Received June 1, 2020, accepted July 12, 2020, date of publication July 17, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010006

Bone Marrow Radiomics of T1-Weighted Lumber Spinal MRI to Identify Diffuse Hematologic Marrow Diseases: Comparison With Human Readings

EO-JIN HWANG, SANGHEE KIM, AND JOON-YONG JUNG
Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, South Korea

Corresponding author: Joon-Yong Jung (jjdragon112@gmail.com)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2019R1C1C1007122.

ABSTRACT We developed a radiomics model to differentiate hematologic marrow diseases and compared the performance with radiologists' readings and a quantitative measurement. Patients were retrospectively analyzed from the diseased (n = 254) and control groups (n = 230). A sagittal T1-weighted lumbar spinal MR image was normalized by an intervertebral disk, and bone marrow was segmented. A hundred features were extracted, and final features were selected using Principle Component Analysis (PCA) and least absolute shrinkage and selection operator (LASSO). Finally, Random forest (RF) and logistic regression (LR) models were trained. Two radiologists with different levels of experience analyzed the images for the presence of bone marrow diseases, independently. The area under the receiver operating characteristic curves (AUC) and decision curve analysis (DCA) was evaluated. Among the subjects, 363 cases were assigned as a training set and 121 as a validation set. The combination of LASSO and RF produced the best results. With the validation set, the sensitivity (SE) was 87.3%, specificity (SP) was 86.2% and AUC was 0.928 ($p < 0.05$). We selected Firstorder -Maximum as the best feature to identify diseased marrows, which achieved SE of 75.0% and AUC of 0.787 ($p < 0.05$). The reader with 11 years of experience yielded SE of 86.5% and AUC of 0.861 ($p < 0.05$). The second reader with 1 year of experience yielded SE of 75.0% and AUC of 0.767 ($p < 0.05$). We demonstrated the advantage of bone marrow radiomics over conventional methods of diagnosing with radiologists' readings and quantitative measurements.

INDEX TERMS Bone marrow, magnetic resonance imaging, radiomics, least absolute shrinkage and selection operator, principal component analysis, random forest, logistic regression.

I. INTRODUCTION

Bone marrow is an important part to be interpreted in spinal magnetic resonance imaging (MRI). However, less experienced physicians have difficulty in determining whether the bone marrow is abnormal on MRI and in warranting further diagnostic work up [1]. The reason that the diagnosis of bone marrow disease is often challenging is because diffuse bone marrow infiltration may not be discernable due to the repetitive patterns across the entire marrow spaces [2]. In addition, bone marrow interpretation is complicated by age-dependent variabilities and marrow reconversion in response to physiological oxygen demands [3], [4].

Radiomics is a novel field in medical imaging that aims to utilize large amount of quantitative features in order to advance decision support. Previously, ^{18}F-FDG PET/CT radiomics from bone marrows were used to differentiate various types of diseases [5]–[7], and radiomics from T1 and T2-weighted MRI were used to discriminate bone chondrosarcoma [8], metastatic diseases [9], and osteoporosis [10]. However, the MRI-based radiomics studies often discriminated diseases with low predictive performance [10] or with performance only comparable with experienced radiologists [8]. We hypothesized that the well-trained model based on bone marrow radiomics would advance clinical

