Neuro-Oncology

20(9), 1251–1261, 2018 | doi:10.1093/neuonc/noy021 | Advance Access date 9 February 2018

Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation

Daesung Kang, Ji Eun Park, Young-Hoon Kim, Jeong Hoon Kim, Joo Young Oh, Jungyoun Kim, Yikyung Kim, Sung Tae Kim, and Ho Sung Kim

Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea (D.K., J.E.P., J.Y.O., J.K., H.S.K.); Department of Neurosurgery, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea (Y.-H., J.H.K.); Department of Radiology, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Korea (Y.K., S.T.K.)

Corresponding Author: Ji Eun Park, M.D., Ph.D., Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 43 Olympic-ro 88, Songpa-Gu, Seoul 05505, Korea (jieunp@gmail.com).

Abstract

Background. Radiomics is a rapidly growing field in neuro-oncology, but studies have been limited to conventional MRI, and external validation is critically lacking. We evaluated technical feasibility, diagnostic performance, and generalizability of a diffusion radiomics model for identifying atypical primary central nervous system lymphoma (PCNSL) mimicking glioblastoma.

Methods. A total of 1618 radiomics features were extracted from diffusion and conventional MRI from 112 patients (training set, 70 glioblastomas and 42 PCNSLs). Feature selection and classification were optimized using a machine-learning algorithm. The diagnostic performance was tested in 42 patients of internal and external validation sets. The performance was compared with that of human readers (2 neuroimaging experts), cerebral blood volume (90% histogram cutoff, CBV90), and apparent diffusion coefficient (10% histogram, ADC10) using the area under the receiver operating characteristic curve (AUC).

Results. The diffusion radiomics was optimized with the combination of recursive feature elimination and a random forest classifier (AUC 0.983, stability 2.52%). In internal validation, the diffusion model (AUC 0.984) showed similar performance with conventional (AUC 0.968) or combined diffusion and conventional radiomics (AUC 0.984) and better than human readers (AUC 0.825–0.908), CBV90 (AUC 0.905), or ADC10 (AUC 0.787) in atypical PCNSL diagnosis. In external validation, the diffusion radiomics showed robustness (AUC 0.944) and performed better than conventional radiomics (AUC 0.819) and similar to combined radiomics (AUC 0.946) or human readers (AUC 0.896–0.930).

Conclusion. The diffusion radiomics model had good generalizability and yielded a better diagnostic performance than conventional radiomics or single advanced MRI in identifying atypical PCNSL mimicking glioblastoma.

Keywords

atypical | diffusion-weighted imaging | magnetic resonance imaging | radiomics | primary central nervous system lymphoma

Early diagnosis of primary central nervous system lymphoma (PCNSL) is important, because its treatment differs substantially from that of other primary CNS tumors, and a correct diagnosis can avoid unnecessary surgical resection.[1] Advanced magnetic resonance imaging (MRI) techniques have been proposed, as PCNSL has lower apparent

Importance of the study

Diagnosing PCNSL mimicking glioblastoma is challenging. Here, a high-dimensional, diffusion radiomics model provided higher diagnostic performance compared with conventional radiomics or individual advanced MRI parameters. This is the first study to validate radiomics analysis with an external dataset obtained using a heterogeneous MRI protocol, which confirmed its robustness. Our results suggest diffusion radiomics could be used across centers for tumor diagnosis.

diffusion coefficient (ADC) values[2] and a lower relative cerebral blood volume (rCBV) ratio[3–5] than glioblastoma (GBM). However, there is some overlap in ADC values between PCNSL and GBM,[6,7] making it difficult to distinguish the 2 entities with the ADC parameter alone. The CBV has demonstrated better diagnostic performance,[5] but it remains difficult to unequivocally distinguish the atypical manifestations with this parameter. The combination of the imaging parameters ADC and rCBV has a diagnostic performance of 84%.[3]

A recently introduced radiomics model is able to extract descriptors using an automated data mining algorithm and to extend MRI data into a high-dimensional feature space.[8,9] Because radiomics models use high-throughput imaging features, they are prone to discover hidden information inaccessible with single-parameter approaches. Radiomics studies in neuro-oncology have reported that this approach can predict the prognosis[10,11] and/or treatment response[12] and is correlated with genetic features[13] in gliomas. Moreover, exploratory studies using radiomics models have shown great promise in differentiating various histopathological types of cancer, such as lung,[14] head and neck,[15,16] and breast cancer.[17]

The conventional types of MRI sequences include T1, contrast-enhanced T1-weighted imaging (CE-T1WI), and fluid-attenuated inversion recovery (FLAIR), which are highly available and the most widely used method in radiomics. However, the biological meaning of conventional radiomics data is often unclear. Moreover, it is difficult to derive physiological biomarkers from the extracted imaging phenotypes. On the other hand, ADC values reflect high cellularity and a high nuclear/cytoplasmic ratio in PCNSL,[18] and radiomics features of ADC may contain useful information and improve diagnostic performance relative to what is currently being utilized as a single parameter. Furthermore, because ADC maps are parametric, they can provide more robust results than conventional radiomics models when tested for different imaging parameters across different institutions. Several investigators have utilized ADC maps to differentiate PCNSL from high-grade gliomas, but no radiomics analyses have been performed. We hypothesized that a radiomics model using ADC maps along with conventional post-contrast T1-weighted imaging would result in distinct combinations of imaging parameters to differentiate atypical PCNSL from GBM. In addition, this approach will extract more information than single-parameter analyses of ADC or rCBV and improve diagnostic performance without contrast bolus injection. In this study, we tested the technical feasibility, generalizability, and diagnostic performance of a radiomics model using ADC for the identification of atypical PCNSL mimicking GBM.

