cases, 23 were dedifferentiated from CCRCC tumors, followed by chromophobe RCC (4 cases), collecting duct carcinoma (1 case), and Xp11.2 translocation RCC (1 case), while among the CCRCC patient cases, the number of Fuhrman I, II, III, and IV was 4, 51, 40, and 4, respectively.

3.2. Interreader Agreement of Subjective CT Findings and Radiomics Features. Venous thrombus showed excellent agreement, with a Kappa value of 0.867 (95% CI (confidence interval): 0.598-1.000). Both peritumoral neovascularity and calcification showed good agreement, with Kappa values of 0.629 (95% CI: 0.489-0.761) and 0.787 (95% CI: 0.653-0.901), respectively. Besides, spread pattern and intratumoral neovascularity showed moderate agreement, with Kappa values of 0.571 (95% CI: 0.391-0.733) and 0.404 (95% CI: 0.270-0.537), respectively.

Table 1: The characteristics of SRCC and CCRCC patient groups.

	SRCC (29)	CCRCC (99)	Whole set (128)	p value
		Gender		
Male	19 (65.51%)	70 (70.71%)	89 (69.53%)	0.593[a]
Female	10 (34.48%)	29 (29.29%)	39 (30.47%)	
Age (yrs, mean ± std)	55.3 ± 14.0	57.6 ± 9.3		0.297[b]
Size (cm, mean ± std)	10.1 ± 3.0	7.7 ± 1.6		<0.001[b]
		T stage		
1b	6 (20.69%)	66 (66.67%)	72 (56.25%)	<0.001[c]
2	11 (37.93%)	19 (19.19%)	30 (23.44%)	
3	11 (37.93%)	13 (13.13%)	24 (18.75%)	
4	1 (3.45%)	1 (1.01%)	2 (1.56%)	

yrs: years; std: standard deviation; $p < 0.05$ is set as significant difference; [a]χ^2 test; [b]Student's t-test; [c]Fisher's exact test.

It was found that 1020 CMP radiomics features and 1023 NP radiomics features were with good interreader agreement, and ICC values, respectively, ranged from 0.786 to 0.999 and 0.765 to 0.999. In addition, 9 CMP radiomics features and 6 NP radiomics features were with ICC values less than 0.75, ranging from 0.148 to 0.748 and 0.102 to 0.696, respectively.

3.3. The Selection of Subjective CT Findings. It was found out that spread pattern ($p < 0.001$), venous thrombus ($p = 0.001$), peritumoral neovascularity ($p = 0.017$), calcification ($p = 0.005$), and diameter ($p < 0.001$) showed significant differences between the SRCC and CCRCC groups, while there was no significant difference of intratumoral neovascularity ($p = 0.073$) and thus, it was not used in model building. Subjective CT findings between the two patient groups are shown in Table 2.

3.4. The Selection of Radiomics Features. Using the regularized regression with the penalty (α is denoted as the weight of penalty term), the number of CMP features was reduced to 6 ($\alpha = 0.074$ and$-\log(\alpha) = 1.13$) and that of NP features was decreased to 29 ($\alpha = 0.028$ and$-\log(\alpha) = 1.55$) with nonzero coefficients. As shown in Figure 3, (a) shows the optimization of the parameter α by using LASSO, and (b) indicates the coefficients of selected CMP radiomics features, while (c) and (d) demonstrate the results of the parameter α and corresponding coefficients of selected NP radiomics features.

Table 2: Subjective CT findings of SRCC and CCRCC patient groups.

Imaging features	SRCC (29)	CCRCC (99)	p value
Spread pattern			
Infiltrative	16 (55.17%)	12 (12.12%)	<0.001[a]
Noninfiltrative	13 (44.83%)	87 (87.88%)	
Venous thrombus			
Present	6 (20.69%)	3 (3.03%)	0.001[a]
Absent	23 (79.31%)	96 (96.97%)	
Intratumoral neovascularity			
Present	14 (48.28%)	30 (30.30%)	0.073[a]
Absent	15 (51.72)	69 (60.70%)	
Peritumoral neovascularity			
Present	24 (82.76%)	58 (58.59%)	0.017[a]
Absent	5 (17.24%)	41 (41.41%)	
Calcification			
Present	13 (44.83%)	19 (19.19%)	0.005[a]
Absent	16 (55.17%)	80 (80.81%)	
Diameter (cm, mean ± std)	10.1 ± 3.0	7.7 ± 1.6	<0.001[b]

std: standard deviation; $p < 0.05$ is set as significant difference; [a]χ^2 test; [b]Student's t-test.

Specifically, the selected CMP features are 3 first-order features, 1 shape feature, and 2 texture features, and the selected NP features include 8 first-order features, 3 shape features, and 18 texture features. The coefficients of selected features are shown in Table 3.

For the combination model, subjective CT findings (spread pattern and calcification) with minimum AIC value were integrated into the selected CMP and NP radiomics features as the input for tumor differentiation. AIC values of subjective CT findings are shown in supplemental S02.

3.5. Development, Diagnostic Performance, and Comparison of Classification Models. Five radiomics approaches were explored via logistic regression. The subjective CT findings model considered 4 features (venous thrombus, peritumoral neovascularity, calcification, and diameter). For radiomics approaches, one utilized 6 CMP features, one used 29 NP features, and one concerned these 35 features (6 CMP features and 29 NP features). The last model contained those 35 radiomics features and 2 subjective CT findings.

The diagnostic performance of the five models is shown in Table 4. The subjective CT findings model and the CMP radiomics model showed inferior values of AUC, sensitivities, specificity, and accuracy when compared to the models using NP features, using CMP and NP features, and using the combined features. The CMP radiomics model showed the worst performance with AUC (0.772, 95% CI: 0.689-0.841), accuracy (78.12%), and sensitivity (65.52%), and the combined model achieved the best AUC (0.974, 95% CI: 0.924-0.992), accuracy (93.75%), and sensitivity (96.55%).

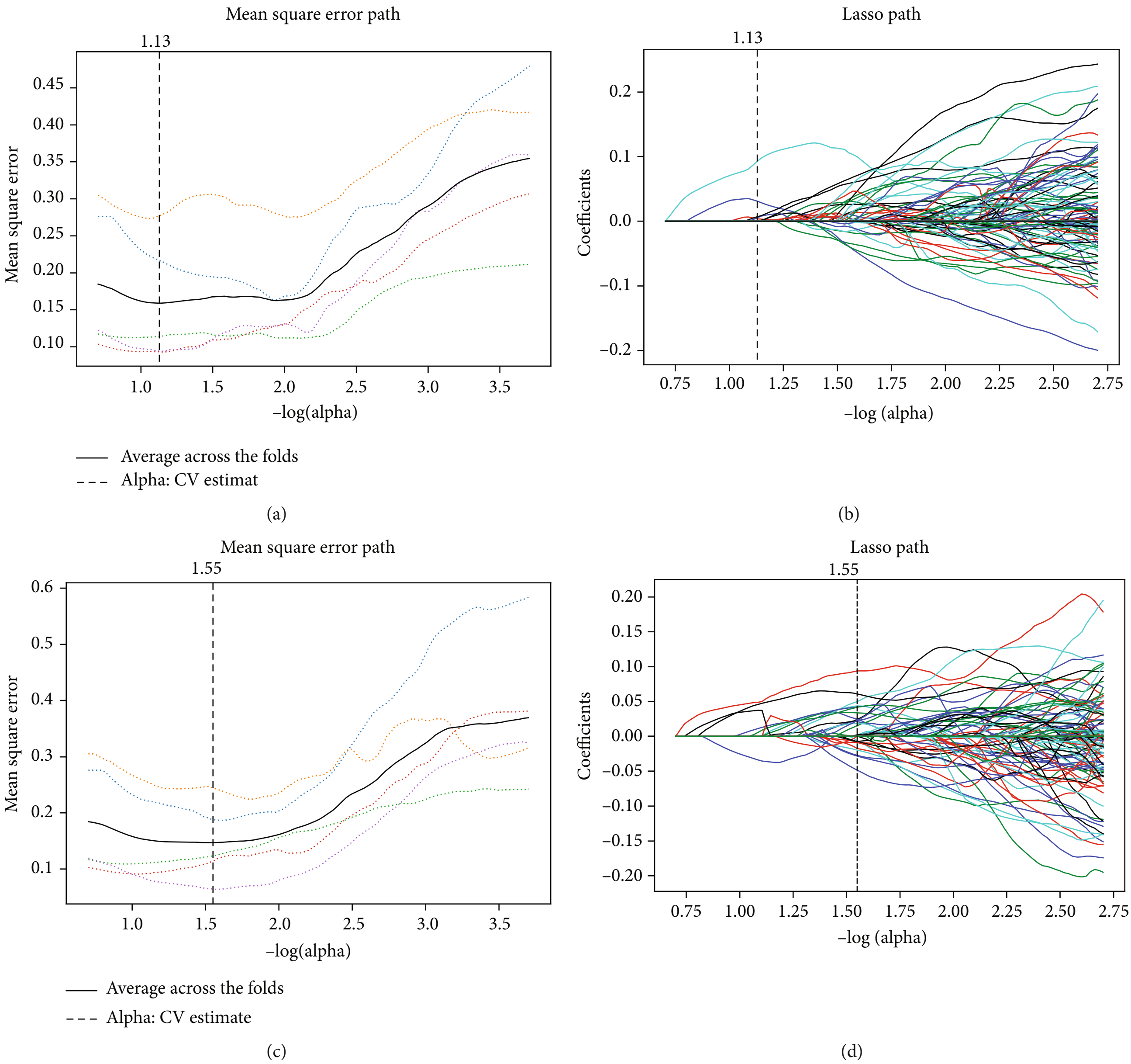
Figure 4 shows ROC curves of the five models. The model using combined features achieved the best AUC, followed by the model using the selected CMP and NP radiomics features, and the model using NP features. Relatively, the model using subjective CT findings or CMP radiomics features obtained relatively worse results. According to the DeLong test, there was no significant difference of the AUC values among the NP radiomics model, the CMP and NP radiomics model, and the combined model ($0.2245 \le p \le 0.6692$), as well as between the CMP radiomics model and the subjective CT findings model ($p = 0.7479$). On the other hand, each of the former three models showed significant improvement compared with each of the latter two models (the CMP model, $0.0001 \le p \le 0.0051$; the subjective CT findings model, $0.0006 \le p \le 0.0079$).

4. Discussion

Sarcomatoid change is believed to be a common pathway of dedifferentiation likely occurring in all subtypes of RCC tumors [4], and preoperative identification of the change is challenging but important in clinic. This study found that the CT-based radiomics approach could help discriminate the SRCC and CCRCC tumors and it also achieved superior performance over the subjective CT findings.

The AUC value using the selected CMP and NP radiomics features was significantly higher than that using the subjective findings, while incorporating the subjective CT findings into the model achieved no incremental predictive value. The AUC value of the NP radiomics model was higher with significant difference than that of the model using the CMP radiomics features. It was slightly lower than that of the CMP and NP radiomics model and that of the combined model with no statistical difference. Such an interesting finding indicated that the NP features are important in radiomics discrimination of SRCC and CCRCC tumors. In addition, the diagnosis power of the NP features better than the CMP features has been reported in machine learning-based CT images [26], which aimed for discriminating fat-poor renal angiomyolipoma from CCRCC. Thus, it might allow the omission of CMP acquisition to reduce the radiation dose in the differentiation of SRCC and CCRCC tumors.