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

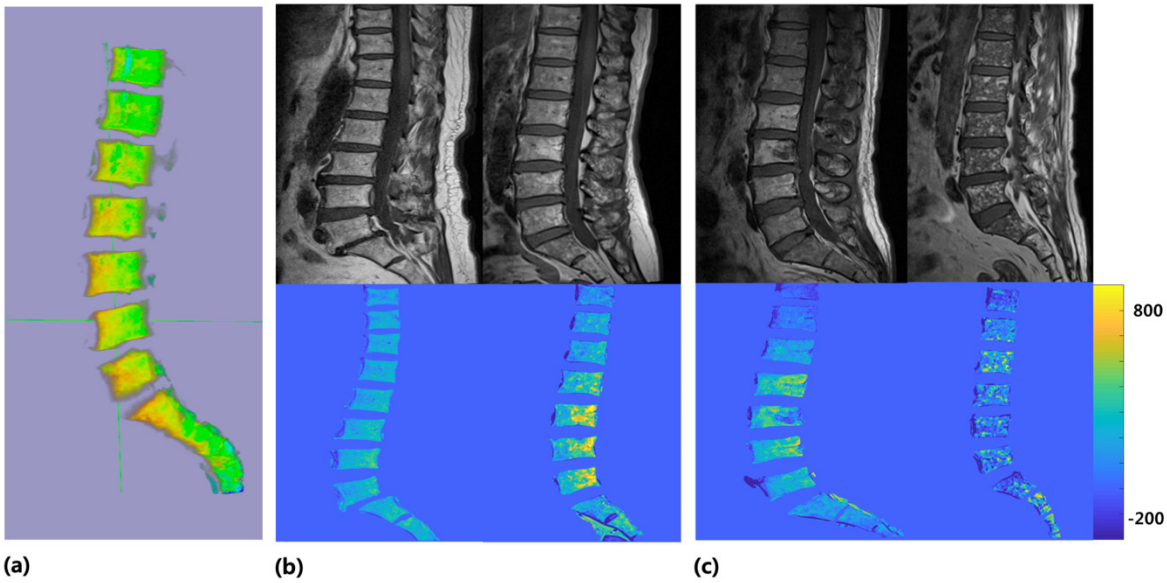

FIGURE 1. Representative slices of the sagittal T1-weighted images of the (a) segmented bone marrows in 3 dimension (3D), (b) normal spines and (c) spines with hematologic diseases (top), and the corresponding normalized and segmented bone marrows in 2D, which were obtained by subtracting the whole image pixels from the annulus fibrosus of non-degenerated intervertebral disks and by segmenting the bone marrows using a 3-dimensional semi-automatic algorithm (bottom).

decision of bone marrow diseases than conventional methods of radiologists' readings and quantitative measurements. The purpose of our study was in two folds: first, we constructed six radiomics models with various types of feature selection methods and machine learning classifiers to identify one model with the best performance; second, we compared the performance of our radiomics model with that of the human readers with different levels of experience and of a quantitative measurement.

II. METHODS AND MATERIALS

A. SUBJECTS

This retrospective study was approved by the institutional review board, and informed consent was waived. Among the 1242 patients who visited the hematology department and received MRI of lumbar spines between March 2010 and June 2019, 254 diseased cases (mean age = 60.3 ± 11.6, male: female = 138:253) who were clinically confirmed to have active hematologic diseases were included as a diseased group. The diseased group consisted of patients with multiple myeloma (n = 166), leukemia (n = 32), lymphoma (n = 26), monoclonal gammopathy of unknown significance (MGUS) (n = 15), myelodysplastic syndrome (MDS) (n = 10), myelofibrosis (n = 4) and hypereosinophilia (n = 1). For a healthy control group, 300 patients were randomly selected from our PACS system among those who received lumbar spines MRI between March 2010 and June 2019. Absence of bone marrow pathology was verified based on the laboratory results. After exclusion of 62 patients who were not completely assured of normal bone marrows and 8 patients who had metallic instruments, a total of 230 subjects (mean age = 65.2 ± 12.4, male: female = 93:137) were included as a healthy control group.

One hundred and twenty-one cases were randomly assigned as a test group (mean age = 62.3 ± 12.0, male: female = 48:73, control: disease = 58:53) and the remaining 363 cases as a training group (mean age = 62.7 ± 12.3, male: female = 183:180, control: disease = 172:192).

B. IMAGE ACQUISITION AND SEGMENTATION

MR images were acquired using 7 MRI scanners in our institution. The T1-weighted images included in this study were heterogeneous in terms of manufacturers, model names, magnetic fields and scanning parameters. The various scanning parameters from the multiple vendors used to acquire images from the 484 patients are summarized in Supplementary material. The acquired T1-weighted images were normalized by subtracting the whole image from the annulus fib3rosus of non-degenerated intervertebral disk of the same subject. The intervertebral disk was partitioned into 5 regions with equal distance from anterior to posterior, and the first and last regions were regarded as annulus fibrosus. All pixels from the image were subtracted from the average of the first and last regions of the intervertebral disk of the same subject. The disk-normalized bone marrows were segmented using a 3-dimensional GrowCut algorithm, which is a semi-automatic way of segmenting the area of interest from multiple slices of an image [11]. Figure 1 illustrates the sagittal T1-weighted images and the processed images of the controls and diseased patients, respectively.

C. RADIOMICS FEATURE EXTRACTION

Figure 2 summarizes the study process from radiomics feature extraction to classification. A total of 100 radiomics features were extracted from the segmented bone marrow images, using an open-source python package for radiomics