Materials and Methods

Patients

Our institutional review board approved this retrospective study, and the requirement for informed consent was waived. We searched the electronic database of the Department of Radiology at our institution, retrospectively reviewed records for patients between March 2011 and March 2017, and identified 208 patients pathologically confirmed to have de novo GBM or PCNSL. Patients were immune competent (n = 208); CE-T1WI was obtained for the patients (n = 200); and preoperative diffusion-weighted imaging (DWI) was obtained for the patients (n = 194). Patients were excluded if no histopathological specimen was available (n = 39) or DWI was unreadable (because of an artifact) (n = 1). These steps yielded 154 consecutive patients (mean age, 62.4 y; male:female ratio, 81:73).

To test the diagnostic performance of our model for atypical PCNSL, we constructed a separate set with atypical PCNSL cases. An independent radiologist (H.S.K., with 15 years experience in neuroradiology) assessed the original radiological reports and assigned an 'atypical PCNSL' diagnosis to each patient when 2 differential diagnoses of GBM or PCNSL were listed in the official radiological reports for patients with pathologically confirmed lymphoma. Atypical PCNSL cases had necrosis, hemorrhage, or heterogeneous contrast-enhancing lesions.[6,19] The same number of patients (n = 21) with GBM were assigned to the set using the random number generation function in Excel. This internal validation set (n = 42) was not included in the construction process of the radiomics model, which was performed with the training set (n = 112; 70 GBMs, 42 PCNSLs).

To further validate our model, a cohort of 42 patients with pathologically confirmed GBM (n = 28) and PCNSL (n = 14) at another tertiary medical center was used for external validation of the model. The PCNSL group included 11 cases of atypical PCNSLs in the external validation set, and the assignment criteria were the same as those in the internal validation set.

Imaging Data

All MRI studies in enrolled patients in both institutions were performed on the same 3T unit (Achieva, Philips Medical Systems), using an 8-channel head coil.

The brain-tumor imaging protocol at our institution includes the following sequences: T2-weighted imaging, FLAIR imaging, T1-weighted imaging, DWI, CE-T1WI, and dynamic-susceptibility contrast (DSC) perfusion MRI. The

external validation set followed the similar brain tumor imaging protocol, except for DSC perfusion MRI. Detailed description and comparison of imaging parameters is shown in the Supplementary Table S1.

Imaging Postprocessing and Tumor Segmentation

The ADC map was calculated using the b values of 0 and 1000 s/mm^2, using a 2-point estimate of signal decay: ADC = –ln ($S[b]/S[0]$)/b, where b indicates the b value and S(0) and S(b) are the signal intensities of images with b values at 0 and 1000, respectively. The postprocessing of DSC imaging was performed using commercial software (NordicICE, NordicNeuroLab). After correction for contrast agent leakage, the whole-brain rCBV was calculated using the numerical integration of the time concentration curve. Then we normalized the rCBV (nCBV) images with the mean intensity of the contralateral normal-appearing cerebral white matter at the corona radiata, which was manually selected by a researcher (with 3 years experience in neuroimaging processing). The diameter of the selected region of interest (ROI) was 10 mm. The nCBV maps were created by dividing each CBV value by the contralateral ROI on a pixel-by-pixel basis.

Calculated ADC and DSC maps were then co-registered to the 3D CE-T1WI using SPM software (www.fil.ion.ucl.ac.uk/spm/). The co-registration process includes generation of a brain mask from 3D CE-T1WI and transformation to ADC and DSC maps for each patient. Images were registered to the brain-extracted 3D CE-T1WI volume using affine transformation with normalized mutual information as a cost function,[20] with 12 degrees of freedom and trilinear interpolation.

Tumor segmentation was performed semi-automatically by a neuroradiologist (with 4 years experience in neuro-oncological imaging) on 3D CE-T1WI to select the contrast-enhancing solid portion of the tumor using a segmentation threshold and region-growing segmentation algorithm that was implemented using MITK software (www.mitk.org, German Cancer Research Center, Heidelberg).[21] All segmented images were checked by one author. Finally, we resampled the ADC images into a uniform voxel size of 1 × 1 × 1 mm across all patients for radiomics construction. The overall process of the radiomics pipeline is shown in Fig. 1.

Extraction of Radiomics Features

Both ADC and 3D CE-T1WI data were subjected to radiomics feature extraction. For 3D CE-T1WI data, signal intensity normalization was used to reduce variance in the T1-based signal intensity of the brain. We applied the hybrid white-stripe method[21] for intensity normalization using the ANTsR and WhiteStripe packages[22,23] in R. This incorporates processes of the statistical principles of image normalization, preserving ranks among tissue and matching the intensity of tissues without upsetting the natural balance of tissue intensities.[23] Before feature extraction, we normalized the image intensities of ADC between $\mu \pm 3\sigma$ where μ and σ were the mean value of the standard deviation inside the ROI, respectively.[24]

Radiomics features were extracted using Matlab R2014b (Mathworks), in accordance with previous studies.[11,12] The 1618 radiomics features comprised 4 feature groups: 17 first-order features, 7 volume and shape features, 162 texture features, and 1432 wavelet features. The first-order, texture, and wavelet features were estimated using signal intensity and volume, and the shape features were obtained from the segmented mask. Further details of the radiomics feature extraction are described in the Supplementary Material S2, and the code for extracting