The selected features showed that the "GrayLevelNonUniformity" of the GLSZM texture feature was the most frequently selected feature (Table 3). The feature quantifies the heterogeneity of a tumor. It appeared in the selected SRCC radiomics features with "squareroot_GrayLevelNonUniformity" (coefficient, 0.0926) and in the CCRCC features with "logarithm_GrayLevelNonUniformity" (coefficient, 0.0938) and with "wavelet-HLL_GrayLevelNonUniformity" (coefficient, -0.0063). One reason might be attributed to necrosis which was extremely highly frequent in tumors, for instance, the component of sarcomatoid carcinoma [15], and showed low or nonenhanced in CT images. Some previous studies [27, 28] also highlighted that low enhancement on CT images could be an independent predictor of the presence of high tumor grade of CCRCC, since CCRCC was more heterogeneous [29, 30]. Moreover, that lesion heterogeneity was a feature of malignancy and potential marker of survival, and the patients having heterogeneous tumors with lower uniformity might be with poorer survival [17]. Since SRCC tumors show heterogeneous appearance in multiphase CT imaging, this



(a)

(b)

(c)

(d)

Figure 3: Radiomics feature selection by using the LASSO regression method. The optimal α was selected using a tenfold crossvalidation via the minimum of average mean square error. To the CMP features, $\alpha = 0.074$ and $-\log(\alpha) = 1.13$ (a) and to the NP features, $\alpha = 0.028$ and $-\log(\alpha) = 1.55$ (c). (b) and (d), respectively, showed the coefficient profiles along the full path of possible α values in the CMP and the NP feature selection. In addition, dashed vertical lines were drawn at the optimal α based on the minimum of average mean square error in (a–d).

kind of gray-level nonuniformity feature is difficult to be quantified by a subjective finding until the radiomics emerged.

To the subjective CT findings, they figured out that infiltrative spread pattern, venous thrombus, peritumoral neovascularity, and calcification were more frequently showed in SRCC tumors, yet the Kappa values of these findings were relatively lower. The subjective CT findings to discriminate between SRCC and CCRCC tumors with poor Kappa values were also reported in [14]. However, there is little focus on the calcification of renal mass. One study [30] found that SRCC contained more calcium than RCC (28.6% vs 10.3%) which presumed that calcification is related to necrosis. Indeed, calcification was highly frequent in SRCC, particularly in the components of sarcomatoid carcinoma [15]. In this study, venous thrombus had the highest Kappa value of 0.867, while only 6 out of 29 SRCC and 3 out of 99 CCRCC manifested this feature, indicating the incidence was too low to be used. In short, the subjective findings were valuable but unstable or identifiable in a small cohort, which might account for unsatisfactory diagnostic performance of the subjective CT findings model.

At present, the main research of radiomics approaches for RCC analysis is the renal mass differentiation and nuclear

Table 3: The selected radiomics features and corresponding coefficients.

Selected CMP features	Coefficients
First-order features	
squareroot_Energy	0.0003
squareroot_Maximum	0.0057
wavelet-LHH_Skewness	0.0069
Shape features	
original_Minoraxis	0.0312
Texture features	
Gray-level run length matrix (GLRLM)	
exponential_RunVariance	0.0037
squareroot_GrayLevelNonUniformity	0.0926
Selected NP features	
First-order features	
wavelet-HLH_Skewness	-0.04831
wavelet-LHH_Median	-0.0272
wavelet-HHH_Median	-0.0076
squareroot_Energy	0.0002
wavelet-LLH_fskewness	0.0019
square_Kurtosis	0.0337
wavelet-LHL_Mean	0.0417
wavelet-LLH_Kurtosis	0.0613
Shape features	
original_SurfaceArea	2.85E-5
original_RunVariance	0.0210
original_SphericalDisproportion	0.0226
Texture features	
Gray-level cooccurrence matrix (GLCM)	
square_Idmn	-0.0196
square_Correlation	-0.0075
wavelet-HHH_ClusterProminence	0.0152
squareroot_DifferenceVariance	0.0422
Gray-level run length matrix (GLRLM)	
wavelet-LLL_ShortRunLowGrayLevelEmphasis	-0.0075
square_ShortRunLowGrayLevelEmphasis	0.0017
wavelet-HHH_RunVariance	0.0225
exponential_RunVariance	0.0242
exponential_RunEntropy	0.0246
exponential_ShortRunLowGrayLevelEmphasis	0.0499
Gray-level size zone matrix (GLSZM)	
square_ZoneVariance	-0.0285
wavelet-HLL_SizeZoneNonUniformityNormalized	-0.0173
wavelet-LLL_LowGrayLevelZoneEmphasis	-0.0086
wavelet-HLL_GrayLevelNonUniformity	-0.0063
wavelet-LLL_ZoneVariance	0.0022
wavelet-HHH_LargeAreaEmphasis	0.0183
logarithm_LargeAreaLowGrayLevelEmphasis	0.0437
logarithm_GrayLevelNonUniformity	0.0938

grade prediction [31], and few studies focus on the differentiation of SRCC and CCRCC tumors. To our knowledge, one study explored CT-based radiomics approaches to classify the SRCC and CCRCC tumors [14]. It involved 20 SRCC and 25 CCRCC cases, and both CT subjective findings and texture features were analyzed through noncontrast images. The study indicated that SRCC tumors (7.1 ± 2.7 cm) were significantly larger than CCRCC tumors (5.0 ± 2.9 cm), peritumoral neovascularity and the size of peritumoral vessels differed between the SRCC and CCRCC tumors in the subjective analysis, and SRCC tumors were with greater values of run length nonuniformity and gray-level nonuniformity features. In addition, the classification performance reached an AUC value of 0.81 ± 0.08 based on the combined textural features. Interestingly, as reported in [14], the current study also figured out SRCC tumors with significantly larger size over CCRCC tumors. Except for peritumoral neovascularity, subjective CT findings of spread pattern, venous thrombus, and calcification showed significant difference. In particular, the current study achieved superior performance on tumor differentiation through the analysis of multiphase CT images. It is worth noting that there are two other studies that concerned SRCC and CCRCC tumors by using MRI. One study [15] involved 11 patients with SRCC dedifferentiated from CCRCC tumors, and preoperative renal T1- and T2-weighted MRI were utilized. Compared to a normal renal cortex, it showed that the presence of the areas showing a hypovascular nature and markedly restricted diffusion might be characteristic findings of SRCC. The other study [32] collected 17 patients with SRCC and 17 patients with CCRCC, and dynamic T1-weighted MRI was analyzed. It indicated that the portion of segmented whole tumor with MRI signal suggestive of sarcomatoid involvement was correlated with histological examination, while the percentage of sarcomatoid differentiation was underestimated. Therefore, the current study differs itself from other studies [14, 15, 32] by using multiphase CT.

Multiphase CT was widely used in RCC analysis, and both CMP and NP have been proved to be important in renal lesion differentiation and staging [29]. CMP, the first phase of contrast enhancement, is between 25 and 70 seconds after the injection of contrast material, and the renal cortex enhances more brightly than the renal medulla. NP is the second phase when the contrast material filters through the glomeruli into the loops of Henle and the collecting tubules. At this time, the renal parenchyma becomes homogeneous, and the difference between a normal renal medulla and masses is well observed [33]. In the current study, 6 CMP features and 29 NP features were retrieved, and the NP radiomics approach achieved a significantly higher AUC value over the CMP radiomics approach. The reasons are manifold. First, various amounts of sarcomatoid differentiation are presented in SRCC tumors, which leads to inconsistent CMP imaging features, and in addition, the identified radiomics features cannot well differ the SRCC and CCRCC tumors. Second, NP image features have been reported as the most sensitive features for characterizing CCRCC from other subtypes of tumors, since the features coincided with the maximum tumor-to-kidney contrast [34]. Unfortunately,

Table 4: The diagnostic performance of the five radiomics approaches.

	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Subjective CT findings	0.792 (0.712-0.859)	78.12	82.76	75.76
CMP features	0.772 (0.689-0.841)	78.12	65.52	82.83
NP features	0.938 (0.881-0.973)	90.62	89.66	91.92
CMP + NP features	0.966 (0.918-0.990)	93.75	89.66	94.95
Combined features	0.974 (0.924-0.992)	93.75	96.55	88.89

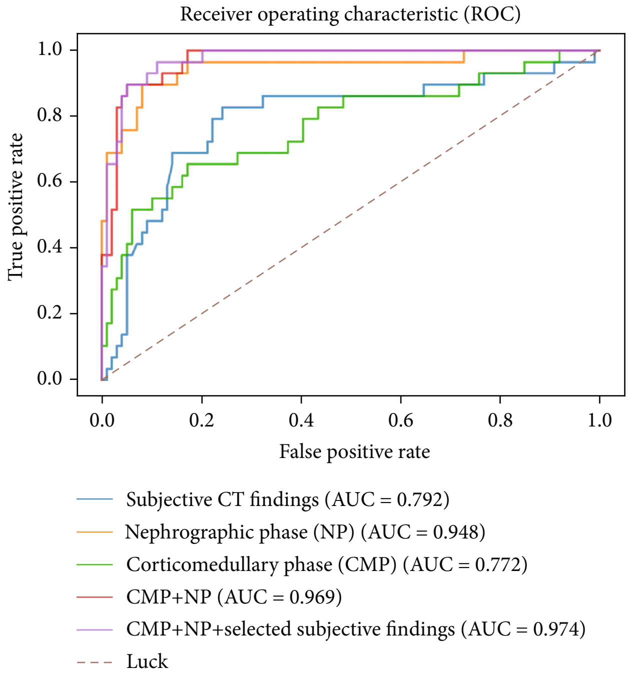


Figure 4: ROC curves of five radiomics approaches for differentiation of SRCC and CCRCC cases. The models are the subjective findings model (blue line), the CMP radiomics model (green line), the NP radiomics model (orange line), the CMP and NP radiomics model (red line), and the combined model (purple line). In addition, the brown dashed line shows the prediction distribution of random inputted features.

due to different purposes and specific data sets, there are conflicts of evidence. For instance, [20] indicated that there was no significant difference when CMP and MP radiomics features were independently used for low- and high-grade CCRCC staging, while [35] showed that CMP features resulted in better performance. Therefore, the exact reason why the selected NP features are better than the selected CMP features in the SRCC and CCRCC differentiation requires further investigation.

In the current study, SRCC and CCRCC tumors are with size larger than 6 cm. Two reasons account for this setting. First, one feature differing SRCC from other tumors is their larger tumor size [14]. During the data collection, it was found that almost all SRCC tumors had a diameter larger than 6 cm. Therefore, to reduce the effect of lesion size on the outcome, this study concerned a large tumor size. Second, a large tumor size benefits manual annotation of lesions, and good interreader agreement can be achieved. It should be noted that several studies concerned small RCC tumors. For instance, multiphase CT of tumor attenuation was explored for the differentiation between renal oncocytomas and CCRCC tumors (size ≤ 5 cm) [34, 36] and for distinguishing subtypes of RCC, angiomyolipoma, and oncocytoma tumors (≤4 cm) [21, 37]. Meanwhile, MR image texture features were also utilized for predicting histologic grade of CCRCC with tumor size ≤ 4 cm [38].

