IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 21, 2020, accepted October 2, 2020, date of publication October 12, 2020, date of current version October 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030060

Deep Longitudinal Feature Representations for Detection of Postradiotherapy Brain Injury at Presymptomatic Stage

LIMING ZHONG[1], XIAO ZHANG[1,2], YUHUA XI[1], ZHOUYANG LIAN[3], QIANJIN FENG[1], (Member, IEEE), WUFAN CHEN[1], (Senior Member, IEEE), SHUIXING ZHANG[4], AND WEI YANG[1]

[1]Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
[2]Zhuhai Precision Medical Center, Zhuhai People's Hospital (Zhuhai Hospital Affiliated With Jinan University), Zhuhai 519000, China
[3]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China
[4]Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou 510630, China

Corresponding authors: Shuixing Zhang (shui7515@126.com) and Wei Yang (weiyanggm@gmail.com)

Liming Zhong and Xiao Zhang contributed equally to this work.

This work was supported in part by the National Natural Science Foundation of China under Grant 81771916, and in part by the Guangdong Provincial Key Laboratory of Medical Image Processing under Grant 2014B030301042.

ABSTRACT Temporal lobe injury (TLI), a form of nervous system damage in the brain, is a major neurological complication after radiation therapy (RT). TLI must be highly valued because of the irreversible brain injury. This article aims to develop a predictive pipeline, called deep longitudinal feature representations (DLFR), to detect TLI at the presymptomatic stage accurately via the learning of effective deep longitudinal feature representations. DLFR characterizes high-level information and developmental changes within and across subjects The DLFR consists of four components: (i) extraction of deep features from a pretrained ResNet50 model; (ii) compression of learned highly representative features by the global max pooling; (iii) fusion of deep longitudinal features for the fully use of all follow-up data; (iv) random forest-based prediction of the diagnostic status. In total, 244 nasopharyngeal carcinoma patients before and after RT with a follow-up period of $0 \sim 9$ years were included for analysis. All patients were divided into four different latency groups, and the current latency was used for training to predict the diagnostic status of the next latency. The AUCs of the predicted three different latency groups using DLFR were 0.64 ± 0.11, 0.76 ± 0.10, and 0.88 ± 0.05, while those of radiomics features were 0.56 ± 0.06, 0.63 ± 0.03, and 0.53 ± 0.04, and those of histogram of oriented gradients features were 0.60 ± 0.09, 0.52 ± 0.03, and 0.58 ± 0.06. Most importantly, the AUCs of the predicted three different latency groups for white matter regions were 0.66 ± 0.10, 0.80 ± 0.09, and 0.78 ± 0.09. Our proposed method can dynamically detect TLI at the presymptomatic stage, which can enable the administration of preventive neurological intervention.

INDEX TERMS Temporal lobe injury, nasopharyngeal carcinoma, deep longitudinal features, white matter.

I. INTRODUCTION

Nasopharyngeal carcinoma (NPC) [1], which develops from the nasopharynx epithelium, is the most common malignant tumor of the head and neck. NPC has a unique pattern of ethnic and geographic distribution. As reported in 2012, 71% of 86500 NPC patients were in the eastern and southeastern parts of Asia, south-central Asia, and north and east Africa [2]. The highest incidence is found among Southern Chinese individuals, especially those of Cantonese origin, with 25∼30 per 100,000 persons per year [3].

Due to the high radiosensitivity of NPC, radiation therapy (RT) is a routine and curative clinical treatment for this cancer [1]. During RT, the medial temporal lobe along the RT pathway and near the tumor is inevitably irradiated, causing progressive and irreversible brain injury after several years [4]. Postradiotherapy temporal lobe injury (TLI), a form of nervous system damage in the brain, is a major neurological complication after RT. The damage affects the ability of recognizing faces and understanding spoken words, along with a

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

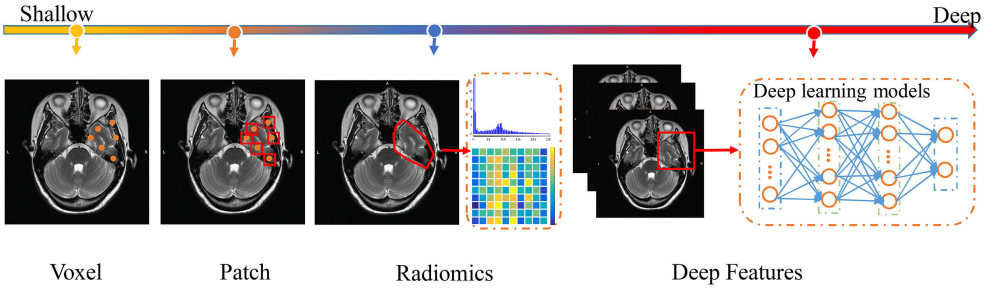

FIGURE 1. Illustration of feature representations for brain disease diagnosis, including voxel-based, patch-based, radiomics-based, and deep feature-based representations.

disturbance with selective attention and the short-term memory loss [5], [6]. TLI is the cause of 65% of irradiation-related deaths from NPC [7]. Most incipient symptoms can be reversed through conservative treatment, whereas late symptoms, including more severe and irreversible symptoms, can only be relieved via active treatment [8].

TLI remains poorly understood because of the complexity of the time-evolving interaction between brain recovery, plasticity, and degeneration. There are two major limitations in the study of TLI in NPC patients. First, the survival time and the less strict follow-up management of patients limit the investigation of TLI [9]. Second, the diagnosis of TLI relies on medical imaging including computed tomography (CT) and magnetic resonance imaging (MRI). No special symptoms occur during the incipient stage, whereas white matter (WM) edema and demyelination generally appear in the late stage [7]. Patients with normal imaging may suffer from nervous system damage or loss of cognitive function. Current studies [4], [7], [10], [11] have focused on MRI-visible TLI in patients with NPC who typically have multiyear incubation periods after RT. However, TLI is irreversible at the MRI-visible WM edema stage due to the limited recovery ability and the impossibility of reversing impaired cognitive function. Thus, developing a method to investigate TLI in patients with NPC at the presymptomatic stage is desired in the clinic. Various types of features or patterns extracted from neuroimaging modalities for brain disease diagnosis with machine learning-based classification methods can be used to achieve this goal.

The previous feature representations (shown in Fig. 1) can be classified into four categories: voxel-based, patch-based, radiomics-based, and deep feature-based methods. Voxel-based methods directly use the voxel intensities as features in classification [12]–[14]. Two significant limitations of the voxel-based method include the high-dimensionality of feature vectors and ignoring of the region information. Patch-based methods partition the images into smaller patches based on different tissues, organs, cortical thickness, or specific regions of interest (ROIs) [15]–[18]. These methods efficiently capture the local detailed heterogeneous structures and handle the high-dimensional features. Generally, the formulation of patches can adapt to the local information using superpixels [19] and descriptors [20]. However, these patch-based methods involve mostly low-level features that are unable to capture the neuropathological heterogeneity of brain tissue associated with the conversion to TLI. Radiomics-based methods extract large amounts of advanced and high-order quantitative features with high-throughput and high-fidelity information using a large number of automated feature extraction algorithms [21]–[24]. These radiomics features effectively depict the in-depth information that are not readily apparent in standard imaging analyses. One major drawback of radiomics features is that the pathological mechanism of TLI is completely different from that of common tumors. Radiomics features are mainly designed to extract tumor information. Another limitation is that they can not provide visible details that reflect local alterations in features associated with TLI. In contrast to the shallow feature representations used in the aforementioned methods, deep feature-based methods yield higher-level and richer feature representations to enhance the accuracy of disease diagnosis [25]–[27].

Recently, deep learning methods have shown excellent performance in classification tasks [28]–[32] when compared to existing radiomics methods. When the sample size is small, it is more suitable to use a neural network to extract the required deep feature representations for classification rather than to train an end-to-end classification model. The deep feature classification framework, which trains classic machine learning methods using features extracted by pre-trained neural network, has been applied for disease diagnosis. Moreover, deep feature classification can improve the performance over conventional machine learning. To the best of our knowledge, early detection of TLI in NPC patients through deep longitudinal features has not been investigated.