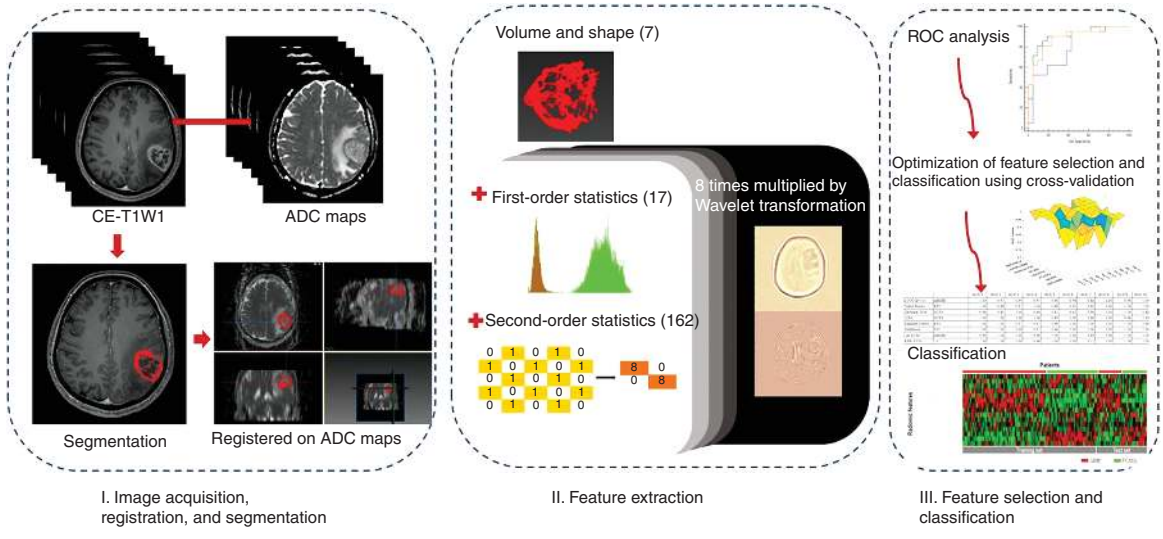

Fig. 1. The radiomics pipeline of our study.

radiomic features is given in the online supplement. The processing time to extract 1618 features was approximately 3 minutes per patient. The entire feature extraction algorithm was fully automated and yielded identical features regardless of operators.

Feature Selection and Classification Methods

Because radiomics has a highly redundant feature space,[14] it is very important to reduce highly correlated features in the selected feature subset to avoid collinearity.[25] Furthermore, a high-dimensional feature space presents potential risks of overfitting or false discovery.[26] To improve system efficiency and accuracy,[27] we computed different combinations of feature selection methods and classifications using machine learning. Methods were chosen largely based on their common use in previous studies and readily available implementation.[14,17,28] Details of each feature selection and classification methods are summarized in the Supplementary Table S3. The feature selection methods included minimum redundancy maximum relevance (mRMR), relief, mutual information, feature selection via concave minimization, recursive feature elimination (RFE), zero-norm minimization, generalized Fisher score, infinite feature selection, eigenvector centrality, unsupervised discriminative feature selection, and correlation-based and local learning-based clustering feature selection. The feature selection methods were carried out using Feature Selection Library (FSLib).[29] For classification, we investigated 8 machine learning classifiers, including k-nearest neighbor, naïve Bayes classifier, decision tree, linear discriminant analysis (LDA), random forest, adaptive boosting, linear support vector machine, and radial basis function support vector machine classifiers, using Statistics and Machine Learning Toolbox in Matlab.

The best performing combinations of different feature selection and classification methods were computed using the training set with 12 feature selection and 8 machine learning classification methods for the analysis. The different feature selection and classification methods were computed using Matlab R2014b (Mathworks).

Measurement of Single Parameters

In the internal validation set, values of the cumulative histogram parameters ADC and DSC were generated for segmented entire contrast-enhancing volumes. For cumulative histogram parameters, the 10th percentile of ADC (ADC10) and the 90th percentile of nCBV (nCBV90) were derived for the entire contrast-enhancing lesion. The *n*th percentile is the point at which *n*% of the voxel values making up the histogram are located left of the point. The number of bins was 100 in the histogram analysis. The cutoffs of 10% for the ADC histogram and 90% for the nCBV histogram were chosen because they are less influenced by random statistical fluctuations and are analogous to the minimum and maximum values that have commonly been used with the hotspot method.[30] In the external validation set, ADC10 was derived for contrast-enhancing lesion.

Comparison of the Diagnostic Performance with That of Human Readers

We tested the diagnostic performance of distinguishing PCNSL from GBM in the internal and external validation set by 2 readers, who were experienced neuroradiologists with 5 years (reader 1) and 20 years experience (reader 2). After anonymization and data randomization, the readers were given 4 image sets for each patient, which included FLAIR, DWI, ADC maps, and CE-T1WI images. The reason for choosing the above sequences was to establish a comparison with the radiomics model, while providing important imaging sequences to simulate the radiology workflow. We did not separately test the imaging sequences to prevent learning effect of readers. The readers indicated the level of confidence in their interpretation for each patient using 5 levels: definitely PCNSL, probably PCNSL, equivocal, probably GBM, and definitely GBM.

Statistical Analysis

Optimization of the Radiomics Model in the Training Set

All radiomics features were *z* transformed for group comparison. To find the best combination between feature selection and classification methods, we tuned hyperparameters with the fashion of grid search. For each of the 12 feature selection methods, we incrementally selected features ranging from 5 to 50, with an increment of 5. The 8 classifiers had different types of hyperparameters, and the detailed parameter search is given in the Supplementary Table S3 for each classifier. The subsets of selected features were evaluated by being paired with each of the 8 classifiers and the diagnostic performance was calculated using area under the receiver operating characteristic curve (AUC). For each combination, we trained the model on the subsampled cohort in the training set (size *n*/10) and cross-validated data using 10-fold cross-validation. The AUC of each combination was compared. Optimization was separately performed for ADC maps and 3D CE-T1WI data.

Measuring Stability of the Radiomics Model

The stability of the combination was quantified empirically using relative standard deviation (RSD).[15] RSD was defined as a percentage equal to: (standard deviation of AUC/mean AUC) × 100, where the AUC values were the 10 obtained from cross-validation in the training set.