There are several limitations of the current study. First, due to the rarity of SRCC, data imbalance occurs. In order to overcome the risk of overfitting, the K-folder crossvalidation strategy was performed and the built radiomics approach was verified on an independent data set [39]. To overcome the issue of data imbalance, potential solutions include multicenter collaboration and nationwide and worldwide data sharing. Second, SRCC samples were not stratified according to the underlying diagnosis and the ratio of sarcomatoid component. RCC tumors with even a small component of sarcomatoid change might have an enormously adverse outcome, whereas the primary histologic appearance of SRCC does not have an impact on the prognosis [3, 11]. RCC that contains a sarcomatoid component is categorized to grade IV in the ISUP system, and there was a consensus that a minimum proportion of sarcomatoid tumor was not required to make a diagnosis of sarcomatoid carcinoma [3, 4]. Third, this study concerned SRCC and CCRCC with diameters larger than 6 cm. Pilot studies explored predict histologic grade of CCRCC less than 4 cm using CT and MRI, and statistically significant features were figured out [38, 40], which inspire our future investigation of small RCC samples. Furthermore, MRI features can be embedded into CT-based radiomics approach for improved differentiation [12, 15]. Last but not the least, novel techniques, such as full-automated image segmentation [41], feature dimension reduction [42], multiobjective optimization [43], and deep learning [44], could be further considered for improving classification performance.

5. Conclusion

Sarcomatoid change is believed to be a common pathway of dedifferentiation likely occurring in all subtypes of renal cell carcinoma, and preoperative identification of SRCC helps determine the therapeutic strategies. This study shows that the CT-based radiomics approaches could help discriminate the SRCC and CCRCC tumors and further improve patient management, treatment, and quality of life.

Data Availability

The clinical CT images used to support the findings of this study are available from the corresponding author upon request, while the radiomics date extracted from CT images is included within the supplementary information file.

Conflicts of Interest

The authors declare no conflict of interest.

Supplementary Materials

S01_Radiomics data: the radiomics data extracted from CT images was included within this supplementary file. S02_ AIC radiomics and subjective CT findings: CT findings were further selected by Akaike information criterion (AIC). This supplementary table shows the selected features and corresponding scores. Furthermore, AIC selected features were integrated into a combined model with selected CMP and NP features for tumor differentiation. *(Supplementary Materials)*

References

[1] G. M. Farrow, E. G. Harrison Jr., D. C. Utz, and W. H. Remine, "Sarcomas and sarcomatoid and mixed malignant tumors of the kidney in adults—part I," *Cancer*, vol. 22, no. 3, pp. 545–550, 1968.

[2] B. Shuch, G. Bratslavsky, W. M. Linehan, and R. Srinivasan, "Sarcomatoid renal cell carcinoma: a comprehensive review of the biology and current treatment strategies," *The Oncologist*, vol. 17, no. 1, pp. 46–54, 2011.

[3] B. Delahunt, J. C. Cheville, G. Martignoni et al., "The International Society of Urological Pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters," *American Journal of Surgical Pathology*, vol. 37, no. 10, pp. 1490–1504, 2013.

[4] H. Moch, A. L. Cubilla, P. A. Humphrey, V. E. Reuter, and T. M. Ulbright, "The 2016 WHO classification of tumours of the urinary system and male genital organs - part a: renal, penile, and testicular tumours," *European Urology*, vol. 70, no. 1, pp. 93–105, 2016.

[5] B. M. Mian, N. Bhadkamkar, J. W. Slaton et al., "Prognostic factors and survival of patients with sarcomatoid renal cell carcinoma," *Journal of Urology*, vol. 167, no. 1, pp. 65–70, 2002.

[6] J. C. Cheville, C. M. Lohse, H. Zincke et al., "Sarcomatoid renal cell carcinoma: an examination of underlying histologic subtype and an analysis of associations with patient outcome," *American Journal of Surgical Pathology*, vol. 28, no. 4, pp. 435–441, 2004.

[7] T. Cangiano, J. Liao, J. Naitoh, F. Dorey, R. Figlin, and A. Belldegrun, "Sarcomatoid renal cell carcinoma: biologic behavior, prognosis, and response to combined surgical resection and immunotherapy," *Journal of Clinical Oncology*, vol. 17, no. 2, pp. 523–528, 1999.

[8] L. C. Pagliaro, N. Tannir, K. Sircar, and E. Jonasch, "Systemic therapy for sarcomatoid renal cell carcinoma," *Expert Review of Anticancer Therapy*, vol. 11, no. 6, pp. 913–920, 2011.

[9] A. Kutikov, R. G. Uzzo, A. Caraway et al., "Use of systemic therapy and factors affecting survival for patients undergoing cytoreductive nephrectomy," *BJU International*, vol. 106, no. 2, pp. 218–223, 2010.

[10] B. Shuch, J. Said, J. C. La Rochelle et al., "Cytoreductive nephrectomy for kidney cancer with sarcomatoid histology—is up-front resection indicated and, if not, is it avoidable?," *Journal of Urology*, vol. 182, no. 5, pp. 2164–2171, 2009.

[11] M. de Peralta-Venturina, H. Moch, M. Amin et al., "Sarcomatoid differentiation in renal cell carcinoma a study of 101 cases," *American Journal of Surgical Pathology*, vol. 25, no. 3, pp. 275–284, 2001.

[12] M. Takeuchi, T. Kawai, T. Suzuki et al., "MRI for differentiation of renal cell carcinoma with sarcomatoid component from other renal tumor types," *Abdominal Imaging*, vol. 40, no. 1, pp. 112–119, 2015.

[13] J. R. Young, J. A. Young, D. J. A. Margolis et al., "Sarcomatoid renal cell carcinoma and collecting duct carcinoma discrimination from common renal cell carcinoma subtypes and benign RCC mimics on multiphasic MDCT," *Academic Radiology*, vol. 24, no. 10, pp. 1226–1232, 2017.

[14] N. Schieda, R. E. Thornhill, M. Al-Subhi et al., "Diagnosis of sarcomatoid renal cell carcinoma with CT: evaluation by qualitative imaging features and texture analysis," *American Journal of Roentgenology*, vol. 204, no. 5, pp. 1013–1023, 2015.

[15] M. Takeuchi, M. Urano, M. Hara, Y. Fujiyoshi, H. Inagaki, and Y. Shibamoto, "Characteristic MRI findings of sarcomatoid renal cell carcinoma dedifferentiated from clear cell renal carcinoma radiological-pathological correlation," *Clinical Imaging*, vol. 37, no. 5, pp. 908–912, 2013.

[16] P. Lambin, E. Rios-Velazquez, R. Leijenaar et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[17] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[18] H. J. W. L. Aerts, E. R. Velazquez, R. T. Leijenaar et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, pp. 4006–4013, 2014.

[19] J. J. M. Van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[20] J. Shu, Y. Tang, J. Cui et al., "Clear cell renal cell carcinoma: CT-based radiomics features for the prediction of Fuhrman grade," *European Journal of Radiology*, vol. 109, pp. 8–12, 2018.

[21] R. Yang, J. Wu, L. Sun et al., "Radiomics of small renal masses on multiphasic CT: accuracy of machine learning-based classification models for the differentiation of renal cell carcinoma and angiomyolipoma without visible fat," *European Radiology*, vol. 30, no. 2, pp. 1254–1263, 2020.

[22] E. Cui, Z. Li, C. Ma et al., "Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics," *European Radiology*, vol. 30, no. 5, pp. 2912–2921, 2020.

[23] Y. Q. Huang, C. H. Liang, L. He et al., "Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer," *Journal of Clinical Oncology*, vol. 34, no. 18, pp. 2157–2164, 2016.

[24] Z. Ma, M. Fang, Y. Huang et al., "CT-based radiomics signature for differentiating Borrmann type IV gastric cancer from primary gastric lymphoma," *European Journal of Radiology*, vol. 91, pp. 142–147, 2017.

[25] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 1, pp. 837–845, 1988.

[26] T. Hodgdon, M. D. McInnes, N. Schieda, T. A. Flood, L. Lamb, and R. E. Thornhill, "Can quantitative CT texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced CT images?," *Radiology*, vol. 276, no. 3, pp. 787–796, 2015.

[27] Z. Feng, P. Rong, P. Cao et al., "Machine learning-based quantitative texture analysis of CT images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma," *European Radiology*, vol. 28, no. 4, pp. 1625–1633, 2018.

[28] Y.-H. Zhu, X. Wang, J. Zhang, Y. H. Chen, W. Kong, and Y. R. Huang, "Low enhancement on multiphase contrast-enhanced CT images: an independent predictor of the presence of high tumor grade of clear cell renal cell carcinoma," *American Journal of Roentgenology*, vol. 203, no. 3, pp. W295–W300, 2014.

[29] B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, and K. Miles, "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival," *European Radiology*, vol. 22, no. 4, pp. 796–802, 2012.

[30] W. W. Daniel Jr., G. W. Hartman, D. M. Witten, G. M. Farrow, and P. P. Kelalis, "Calcified renal masses: a review of ten years experience at the Mayo Clinic," *Radiology*, vol. 103, no. 3, pp. 503–508, 1972.

[31] R. Suarez-Ibarrola, M. Basulto-Martinez, A. Heinze, C. Gratzke, and A. Miernik, "Radiomics applications in renal tumor assessment: a comprehensive review of the literature," *Cancers*, vol. 12, no. 6, p. 1387, 2020.

[32] D. Jeong, D. H. Natarajan Raghunand, M. Poch, K. Jeong, B. Eck, and J. Dhillon, "Quantification of sarcomatoid differentiation in renal cell carcinoma on magnetic resonance imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 4, pp. 373–382, 2018.

[33] S. Sheth, J. C. Scatarige, K. M. Horton, F. M. Corl, and E. K. Fishman, "Current concepts in the diagnosis and management of renal cell carcinoma: role of multidetector CT and three-dimensional CT," *Radiographics*, vol. 21, suppl_1, pp. S237–S254, 2001.

[34] G. Gakis, U. Kramer, D. Schilling, S. Kruck, A. Stenzl, and H. P. Schlemmer, "Small renal oncocytomas: differentiation with multiphase CT," *European Journal of Radiology*, vol. 80, no. 2, pp. 274–278, 2011.

[35] J. Shu, D. Wen, Y. Xi et al., "Clear cell renal cell carcinoma: machine learning-based computed tomography radiomics analysis for the prediction of WHO/ISUP grade," *European Journal of Radiology*, vol. 121, article 108738, 2019.

[36] F. Gentili, I. Bronico, U. Maestroni et al., "Small renal masses (≤ 4 cm): differentiation of oncocytoma from renal clear cell carcinoma using ratio of lesion to cortex attenuation and aorta-lesion attenuation difference (ALAD) on contrast-enhanced CT," *La Radiologia Medica*, 2020.

[37] P. M. Pierorazio, E. S. Hyams, S. Tsai et al., "Multiphasic enhancement patterns of small renal masses (≤4 cm) on preoperative computed tomography: utility for distinguishing subtypes of renal cell carcinoma, angiomyolipoma, and oncocytoma," *Urology*, vol. 81, no. 6, pp. 1265–1272, 2013.