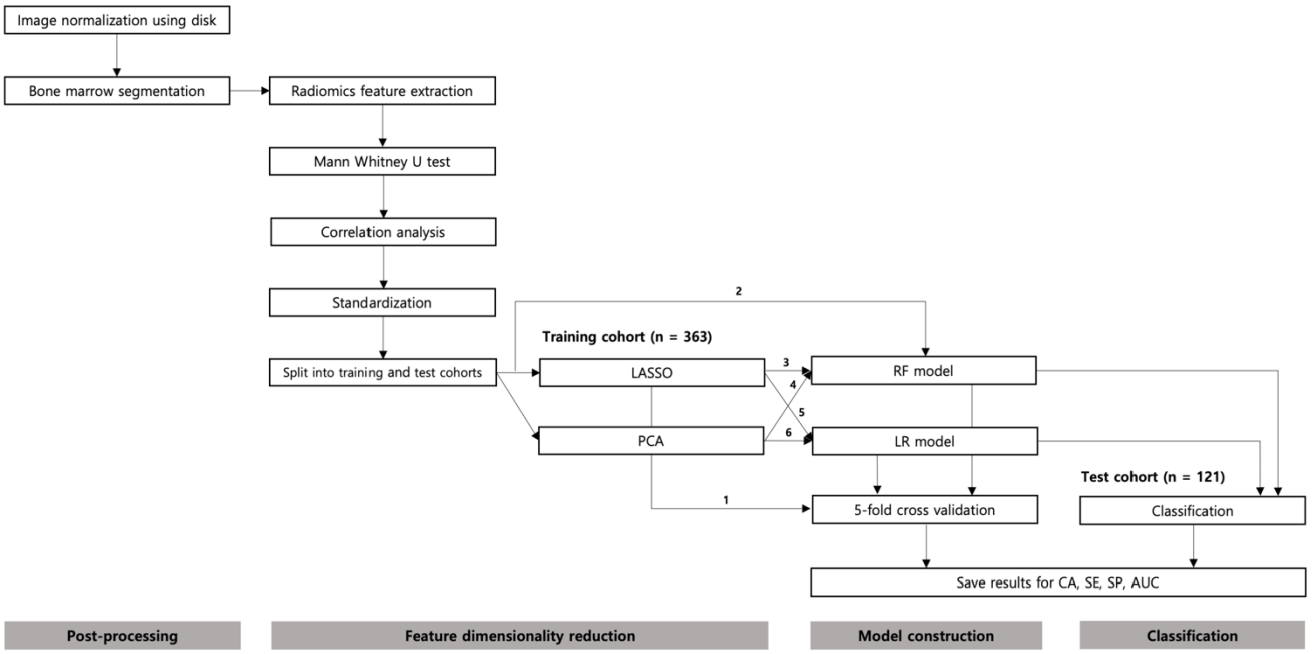

FIGURE 2. Flowchart showing the study process from post-processing to classification using different feature dimensionality reduction methods and model constructions. Abbreviations: LASSO = least absolute shrinkage and selection operator, PCA = principal component analysis, RF = random forest, LR = logistic regression, CA = classification accuracy, SE = sensitivity, SP = specificity, AUC = area under the receiver operating characteristics (ROC) curve.

feature extraction from medical images called ''pyradiomics'' [12]. Radiomics features included 18 first order features, 16 3-dimensional shape-based features, 22 Gray-level co-occurrence matrix (GLCM) based features, 16 Gray-level run-length matrix (GLRLM) based features, 16 Gray-level size zone matrix (GLSZM) based features and 14 Gray-level dependence matrix (GLDM) based features. For the purpose of our study, we excluded all shape-related features. The detailed information about the radiomics features are explained in Supplementary Data.

D. FEATURE DIMENSIONALITY REUCTION AND MODEL CONSTRUCTION

The feature dimension reduction methods used to select final features were implemented as follows: First, a Mann Whitney-U test, which is a univariate filter feature selection method, was performed to each radiomics feature and eliminated if there was no statistically significant difference between the control and diseased groups ($p > 0.05$). Second, a correlation analysis was performed among the remaining features to eliminate one of the features whose spearman coefficients were greater than 0.9. The remaining features were then standardized by centering the mean to zero and scaling to unit variance. Finally, a number of different feature selection methods and classifiers were applied to build a final model.

The least absolute shrinkage and selection operator (LASSO) and random forest (RF) were used to select final features and build a classification model simultaneously. Among many feature selection methods, LASSO and RF are embedded methods that perform feature selection as a part of the training process. Since they process feature selection and training algorithms simultaneously, LASSO and RF reduce a burden of having to choose separate methods for feature selection and classification. LASSO is a penalty-based embedded method that uses a linear regression model. A typical linear regression aims to minimize a cost function, or a mean square error (MSE), which is defined as an average squared difference between the estimated values (Y_i) and the actual values (Y) (Equation 1). LASSO operates L1 regularization by adding a penalty term equal to the sum of the absolute value of the magnitude of coefficients (β_j), which results in a sparse model with few coefficients (Equation 2). The L1 regularization eliminates some coefficients that converge to zero, making the final model much simpler than the model with the initial number of features. The tuning parameter, lambda (λ), controls the strength of L1 penalty. The goal of LASSO is to find the most optimal λ that yields the minimum MSE between the actual and estimated values, which eventually decides β_j and the number of non-zero features used to build a final model [13].

$$\mathrm{MSE} = \frac{1}{n}\left[\sum_{i=1}^{n}(Y - Y_i)^2\right] \tag{1}$$

$$\mathrm{MSE}_{\mathrm{LASSO}} = \frac{1}{n}\left[\sum_{i=1}^{n}(Y - Y_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right] \tag{2}$$

The ''lassocv'' function of a linear model package in Scikit-learn version 0.23.1, which is an open source machine learning library in python [14], was used to build a LASSO model.

TABLE 1. Summary of subject characteristics. Abbreviation: STD = standard deviation.