Comparison of Diagnostic Performance

Based on the radiomics construction in the training set, the best combination of feature selection and classification methods was used on ADC and CE-T1WI data in the validation set. Also, the best ADC and CE-T1WI radiomics features were combined and the diagnostic performance was calculated. The ADC radiomics model, the CE-T1WI radiomics model, combined ADC and CE-T1WI radiomics model, nCBV90, and ADC10 were compared in their diagnostic performance using AUC. The accuracy, sensitivity,

and specificity are defined as follows, in correctly diagnosing PCNSLs (TP: true positive, TN: true negative, FP: false positive, FN: false negative):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$
$$Sensitivity = \frac{TP}{TP + FN},$$
$$Specificity = \frac{TN}{TN + FP}$$

P-values less than 0.05 were considered to indicate significant differences. Statistical analyses were performed using R v3.3.2 statistical software and Matlab R2014b with Windows 10.

Measuring Effect of Adding Atypical PCNSLs in the Training Set

Additionally, using the best model in each radiomics analysis, the effect of adding atypical PCNSLs in the training set was tested. The lymphoma dataset was randomly divided without selecting for atypical feature and the diagnostic performance was tested in both validation sets.

Results

Clinical Characteristics of the Study Patients

The baseline demographics and clinical characteristics of the patients are summarized in Table 1. There were 112 patients in the training set (70 GBMs, 42 PCNSLs), 42 patients in the internal validation set (21 GBMs, 21 atypical PCNSLs), and 42 patients in the external validation set (28 GBMs, 14 atypical PCNSLs). There were no significant differences in sex between patients with GBM and PCNSL in either the training or the validation set. The PCNSL group showed a smaller tumor size than the GBM group on CE-T1WI in both the training ($P = 0.0002$) and the internal ($P = 0.02$) and external validation sets ($P = 0.04$).

Determination of the Best Radiomics Model in the Training Set

Using radiomics features, each combination of the 12 feature selection and 8 classification methods was trained and diagnostic performance was calculated with a 10-fold cross-validated threshold. Table 2 summarizes the results of the diagnostic performance using different combinations for each feature selection and classification method, according to the feature numbers.

ADC Radiomics

In the ADC radiomics model, the combination of the RFE feature selection and the random forest classification method (model 5, mean AUC = 0.983) showed the best diagnostic performance. The stability of this combination was 2.52% (RSD), and the optimal number of radiomics features was 15. Significant radiomics features are given in the Supplementary Table S4. Fig. 2 demonstrates different combinations of each feature selection and classification method with 15 ADC radiomics features.

CE-T1WI Radiomics

The combination of the relief feature selection method and LDA classifier showed the best diagnostic performance (model 4, mean AUC = 0.976), with 40 optimal radiomic features. The stability of this combination was 1.73%. Significant radiomic features of CE-T1WI are presented in Supplementary Table S6.

Generalizability of the Radiomics Model in the Validation Sets

Table 3 summarizes the diagnostic performance of the radiomics model in internal and external validation sets. In all, 15 significant ADC radiomics features and combinations of RFE feature selection and random forest classifier were trained in the validation sets. In an independent internal validation, the ADC radiomics model demonstrated an AUC of 0.984 (95% CI: 0.945–1), with a sensitivity of 80.9%, a specificity of 100%, and an accuracy of 90.5%. The 40 significant CE-T1WI radiomics features and the combination of relief feature selection with LDA classification were used in the validation set. The CE-T1WI radiomics model had an AUC of 0.968 (95% CI: 0.913–1), with a sensitivity of 85.7%, a specificity of 95.2%, and an accuracy 85.7%. The combined ADC and CE-T1WI radiomics features (using the 15 ADC and 40 CE-T1WI radiomics features) showed an AUC of 0.984 (95% CI: 0.945–1), a sensitivity of 85.7%, a specificity of 100%, and an accuracy of 92.9%.

Table 1. Baseline demographics and clinical characteristics of patients

Pathology	Training Set		*P*-value	Internal Validation Set		*P*-value	External Validation Set		*P*-value
	PCNSL	GBM		PCNSL	GBM		PCNSL	GBM	
Patients (*N*)	42	70		21	21		14	28	
Number of females (*N*)	23	29	.47	11	10	.51	12	11	.53
Age, y	64.1 ± 12.9	61.1 ± 11.5	.19	67.8 ± 10.6	62.3 ± 11.4	.01	57.9 ± 12.1	57.4 ± 11.4	.43
Tumor size (cm^2)	10.9 ± 7.9	16.1 ± 8.4	.002	9.9 ± 7.7	16.2 ± 7.9	.02	11.9 ± 8.7	16.2 ± 7.9	.04

Note: *P*-values apply to the differences between the PCNSL and GBM groups. Age and tumor size are expressed as means ± standard deviations.

Table 2. Diagnostic performance and stability of the ADC and post-contrast T1 radiomics model, using different combinations of feature selection and classification methods in the training set

Model	Classification Method	Best Feature Selection Method	Optimum Feature Number	Mean AUC Value	RSD (%)
ADC					
1	k-NN	mRMR	45	0.968	3.99
2	Naïve Bayes	RFE	45	0.955	6.88
3	Decision tree	RFE	10	0.910	9.51
4	LDA	mRMR	15	0.982	3.53
5	Random forest	RFE	15	0.983	2.52
6	Adaboost	FSV	20	0.979	5.76
7	Lin-SVM	L0	35	0.979	3.53
8	RBF-SVM	L0	15	0.968	9.28
Post-contrast T1					
1	k-NN	FSV	20	0.940	5.77
2	Naïve Bayes	L0	5	0.940	7.66
3	Decision tree	Relief	40	0.927	10.61
4	LDA	Relief	40	0.976	1.73
5	Random forest	FSV	45	0.954	5.33
6	Adaboost	Inf.FS	10	0.941	8.95
7	Lin-SVM	Relief	15	0.958	5.23
8	RBF-SVM	RFE	50	0.937	5.24

Abbreviations: RSD = relative standard deviation; k-NN = k-nearest neighbor; mRMR = minimum redundancy maximum relevance; RFE = recursive feature elimination; LDA = linear discriminant analysis; Adaboost = adaptive boosting; FSV = feature selection via concave minimization; Lin-SVM = linear support vector machine; L0 = zero-norm minimization; RBF-SVM = radial basis function support vector machine; Inf.FS = infinite feature selection

In the external validation set, the ADC radiomics model demonstrated an AUC of 0.944 (95% CI: 0.856–1), with a sensitivity of 85.7%, specificity of 75%, and an accuracy of 88.6%. The CE-T1WI radiomics model had an AUC of 0.819 (95% CI: 0.671–0.967), with a sensitivity of 71.4%, specificity of 82.1%, and an accuracy of 78.6%. Meanwhile, the combined ADC and CE-T1WI radiomics features had an AUC of 0.946 (95% CI: 0.861–1), with a sensitivity of 85.7%, a specificity of 82.1%, and an accuracy of 83.3%. The combined radiomics showed similar diagnostic performance with ADC radiomics in both internal and external validation sets.