[38] S. Haji-Momenian, Z. Lin, B. Patel et al., "Texture analysis and machine learning algorithms accurately predict histologic grade in small (< 4 cm) clear cell renal cell carcinomas: a pilot study," *Abdominal Radiology*, vol. 45, no. 3, pp. 789–798, 2020.

[39] B. Kocak, E. S. Durmaz, C. Erdim, E. Ates, O. K. Kaya, and O. Kilickesmez, "Radiomics of renal masses: systematic review of reproducibility and validation strategies," *American Journal of Roentgenology*, vol. 214, no. 1, pp. 129–136, 2020.

[40] K. Moran, J. Abreu-Gomez, S. Krishna et al., "Can MRI be used to diagnose histologic grade in T1a (< 4 cm) clear cell renal cell carcinomas?," *Abdominal Radiology*, vol. 44, no. 8, pp. 2841–2851, 2019.

[41] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.

[42] L. H. Nguyen and S. Holmes, "Ten quick tips for effective dimensionality reduction," *PLoS Computational Biology*, vol. 15, no. 6, article e1006907, 2019.

[43] Z. Zhou, S. Li, G. Qin, M. Folkert, S. Jiang, and J. Wang, "Multi-objective-based radiomic feature selection for lesion malignancy classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 194–204, 2020.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

Hindawi
BioMed Research International
Volume 2020, Article ID 5615371, 11 pages
https://doi.org/10.1155/2020/5615371

Research Article

To Align Multimodal Lumbar Spine Images via Bending Energy Constrained Normalized Mutual Information

Shibin Wu,[1] Pin He,[2,3] Shaode Yu,[4] Shoujun Zhou,[1] Jun Xia,[2] and Yaoqin Xie[1]

[1] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[2] *Department of Radiology, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen 518035, China*
[3] *National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen 518116, China*
[4] *Department of Radiation Oncology, University of Texas, Southwestern Medical Center, Dallas, TX 75390, USA*

Correspondence should be addressed to Yaoqin Xie; yq.xie@siat.ac.cn

Received 9 April 2020; Accepted 15 June 2020; Published 11 July 2020

Academic Editor: Tobias De Zordo

To align multimodal images is important for information fusion, clinical diagnosis, treatment planning, and delivery, while few methods have been dedicated to matching computerized tomography (CT) and magnetic resonance (MR) images of lumbar spine. This study proposes a coarse-to-fine registration framework to address this issue. Firstly, a pair of CT-MR images are rigidly aligned for global positioning. Then, a bending energy term is penalized into the normalized mutual information for the local deformation of soft tissues. In the end, the framework is validated on 40 pairs of CT-MR images from our in-house collection and 15 image pairs from the SpineWeb database. Experimental results show high overlapping ratio (in-house collection, vertebrae 0.97 ± 0.02, blood vessel 0.88 ± 0.07; SpineWeb, vertebrae 0.95 ± 0.03, blood vessel 0.93 ± 0.10) and low target registration error (in-house collection, $\leq 2.00 \pm 0.62$ mm; SpineWeb, $\leq 2.37 \pm 0.76$ mm) are achieved. The proposed framework concerns both the incompressibility of bone structures and the nonrigid deformation of soft tissues. It enables accurate CT-MR registration of lumbar spine images and facilitates image fusion, spine disease diagnosis, and interventional treatment delivery.

1. Introduction

Spine is the backbone of body trunk. It protects the most significant nerve pathway in the spinal cord and the body. On the other hand, spine injury and disorders affect up to 80% world population and may cause deformity and disability, which become a major health and social problem [1–3]. For instance, the lumbar degenerative disease accompanied by pathological changes might result in lumbocrural pain, neural dysfunction, instability of facet joints, and spino-pelvic sagittal imbalance, and thus, the quality of life decreases dramatically. In addition, due to the aging population, the global burden relating to spinal disease remedy is expected to raise significantly in the next decades.

To align intrapatient multimodal images, such as computerized tomography (CT) and magnetic resonance (MR), benefits clinical diagnosis, treatment planning, and delivery for lumbar spinal diseases [4, 5]. However, few methods were dedicated to matching lumbar spine images. Panigrahy et al. developed a method for CT-MR cervical spine images which needed anatomical landmarks to guide image registration [6]. Palkar and Mishra combined different orthogonal wavelet transforms with various transform sizes for CT-MR spine image fusion, while interactive localization of control points was required [7]. Tomazevic et al. implemented an approach for rigid alignment of volumetric CT or MR to X-ray images [8]. To simplify the registration problem in real-world scenarios, images were acquired from a cadaveric lumbar spine phantom and three-dimensional (3D) images contained only one of the five vertebrae. Otake et al. proposed a registration method for 3D and planar images which was used for spine intervention and vertebral labeling in the presence of anatomical deformation [9]. Harmouche et al. designed an articulated model for MR and X-ray spine images [10]. Hille et al. presented an interactive framework, and rough annotation of the center regions in different modalities was used to guide the registration [11].

Accurate alignment of intrapatient CT-MR images is challenging. From the anatomy, human spine consists of inflexible vertebrae surrounded by soft tissues, such as nerves, vessels, and muscles. Moreover, the vertebrae of lumbar spine are connected by facet joints in the back, which allows for forward and backward extension and twisting movements. Moreover, spinal deformity imposes difficulties on multimodal image registration. Specifically, during image acquisition, patients can lay flatly for a short time due to pain, and subsequently, motion becomes unavoidable. Last but not the least, there are intrinsic differences between CT and MR imaging.

Figure 1 shows a pair of intrapatient CT-MR images. It is found that in CT images, the lumbar spine region easily highlights itself from the rest of soft tissues (the top row), while in MR images, soft tissues show various intensities and in particular, it might be hard to distinguish rigid bones from soft tissues (the bottom row). In the figure, soft tissues in MR images are with various contrast than those in CT images (red arrows), undesirable artifacts caused by the bias field are observed in MR images (green arrows), and these pairs of images show different imaging field of views. It is obvious that these facets pose difficulties in image registration.

2. Related Works

Image registration is important in medical image analysis [12, 13, 14]. Based on similarity metrics, registration methods could be generally classified into intensity- and feature-based methods. Among the intensity-based methods, mutual information (MI) is well known, and it was primarily presented for MR breast image alignment [15]. Afterwards, the metric is used in multimodal medical image registration [16]. For specific applications, MI has been modified to enhance the performance of image registration. For instance, normalized MI (NMI) was proposed for invariant entropy measure [17], regional MI was implemented to capture volume changes when local tissue contrast varied in serial MR images [18], localized MI was designed for atlas matching and prostate segmentation [19], conditional MI was developed to incorporate joint histogram and intensity distribution for image description [20], self-similarity weighted αMI was presented for handheld ultrasound and MR image alignment [21], and MI was also advanced with spatially encoded information [22].

Feature-based methods aim to quantify detected landmarks with features for image registration. Ou et al. collected multiscale multiorientation Gabor features to weight mutual-saliency points for matching [23]. Zhang et al. used scale-invariant features and corner descriptors for lung image registration [24]. Heinrich et al. designed modality independent neighborhood descriptor (MIND) which extracted the distinctive structure in small image patches for multimodal deformation registration [25]. Via principal component analysis of deformation, a low-dimension statistical model was learned [26]. Toews et al. combined invariant features of volumetric geometry and appearance for image alignment [27]. Determined by the moments of image patches, a self-similarity inspired local descriptor was presented [28]. Jiang et al. designed a discriminative local derivative pattern which encoded images of different modalities into similar representation [29]. Woo et al. combined spatial and geometric context of detected landmarks [30], and Carvalho et al. considered intensity and geometrical features [31] into a similarity metric. Weistrand and Svensson constrained image registration with anatomical landmarks for local tissue deformation [32].

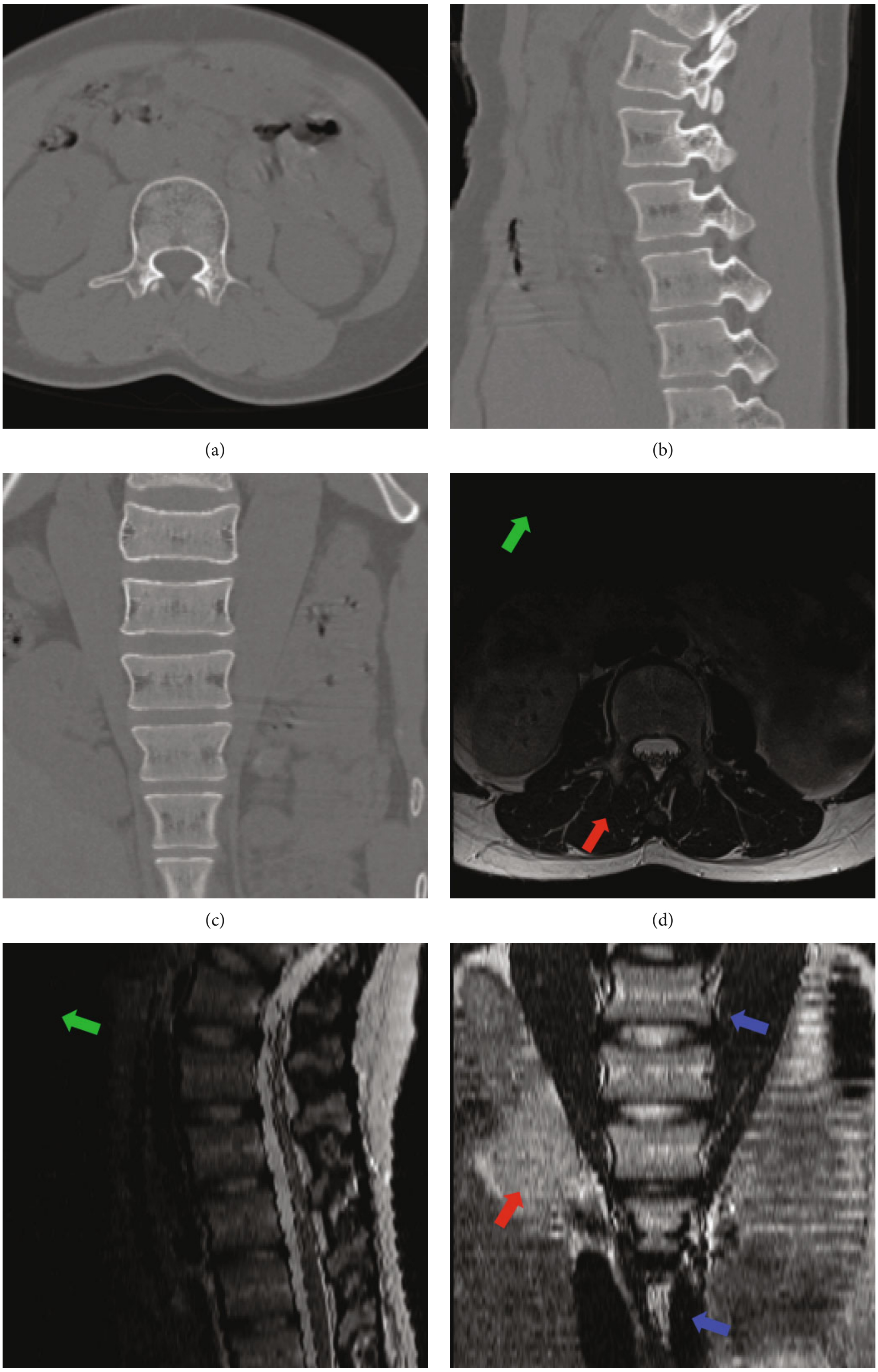
Embedding a proper penalty term into a similarity metric is helpful in specific applications. Rueckert et al. used a term to regularize the local deformation to be smooth in breast MR image registration [33]. Rohlfing et al. designed a local volume preservation constraint, assuming the soft tissues incompressible in small deformation [34]. Staring et al. proposed a rigidity penalty and modeled the local transform when thorax images with tumors were aligned [35]. To model fetal brain motion, Chen et al. utilized the total-variation regularization and a penalty was adopted toward piece-wise convergence [12]. Due to local tissue rigidity characteristics, Ruan et al. added a regularization term for aligning inhale-exhale CT lung images [36]. Fiorino et al. designed the Jacobian-volumehistogram of deforming organs to evaluate the parotid shrinkage [37].