In this article, by considering the difficulty in the strict management of the follow-up of NPC patients after RT, an effective deep longitudinal feature representation (DLFR) method for automatic diagnosis of TLI was developed. Deep learning methods [33], which learn image filters for extracting latent feature representations by optimizing their discriminative performance, have been successfully applied to a variety of image analysis problems. Transfer learning has been proven to obtain superior performance,

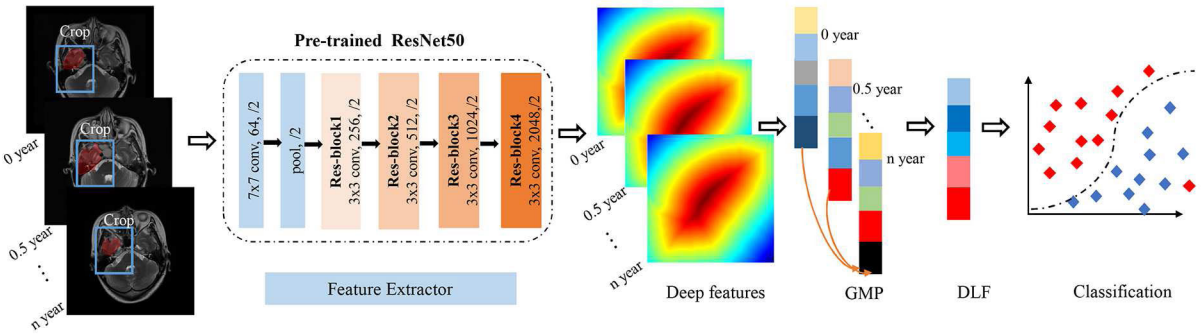

FIGURE 2. The detailed steps of DLFR for early detection of TLI in NPC patients.

particularly when with domains with limited data [34]. The proposed method, the DLFR pipeline, extracts high-dimensional deep features from a pretrained model. Global max pooling (GMP) [35], a method in which no additional parameters are needed for optimization, is used to compress the extracted high-dimensional features. After the extraction of deep features, we propose the fusion of deep longitudinal features (DLFs) to integrate the learned features in one subject with follow-up data obtained at different times. Finally, random forest (RF) [36] is performed to detect TLI at the presymptomatic stage. No additional training model is required, and the extraction of longitudinal feature representations with follow-up data obtained at different times distinguishes our method from other deep feature-based disease diagnosis methods. In the experiments, five pretrained deep models including VGG16 [37], ResNet50 [38], Densenet121 [39], Xception [40], and InceptionV3 [41], and five machine learning methods including RF, k-nearest neighbors (KNN) [42], Adaboost (AB) [43], generalized linear regression (GLR) [44], and support vector machines (SVM) [45] were respectively compared and analyzed. ResNet50 and RF achieved the best performance. The detailed steps are shown in Fig. 2.

II. METHOD

A. DATA PREPARATION

Two hundred and forty-four NPC patients (mean age 48.61 $\pm$ 10.4 years, range 17 $\sim$ 76, 171 males, and 73 females), treated with RT with or without chemotherapy, were retrospectively enrolled with follow-up periods of 0 $\sim$ 9 years and with an average of 7.28 $\pm$ 3.03 examinations. Among them, two hundred patients, based on the clinical manifestations of brain lesions and related imaging studies, were diagnosed with TLI. These patients were subsequently followed up every half a year to one year. The latency of TLI was measured from the day after RT to the day diagnosed with TLI, which ranged from several months to several years. In our longitudinal study, due to the irreversible characteristics of TLI, it was meaningless to add the last follow-up data diagnosed with TLI into the analysis. The demographics of the patient population are listed in Table 1.

TABLE 1. The demographics of the NPC patients.

Characteristic	Value
Number of patients	244
Age(Mean)	48.61 $\pm$ 10.4
Age(Range)	17$\sim$76
Gender(Male)	171
Gender(Female)	73
Stages I-II	16
Stages III-IV	228
WHO type I-II	25
WHO type III	219
Radiation dose (Gy)	32 $\pm$ 5.39
Chemotherapy (Yes)	235
Chemotherapy (No)	9

Two types of MR image protocols, T1-weighted contrast-enhanced (T1c) and T2-weighted (T2w), were used for an auxiliary diagnosis of TLI. MR images were acquired on a 1.5T or 3T MR scanner (GE Medical Systems Signa Excite, Philips Medical Systems Achieva, Philips Medical Systems Gyroscan NT, and Siemens Espree). The acquisition parameters of T1c images were as follows: flip angle 69° $\sim$ 126°, echo time (TE) 5.41$\sim$18 ms, repetition time (TR) 205.67$\sim$716.69 ms, and voxel size 0.43 $\times$ 0.43 $\times$ 5 mm^3 / 0.47 $\times$ 0.47 $\times$ 5 mm^3. The acquisition parameters of T2w images were as follows: 83° $\sim$ 142°, TE 80$\sim$197 ms, TR 2470 $\sim$ 6440 ms, and voxel size 0.39 $\times$ 0.39 $\times$ 5 mm^3 / 0.43 $\times$ 0.43 $\times$ 5 mm^3.

B. MRI DATA PREPROCESSING

The MR images were processed by applying typical procedures, including bias correction, intensity normalization, skull stripping, and segmentation of different tissue types. N4 bias correction [46] was performed to correct nonuniform tissue intensities. Then, intensity normalization [47] was used to reduce inhomogeneity across different patients, different follow-up examinations, and MR images obtained with different protocols. Due to the tight relation between WM and TLI, the analysis of WM is needed at the presymptomatic stage. Thus, three tissue types, including WM, gray matter (GM), and cerebroSpinal fluid (CSF), were obtained by using BET in the FSL package[1] for skull stripping and

[1] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki

FAST in the FSL package for segmentation. For accurate analysis of TLI, manual segmentation of the temporal lobe was performed by a radiologist (with 10 years of experience) using ITK-SNAP [48].

Due to a follow-up period of $0 \sim 9$ years for the patients after RT, we divided all the presymptomatic patients into four groups, including latency ≤ 1 year, $1 <$ latency ≤ 2 years, $2 <$ latency ≤ 3 years, and latency > 3 years. The division of the TLI groups was based on the diagnostic status of the last follow-up examinations. Patients who were not included in the TLI group were classified into the No-TLI group. The details of the patient partitions are illustrated in Table 2. Some individuals missed or advanced their scheduled date of observation. Therefore, patients who missed the strict follow-up time will temporarily not be included in the specific No-TLI groups.

TABLE 2. The details of patient partitions at different periods.

Groups	No-TLI	TLI	Total	Average examinations
latency ≤ 1 year	146	61	207	2.67 ± 1.10
$1 <$latency≤ 2 years	98	53	151	3.86 ± 1.55
$2 <$ latency≤ 3 years	53	46	99	5.09 ± 1.99
latency > 3 years	33	21	54	7.28 ± 3.03

C. EARLY DETECTION OF TLI

A schematic diagram of our DLFR framework for the detection of TLI is presented in Fig. 2. Our longitudinal study consists of four major steps, including extraction of deep features, compression of deep features, fusion of longitudinal features, and detection of TLI.

1) EXTRACTION OF DEEP FEATURES

Given an NPC patient I after RT, we aim to obtain the diagnostic status C through all training MR images $T = \{I_1^1, I_2^2, \ldots, I_n^N\}$ with their diagnostic status $\psi = \{C_1^1, C_2^2, \ldots, C_n^N\}$, where N denotes the number of subjects, and n is the number of follow-up examinations for each patient. The likelihood of the diagnostic status $C \in \{0, 1\}$ of a new arrival patient can be formulated as:

$$y_i = P(C(i)|I; T, \psi), \tag{1}$$

where the diagnostic status 0/1 denotes the NPC patient as No-TLI/TLI, respectively.