Fig. 3 shows the heatmap of GBM and PCNSL in the training, internal validation, and external validation sets.

Comparison of the Diagnostic Performance of the Radiomics Model and Single Parameters in the Validation Set

ADC10 was obtained in both internal and external validation sets, whereas nCBV was obtained in only the internal validation set. Using the optimal cutoff, ADC10 had an AUCs of 0.79 (95% CI: 0.633–0.898) and 0.81 (95% CI: 0.683–0.901) in the internal and external validation sets, respectively, with a sensitivity of 95.2% and 75.9%, a specificity of 57.1% and 82.1%, and an accuracy of 76.2% and 84.9%, respectively. Using the optimal cutoff, nCBV90 had an AUC of 0.905 (95% CI: 0.774–0.973), with a sensitivity of 80.9%, a specificity of 90.5%, and an accuracy of 85.7%. Both ADC radiomics and CE-T1WI radiomics had better diagnostic performance than the single parameters ADC10 and nCBV90 (Table 3). A comparison of AUCs among the ADC radiomics model, CE-T1WI radiomics model, and single parameters (ADC10 or nCBV90) did not show statistically significant differences, but the trend of higher diagnostic performance of the radiomics model was consistent in the training and validation sets.

Diagnostic Performance of the Human Readers

Table 3 summarizes the results. In the internal validation set, reader 1 had an AUC of 0.825 (95% CI: 0.755–0.881), with a sensitivity of 69.4%, a specificity of 95.6%, and an accuracy of 86.7%, while reader 2 had an AUC of 0.908 (95% CI: 0.851–0.949), with a sensitivity of 83.9%, a specificity of 97.8%, and an accuracy of 92.8%. In the external validation set, reader 1 had an AUC of 0.896 (95% CI: 0.836–0.940), with a sensitivity of 82.5%, a specificity of 96.7%, and an accuracy of 90.8%, while reader 2 had an AUC of 0.930 (95% CI: 0.831–0.981), with a sensitivity of 89.7%, a specificity of 96.4%, and an accuracy of 93.0%.

Effect of Incorporating Atypical PCNSL in the Training Set

When the training set included atypical PCNSLs, the ADC radiomics model showed an AUC of 0.971 (95% CI: 0.917–1), with a sensitivity of 85.7%, a specificity of 95.2%, and an accuracy 90.5% in the internal validation set, and