This study proposes a coarse-to-fine framework to address the registration of intrapatient CT-MR images of lumbar spine. It develops a similarity metric that penalizes a bending energy term into NMI for local deformation of soft tissues. The most similar work is from the comparison of bending energy penalized global and local MI metrics in aligning positron emission tomography and MR images [38], while this study differs itself from the proposed coarse-to-fine registration framework, the bending energy penalized NMI (BEP-NMI) and the application to CT-MR lumbar spine images.

3. Materials and Methods

3.1. Data Collection. Two data sets were analyzed. One is our in-house collection which contains 40 pairs of lumbar spine images from the Department of Radiology, Shenzhen Second People's Hospital, the First Affiliated Hospital of Shenzhen University. CT images were acquired through SIEMENS SOMATO. The voxel resolution is $0.35 \times 0.35 \times 1.00$ mm3, and the matrix size is 512×512 with 180 ± 25 slices. T_2-weighted MR images were acquired using a 1.5 Tesla scanner (SIEMENS Avanto). The physical resolution is $0.7 \times 0.7 \times 3$ mm3, the matrix size is 256×256, and the slice number ranges between 60 and 75.

The other data set is accessible online, namely SpineWeb (http://spineweb.digitalimaginggroup.ca). It includes 15 image pairs of lumbar spine. The physical resolution of CT images is



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Perceived visual difference between CT and MR images of lumbar spine from three perspective views. The difference of imaging characteristics, fields of view, and unavoidable motion make the registration challenging. Red arrows show different imaging contrast, green arrows direct to the undesirable artifact of bias field in MR images, and blue arrows indicate different field of views. Note that images are cropped and scaled for display purpose.

$0.27 \times 0.27 \times 2.50$ mm3, the image size is 512×512, and the slice number is 77 per volume. The resolution of T_1-weighted MR images is $0.39 \times 0.39 \times 5.00$ mm3, the image size is 512×512, and each volume contains 42 slices.

3.2. The Proposed Framework. The proposed framework consists of two steps both of which use intensity-based image registration methods. An intensity-based registration method can be treated as an optimization problem, and the similarity metric S performs as the cost function. Given a fixed image IF : $\Omega 1 \in R3$ and a moving image $IM : \Omega 2 \in R3$ in 3D space, image registration aims for mapping the moving image IM to the space of the fixed image IF guided by the metric S. When an additional regularization term of P is penalized into S, the registration problem can be formulated as,

$$\begin{gathered}\widehat{\mathrm{T}} = \arg \min_{T} \mathrm{C}(\mathrm{T}\,; \mathrm{I_F}\,; \mathrm{I_M}) \ \text{w.r.t.} \mathrm{C}(\mathrm{T}\mu\,; \mathrm{I_F}\,; \mathrm{I_M}) \\ \mathrm{C}(\mathrm{T}_\mu\,; \mathrm{I_F}, \mathrm{I_M}) = \mathrm{S}(\mathrm{T}_\mu\,; \mathrm{I_F}, \mathrm{I_M}) + \lambda \mathrm{P}(\mathrm{T}_\mu)\end{gathered} \qquad (1)$$

where T is a transform model, λ compromises the metric S and the regularity term P, μ is the transform coefficients, and $T\mu$ is the initialized model by μ.

Figure 2 illustrates the proposed framework. It indicates a rigid registration stage and a hierarchical deformation stage, and NMI and BEP-NMI, respectively, perform as the similarity metric. Moreover, adaptive stochastic gradient descent (ASGD) [39] is applied for hyperparameter optimization. Specifically, an affine transformation with 12 degrees of freedom is employed in the first stage, and a B-spline elastic model is used for free-form deformation in the second stage.

3.2.1. Rigid Registration. An affine transform model is used here. The transform $T : \Omega 2 \rightarrow \Omega 1$ can be formulated by

$$\mathrm{T}_\mu(\mathrm{x}) = \mathrm{R}(\mathrm{x} - \mathrm{c}) + \mathrm{t} + \mathrm{c} \qquad (2)$$

where R is a matrix that contains the rotation, scale, and shear coefficients, c is the center of rotation, t is a translation vector, and μ is a vector of 12 degrees of freedom in volumetric image registration.

Rigid registration attempts for global positioning of the whole body, and thus, an initial alignment of lumbar spine. A 3-level recursive pyramid denotes smoothing that downsamples the source volumes by a factor of 2. Besides, the metric NMI and the affine transform are employed in each scale.

3.2.2. Hierarchical Deformation. Hierarchical deformation is a coarse-to-fine adjustment procedure [40]. This setup utilizes Gaussian pyramid without downsampling to match images from the global structures toward the fine details.

B-spline transform. The B-splines are used to depict the local shape difference between the lumbar vertebrae. To construct the B-spline based free-form deformation model, let $\Omega = \{(x, y, z) \mid 0\,6\,x < X, 0\,6\,y < Y, 0\,6\,z < Z\}$ be a spatial domain of a 3D image. A lattice $(px \times py \times pz)$ of control points is denoted as Ψ, spanning the integer grid in Ω, and Φijk denotes the control point at (i, j, k) on the mesh Ψ. Then, the elastic model can be expressed as a 3D tensor product of the uniform B-spline of order 3 as below,

$$T_I(x, y, z) = \sum_{I=0}^{3} \sum_{m-0}^{3} \sum_{n=0}^{3} B_I(u_1) B_m(u_2) B_n(u_3) \Phi_{i+l, j+m, k+n} \qquad (3)$$

where $i = \lfloor x/P_x \rfloor - 1, j = \lfloor y/P_Y \rfloor - 1, k = \lfloor z/P_z \rfloor - 1, u_1 = x/P_x - \lfloor x/P_x \rfloor, u_2 = (y/P_y)\lfloor y/P_y \rfloor, u_3 = z/P_Z - \lfloor z/P_z \rfloor$, and Bl repents the 1^{th} basis function of the B-spline,

$$\begin{cases} B_0(u) = (1-u)^3/6, \\ B_1(u) = \left(3u^3 - 6u^2 + 4\right)/6 \\ B_2(u) = \left(-3u^3 + 3u^2 + 3u + 1\right)/6 \\ B_3(u) = u^3/6 \end{cases} \qquad (4)$$

where $0\,6\,u < 1$. The basic functions weigh the contribution of each control point to $Tl(x, y, z)$ based on its distance to the point (x, y, z).

Since the B-splines can be locally controlled, it makes the computation efficient for a large number of control points. In particular, changing a control point affects only the transforms of its local neighborhood.

BEP-NMI. The metric MI is preferred in multimodal image registration. Given IF and IM with intensity bins of f and m, MI is quantified from a joint probability function $p($ IF, $IM)$ and marginal probability distribution functions.

of $p(f) = Pf\,\{p(f, m)\}$ and $p(m) = Pm\,\{p(f, m)\}$. The metric MI between a pair of images, IF and IM, can be described as

$$\begin{aligned}\mathrm{MI}(\mathrm{I_F}\,; \mathrm{I_M}) &= \mathrm{H}(\mathrm{I_F}\,) + \mathrm{H}(\mathrm{I_M})..\mathrm{H}(\mathrm{I_F}\,; \mathrm{I_M}) \\ &= \sum_{f \in F} \sum_{m \in M} p(f, m) \log \left\langle \frac{p(f, m)}{p(f)p(m)} \right\rangle\end{aligned} \qquad (5)$$

where $H(\mathrm{IF})$ and $H(IM)$ are the marginal entropy and the $H(\mathrm{IF}, IM)$ is the joint entropy of IF and IM.

The metric NMI is more robust to the change of overlapped tissue regions. It uses a Parzen-window approach to estimate the probability density function. The entropy of a fixed image IF is defined as $H(\mathrm{IF}) = -Pf \in F p(f) logp(f)$, where $p(f)$ is a probability distribution estimated by using Parzen-windows. The entropy of a moving image IM can be computed in a similar way. And subsequently, the NMI between IF and IM can be presented as

$$\mathrm{NMI}(\mathrm{I_F}\,; \mathrm{I_M}) = \frac{\mathrm{H}(\mathrm{I_F}\,) + \mathrm{H}(\mathrm{I_M})}{\mathrm{H}(\mathrm{I_F}\,; \mathrm{I_M})} \qquad (6)$$

In order to regularize the B-spline deformation and to prevent the rigid structures from being smoothed, a BEP term $P(u)$ is added to the NMI. The new cost function, BEP-NMI, is formulated as

$$\mathrm{C}(\mu) = \mathrm{y}_1 S(\mu) + y_2 P(\mu) \qquad (7)$$

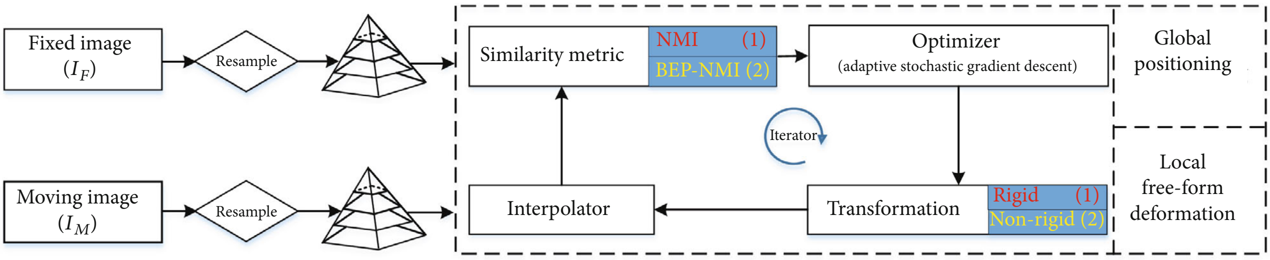


Figure 2: The proposed coarse-to-fine framework for aligning CT-MR lumbar spine images. It consists of two stages. The first stage is for global positioning via NMI based rigid registration (highlighted in red), and the second stage is for the local deformation of soft tissues via the bending energy penalized NMI (highlighted in yellow). Both stages utilize the same workflow for iterative optimization.

where $\gamma 1$ and $\gamma 2$ are predefined constants to weigh between global similarity and local regularity. In this study, off-line experiments indicated that $\gamma 1 = \gamma 2 = 1$ was a good choice.