	Training cohort		Test cohort		*P*-value
	Control	Diseased	Control	Diseased	
Number of subjects	172	192	58	53	-
Age (mean±STD)	65.6±12.1	60.1±12.1	64.0±13.4	60.7±10.4	0.746
Gender (male:female)	74:98	109:83	19:39	29:24	0.040

With an iterative fitting along a regularization path, the linear model searches for λ, and the best model was chosen by a 3-fold cross-validation. LASSO can be used to select features, or to differentiate classes by calculating a radiomics score (RS) for each subject, by summing the multiplications of each radiomics feature (X_j) with the corresponding coefficient (β_j) LR estimated by LASSO (Equation 3) [15]. For our purpose of the study, we used LASSO twice for feature selection and for feature selection and classifier using RS, respectively.

$$\mathrm{RS} = Y_j = \sum_{k=1}^{p} \beta_j X_j \tag{3}$$

RF is a tree-based embedded method that uses multitude of decision trees during training and determines class based on the ''mode'' of the classes. Each decision tree consists of a hierarchical construction of nodes and edges, which takes randomly extracted observations and features from the dataset and determines class by passing each node with a split function between 0 and 1 based on feature values. The data reaching each node is passed onto the next node based on the split function score. Each decision tree has low bias but high variance, which causes a problem of overfitting. Therefore, the random forest is built through bagging, amplifying the number of decision trees with different subsets of training samples and features. This process guarantees that each tree is not correlated with each other and is less prone to overfitting by maintaining low bias but lessening high variance [16].

The tree-based strategies used by RF naturally rank features by how well features improve purity of each node. Gini-index of a node is a probability of a randomly chosen sample in a node would be incorrectly labelled if it was labelled by the distribution of samples in the node. At each node, the decision tree searches through features that results in the greatest reduction in Gini-index. The importance of features could thus be determined by measuring how much given features reduce impurity. In short, RF naturally involves feature selection as part of the training algorithm.

For our purpose of the study, we used RF for a classifier and for both feature selection and classifier, respectively. The ''GridSearchCV'' function in Scikit-learn was used to optimize parameters, and ''RandomForestClassifier'' function in Scikit-learn was used to build the model using the optimized parameters. The parameters used for RF classification are summarized in Supplementary materials. For evaluating performances of the two embedded methods, we examined a separate feature selection method and classifier. A principle component analysis (PCA) was applied, which is a dimensionality reduction method used to determine the most valuable variables to cluster data. It uses an orthogonal transformation to convert a set of values with possibly correlated variables into a set of values with linearly uncorrelated variables called principal components (PC). Each PC is a linear combination of multiple features, which reveals the ratio of features that contribute to the PC variation. PCA is performed first by calculating the mean of all radiomics features and subtracting all features from the calculated means. The covariance matrix is constructed, which decomposes into the eigenvectors and eigenvalues. The eigenvectors are sorted by a decreasing order, from which the number of eigenvectors are selected to explain a certain amount of variance with a reduced dimensionality [17]. The ''pca'' function in Scikit-learn was used, which used Singular Value Decomposition (SVD) to project the data to a lower dimensional space. The sum of variance was set to 0.95.

Finally, a LR model was examined as a separate classifier to be compared with RF, which uses a logistic function, instead of a linear function, to model a binary dependent variable [18]. The ''LogisticRegressionCV'' function in Scikit-learn was used with L2 penalty. For each classifier, a 5-fold stratified k-fold cross validation was performed on a training cohort to estimate overall performance, respectively. The constructed models were applied to a separate test set, and classification was performed. In summary, we examined a total of six radiomics models using different combinations of feature selection methods and classifiers to choose the best model with the highest diagnostic performances. We chose the radiomics model that produced the highest AUC on the test set.

III. RESULTS

A. SUBJECTS

Table 1 summarizes the subject characteristics of the training and test cohorts. There was no statistically significant difference between the ages ($p = 0.746$), but there was a statistically significant difference between gender ($p = 0.041$) of the training and test cohorts.

B. FEATURE DIMENSIONALITY REDUCTION RESULTS USING LASSO AND PCA

Table 2 summarizes the results of our feature dimensionality reduction methods using LASSO and PCA. The number of features reduced as each method of feature dimensionality

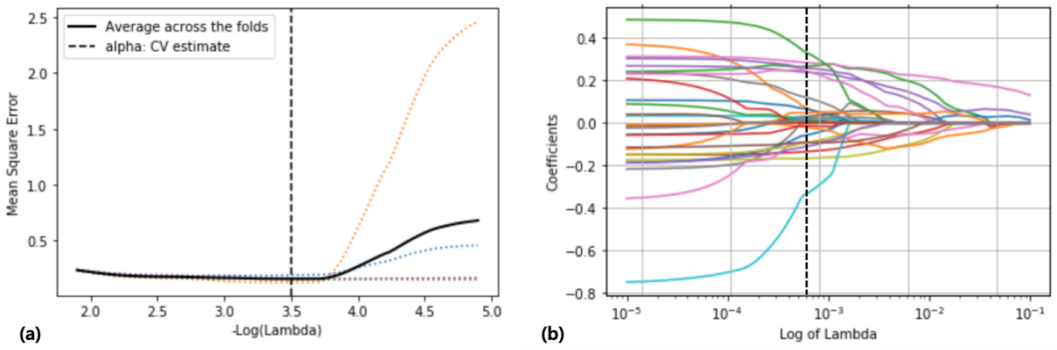

FIGURE 3. LASSO Regularization results. (a) A tuning parameter (λ) search after 3-fold cross validation and (b) LASSO coefficient profiles using a LASSO regression model.