The likelihood can be obtained by a classifier with learned features. Thus, a pretrained convolutional neural network model (ResNet50 [38]) was used to extract deep features from the MR images. The ResNet50 model was pretrained on the ImageNet LSVRC-2010[2] database, which contains 1,000 classes and 1.2 million 3-channel images for image classification. ResNet50 contains five stages of blocks, including one conventional convolution block (convolution, batch normalization, rectified linear unit, and max pooling) in stage one, and a convolution block without max pooling followed by an identity block in stages $2 \sim 5$. The output convolution feature maps are compressed by a fully-connected layer, and connected with a final 1000-way softmax. The fully-connected layer, which requires a pre-defined fixed-size/length input, limits the flexibility in extracting deep features. The temporal lobe in each slice of the MR images was cropped out and duplicated into three channels. The pretrained ResNet50 model was used to extract the convolutional feature maps of the reshaped images, generating the feature representation $x \in k \times k \times f$ (where k is the size of the kernel, and f is the number of filters). However, due to the high dimensionality of the extracted feature x, it is inconvenient to directly adopt the deep features into a classifier. Therefore, it is desirable to develop compressed deep features.

[2] http://www.image-net.org/challenges/LSVRC/2010

2) COMPRESSION OF THE DEEP FEATURES

Compression of the deep features is needed to improve the efficiency of a classifier due to the high dimensionality of the extracted deep features. In ResNet50 [38], the fully-connected layer limits the input size of the images [49], and the fixed size of input images restricts the flexibility of the deep model. Moreover, the fully connected layer hampers the generalization ability of the network. Therefore, in our study, we used global max pooling [35] to replace the fully connected layer for compressing the learned features. Each output feature representation $x \in k \times k \times f$ can be compressed through the maximum of each feature map, thus generating a compressed feature vector $\tilde{x} \in 1 \times f$.

3) FUSION OF DEEP LONGITUDINAL FEATURES

Given that patients may miss or advance their scheduled date of observation, the number of observation times for the different patients is not uneven. One simple way to early detect the diagnostic status of each patient, is to use the specific MR data in current latency to predict the diagnostic status of next latency. However, it is not recommended to use this simple method because it ignores a lot of useful information. Therefore, we proposed the fusion of DLFs to fully use all follow-up data. Due to the need of 2D input for the pretrained ResNet50 [38] model, the deep features for each patient with different follow-up data can be fused via:

$$X_i = \tilde{X}_i^n + \sum_{j=1}^{n-1}(\tilde{X}_i^n - \tilde{X}_i^j), \quad i = 1, \ldots, N, \tag{2}$$

where $\tilde{X} = \frac{\sum_{k=1}^{M} \tilde{x}}{M}$ is the mean of the deep features for M slices in one MR volume.

4) CLASSIFICATION OF TLI

Given the output of the DLFs, we built three classified models to predict the diagnostic status of the patients. In this study, we used the random forest method for classification [36], which is a well-known machine learning method by building a set of weak learners of decision trees to improve the

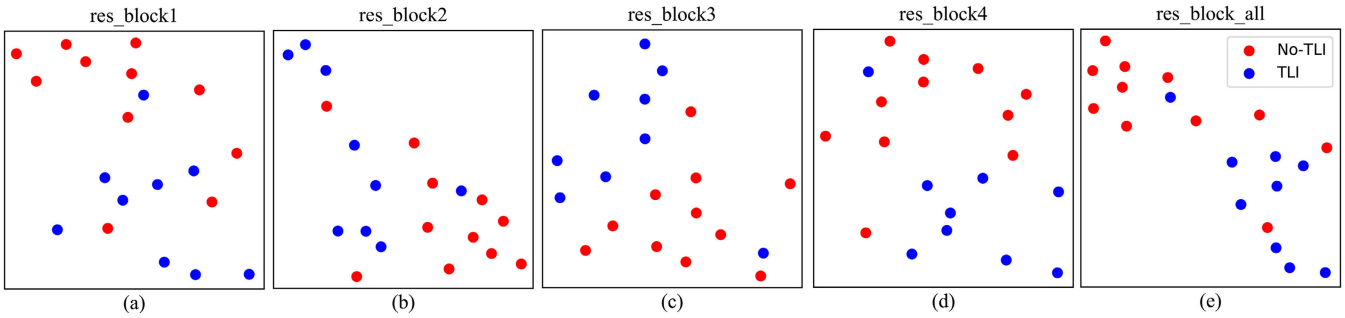

FIGURE 3. Visualization of deep features extracted by our DLFR pipeline for patients with TLI and patients without TLI for one randomly selected fold results in 2 < latency ≤ 3 years group, by t-SNE projection in different layers including (a) res-block1, (b) res-block2, (c) res-block3, (d) res-block4, (e) res-block all. The third res-block shows the greatest discriminative power between the patients with TLI and patients without TLI in comparison with other res-blocks.

generalization ability of the classifier. A fixed seed for initialization was set for random forest to avoid the uncertainty in the results. For the training data, we used a bootstrapped version (bagging) to train the models, and then averaged all the outputs as the final value:

$$y = \frac{1}{B}\sum_{b=1}^{B}\tilde{y}(X, C), \tag{3}$$

where B is the number of bootstrapped times.

The detailed implementation of our proposed method is provided in Table 3.

TABLE 3. Algorithm DLFR.

Input: The cropped test image I^n_{test} , the cropped training set $T = \{I^1, ..., I^{n-1}\}^N$;
Output: The predicted diagnostic status y;
1. **for** each $i \in M$ slices
2. Extract deep feature x of the fourth output of ResNet50 from each slice of test image I^n_{test} and training images $T = \{I^1, ..., I^{n-1}\}^N$;
3. Deep feature x is compressed by GAP, generating compressed feature $\tilde{x}$;
4. **end for**
5. Calculate the mean of deep features $\tilde{X} = \frac{\sum_{k=1}^{M}\tilde{x}}{M}$ for M slices in each MR volume;
6. Calculate deep longitudinal features using Eq. 2;
7. Train classification models using random forest with a fixed seed;
8. Obtain the predicted diagnostic status y using Eq. 3.

D. STATISTICAL ANALYSIS

All deep features extracted by the pretrained models were implemented using Keras with TensorFlow backend framework and obtained with 1 NVIDIA TITAN Xp GPU. The data analysis was performed on MATLAB 2015b and Python 3.6. The performance of our longitudinal analysis was evaluated through the area under the curve (AUC), sensitivity (SEN), specificity (SEP), and accuracy (ACC), defined as follows:

$$SEN = \frac{TP}{TP + FN}, SEP = \frac{TN}{TN + FP},$$
$$ACC = \frac{TP + TN}{TP + FP + FN + TN}, \tag{4}$$

where TP, TN, FP, and FN are the abbreviation of true positives, true negatives, false positives and false negatives, respectively. The prediction accuracy was computed by using patient-wise five-fold cross-validation. In each fold, 20% of patients were left out for testing, and the random forest created an optimal classifier using the remaining 80% of patients. Due to that the random forest classifier with random seeds could result in random results, a fixed seed for initialization was set for random forest. As illustrated in Table 2, all the presymptomatic patients were partitioned into four groups. In each fold, the random forest created an optimal classifier using 80% of patients with latency $<$ 1 years for training, and the rest of patients with 1 $<$ latency $\leq$ 2 years for validation. The remaining models for the other groups were trained and validated in the same manner. Unless otherwise stated, all results were obtained by using T2w MR data.

III. RESULTS

In this section, we first show the effective layers of ResNet50 [38]. Then, we report the experimental results for TLI classification using different dimensional reduction methods and machine learning-based classification methods. Comparisons between different well-known deep learning pretrained models and between different feature representations are also included. Most importantly, comparisons between different modalities of MR images and between different tissues are reported.