The penalty terms are commonly based on the first or second-order spatial derivatives of the transform [35, 36]. In this study, the BEP term is composed of the second-order derivatives [35, 40] in the volumetric space,

$$P_{BEP}(\mu) = \int_V \left\{ \left(\frac{\partial^2 T}{\partial x^2}\right)^2 + \left(\frac{\partial^2 T}{\partial y^2}\right)^2 + \left(\frac{\partial^2 T}{\partial z^2}\right)^2 + 2\left(\frac{\partial^2 T}{\partial x \partial y}\right)^2 + 2\left(\frac{\partial^2 T}{\partial y \partial z}\right)^2 + 2\left(\frac{\partial^2 T}{\partial z \partial x}\right)^2 \right\} dxdydz \quad (8)$$

where V is a 3D image. The Equation (8) can be approximated as a discretely sampled sum over the volume V as below,

$$\mathrm{P}_{\mathrm{BEP}} = \frac{1}{N_V} \sum_{x \in V} \Phi T(x, y, z) \quad (9)$$

where N is the number of voxels in V, and Φ denotes a sum of the squared second-order derivatives of T inside the integral part in Equation (8) at a voxel location (x, y, z). Specially, the derivative approximation with finite differences can be restricted to the local neighborhood of the control point.

Optimization. Given an initial parameter μ, an optimization algorithm updates an incremental $\Delta\mu$ to reduce the cost function C iteratively. ASGD is used in the study, since it runs faster and less likely to get trapped in the local minima when compared to other gradient-based optimization algorithms [39]. Notably, ASGD implemented in the elastix package (http://elastix.isi.uu.nl) is used for adaptive step size prediction and the initial parameters are set as those in [39, 40].

3.3. A Comparison Method. The MIND is a feature-based method and it has been widely used in multimodal deformable registration [25, 41]. It aims to represent the distinctive image structure in a local neighborhood and explore the similarity of small image patches by using Gaussian-weighted patch distances [25].

MIND can be formulated by a distance Dp, a spatial search region R and a variance estimate V as below,

$$\mathrm{MIND}(\mathrm{I}, \mathrm{x}, \mathrm{r}) = \frac{1}{n} \exp\left(\frac{D_p(I, x, x + r)}{V(I, x)}\right) r \in R \quad (10)$$

$$\mathrm{Dp}\ (\mathrm{I}, \mathrm{x.x} + \mathrm{r}) = \mathrm{C} * (I - I(r))^2 \quad (11)$$

where n is a normalization constant, r the search region, C a convolution filter of size $(2p + 1)d$, $*$ a convolution filter, and $I0(r)$ a dense sampling on r. As such, an image can be represented by a vector of size $|R|$ at each location x. Moreover, $V\ (I, x)$ can be computed based on a mean of the patch distances within a small neighborhood $n\ (n \in N)$

$$V(I, x) = \frac{1}{N} \sum_{n \in N} D_p(I, x, x + n), \quad (12)$$

In Equation (10) to Equation (12), $n = 6$ denotes a six-connected neighborhood and $p = 1$ indicates a $3 \times 3 \times 3$ volume block.

The similarity metric used in MIND comes from the sum of absolute difference. To the fixed image (IF) and the moving image (IM), the local difference at a voxel x is

$$LD(x) = \frac{1}{|R|} \sum_{r \in R} \left| MIND(I_{F,x,r}) - MIND(I_{M,x,r}) \right| \quad (13)$$

The default value of $|R|$ is 6 and it means 6-connected neighbors are taken into computation.

3.4. Performance Evaluation

3.4.1. Tissue Overlapping. Tissue overlapping quantifies the overlapping ratio of outlined tissue regions in the fixed and its aligned image, which can distinguish the reasonable from the poor registration [42, 43]. This study focuses on the region of lumbar vertebrae and blood vessels. Assuming the outlined tissues in the fixed and aligned image are, respectively, denoted as OF and OA, the voxel-wise Jaccard (J) index and

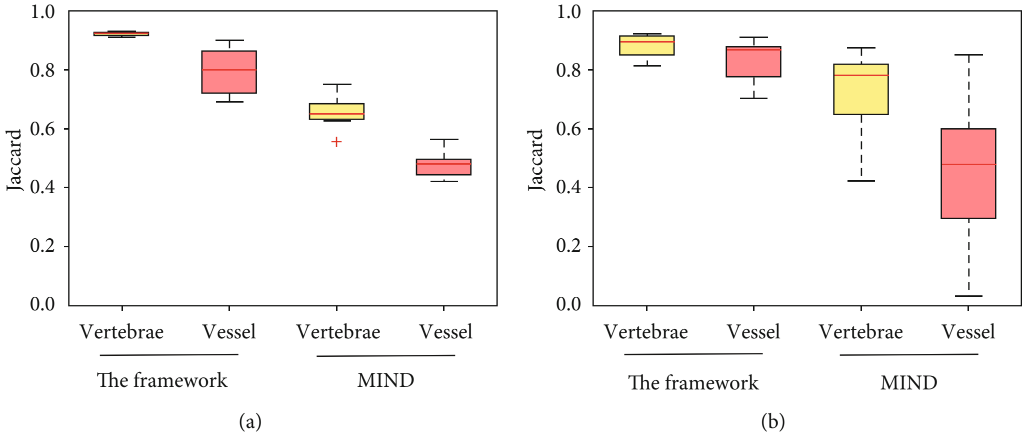


FIGURE 3: Jaccard index of the vertebrae and blood vessel overlapping on the in-house dataset (a) and the online dataset (b). Box-and-whisker plots represent the median Jaccard index (horizontal line) and total range (whiskers). The red $^+$ indicates an outlier that causes failure in image registration.

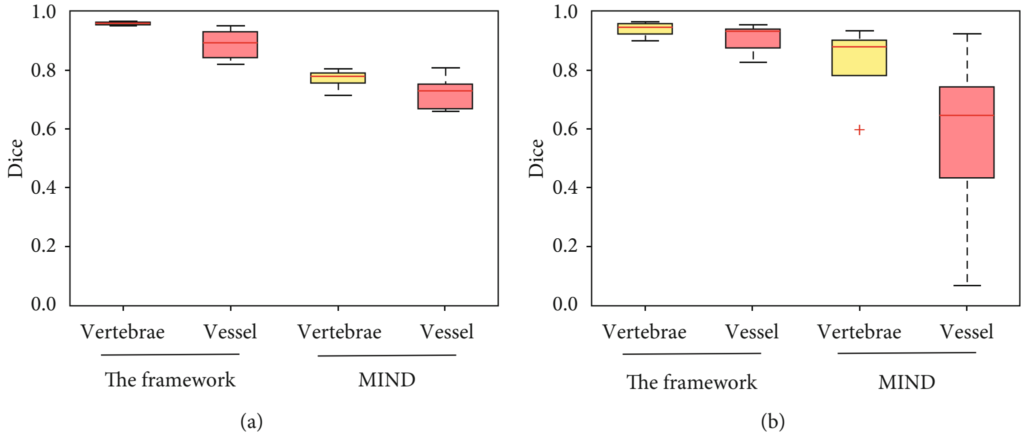


FIGURE 4: Tissue overlapping metric of Dice coefficient of the vertebrae and blood vessel on the in-house dataset (a) and the online dataset (b). Box-and-whisker plots show the median coefficient (horizontal line) and total range (whiskers). The red $^+$ indicates a failure case.

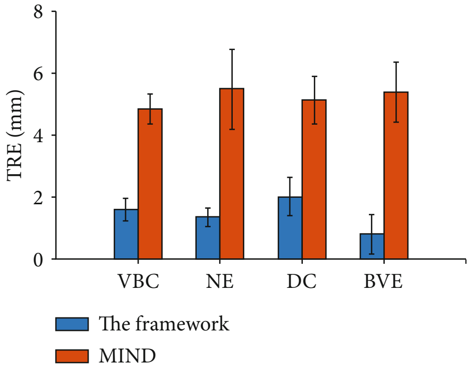


FIGURE 5: TRE values of anatomical landmarks on the in-house collection.

Dice (D) coefficient can be, respectively, described as

$$J = \left| \frac{O_F \cap O_A}{O_F \cup O_A} \right|,$$
$$D = 2\frac{|O_F \cap O_A|}{|O_F| + |O_A|} \quad (14)$$

where $|\cdot|$ indicates the number of voxels per volume.

3.4.2. Target Registration Errors. As for landmark annotation, ImageJ (http://imagej.nih.gov/ij/) was used. A pair of CT-MR images are displayed side-by-side. Then, landmarks are identified and manually annotated by an imaging radiologist (3+ year experience) and further confirmed by a senior radiologist (10+ year experience). Once landmarks are annotated, their locations in 3D space are recorded. In this study, anatomical landmark points are localized on the vertebral body center (VBC), neural edge (NE), disc center (DC), and blood vessel edge (BVE).

Target registration error (TRE) evaluates the distance between anatomical point pairs in the fixed and moving

Table 1: TRE values (mean ± SD) on the in-house collection images.

	The framework (mm)	MIND (mm)
VBC	1.52 ± 0.33	5.02 ± 3.76
NE	1.38 ± 0.29	5.07 ± 4.06
DC	2.01 ± 0.62	5.11 ± 3.69
BVE	0.78 ± 0.64	3.77 ± 4.21

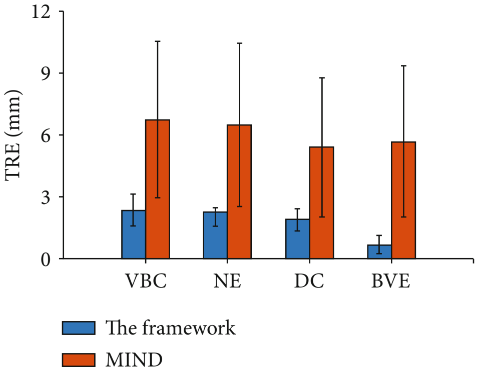


Figure 6: TRE values of anatomical landmarks on the SpineWeb dataset.

Table 2: TRE values (mean ± SD) on the SpineWeb images.

	The framework (mm)	MIND (mm)
VBC	2.37 ± 0.76	6.75 ± 3.80
NE	1.91 ± 0.55	5.41 ± 3.38
DC	2.26 ± 0.98	6.49 ± 3.95
BVE	0.66 ± 0.46	5.71 ± 3.65

image. Here, assuming *li* and , respectively, denotes the corresponding landmark point pairs in the fixed and moving image, the mean *TRE* for a given *T* is defined as

$$TRE = \frac{1}{n}\sum_{i}^{n} \left\| l_i - T\left(l_i'\right) \right\| \qquad (15)$$

where n is the number of pairs of landmark, and $k \cdot k$ indicates Euclidian distance in 3D space.

3.5. Software and Platform. The whole framework is implemented with Insight Segmentation and Registration Toolkit (http://www.itk.org) and the elastix package [40]. Experiments are performed on a desktop computer equipped with dual-core Intel i7 CPU (3.70 GHz) and 16 GB RAM memory.