TABLE 2. Feature dimensionality reduction by LASSO and PCA. Abbreviation: GLCM – Grey Level Co-occurrence Matrix.

		LASSO (coefficients)	PCA (variance ratio)
Mann Whitney U test		80	80
Correlation analysis (< 0.9)		28	28
LASSO	Lambda	0.000312	-
	Mean squared error	0.135	-
PCA	Sum of variance	-	0.95
Final number of features		12	8
Firstorder – Maximum		o (0.228)	o (0.0625)
Firstorder - Minimum		o (-0.104)	
Firstorder – Energy		o (0.200)	
Firstorder – Entropy		o (-0.0552)	o (0.0436)
Firstorder – Interquartile Range		o (0.150)	
Firstorder – Kurtosis		o (0.0440)	
Firstorder – Root Mean Squared		o (-0.110)	o (0.116)
Firstorder – Uniformity		o (0.118)	
Firstorder – Range			o (0.0326)
GLCM – Autocorrelation		o (-0.0544)	o (0.194)
GLCM – Contrast		o (0.0439)	
GLCM – Difference Variance		o (-0.0325)	
GLCM – Difference Average			o (0.0200)
GLCM – Joint Energy		o (0.0619)	
GLCM – Cluster Shade			o (0.454)
GLCM – Cluster Tendency			o (0.0129)

reduction was performed. After Mann Whitney U test, the remaining feature number was 80, which was further reduced to 28 after correlation analysis. LASSO regularization selected 12 final features, while PCA selected 8 features to be used as inputs to machine learning models. It is to be noted that the collections of features selected by LASSO and PCA were not the same; however, there were four features that were chosen identically between the two methods: Firstorder – Maximum, Entropy, Root Mean Squared and GLCM – Auto-correlation. The remaining 8 features from LASSO and 4 features from PCA differed from one another.

Figure 3 summarizes the LASSO regularization results of a lambda and LASSO coefficient profiles, and Figure 4 summarizes the feature selection results by LASSO, PCA and RF. It is to be noted that important features selected by RF alone, and RF after LASSO, PCA were different.

C. DIAGNOSTIC PERFORMANCES OF THE RADIOMICS MODELS ON DIFFERENTIATING DISEASED MARROWS FROM THE NORMAL CONTROLS

Table 3 summarizes CA, SE, SP and AUC with the 95% confidence intervals of the training and test cohorts, separately, after using six radiomics models. Overall, those with RF yielded higher CA and SE than without RF both in training and test sets. For the training set, the combination of PCA and RF yielded the highest AUC of 0.900 after 5-fold cross validation. The highest CA of 83.9% and SE of 88.5% were yielded when LASSO and RF were used. For the test set, the overall CA, SE, SP and AUC were slightly higher than those of the training set for all six radiomics models. The models involving RF yielded the highest CA of 86.8%, and the highest SE of 88.9% was achieved with a RF model; the AUC of 0.928 was yielded when LASSO and RF were used. Figure 5 illustrates the ROC curves of all six models on the training (Figure 5a) and test sets (Figure 5b). The models with LR yielded less effective results than those with RF. However, the performances of the six radiomics models did not significantly differ from one another.

D. COMPARISON OF PERFORMANCES AMONG A SINGLE RADIOMICS FEATURE, RADIOMICS MODEL AND RADIOLOGIST READINGS

Among the 100 radiomics features extracted from "pyradiomics", the highest AUC was produced by maximum from the first-order statistics, which yielded 0.695 for the training cohort and 0.787 for the test cohort, respectively. The diagnostic performance of Firstorder – Maximum was compared with that of the combination of LASSO and RF, which yielded the highest accuracy and AUC among the other radiomics models we constructed, and that of the two radiologists with different years of experience. Table 4 summarizes the SE, SP, number of mis-classifications and AUC with 95% confidence level results of all performances in the test cohort. The radiomics model showed the better performance than Reader 1, which yielded the highest SE of 87.3%, SP of 86.2% and AUC of 0.928. The performance of the single best radiomics feature was only slightly better than