A. EFFECTIVE FEATURE REPRESENTATIONS

It is worth noting that the depth of the pretrained models affects the classification performance. To analyze the discriminating ability of the learned features with networks of different depths, we visualized the extracted features at the output of four different res-blocks. Extracted features, generated by one randomly selected fold for detecting TLI in the group of 2 $<$ latency $\leq$ 3 years, were projected down to 2 dimensions using the t-SNE dimension reduction algorithm [50]. As shown in Fig. 3 (a-e), the third res-block showed the greatest discriminative power between the patients with TLI and patients without TLI in comparison

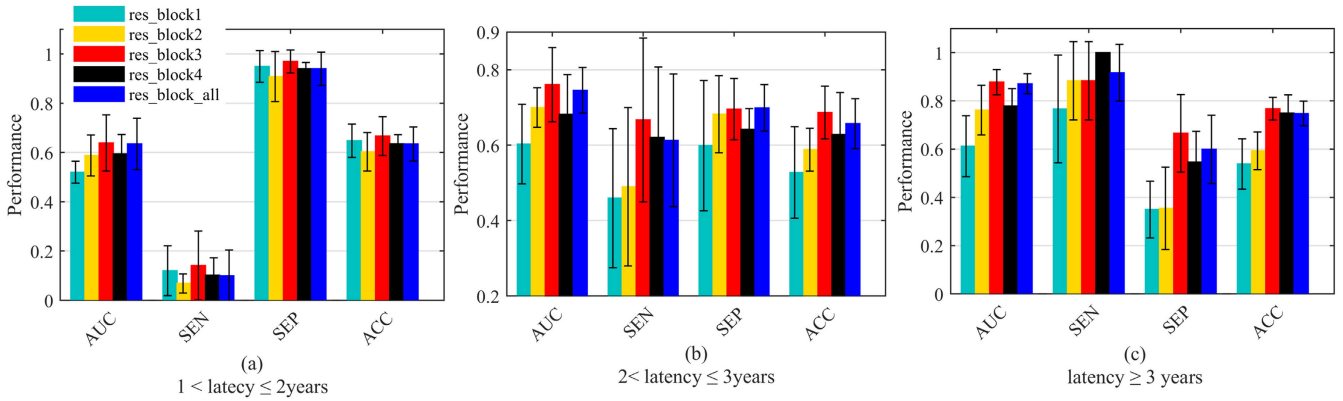

FIGURE 4. Five-fold cross validation results using our DLFR pipeline with different depth of the model for patients with TLI vs. patients without TLI. The features generated by the third res-block achieve the best performance for early detection in the groups of 1 < latency ≤ 2 years, 2 < latency ≤ 3 years, and latency > 3 years.

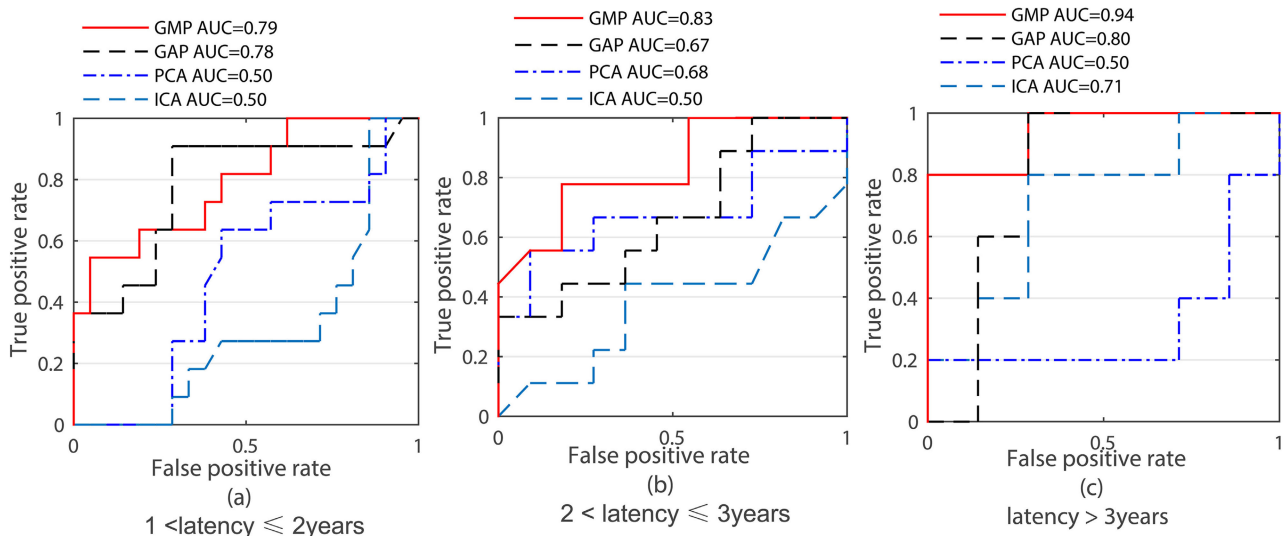

FIGURE 5. The ROC plot for randomly selected folds of three different latencies obtained by different feature compressed methods.

with other res-blocks. To further evaluate the performance of these res-blocks, we summarized the AUC, SEN, SEP, and ACC results for three latency groups in Fig. 4. As presented in Fig. 4, the features generated by the third res-block achieved the best performance for early detection in the groups of 1 < latency ≤ 2 years, 2 < latency ≤ 3 years, and latency > 3 years. Thus, we summarized the specific results of the third res-block in Table 4. As presented in Fig. 3-4 and Table 4, in our DLFR pipeline, the deep features extracted via the third res-block of the pretrained ResNet50 model were suitable for early detection of TLI at the presymptomatic stage.

B. COMPARISONS OF DIFFERENT DIMENSIONAL REDUCTION METHODS

In our DLFR pipeline, for the task of compressing the high-dimensional features, we compared GMP with other different dimensional reduction methods, including global average pooling (GAP), principal components analysis (PCA) [51], and independent components analysis (ICA) [52], to evaluate the performance in predicting the diagnostic status. The GAP/GMP method compressed the output of the third res-block of ResNet50 through average/maximum each feature map, and generated a vector of size 512. In both the PCA and ICA methods, the top 512 main components of features extracted from the third res-block of ResNet50 were selected to maintain the fair comparisons. Fig. 5 shows the receiver operating characteristic curve (ROC) for the results achieved

TABLE 4. The specific early detection of TLI results using our DLFR pipeline with res-block3 for three different latency.

Groups	1 < latency ≤ 2 years	2 < latency ≤ 3 years	latency > 3 years
AUC	0.64 ± 0.11	0.76 ± 0.10	0.88 ± 0.05
SEN	0.14 ± 0.14	0.67 ± 0.22	0.88 ± 0.16
SEP	0.97 ± 0.05	0.70 ± 0.08	0.67 ± 0.16
ACC	0.67 ± 0.08	0.69 ± 0.07	0.77 ± 0.05

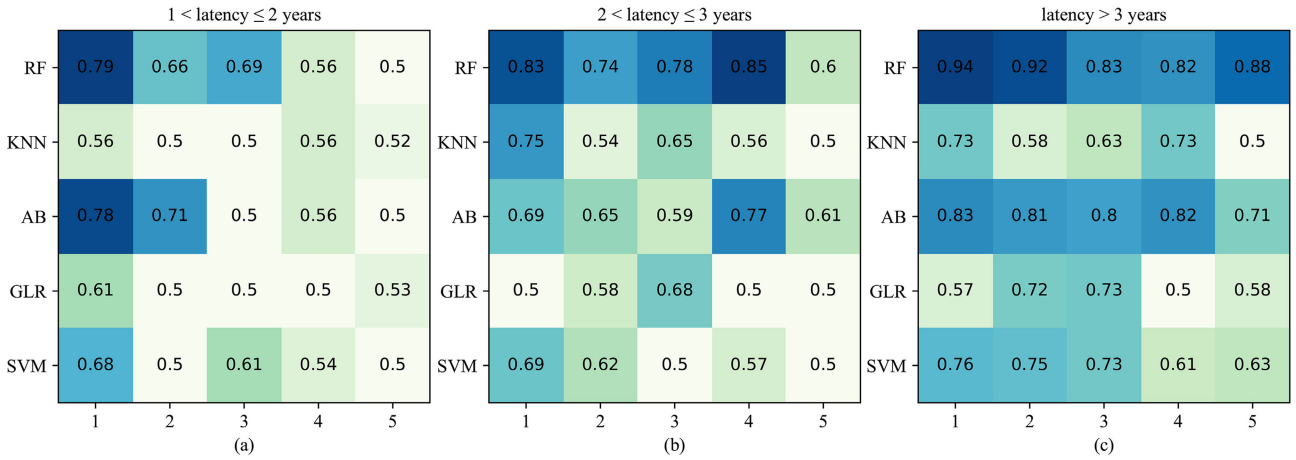

FIGURE 6. Heatmap depicting the performance (AUC) of different classification methods including random forest (RF), k-nearest neighbors (KNN), Adaboost (AB), generalized linear regression (GLR), and support vector machines (SVM).