4. Results

4.1. Tissue Overlapping. Figure 3 illustrates the tissue overlapping measure J of CT-MR image registration on the in-house collection (left) and the SpineWeb (right). The left shows that the proposed framework outperforms the MIND method on the vertebrae (0.93 ± 0.02 versus 0.69 ± 0.06) and blood vessel (0.81 ± 0.10 versus 0.48 ± 0.07) overlapping. In the right figure, the framework achieves higher values (vertebrae, 0.89 ± 0.05; blood vessel, 0.81 ± 0.12) than the MIND method (vertebrae, 0.75 ± 0.12; blood vessel, 0.52 ± 0.33), and thus, it leads to better performance.

Figure 4 shows the overlapping ratio D of multimodal image registration on the in-house collection (left) and the SpineWeb (right). The left figure indicates that the coarse-to-fine registration framework obtains better results than the MIND method on the vertebrae (0.97 ± 0.02 versus 0.77±0.05) and blood vessel (0.88 ± 0.07 versus 0.74 ± 0.07) overlapping. In the right figure, the MIND method (vertebrae, 0.86 ± 0.12; blood vessel, 0.61 ± 0.33) obtains inferior performance than the proposed framework (vertebrae, 0.95 ± 0.03; blood vessel, 0.93 ± 0.10).

4.2. Target Registration Errors. Figure 5 demonstrates the mean TRE value of anatomical landmark points between the proposed framework and the MIND algorithm on the in-house collection. The error-bar plot indicates that the TRE of the proposed framework is less than 3.00 mm (DC), while that of the MIND algorithm is larger than 4.00 mm (VBC) on average. In addition, statistical analysis indicates that the proposed framework significantly outperforms the MIND algorithm in each of the four sets of landmarks ($p < 0.005$, two-sample t-test).

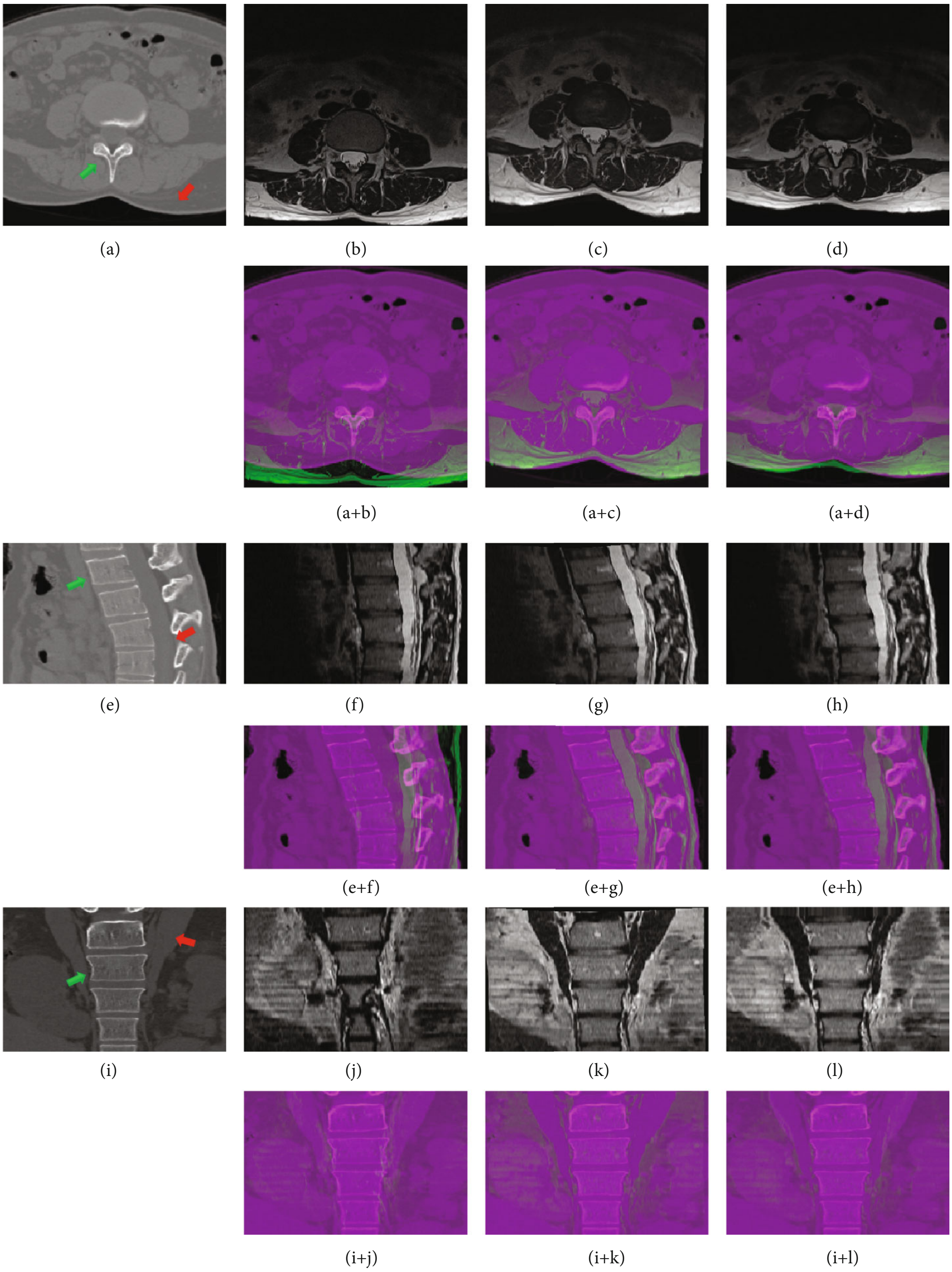
Table 1 shows the TRE values (mean ± standard deviation, mean ± SD) with respect to different landmark sets. The coarse-to-fine framework achieves TRE between 0.78 ± 0.64 mm (BVE) and 2.01 ± 0.62 mm (DC), while the TRE of the MIND method ranges from 3.77 ± 4.21 mm (BVE) to 5.11 ± 3.69 mm (DC), correspondingly larger than that from the proposed framework.

The mean TRE on the SpineWeb dataset is shown in Figure 6. It is observed that the TRE value of the proposed framework is less than 3.00 mm (VBC and NE), while the MIND algorithm leads to the TRE values larger than 5.00 mm.

Statistical analysis indicates significant difference of the TRE values between the proposed framework and the MIND algorithm on aligning the pairs of VBC and BVE landmarks ($0.01 < p < 0.05$, two-sample t-test).

Table 2 summarizes the mean TRE values on different sets of landmark pairs. The proposed framework achieves the TRE values between 0.66 ± 0.46 mm (BVE) to 2.37 ± 0.76 mm (VBC), and the TRE values of the MIND algorithm ranges from 5.71 ± 3.65 mm (BVE) to 6.75 ± 3.80 mm (VBC).

4.3. Perceived Quality of Image Alignment. Visual assessment of registration quality is perceived from the fusion of CT and MR images and observed from three perspective views in Figure 7, where (A, E, I) are the CT image, (B, F, J) are the MR image, (C, G, K) are the aligned image from the proposed framework, and (D, H, L) are the aligned image from the MIND algorithm. Red arrows directing to the soft tissue regions and green arrows directing to the bone regions are used for comparison. Before registration, both bones and tissues are



(a) (b) (c) (d)

(a+b) (a+c) (a+d)

(e) (f) (g) (h)

(e+f) (e+g) (e+h)

(i) (j) (k) (l)

(i+j) (i+k) (i+l)

Figure 7: Perceived visual difference of CT-MR images before and after image registration. The regions directed by the arrows are for comparison before and after registration. In addition, images are cropped for display purpose.

misaligned, such as acantha $(A + B)$, bones and nerves $(E + F)$ and muscles $(I + J)$. After image registration, the proposed framework aligns these parts in the MR images with fine deformation to the CT images. Specifically, both rigid bones and soft tissues are well matched, and the anatomical textures shows consistent distributions in the aligned image. On contrary, the MIND algorithm fails to overlap the acantha $(A + D)$, bones and nerves $(E + H)$ and muscles $(I + L)$ accurately.

4.4. Computation Time. Based on the software and platform, it took about 62 seconds to complete the affine registration and 427 seconds to complete the deformable registration. And thus, it required a total of 8.15 minutes to fulfill the coarse-to-fine registration for a pair of CT-MR lumbar spine image.

5. Discussion

Intrapatient multimodal image registration can fuse multisource information that benefits disease diagnosis and treatment delivery. This study develops a coarse-to-fine framework and aligns intrapatient CT-MR lumbar spine images. It first utilizes the similarity metric NMI for global positioning, and then, bending energy penalized NMI for local deformation of soft tissues. The proposed framework achieves high tissue overlapping ratio and low target registration error. It not only preserves the incompressibility of vertebrae but also well matches local soft tissues that provide accurate elastic registration of lumbar spine images for clinical applications.

The proposed framework is a coarse-to-fine approach for multimodal image registration. It aligns anatomical structures and addresses the potential difference on the fields of view and the intrinsic differences between medical imaging. The metric NMI is used, since it is a robust and accurate measure in multimodal image registration [17, 44]. After global positioning, a new similarity metric that integrates a bending energy term into NMI is used for local deformation and registration of soft tissues in medical images. It is worth of note that the term encourages smooth displacements in registration [33]. Ceranka et al. embedded the term to improve multiatlas segmentation of the skeleton from whole-body MR images [45], and de Vos et al. integrated the term into unsupervised affine and deformable image registration by using a convolutional neural network [46]. Both works [45, 46] figured out that the term caused significantly less folding in image registration.

The framework takes the incompressibility of vertebrae into account. Vertebrae are bony structures which are connected to each other by the ligamentum flavum at the neural arch [47]. The proposed framework enables global and local image structures well matched, and inflexible bones and soft tissues properly deformed. Its superior performance has been verified on the in-house collection and the SpineWeb database. Experiential results demonstrate that the overlapping ratio of annotated vertebrae and blood vessels are larger than 0.85, and the target registration error is less than 2.40 mm on average. It outperforms the MIND algorithm partly due to its proper deformation of local soft tissues and incompressible lumbar vertebrae. The registration quality is further perceived in a CT-MR image pair. It is found that the marked tissues keep relative location after image registration by using the proposed framework, since it not only well tackles the local soft tissue deformation but also conserves the rigid lumbar vertebrae.

Even if the proposed framework achieves superior performance on aligning CT-MR lumbar spine images, there is still room for further improvement. One way to enhance registration accuracy is by transferring multimodal image registration into mono-modal image registration. Wachinger and Navab developed structural representations, such as Entropy and Laplacian images, which could represent the images in a third space where the images showed close intensity or gradient distribution [48]. Moreover, deep networks have been explored to estimate CT images from MR images directly and in particular, the mapping between CT and MR images was learned without any patch-level pre- or postprocessing [49]. Another straightforward way is to utilize deep networks to learn the deformation field between different imaging modalities [50]. In addition, interactive image registration is admirable in interventional surgery and a doctor user could localize landmarks to guide and to update the registration procedure [51].