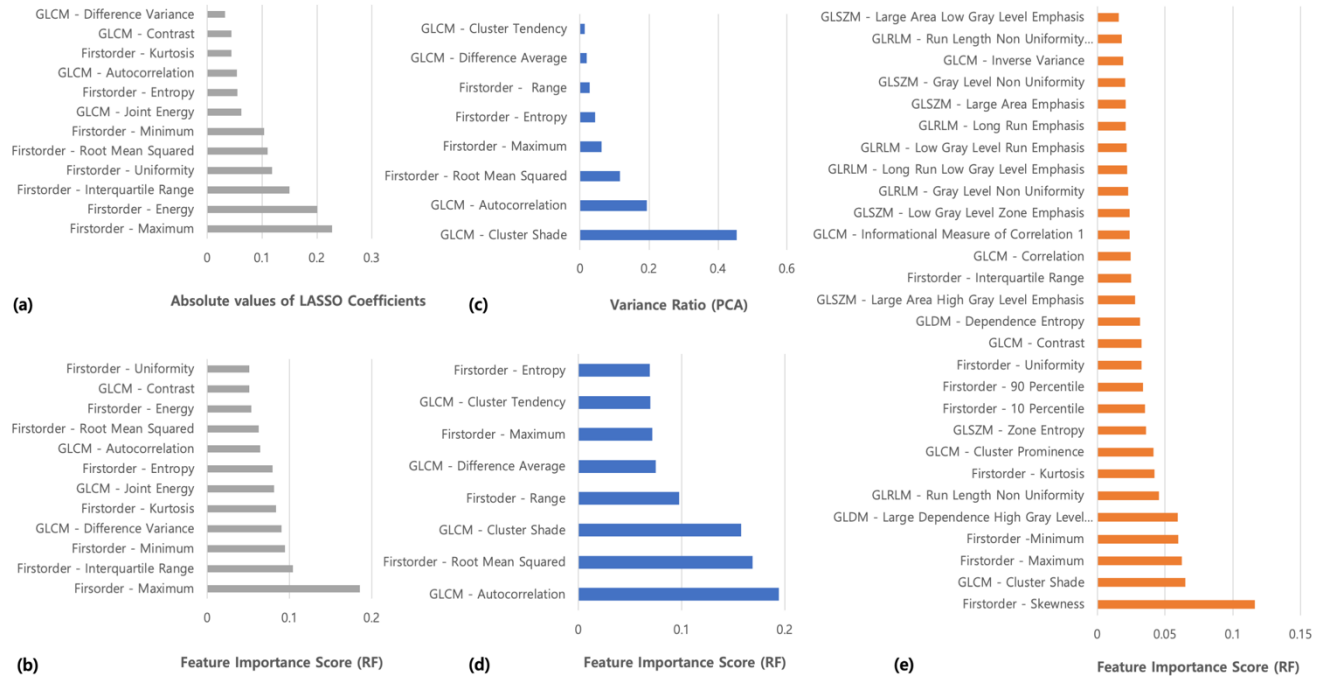

FIGURE 4. Feature selection results by (a) LASSO, (b) feature rank by random forest after LASSO, (c) PCA, (d) feature rank by RF after PCA, (e) feature rank by RF only. Abbreviations: GLCM = Grey Level Co-occurrence Matrix, GLRLM = Grey Level Run Length Matrix, GLSZM = Gray Level Size Zone Matrix, GLDM = Grey Level Dependence Matrix.

TABLE 3. The classification results of the six radiomics models on the training set after 5-fold cross validation and the test set. All *p*-values for AUC were < 0.05. Abbreviations: CA = classifying accuracy, SE = sensitivity, SP = specificity, CI = confidence interval.

		LASSO	RF	LASSO + RF	PCA + RF	LASSO + LR	PCA + LR
Training (n = 363)	**CA (%)**	78.0	82.1	83.9	82.6	80.2	74.6
	SE (%)	78.9	85.3	88.5	85.9	80.1	74.4
	SP (%)	77.3	78.5	76.7	79.1	80.2	75.0
	AUC [95% CI]	0.847 [0.807-0.887]	0.893 [0.859-0.926]	0.891 [0.857-0.925]	0.900 [0.866-0.931]	0.849 [0.810-0.889]	0.786 [0.739-0.832]
Test (n = 121)	**CA (%)**	86.0	86.8	86.8	86.8	83.5	83.5
	SE (%)	85.7	88.9	87.3	84.1	85.7	81.0
	SP (%)	86.2	84.5	86.2	89.7	81.0	86.2
	AUC [95% CI]	0.902 [0.846-0.958]	0.926 [0.805-0.927]	0.928 [0.880-0.976]	0.914 [0.861-0.966]	0.901 [0.833-0.950]	0.887 [0.826-0.946]

TABLE 4. Diagnostic performance of T1-weighted image on differentiating diseased marrows from normal controls using different methods: single radiomic feature, radiologist readings and machine learning model. All *p*-values for AUC were < 0.05.