by different dimensional reduction methods of randomly selected folds of three different latency groups. The GMP method yielded mean AUCs of 0.64 ± 0.11 ($1 <$ latency ≤ 2 years), 0.76 ± 0.10 ($2 <$ latency ≤ 3 years), and 0.88 ± 0.05 (latency > 3 years), which were higher than those of GAP (0.56 ± 0.12, 0.67 ± 0.04, and 0.76 ± 0.12), PCA (0.55 ± 0.08, 0.59 ± 0.10, and 0.69 ± 0.12), and ICA (0.61 ± 0.15, 0.56 ± 0.07, and 0.59 ± 0.08). As presented in Fig. 5, GMP achieved much better performance in predicting the diagnostic status.

C. COMPARISONS OF DIFFERENT CLASSIFICATION METHODS

In our DLFR pipeline, the capability of the selected classifier plays important roles in the task of prediction of diagnostic status. For classification, we used RF [36], trained with 150 trees (detailed setting of trees is illustrated in Table S4), a fixed seed of 161, and a minimum leaf size of 1 (minimum number of observations per tree leaf). For comparison, adaboost (AB) [43] was trained with 150 learning cycles; k-nearest neighbor (KNN) [42] was trained with a single nearest neighbors classifier using the Euclidean distance and exhaustive NSMethod; generalized linear regression (GLR) [44] was trained with a normal distribution; a support vector machine (SVM) classifier [45] was trained with a linear kernel, the parameters of which were trained by automatically adjusting the maximum-margin hyperplane and correctly separating the instance hyperplane. Fig. 6 shows a heatmap depicting the comparisons of different classification methods. According to the AUCs, the RF method had the highest performance in early detecting the TLI at the presymptomatic stage with three latency groups (0.64 ± 0.11, 0.76 ± 0.10, and 0.88 ± 0.05), followed by AB (0.61 ± 0.11, 0.66 ± 0.06, and 0.79 ± 0.04). KNN (0.53 ± 0.03, 0.60 ± 0.09, and 0.63 ± 0.09), GLR (0.53 ± 0.04, 0.55 ± 0.07, and 0.62 ± 0.09), and SVM (0.57 ± 0.07, 0.58 ± 0.07, and 0.70 ± 0.06), were reported with even lower AUCs.

D. COMPARISONS OF DIFFERENT DEEP LEARNING METHODS

We tested five well-known pretrained deep learning models, including VGG16 [37], ResNet50 [38], Densenet121 [39], Xception [40], and InceptionV3 [41], by considering that the choice of pretrained deep models affects the classification performance in our DLFR pipeline. To ensure fair comparison, we used the output of the fourth downsampling block for all the models. The input size was based on the size cropped by the segmentation results for the temporal lobe. Due to the differences in manual segmentation of the temporal lobe by experts and MRI resolutions, the sizes of the cropped images could vary. Table 5 and Table S5 (supplementary material) illustrate the AUC, SEN, SEP, and ACC results of these five pretrained models for three latency groups. As reported in Table 5, ResNet50 outperformed the other methods for the three latency groups except Xception. AUCs of the Xception model were slightly higher than those of ResNet50 in the groups of $1 <$ latency ≤ 2 years and $2 <$ latency ≤ 3 years. Accordingly, the p values of the paired t-test with Bonferroni correction between ResNet50 and Xception across all measures for the three latencies were 0.10, 0.48, and 0.004 ($p < 0.05$), respectively. Thus, ResNet50 is more suitable for the extraction of deep features.

E. EFFECTIVENESS OF OUR DLFR PIPELINE

Considering the predictive values of the selected feature representations, we evaluated our DLFR pipeline with four high-level features including deep features not in combination with longitudinal data (DFs), output representation of the fine-tuned ResNet50, radiomics features [22], and the histogram of oriented gradients (HOG) features [20]. The DFs were extracted from the specific MR data in current latency without using any longitudinal data. The fine-tuned ResNet50 model was trained with the specific MR data in current latency, and further tested for predicting the diagnosis

TABLE 5. The comparisons between different deep features for developing our DLFR pipeline.

Methods	VGG16 [37]	ResNet50 [38]	Densenet121 [39]	Xception [40]	InceptionV3 [41]
$1 <$ latency ≤ 2 years					
AUC	0.57 ± 0.04	$\mathbf{0.64 \pm 0.11}$	0.55 ± 0.06	0.65 ± 0.07	0.58 ± 0.08
SEN	0.02 ± 0.04	0.14 ± 0.14	0.05 ± 0.05	0.05 ± 0.07	0.05 ± 0.05
SEP	0.96 ± 0.04	0.97 ± 0.05	1	0.95 ± 0.04	0.97 ± 0.06
ACC	0.62 ± 0.04	0.67 ± 0.08	0.65 ± 0.02	0.62 ± 0.02	0.64 ± 0.05
p-value	0.004	-	0.1325	0.1037	0.1325
$2 <$ latency ≤ 3 years					
AUC	0.68 ± 0.09	$\mathbf{0.76 \pm 0.10}$	0.63 ± 0.07	0.78 ± 0.12	0.64 ± 0.09
SEN	0.50 ± 0.19	0.67 ± 0.22	0.49 ± 0.27	0.61 ± 0.21	0.53 ± 0.09
SEP	0.77 ± 0.13	0.70 ± 0.08	0.73 ± 0.11	0.81 ± 0.16	0.71 ± 0.14
ACC	0.65 ± 0.10	0.69 ± 0.07	0.62 ± 0.09	0.72 ± 0.11	0.62 ± 0.05
p-value	0.0531	-	0.0100	0.4751	0.0100
latency > 3 years					
AUC	0.79 ± 0.16	$\mathbf{0.88 \pm 0.05}$	0.64 ± 0.19	0.72 ± 0.13	0.59 ± 0.13
SEN	0.93 ± 0.09	0.88 ± 0.16	0.77 ± 0.22	0.76 ± 0.19	0.65 ± 0.25
SEP	0.47 ± 0.33	0.67 ± 0.16	0.39 ± 0.08	0.59 ± 0.15	0.29 ± 0.14
ACC	0.68 ± 0.17	0.77 ± 0.05	0.56 ± 0.14	0.67 ± 0.15	0.44 ± 0.12
p-value	0.0758	-	<0.0001	0.0042	<0.0001

TABLE 6. The mean AUC, SEN, SEP, and ACC values of different feature representations, including DLFR, DFs, Fine-tuned ResNet50, Radiomics, and HOG.