There are several limitations in this study. One limitation comes from no comparison of NMI and BEP-NMI on deformable image deformation, since our off-line experimental results show that the NMI based deformable registration is prone to distortion of lumbar spine and unnatural deformation of soft tissues. Moreover, demons and its variants [52, 53, 54] failed in the registration of lumbar spine images. Thus, this study reports the performance of the proposed framework and the MIND method. In addition, how to properly balance the BEP term and the NMI is always a problem and no existing methods could well tackle this issue, while prior knowledge [35, 37] could be employed for further improvement of the registration accuracy.

6. Conclusions

This paper presents a coarse-to-fine framework for the registration of intrapatient CT-MR lumbar spine images. It integrates the bending energy term into normalized mutual information for fine deformation of soft tissues around the incompressible vertebrae. Its high performance benefits multisource information fusion for accurate spine disease diagnosis, treatment planning, interventional surgery, and radiotherapy delivery.

Data Availability

The in-house collection of MR-CT image pairs used to support the findings of this study are restricted by the Medical Ethics Committee of Shenzhen Second People's Hospital in order to protect patient privacy. The SpineWebdata set of MR-CT images used to support the findings is freely available online (https://spineweb.digitalimaginggroup.ca/spineweb/index.php?n=Main.Datasets). If interested, requests for access to these data can be made to the author Shibin Wu (https://sb.wu@siat.ac.cn). Since the database is freely available, requests for access to these data can also be made to the author Shibin Wu (https://sb.wu@siat.ac.cn).

Conflicts of Interest

The authors declare there is no conflict of interest. The founding sponsors had no role in the design of this study, in the collection, analysis or interpretation of data, in the writing of this manuscript, nor in the decision to publish the experimental results.

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for their invaluable comments that have helped to improve the paper quality. This work is supported in part by grants from Shenzhen matching project (GJHS20170314155751703), National Key Research and Develop Program of China (2016YFC0105102); the National Natural Science Foundation of China (61871374); the Leading Talent of Special Support Project in Guangdong (2016TX03R139); the Science Foundation of Guangdong (2017B020229002, 2015B02023301); and CAS Key Laboratory of Health Informatics.

References

[1] G. Zheng and S. Li, "Medical image computing in diagnosis and intervention of spinal diseases," *Computerized Medical Imaging and Graphics*, vol. 45, pp. 99–101, 2015.

[2] F. Raciborski, R. Gasik, and A. Klak, "Disorders of the spine. A major health and social problem," *Annals of Physics*, vol. 4, no. 4, pp. 196–200, 2016.

[3] G. Zhang, Y. Yang, Y. Hai, J. Li, X. Xie, and S. Feng, "Analysis of Lumbar Sagittal Curvature in Spinal Decompression and Fusion for Lumbar Spinal Stenosis Patients under Roussouly Classification," *BioMed Research International*, vol. 2020, 8 pages, 2020.

[4] A. Toussaint, A. Richter, F. Mantel et al., "Variability in spine radiosurgery treatment planning - results of an international multi-institutional study," *Radiation Oncology*, vol. 11, no. 1, p. 57, 2016.

[5] P. A. Helm, R. Teichman, S. L. Hartmann, and D. Simon, "Spinal navigation and imaging: history, trends, and future," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1738–1746, 2015.

[6] A. Panigrahy, S. Caruthers, J. Krejza et al., "Registration of three-dimensional MR and CT studies of the cervical spine," *American Journal of Neuroradiology*, vol. 21, no. 2, pp. 282–289, 2000.

[7] B. Palkar and D. Mishra, "Fusion of multi-modal lumbar spine images using Kekre's hybrid wavelet transform," *IET Image Processing*, vol. 13, no. 12, pp. 2271–2280, 2019.

[8] D. Tomazevic and F. Pernus, "Robust gradient-based 3-D/2-D registration of C-T and MR to X-ray images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 12, pp. 1704–1714, 2008.

[9] Y. Otake, A. S. Wang, J. W. Stayman et al., "Robust 3D-2D image registration: application to spine interventions and vertebral labeling in the presence of anatomical deformation," *Physics in Medicine and Biology*, vol. 58, no. 23, pp. 8535–8553, 2013.

[10] R. Harmouche, F. Cheriet, H. Labelle, and J. Dansereau, "3D registration of MR and X-ray spine images using an articulated model," *Computerized Medical Imaging and Graphics*, vol. 36, no. 5, pp. 410–418, 2012.

[11] G. Hille, S. Saalfeld, S. Serowy, and K. Tonnies, "Multi-segmental spine image registration supporting image-guided interventions of spinal metastases," *Computers in Biology and Medicine*, vol. 102, pp. 16–20, 2018.

[12] L. Chen, H. Zhang, S. Wu, S. Yu, and Y. Xie, "Estimating fetal brain motion with total-variation-based magnetic resonance image registration," in *World Congress on Intelligent Control and Automation (WCICA)*, pp. 809–813, Guilin, China, June 2016.

[13] A. Khalil, S. Ng, Y. M. Liew, and K. W. Lai, *An Overview on Image Registration Techniques for Cardiac Diagnosis and Treatment, Cardiology Research and Practice*, Hindawi, 2018.

[14] J. Li and Q. Ma, *A Fast Subpixel Registration Algorithm Based on Single-Step DFT Combined with Phase Correlation Constraint in Multimodality Brain Image, Computational and Mathematical Methods in Medicine*, Hindawi, 2020.

[15] P. Viola and W. Wells-III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[16] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[17] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[18] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, "Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change," *IEEE Transactions on Medical Imaging*, vol. 25, no. 5, pp. 626–639, 2006.

[19] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. W. Pluim, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Medical Physics*, vol. 35, no. 4, pp. 1407–1417, 2008.

[20] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using conditional mutual information," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 19–29, 2010.

[21] H. Rivaz, Z. Karimaghaloo, and D. L. Collins, "Self-similarity weighted mutual information: a new nonrigid image registration metric," *Medical Image Analysis*, vol. 18, no. 2, pp. 343–358, 2014.

[22] X. Zhuang, S. Arridge, D. J. Hawkes, and S. Ourselin, "A nonrigid registration framework using spatially encoded mutual information and free-form deformations," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1819–1828, 2011.

[23] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, "Dramms: deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, no. 4, pp. 622–639, 2011.

[24] R. Zhang, W. Zhou, Y. Li, S. Yu, and Y. Xie, "Nonrigid registration of lung CT images based on tissue features," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 834192, 2013.

[25] M. P. Heinrich, M. Jenkinson, M. Bhushan et al., "MIND: modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423–1435, 2012.

[26] J. A. Onofrey, L. H. Staib, and X. Papademetris, "Learning nonrigid deformations for constrained multi-modal image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Lecture Notes in Computer Science, pp. 171–178, 2013.

[27] M. Toews, L. Zöllei, and W. M. Wells, "Feature-based alignment of volumetric multimodal images," in *Information*

Processing in Medical Imaging, Lecture Notes in Computer Science, pp. 25–36, 2013.

[28] F. Zhu, M. Ding, and X. Zhang, "Self-similarity inspired local descriptor for non-rigid multi-modal image registration," *Information Sciences*, vol. 372, pp. 16–31, 2016.

[29] D. Jiang, Y. Shi, X. Chen, M. Wang, and Z. Song, "Fast and robust multimodal image registration using a local derivative pattern," *Medical Physics*, vol. 44, no. 2, pp. 497–509, 2017.

[30] J. Woo, M. Stone, and J. L. Prince, "Multimodal registration via mutual information incorporating geometric and spatial context," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 757–769, 2015.

[31] D. D. B. Carvalho, S. Klein, Z. Akkus et al., "Joint intensity-andpoint based registration of free-hand B-mode ultrasound and MRI of the carotid artery," *Medical Physics*, vol. 41, no. 5, pp. 052904–052912, 2014.

[32] O. Weistrand and S. Svensson, "The ANACONDA algorithm for deformable image registration in radiotherapy," *Medical Physics*, vol. 42, no. 1, pp. 40–53, 2015.

[33] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[34] T. Rohlfing, C. Maurer, D. Bluemke, and M. Jacobs, "Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint," *IEEE Transactions on Medical Imaging*, vol. 22, no. 6, pp. 730–741, 2003.

[35] M. Staring, S. Klein, and J. Pluim, "A rigidity penalty term for nonrigid registration," *Medical Physics*, vol. 34, no. 11, pp. 4098–4108, 2007.

[36] D. Ruan, J. Fessler, M. Roberson, J. Balter, and M. Kessler, "Nonrigid registration using regularization that accommodates local tissue rigidity," *Computers in Biology and Medicine*, vol. 42, no. 1, pp. 123–128, 2012.

[37] C. Fiorino, E. Maggiulli, S. Broggi et al., "Introducing the jacobian-volume-histogram of deforming organs: application to parotid shrinkage evaluation," *Physics in Medicine and Biology*, vol. 56, no. 11, pp. 3301–3312, 2011.

[38] S. Leibfarth, D. Monnich, S. Welz et al., "A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning," *Acta Oncologica*, vol. 52, no. 7, pp. 1353–1359, 2013.

[39] S. Klein, J. P. W. Pluim, M. Staring, and M. A. Viergever, "Adaptive stochastic gradient descent Optimisation for image registration," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 227–239, 2009.

[40] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[41] S. Reaungamornrat, T. D. Silva, A. Uneri et al., "MIND demons: symmetric diffeomorphic deformable registration of MR and CT for image-guided spine surgery," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2413–2424, 2016.

[42] S. Yu, S. Wu, L. Zhuang et al., "Efficient segmentation of a breast in B-mode ultrasound tomography using threedimensional GrabCut (GC3D)," *Sensors*, vol. 17, no. 8, p. 1827, 2017.

[43] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 153–163, 2012.

[44] T. Veninga, H. Huisman, R. van der Maazen, and H. Huizenga, "Clinical validation of the normalized mutual information method for registration of CT and MR images in radiotherapy of brain tumors," *Journal of Applied Clinical Medical Physics*, vol. 5, no. 3, pp. 66–79, 2004.

[45] J. Ceranka, S. Verga, M. Kvasnytsia et al., "Multi-atlas segmentation of the skeleton from whole-body MRI - impact of iterative background masking," *Magnetic Resonance in Medicine*, vol. 83, no. 5, pp. 1851–1862, 2019.

[46] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[47] R. Wang and M. M. Ward, "Arthritis of the spine," in *Spinal Imaging and Image Analysis*, pp. 31–66, Springer, Cham, 2015.

[48] C. Wachinger and N. Navab, "Entropy and laplacian images: structural representations for multi-modal registration," *Medical Image Analysis*, vol. 16, no. 1, pp. 1–17, 2012.

[49] W. Li, Y. Li, W. Qin et al., "Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy," *Quantitative Imaging in Medicine and Surgery*, vol. 10, no. 6, pp. 1223–1236, 2020.

[50] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1-2, p. 8, 2020.

[51] A. Herline, J. Herring, J. Stefansic, W. Chapman, R. Galloway, and B. Dawant, "Surface registration for use in interactive, image-guided liver surgery," *Computer Aided Surgery*, vol. 5, no. 1, pp. 11–17, 2000.

[52] J. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[53] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[54] H. Lombaert, L. Grady, X. Pennec, and N. Ayache, "F. Cheriet-Spectral logdemons: Diffeomorphic image registration with very large deformations," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 254–271, 2014.