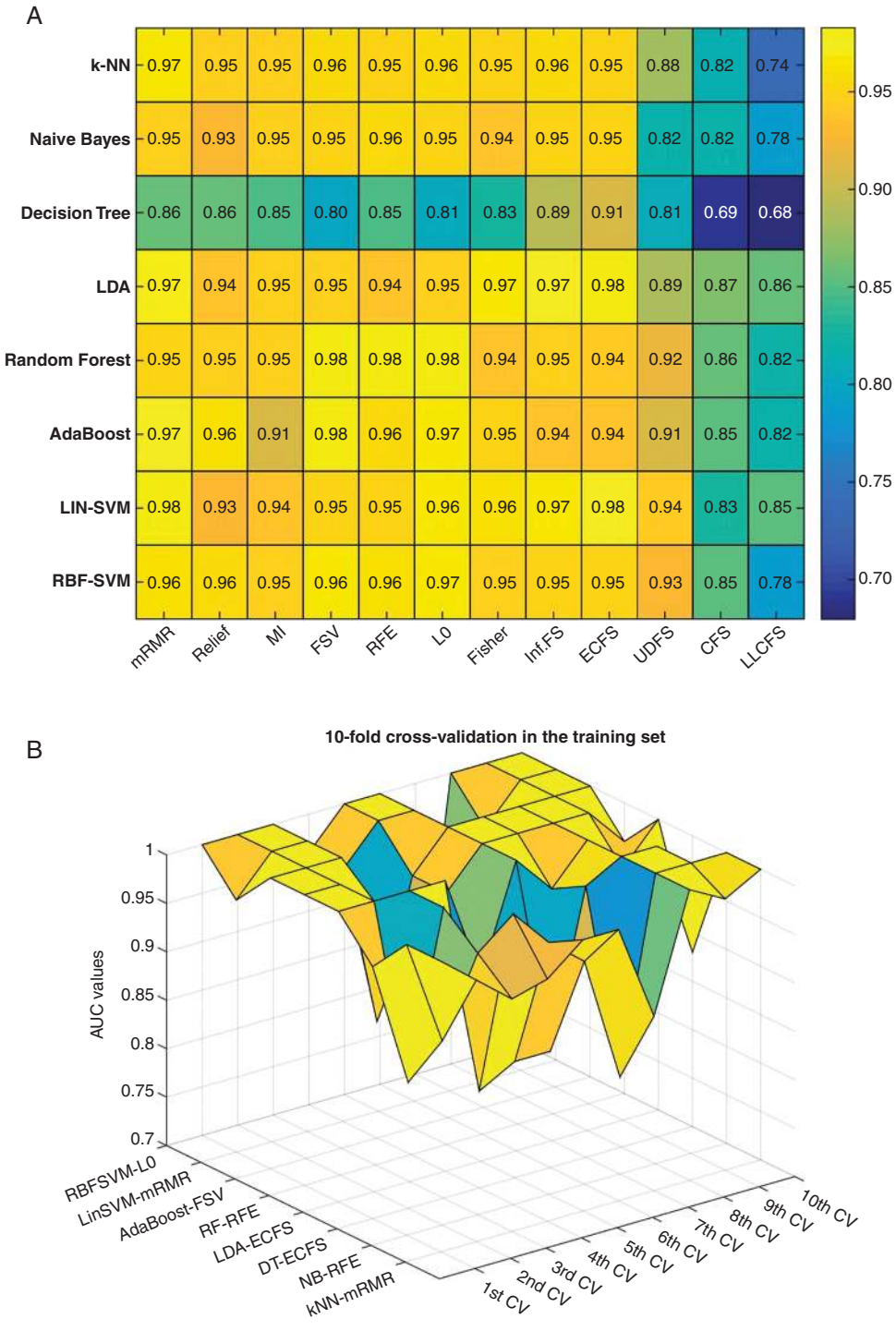

Fig. 2 (A) Heatmap depicting the diagnostic performance (AUCs) of 12 feature selection (columns) and 8 classification (rows) methods in the training set. (B) Stability of the AUCs using 10-fold cross-validation (CV) in the training set. Color scale: expressed from yellow (AUC 1.00) to blue (AUC 0.65).

an AUC of 0.911 (95% CI: 0.802–1), with a sensitivity of 85.7%, a specificity of 75%, and an accuracy of 85.6% in the external validation set. In CE-T1WI radiomics, the model showed an AUC of 0.938 (95% CI: 0.862–1), with a sensitivity of 90.5%, a specificity of 90.5%, and an accuracy of 90.48% in the internal validation; and an AUC of 0.773 (95% CI: 0.611–0.934), with a sensitivity of 71.4%, a specificity of 79%, and an accuracy of 76.2% in the external

Table 3. Comparison of diagnostic performance of the radiomics model, combined radiomic features, histogram analyses of ADC and CBV, and human readers in the validation set

Parameter	AUC Value	Cutoff	Sensitivity (%)	Specificity (%)
Internal validation set				
ADC radiomics	0.984 (0.945, 1)	NA	80.9	100
CE-T1WI radiomics	0.968 (0.913, 1)	NA	85.7	95.2
ADC+ CE-T1WI radiomics	0.984 (0.945, 1)	NA	85.7	100
ADC10	0.787 (0.633, 0.898)	0.92	95.2	57.1
nCBV90	0.905 (0.774, 0.973)	4.27	80.9	90.5
Human readers	0.825–0.908 (0.755, 0.949)	NA	69.4–83.9	95.6–97.8
External validation set				
ADC radiomics	0.944 (0.856, 1)	NA	85.7	75.0
CE-T1WI radiomics	0.819 (0.671, 0.967)	NA	71.4	82.1
ADC+ CE-T1WI radiomics	0.946 (0.861, 1)	NA	85.7	82.1
ADC10	0.809 (0.683,0.901)	0.80	75.9	82.1
Human readers	0.896–0.930 (0.831, 0.981)	NA	82.5–89.7	96.4–96.7

Note. Data in parentheses are 95% CIs. ADC is shown in $\times 10^{-3}$ mm^2 sec^{-1}.

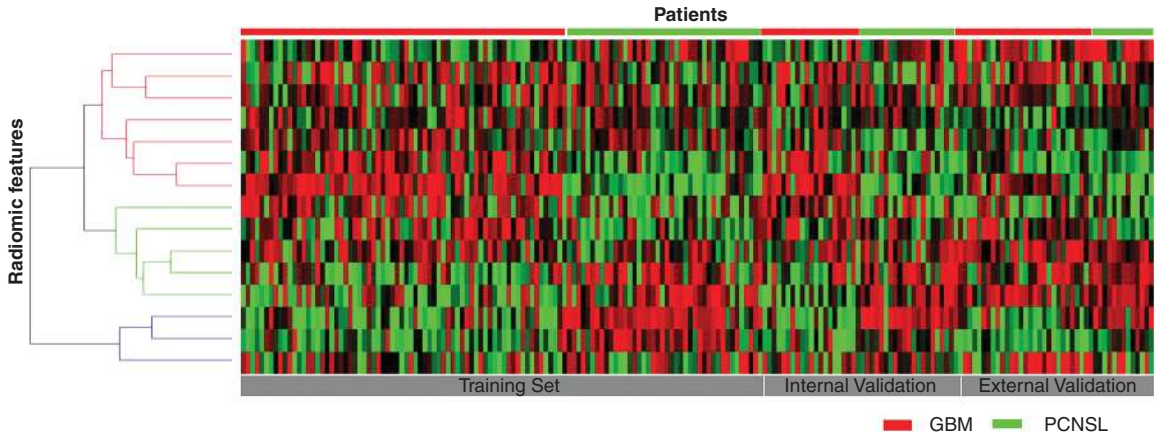

Fig. 3 Heatmap of radiomic features selected on the basis of recursive feature elimination and a random forest classifier in the diffusion radiomics model. Each row corresponds to a *z*-score of normalized radiomics features, and each column corresponds to one patient. The heatmap is grouped for the training, internal validation, and external validation sets and the GBM versus PCNSL group by means of radiomics analysis.

validation set. Compared with the radiomics model trained with typical PCNSLs, the diagnostic performance of the ADC radiomics model was similar in both validation sets, while CE-T1WI radiomics dropped in the external validation comprising typical PCNSLs mostly.

Discussion

Here, we created a high-dimensional feature space from a radiomics model by extracting 1618 phenotypic features and computing different feature selection and classification algorithms, which yielded high diagnostic performance in identifying atypical PCNSL mimicking GBM. We found that the ADC radiomics model was particularly useful in showing better diagnostic performance than human readers and single-parameter ADC or even nCBV, with robust results in external validation. Also, the ADC radiomics model gave similar diagnostic performance with that from combined ADC and CE-T1WI radiomics features in both validation sets. Particularly, the ADC radiomics model provided more robust results than the CE-T1WI radiomics model across different imaging acquisitions, supporting its use as a multicenter imaging biomarker in radiomics research.