Test	**Single Best Radiomics Feature**	**Radiologists' Readings**		**Best Radiomics Model**
(n = 121)	**Firstorder – Maximum**	**Reader 1**	**Reader 2**	**LASSO + RF**
SE (%)	75.0	86.5	75.0	87.3
SP (%)	77.4	73.9	67.7	86.2
No. of FP	17	7	14	6
No. of FN	12	18	21	10
AUC	0.787[0.706-0.807]	0.861[0.798-0.928]	0.767[0.683-0.858]	0.928[0.805-0.927]

Reader 2 with 1 year of experience. Figure 6 illustrates the AUC (Figure 6a) and DCA curves (Figure 6b) of the best single radiomics feature, radiomics model and the two radiologist readings. The DCA results show that the net benefit of the radiomics model surpassed the human readings and single radiomics feature over the threshold probabilities less than around 0.8. For the threshold probabilities equal to or greater than 0.8, the net benefit of the Reader 1 and 2 were greater than that of the radiomics model and single radiomics feature.

IV. DISCUSSION

We constructed a bone marrow radiomics model using T1-weighted MRI to differentiate diseased bone marrows from the normal marrows. We extracted radiomics features

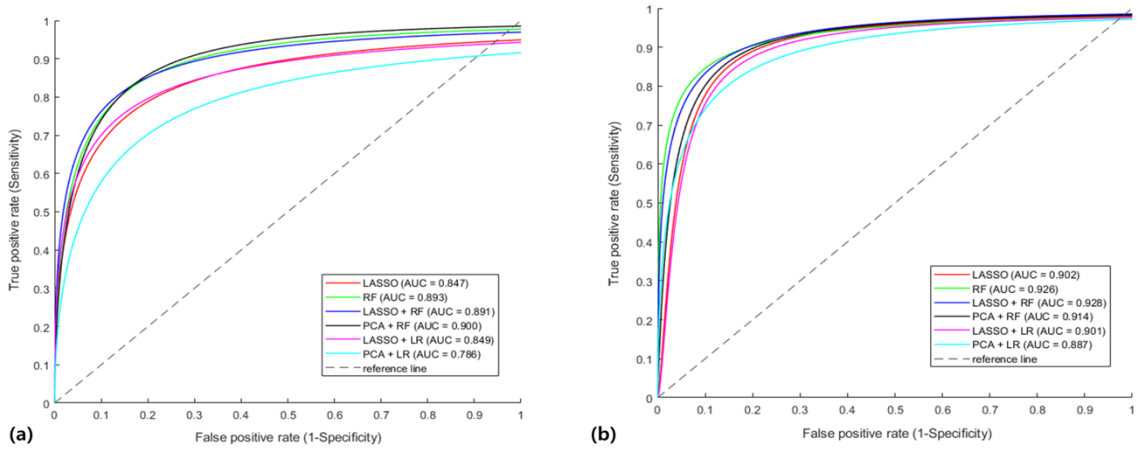

FIGURE 5. ROC curves showing performances of the six radiomics models on the (a) training set after 5-fold cross validation and on the (b) test set.

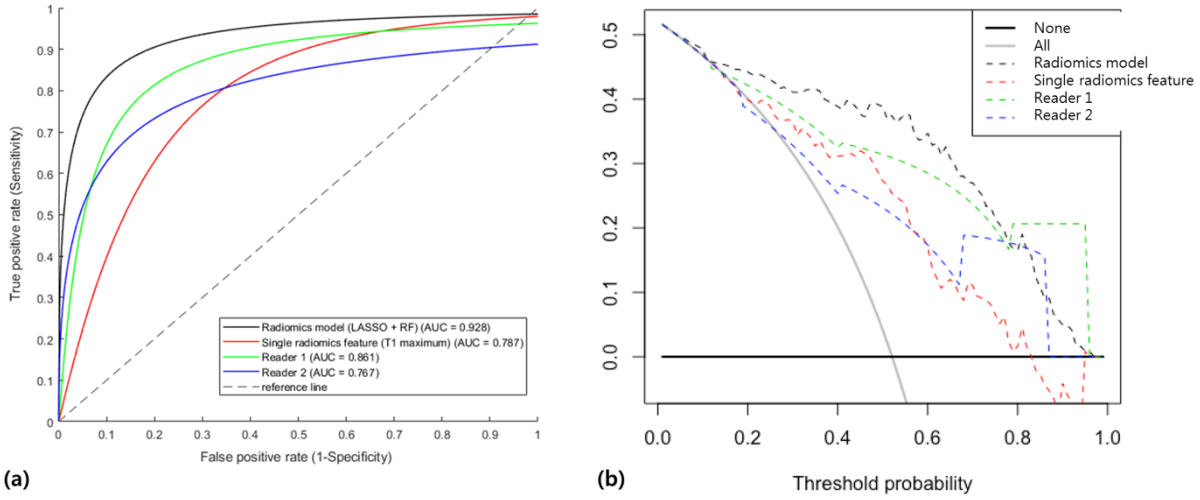

FIGURE 6. (a) ROC curves showing the performances of the best performing radiomics model, two radiologists with different levels of experience and the best single radiomics feature and (b) DCA results showing net benefit over all range of threshold probabilities.

from the segmented bone marrows, reduced feature dimensionalities using a univariate filter method and correlation analysis. We then chose LASSO and RF, which involve feature selection as part of a training algorithm reducing the burden of choosing a separate feature selection algorithm and classifier. We used LASSO and RF to build radiomics models to differentiate bone marrow diseases, and PCA and LR were used as comparison. We constructed a total of six models with various combinations of feature selection methods and classifiers including LASSO, RF, PCA and LR. We finally chose one model with the best performance and compared it with the conventional methods of diagnosing with radiologists' readings and a quantitative measurement.