Methods	DLFR	DFs	Fine-tuned ResNet50	Radiomics [22]	HOG [20]
$1 <$ latency ≤ 2 years					
AUC	$\mathbf{0.64 \pm 0.11}$	0.62 ± 0.14	0.59 ± 0.09	0.56 ± 0.06	0.60 ± 0.09
SEN	0.14 ± 0.14	0.12 ± 0.10	0.20 ± 0.45	0.31 ± 0.09	0.19 ± 0.08
SEP	0.97 ± 0.05	0.94 ± 0.06	0.80 ± 0.45	0.84 ± 0.12	0.94 ± 0.07
ACC	0.67 ± 0.08	0.64 ± 0.04	0.57 ± 0.13	0.65 ± 0.07	0.67 ± 0.06
p-value	-	0.2150	0.3765	0.6978	0.8914
$2 <$ latency ≤ 3 years					
AUC	$\mathbf{0.76 \pm 0.10}$	0.61 ± 0.11	0.58 ± 0.11	0.63 ± 0.03	0.52 ± 0.03
SEN	0.67 ± 0.22	0.32 ± 0.24	0	0.59 ± 0.13	0.33 ± 0.11
SEP	0.70 ± 0.08	0.73 ± 0.11	1	0.65 ± 0.06	0.86 ± 0.11
ACC	0.07 ± 0.07	0.55 ± 0.13	0.55 ± 0.13	0.62 ± 0.03	0.62 ± 0.10
p-value	-	0.0017	0.0597	0.0393	0.0471
latency > 3 years					
AUC	$\mathbf{0.88 \pm 0.05}$	0.73 ± 0.12	0.63 ± 0.10	0.53 ± 0.04	0.58 ± 0.06
SEN	0.88 ± 0.16	1	0.80 ± 0.45	0.51 ± 0.19	0.92 ± 0.18
SEP	0.67 ± 0.16	0.34 ± 0.22	0.20 ± 0.36	0.28 ± 0.16	0.29 ± 0.14
ACC	0.77 ± 0.05	0.64 ± 0.10	0.45 ± 0.07	0.39 ± 0.15	0.57 ± 0.14
p-value	-	0.0310	<0.0001	<0.0001	<0.0001

status of the next latency. Considering the insufficient training data as well as differences in follow-up data, fine-tuning the entire ResNet50 model requires huge amounts of parameters to train and can easily lead to overfitting.[3] Thus, the weights of the convolution layers were frozen, and only the last full connection layers were fine-tuned. The fine-tuning network was trained by minimizing the binary cross-entropy loss function using a stochastic gradient descent (SGD) optimizer with the momentum of 0.9, batch size of 8, weight decay of 1×10^{-6}, an initial learning rate of 0.001 (detailed setting of batch size and learning rate are illustrated in Table S7). A predefined number of epochs was 100, which ended when the network showed no significant reduction of loss on the training set. The radiomics method, which was used in our previous work [53], was applied to extract 2068 radiomics features, including 4 non-texture features and 2064 textual features. Table 6 and Table S6 (supplementary material) illustrate the AUC, SEN, SEP, and ACC results of the four different feature representations. For the HOG method, the features were extracted by using the VLFeat[4] toolbox with a cell size of 16. As seen from Table 6, our proposed DLFR method achieved significant performance over the other high-level features.

[3]https://cs231n.github.io/transferlearning/

[4]https://www.vlfeat.org/matlab/vl_hog.html

F. IMPORTANCE OF WHITE MATTER DETECTION

In this section, we compared different MR modalities and tissues for the early detection of TLI in consideration of the invasiveness of the white matter. Fig. 7 (a-c) shows the results of five-fold cross-validation using T2w, T1c, and the combination of T1c and T2w. The AUCs obtained by T2w in three latency groups were 0.64 ± 0.11, 0.76 ± 0.10, and 0.88 ± 0.05, respectively, which were higher than those of T1w (0.62 ± 0.10, 0.73 ± 0.05, and 0.75 ± 0.12) and similar to those from the combination of T1c and T2w (0.64 ± 0.05, 0.76 ± 0.07, and 0.87 ± 0.01). It is obvious that T2w plays the most important role in detecting the changes in NPC patients. To further test the performance of different tissues, Fig. 7 (d-i) reports the AUC, SEN, SEP, and ACC results of the gray matter (GM) and white matter (WM) of T1c, T2w, and the combination of T1c and T2w. As shown in Fig. 7 (d-i), the results obtained from GM and WM of T2w were best. The AUCs of GM in three latency groups were 0.58 ± 0.07,

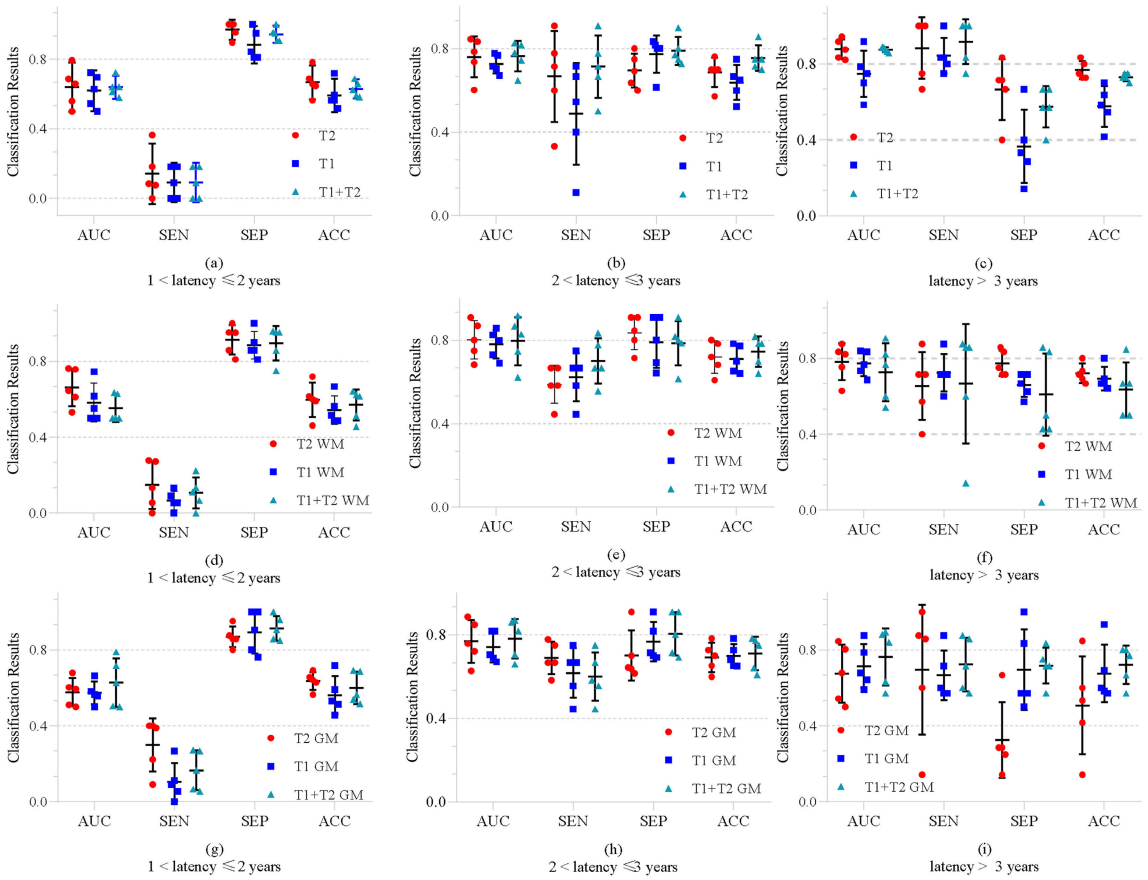

FIGURE 7. Comparisons of common, GM, and WM tissues using T2w, T1c, and the combination of T1c and T2w.

0.77 ± 0.10, and 0.67 ± 0.15, respectively, whereas those of WM were 0.66 ± 0.10, 0.80 ± 0.09, and 0.78 ± 0.09.

IV. DISCUSSION

TLI is one of the most serious complications in postradiotherapy NPC patients. This complication must be highly considered because of the irreversible brain injury. In this study, we proposed the extraction of deep longitudinal feature representations for early detection of TLI in NPC patients. To the best of our knowledge, the early detection of TLI in NPC patients using deep features has yet to be investigated.

Ideally, the total number of patients in the group of latency $<$ 1 year (illustrated in Table 2) should be the same as the number of patients we reported in Table 1. However, due to the differences in subjective initiative of patients, some patients may miss the recommended follow-up time. Thus, the summary of these four latency groups was only based on the specific follow-up time. This will result in the situation that certain patients are not included in current latency but occur in the next latency.

Recently, functional imaging techniques, including functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI), have been used to explore the invisible WM changes in the temporal lobes [9], [54]–[56]. However, three major drawbacks of these methods include time consumption for collecting additional nonstandard MRI sequences, unsatisfactory precision of current neck and head registration techniques, and limitation of the low spatial resolution for tract-based statistical analysis. Thus, it would be more efficient to use standard MRI sequences to detect TLI at the presymptomatic stage.