Atypical PCNSL is still a challenge in radiological diagnosis when it manifests ring-like enhancement, necrosis, internal hemorrhage, or calcification mimicking GBM. As

evidence, the readers (comprising 4 and 20 years of experience in neuroradiology) showed diagnostic performance of AUC 0.825 to 0.908 in diagnosing atypical PCNSLs. On the other hand, a number of studies have found that the mean, minimum/maximum, or histogram value of ADC and CBV extracted from physiological imaging biomarkers can improve diagnostic performance to differentiate PCNSL[4,7,31] from GBM, with better diagnostic performance reported for CBV.[4,5] However, the diagnostic performance of this single-parameter approach drops in atypical manifestations of PCNSL, with the minimum ADC showing an AUC of 0.71–0.73[6] and with the combination of mean ADC and rCBV showing an accuracy of 84%.[3] The single-parameter approach is inherently limited, in that it does not characterize tumor heterogeneity or quantify comprehensive phenotypic information from imaging data. Using a radiomics model, we expanded the feature dimensions of conventional and DW imaging to address an important clinical question of the diagnosis of atypical PCNSL mimicking GBM. Our model yielded better diagnostic performance than the human readers and single-parameter ADC and CBV in the validation sets. Furthermore, the radiomics pipeline was automated and the extraction of 1618 features took 3 minutes per patient. This further emphasized the potential clinical utility of radiomics analysis.

Our radiomics model utilized ADC values for all enhanced tumors. However, to date, they have not been used as a useful source of a radiomics model. Recent studies that have focused on performing analyses of conventional MRI have used T1, FLAIR, and CE-T1WI protocols, but signal intensities are arbitrary units and lack biological information. The ADC radiomics model promises to be able to examine the tumor microenvironment and can enrich existing imaging features, since ADC values can contain biological information. Significant ADC radiomics features included minimum ADC and histogram skewness, which may indicate high cellularity as well as a homogeneous tumor microenvironment compared with glioblastoma. The most relevant imaging feature was the Gray-level co-occurrence matrix (GLCM) in both the ADC radiomics. The GLCM is a texture-analysis method that calculates how often pairs of pixels with specific values and in a specified spatial relationship occur in an image.[32] Since GLCMs characterize tumor heterogeneity, obtaining statistical measures and creating radiomics models from GLCM are important to distinguish PCNSL from GBM. This is further supported by a recent pixelwise study for prostate cancer,[33] which found that increases in ADC values correlated positively with extracellular spaces and nuclear sizes, but negatively with nuclear counts. In addition, ADC maps allow reliable feature extraction using absolute values within the same protocol as in our present study, whereas contrast-enhanced T1 signals vary even within the same subject,[22] and different preprocessing methods of intensity normalization may give different results.[34]

Notably, the diagnostic performance of CE-T1WI radiomics decreased in the external validation set. Also, diagnostic performance of combined CE-T1WI and ADC radiomics features was similar in both internal and external validation sets. In contrast to a highly homogeneous imaging acquisition scheme in the training set, the external validation set had highly heterogeneous imaging protocols for both DWI and CE-T1WI. A previous CT study[35] showed that radiomics features are reproducible over a wide range of imaging settings, unless smooth and sharp reconstruction algorithms are used. Our data suggest that CE-T1WI radiomics features are more vulnerable to changes in acquisition parameters, wherein margin, gadolinium contrast media, and signal-to-noise ratio can be easily varied across imaging protocols. However, ADC maps are parametric and are likely to be robust across the different acquisition schemes. Moreover, ADC maps are derived from non-enhanced MRI data and their clinical utility can be further enhanced.

Further, we tested the effect of including atypical PCNSLs in the training set. Compared with the original training set, the performance of CE-T1WI radiomics dropped in the external validation, while ADC radiomics showed similar results. This is probably because the radiomics analysis is focused on the solid enhancing portion, while "atypical" features of PCNSL usually come from necrosis or hemorrhage, apart from the solid portion. This is partly supported by a histopathological study of PCNSL showing that the macroscopic features of PCNSLs are similar in immunocompetent and immunodeficient patients[36] and that the tumor cells are diffusely compact with a characteristic angiocentric growth pattern in PCNSL, which may result in a similar diffusion restriction pattern. This needs further investigation.

Machine learning methods with high statistical power and stability are desired for radiomics analysis.[15,28] In addition, a robust algorithm for feature selection[37] is important to handle enormous amounts of radiomics features and to reduce the curse of dimensionality.[38] We computed 12 feature selection and 8 classification methods to find an optimal method yielding the best diagnostic performance and model stability. Among them, RFE feature selection with random forest classification and relief feature selection using LDA showed the best diagnostic performance in the ADC and CE-T1WI radiomics models, respectively. In addition, the ADC radiomics model was optimized with 15 features, whereas the CE-T1WI radiomics model needed 40 features. Our current results indicate that the radiomics model must be optimized based on different imaging data and different outcomes, including diagnosis, histopathological subtypes, and survival, to provide a reliable algorithm and automated analysis platform.

This study has several limitations. The first is its retrospective design and the small number of patients in the validation set. We attempted to overcome this issue by constructing a separate model with the training set and demonstrating robust results using the external validation set. Furthermore, results showed robustness regarding the differed ratio of GBM and PCNSL in internal validation (1:1) and external validation (2:1), or different combinations of training set including atypical features of PCNSLs. Second, comparison of diagnostic performance radiomics with CBV was performed using only internal validation. This was due to the lack of DSC imaging in the external validation set. To expand radiomics research and validate different radiomics models, larger cohort studies and standardization of imaging acquisition will be necessary. Third, the preprocessing of DWI is still not well established. Although we tested the ADC radiomics using different scanning parameters on 3T, testing with a 1.5T system will be an important

task that must be completed before ADC radiomics is interchangeably used as a multicenter imaging biomarker.