Due to the difficulty in following up the NPC patients after RT, the number of collected TLI-related patients is limited. It is problematic to directly train or to fine-tune a deep learning model due to the need of sufficient data in the training stage of deep classification models. To deal with the limited data, similar to [57], five-fold-cross validation is used to choose the most suitable model for robustly predicting the diagnostic of new arrival patients and new examinations. Moreover, the pretrained model can be directly used to extract highly representative features for the early detection of TLI. Regarding the choice of feature compression methods, roipooling [58], designed for the input images with random size/length, was not added into our comparison. One major limitation of adding the roipooling method would be the need to determinate the parameters (including the number and size of bounding boxes, and the size of kernels). Note that feature selection methods can improve the classification efficiency [59], however, the differences in selection methods can affect the priority of effective features for classification. Compression of the deep features results in a relatively high number of features, but keeps

more information and maintains high stability. Similar to our method, Antropova *et al.* adopted an average-pooling strategy along spatial dimensions to obtain the deep feature vectors and then performed modeling. Without feature selection, the performance of deep features across 5-fold cross-validation was still better than that of the radiomics model. In the future, more optimization work will be tried.

As illustrated in Table 6, the comparisons were evaluated among our proposed DLFR, DFs, fine-tuned ResNet50, radiomics [22], and HOG [20]. The HOG features are patch-based representations that count occurrences of gradient orientation in localized portions of an image. Comparisons with the voxel-based representations could not be done due to the difficulty in registration of all low contrast MR images across different patients over a follow-up period of 9 years. The performance of our proposed method was better than that of fine-tuned ResNet50 model. Two major drawbacks of the fine-tuned ResNet50 model include easy overfitting by limited samples and the inability to learn the longitudinal-related information. In the future, we will develop a long short-term memory (LSTM) neural network to learn the longitudinal-related information for the early detection of TLI.

As shown in the latency $\leq$ 1 year group in Fig. 6, the AUC results obtained by different classification methods are very small. Besides, it can be seen from the p values in Table 5 and 6, few significant differences can be found in the first two latency. The difference and diversity of samples is the key factor affecting the results. As illustrated in Table 2, the average examinations are increased from 2.67 ± 1.10 in latency $\leq$ 1 year group to 7.28 ± 3.03 in latency $>$ 3 years group. Especially in latency $\leq$ 1 year group, most of patients have only one examination, so only limited image information can be extracted by deep models for training. Due to the increased examinations, the longitudinal information for each patient are more abundant. The increased information can ensure the stability of the trained models. As presented in Table 5 and 6, the p values gradually decreased in three latency groups for most methods to be compared. It is worth noting that the p value is not the only measurement to evaluate the performance of models. We aim to use all measurements to choose the most suitable model for robustly predicting the diagnostic of new arrival patients and new examinations.

In this study, T1c and T2w MR images were collected in our dataset. As shown in Fig. 7, the T2w MR images are more suitable for the early detection of TLI than T1c or the combination of T1c and T2w images. The major reason may be that radiation damage commonly occurs in the white matter and T2w images can better detect white matter lesions. The increased heterogeneity in T2w signal intensity of a white matter lesion is believed to represent demyelination, gliosis, and edema [7]. Although inevitable errors existed in the segmentation of GM and WM, the performance of WM in detecting TLI was better than that of GM.

Several limitations exist in our work. First, the manual segmentation of the temporal lobe relied on a senior radiologist and may be accompanied by the intra-observer variation. The segmentation of the temporal lobe was coarse because low precision was needed for the delineation. It will be more efficient to develop a method to automatically segment the temporal lobe. Second, in this study, we only evaluated the brain injury using the temporal lobe. We plan to include more brain anatomical structures to fully explore the brain injury. Third, due to the limited retrospective data with an irregular follow-up period of 0 9 years of patients, only one single cross-validation loop is applicable for both suitable hyperparameter selection and the "best" model identification. This strategy might result in a bit inflated output because of the hyperparameter skew. Later, after adding enough data, we will apply a nested cross validation loop or separate an independent validation set for more robust hyperparameter selection to make a comprehensive assessment.

V. CONCLUSION

In this study, we proposed a deep longitudinal feature learning method for the early detection of postradiotherapy brain injury in NPC patients with follow-up period of $0 \sim 9$ years. The pretrained ResNet50 model was used to extract the high profile features, global max pooling was used to compress the high-dimensional features, and the longitudinal features were fused to combine all follow-up information. Experimental results demonstrated the effectiveness of the proposed method in the task of predicting the diagnosis at the presymptomatic stage. In the future, we will develop a method to automatically segment the temporal lobe.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This work was approved by the institutional review board, the need for informed patient consent for inclusion was waived.

REFERENCES

[1] M. L. Chua, J. T. Wee, E. P. Hui, and A. T. Chan, "Nasopharyngeal carcinoma," *Lancet*, vol. 387, no. 10022, pp. 1012–1024, 2016.

[2] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *Int. J. cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[3] K. W. Lo, K. F. To, and D. P. Huang, "Focus on nasopharyngeal carcinoma," *Cancer Cell*, vol. 5, no. 5, pp. 423–428, May 2004.

[4] L. Zeng, S.-M. Huang, Y.-M. Tian, X.-M. Sun, F. Han, T.-X. Lu, and X.-W. Deng, "Normal tissue complication probability model for radiation-induced temporal lobe injury after intensity-modulated radiation therapy for nasopharyngeal carcinoma," *Radiology*, vol. 276, no. 1, pp. 243–249, Jul. 2015.

[5] G. A. Mollet, "Fundamentals of human neuropsychology," *J. Undergraduate Neurosci. Educ.*, vol. 6, no. 2, p. R3, 2008.

[6] B. Milner, "Visual recognition and recall after right temporal-lobe excision in man," *Neuropsychologia*, vol. 6, no. 3, pp. 191–209, Sep. 1968.

[7] Y.-L. Chan, S.-F. Leung, A. D. King, P. H. K. Choi, and C. Metreweli, "Late radiation injury to the temporal lobes: Morphologic evaluation at MR imaging," *Radiology*, vol. 213, no. 3, pp. 800–807, Dec. 1999.