In conclusion, our radiomics model utilizing ADC maps had good generalizability and showed a better diagnostic performance than single-parameter measurements in identifying atypical PCNSL mimicking GBM by providing robust high-dimensional analyses of conventional and physiological imaging features.

Supplementary material

Supplementary material is available at *Neuro-Oncology* online.

Funding

This work was supported by the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (1720030).

Conflict of interest statement. None to declare.

References

1. Hunt MA, Jahnke K, Murillo TP, Neuwelt EA. Distinguishing primary central nervous system lymphoma from other central nervous system diseases: a neurosurgical perspective on diagnostic dilemmas and approaches. *Neurosurg Focus*. 2006;21(5):E3.
2. Stadnik TW, Chaskis C, Michotte A, et al. Diffusion-weighted MR imaging of intracerebral masses: comparison with conventional MR imaging and histologic findings. *AJNR Am J Neuroradiol*. 2001;22(5):969–976.
3. Kickingereder P, Wiestler B, Sahm F, et al. Primary central nervous system lymphoma and atypical glioblastoma: multiparametric differentiation by using diffusion-, perfusion-, and susceptibility-weighted MR imaging. *Radiology*. 2014;272(3):843–850.
4. Lu S, Gao Q, Yu J, et al. Utility of dynamic contrast-enhanced magnetic resonance imaging for differentiating glioblastoma, primary central nervous system lymphoma and brain metastatic tumor. *Eur J Radiol*. 2016;85(10):1722–1727.
5. Xu W, Wang Q, Shao A, Xu B, Zhang J. The performance of MR perfusion-weighted imaging for the differentiation of high-grade glioma from primary central nervous system lymphoma: a systematic review and meta-analysis. *PLoS One*. 2017;12(3):e0173430.
6. Suh CH, Kim HS, Lee SS, et al. Atypical imaging features of primary central nervous system lymphoma that mimics glioblastoma: utility of intravoxel incoherent motion MR imaging. *Radiology*. 2014;272(2):504–513.
7. Haldorsen IS, Espeland A, Larsson EM. Central nervous system lymphoma: characteristic findings on traditional and advanced imaging. *AJNR Am J Neuroradiol*. 2011;32(6):984–992.
8. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
9. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234–1248.
10. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol*. 2017;19(6):862–870.
11. Kickingereder P, Burth S, Wick A, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*. 2016;280(3):880–889.
12. Kickingereder P, Götz M, Muschelli J, et al. Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. *Clin Cancer Res*. 2016;22(23):5765–5771.
13. Hu LS, Ning S, Eschbacher JM, et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro Oncol*. 2017;19(1):128–137.
14. Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol*. 2016;6:71.
15. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272.
16. Zhang B, Tian J, Dong D, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clin Cancer Res*. 2017;23(15):4259–4269.
17. Li H, Zhu Y, Burnside ES, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *NPJ Breast Cancer*. 2016;2.
18. Barajas RF Jr, Rubenstein JL, Chang JS, Hwang J, Cha S. Diffusion-weighted MR imaging derived apparent diffusion coefficient is predictive of clinical outcome in primary central nervous system lymphoma. *AJNR Am J Neuroradiol*. 2010;31(1):60–66.
19. Savage J, Quint D. Atypical imaging findings in an immunocompetent patient. Primary central nervous system lymphoma. *JAMA Oncol*. 2015;1(2):247–248.
20. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging*. 1997;16(2):187–198.
21. Nolden M, Zelzer S, Seitel A, et al. The medical imaging interaction toolkit: challenges and advances: 10 years of open-source development. *Int J Comput Assist Radiol Surg*. 2013;8(4):607–620.
22. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*. 2011;54(3):2033–2044.
23. Shinohara RT, Sweeney EM, Goldsmith J, et al; Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin*. 2014;6:9–19.
24. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging*. 2004;22(1):81–91.
25. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36(1):27–46.
26. Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Min Knowl Disc*. 1997;1(1):55–77.
27. Duangsoithong R, Windeatt T. Relevant and redundant feature analysis with ensemble classification. Paper presented at: Advances in Pattern Recognition, 2009. ICAPR '09. Seventh International Conference ; 2009.
28. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*. 2015;5:13087.
29. Roffo G, Melzi S, Cristani M. Infinite feature selection. Paper presented at: Proceedings of the IEEE International Conference on Computer Vision; 2015.

30. Chung WJ, Kim HS, Kim N, Choi CG, Kim SJ. Recurrent glioblastoma: optimum area under the curve method derived from dynamic contrast-enhanced T1-weighted perfusion MR imaging. *Radiology*. 2013;269(2):561–568.
31. Toh CH, Castillo M, Wong AM, et al. Primary cerebral lymphoma and glioblastoma multiforme: differences in diffusion characteristics evaluated with diffusion tensor imaging. *AJNR Am J Neuroradiol*. 2008;29(3):471–475.
32. Materka A, Strzelecki M. *Texture analysis methods–a review*. COST B11 report. Brussels: Technical University of Iodz, Institute of Electronics; 1998:9-11.
33. Lin YC, Lin G, Hong JH, et al. Diffusion radiomics analysis of intratumoral heterogeneity in a murine prostate cancer model following radiotherapy: Pixelwise correlation with histology. *J Magn Reson Imaging*. 2017;46(2):483–489.
34. Fortin JP, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT; Alzheimer's Disease Neuroimaging Initiative. Removing inter-subject technical variability in magnetic resonance imaging studies. *Neuroimage*. 2016;132:198–212.
35. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428.
36. Bhagavathi S, Wilson JD. Primary central nervous system lymphoma. *Arch Pathol Lab Med*. 2008;132(11):1830–1834.
37. Zhang Y, Ding C, Li T. Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics*. 2008;9(Suppl 2):S27.
38. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.