[8] A. Perry and R. E. Schmidt, "Cancer therapy-associated CNS neuropathology: An update and review of the literature," *Acta Neuropathol.*, vol. 111, no. 3, pp. 197–212, Mar. 2006.
[9] C. H. Chapman, V. Nagesh, P. C. Sundgren, H. Buchtel, T. L. Chenevert, L. Junck, T. S. Lawrence, C. I. Tsien, and Y. Cao, "Diffusion tensor imaging of normal-appearing white matter as biomarker for radiation-induced late delayed cognitive decline," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 82, no. 5, pp. 2033–2040, Apr. 2012.
[10] Y. Tang, X. Rong, W. Hu, G. Li, X. Yang, J. Yang, P. Xu, and J. Luo, "Effect of edaravone on radiation-induced brain necrosis in patients with nasopharyngeal carcinoma after radiotherapy: A randomized controlled trial," *J. Neuro-Oncol.*, vol. 120, no. 2, pp. 441–447, Nov. 2014.
[11] X. Zhou, X. Ou, T. Xu, X. Wang, C. Shen, J. Ding, and C. Hu, "Effect of dosimetric factors on occurrence and volume of temporal lobe necrosis following intensity modulated radiation therapy for nasopharyngeal carcinoma: A case-control study," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 90, no. 2, pp. 261–269, 2014.
[12] J. C. Baron, G. Chételat, B. Desgranges, G. Perchey, B. Landeau, V. de la Sayette, and F. Eustache, "*In vivo* mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease," *NeuroImage*, vol. 14, no. 2, pp. 298–309, Aug. 2001.
[13] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, Jan. 2013.
[14] S. Price, D. Paviour, R. Scahill, J. Stevens, M. Rossor, A. Lees, and N. Fox, "Voxel-based morphometry detects patterns of atrophy that help differentiate progressive supranuclear palsy and Parkinson's disease," *NeuroImage*, vol. 23, no. 2, pp. 663–669, Oct. 2004.
[15] M. Liu, D. Zhang, and D. Shen, "Ensemble sparse classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106–1116, Apr. 2012.
[16] F. Zhang, Y. Song, W. Cai, M.-Z. Lee, Y. Zhou, H. Huang, S. Shan, M. J. Fulham, and D. D. Feng, "Lung nodule classification with multilevel patch-based context analysis," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 4, pp. 1155–1166, Apr. 2014.
[17] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, and Adni, "Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1132–1140, Apr. 2015.
[18] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, Jul. 2017.
[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," EPFL, Tech. Rep., 2010, pp. 155–162, vol. 149300.
[20] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.
[21] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. P. M. van Stiphout, P. Granton, C. M. L. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. W. L. Aerts, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012.
[22] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, no. 1, p. 4006, Sep. 2014.
[23] B. Zhang, J. Tian, D. Dong, D. Gu, Y. Dong, L. Zhang, Z. Lian, J. Liu, X. Luo, S. Pei, X. Mo, W. Huang, F. Ouyang, B. Guo, L. Liang, W. Chen, C. Liang, and S. Zhang, "Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma," *Clin. Cancer Res.*, vol. 23, no. 15, pp. 4259–4269, Aug. 2017.
[24] M. Vallières, 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," *Phys. Med. Biol.*, vol. 60, no. 14, p. 5471, 2015.
[25] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.
[26] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 173–183, Jan. 2018.
[27] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia," *NeuroImage*, vol. 124, pp. 127–146, Jan. 2016.
[28] M. Buty, Z. Xu, M. Gao, U. Bagci, A. Wu, and D. J. Mollura, "Characterization of lung nodule malignancy using hybrid shape and appearance features," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 662–670.
[29] M. J. Nyflot, P. Thammasorn, L. S. Wootton, E. C. Ford, and W. A. Chaovalitwongse, "Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks," *Med. Phys.*, vol. 46, no. 2, pp. 456–464, Feb. 2019.
[30] W. Hua, T. Xiao, X. Jiang, Z. Liu, M. Wang, H. Zheng, and S. Wang, "Lymph-vascular space invasion prediction in cervical cancer: Exploring radiomics and deep learning multilevel features of tumor and peritumor tissue on multiparametric MRI," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101869.
[31] X. Jiang, J. Li, Y. Kan, T. Yu, S. Chang, X. Sha, H. Zheng, Y. Luo, and S. Wang, "MRI based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 3, 2020, doi: 10.1109/TCBB.2019.2963867.
[32] T. Xiao, W. Hua, C. Li, and S. Wang, "Glioma grading prediction by exploring radiomics and deep learning features," in *Proc. 3rd Int. Symp. Image Comput. Digit. Med. (ISICDM)*, 2019, pp. 208–213.
[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[34] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
[35] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400
[36] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[42] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural Comput.*, vol. 4, no. 6, pp. 888–900, Nov. 1992.
[43] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Mach. Learn.*, vol. 42, no. 3, pp. 287–320, 2001.
[44] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Boca Raton, FL, USA: CRC Press, 1993.
[45] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1055–1064, Sep. 1999.
[46] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: Improved N3 bias correction," *IEEE Trans. Med. Imag.*, vol. 29, no. 6, p. 1310, Jan. 2010.
[47] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.
[48] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[50] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[51] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[52] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
[53] X. Zhang, L. Zhong, B. Zhang, L. Zhang, H. Du, L. Lu, S. Zhang, W. Yang, and Q. Feng, "The effects of volume of interest delineation on MRI-based radiomics analysis: Evaluation with two disease groups," *Cancer Imag.*, vol. 19, no. 1, p. 89, 2019.
[54] Z. Ding, H. Zhang, X.-F. Lv, F. Xie, L. Liu, S. Qiu, L. Li, and D. Shen, "Radiation-induced brain structural and functional abnormalities in presymptomatic phase and outcome prediction," *Hum. Brain Mapping*, vol. 39, no. 1, pp. 407–427, 2018.
[55] W. Chen, S. Qiu, J. Li, L. Hong, F. Wang, Z. Xing, and C. Li, "Diffusion tensor imaging study on radiation-induced brain injury in nasopharyngeal carcinoma during and after radiotherapy," *Tumori J.*, vol. 101, no. 5, pp. 487–490, Sep. 2015.
[56] W. F. Xiong, S. J. Qiu, H. Z. Wang, and X. F. Lv, "1H-MR spectroscopy and diffusion tensor imaging of normal-appearing temporal white matter in patients with nasopharyngeal carcinoma after irradiation: Initial experience," *J. Magn. Reson. Imag.*, vol. 37, no. 1, pp. 101–108, Jan. 2013.
[57] I. Cetin, S. E. Petersen, S. Napel, O. Camara, M. A. G. Ballester, and K. Lekadir, "A radiomics approach to analyze cardiac alterations in hypertension," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 640–643.
[58] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[59] J. Nalepa, G. Mrukwa, and M. Kawulok, "Evolvable deep features," in *Applications of Evolutionary Computation* (Lecture Notes in Computer Science), vol. 10784, K. Sim and P. Kaufmann, Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-77538-8_34.

LIMING ZHONG received the B.S. and Ph.D. degrees from the Department of Biomedical Engineering, Southern Medical University, Guangzhou, China, in 2013 and 2019, respectively. Her research interests include medical image analysis, machine learning, deep learning, computerized-aid diagnosis, and medical image reconstruction.

XIAO ZHANG received the bachelor's degree in biomedical engineering from Southern Medical University, Guangzhou, China, in 2017, where she is currently pursuing the Master of Engineering degree with the Department of Biomedical Engineering. Her research interests include medical images analysis and radiomics.

YUHUA XI received the bachelor's degree in biomedical engineering from Southern Medical University, Guangzhou, China, in 2018, where she is currently pursuing the Master of Engineering degree with the Department of Biomedical Engineering. Her research interests include segmentation of medical images and bone suppression in CXRs.

ZHOUYANG LIAN received the Ph.D. degree from Southern Medical University, Guangzhou, China. She is currently an attending Doctor with the Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou. Her research interest includes image diagnosis of head and neck.

QIANJIN FENG (Member, IEEE) received the M.S. and Ph.D. degrees in biomedical engineering from First Military Medical University, China, in 2000 and 2003, respectively. From 2003 to 2004, he was a Faculty Member with the School of Biomedical Engineering, First Military Medical University. Since 2004, he has been with Southern Medical University, where he is currently a Professor and the Dean of the School of Biomedical Engineering. His research interests include medical image analysis, pattern recognition, and computerized-aided diagnosis.

WUFAN CHEN (Senior Member, IEEE) received the B.S. and M.S. degrees in applied mathematics and computational fluid dynamics from the Peking University of Aeronautics and Astronautics, China, in 1975 and 1981, respectively. From 1981 to 1987, he was with the School of Aerospace, National University of Defense Technology, China. From 1987 to 2004, he was with the Department of Training, First Military Medical University, China. Since 2004, he has been with Southern Medical University, where he currently holds the rank of a Professor with the School of Biomedical Engineering and the Director of the Key Laboratory for Medical Image Processing of Guangdong province. His research interests include medical imaging and medical image analysis.

SHUIXING ZHANG received the B.S., M.S., and Ph.D. degrees from the Department of Radiology, First Military Medical University, Guangzhou, China, in 1993, 2004, and 2007, respectively. He is currently a Professor and the Chair of the Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou.

WEI YANG received the B.Sc. degree in automation from the Wuhan University of Science and Technology, Wuhan, China, in 2001, the M.Sc. degree in control theory and control engineering from Xiamen University, Xiamen, China, in 2005, and the Ph.D. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Professor with the School of Biomedical Engineering, Southern Medical University, Guangzhou, China. His main research interests include medical image analysis, machine learning, and computerized-aid diagnosis.