We successfully constructed a radiomics model to discriminate bone marrow diseases with a predictive performance greater than that of the previously developed models using magnetic resonance images. Previously, radiomic features from T1 and T2-weighted MRI were used to differentiate bone chondrosarcoma, but the accuracy and corresponding AUC only reached 75% and 0.780, respectively [8]. The same images were used to detect osteoporosis, but the highest AUC was only 0.810 [7]. Our predictive performance was also higher than the recent study of discriminating diseased bone marrows with a support vector machine texture classifier, which gave the AUC of 0.895 [20] and another study of differentiating bone marrow metastatic diseases with the highest AUC of 0.912 [9]. Building a radiomics model involved various steps of post-processing, feature selection and applying an appropriate classifier. Our post-processing step involved normalization with the annulus fibrosus of non-degenerated intervertebral disk in order to minimize the effect of image heterogeneity on radiomics features, which might have added consistency among the images acquired from multiple manufacturers, models, magnetic fields and

scanning parameters. Identifying important features to differentiate diseased marrows was difficult, since LASSO, PCA and RF selected different feature groups; however, the six radiomics models differentiated diseased marrows with high CA, SE and AUC without much variance.

We showed the advantage of RF both as a classifier and as feature selection over other methods we evaluated. First, RF yielded better CA, SE and AUC than LR and LASSO did. Second, the choice of feature selection did not significantly affect the overall results if RF was used as a classifier. Third, although the combination of LASSO and RF yielded the highest AUC in the test set, using RF alone yielded CA and AUC very close to those with LASSO and RF or with PCA and RF combined. Our results demonstrated that RF is sufficient to train radiomics features to differentiate bone marrow diseases without additional feature selection process, since feature selection embedded in RF effectively built a training model to differentiate bone marrow diseases.

We also demonstrated a higher predictive performance of the radiomics model over conventional methods of diagnosis such as radiologists' readings and a quantitative measurement. Although we showed that RF in general yielded better results than the other classifiers, the diagnostic performances of the six models did not significantly differ from one another. All radiomics models we constructed performed better than the two radiologists' readings with different levels of experience and a quantitative measurement of a single radiomics feature with the highest AUC selected among the initially extracted radiomics features. While the radiologists' performances depended largely on the years of experience, the six models we evaluated gave results that were quite similar to one another and without large dependence on the types of feature selection methods or classifiers. For the purpose of our study, we chose the combination of LASSO and RF as the best radiomics model to be compared with other diagnostic methods. However, the model with high SE may benefit in the clinical settings where bone marrows of suspicious diseases are screened without loss. In that sense, the RF model can be useful for the purpose of initial screenings, which produced the highest SE among the six models. Therefore, our study demonstrated that a properly trained machine learning model with magnetic resonance image based radiomics features can successfully differentiate diseased bone marrows from the normal ones with high predictive performances.

The limitations were the following. First, there was a statistically significant difference in gender between training and test sets. Liney et al have reported more age-related fat content increase in male than female marrows [21]. Although there was no age difference between the training and test sets, the significantly different gender ratio might have affected overall results as all of our models demonstrated slightly better performances in the test sets than in the cross-validated training sets. However, the fact that there is gender variation in normal marrow composition is still controversial, and marrow composition is usually more affected by age and osteoporosis than is by gender. We therefore insist that gender difference between the training and test sets would not have largely affected our results. Second, the reliability and reproducibility of the radiomics features were not tested, which should be sensitive at various degrees to image acquisition settings and parameters [22], [23]. In fact, the features selected from various feature selection methods all differed from one another and involved more First-order features than textural features, which are known to be more sensitive to image acquisition and processing methods [22]. The reliability and reproducibility of the features should thus be evaluated to ensure robustness of our model. Third, multimodal images in addition to T1-weighted images should increase the model performances. Finally, external validation using geographically different data set should be considered to ensure generalizability.

In summary, we constructed a radiomics model to differentiate bone marrow diseases using T1-weighted magnetic resonance images and demonstrated the advantage of bone marrow radiomics over conventional methods of diagnosing with radiologists' readings and quantitative measurements.

REFERENCES

[1] R. Chou, A. Qaseem, V. Snow, D. Casey, J. T. Cross, P. Shekelle, and D. K. Owens, "Diagnosis and treatment of low back pain: A joint clinical practice guideline from the American college of physicians and the American pain society," *Ann. Internal Med.*, vol. 147, pp. 478–491, Oct. 2007.

[2] L. M. Shah and C. J. Hanrahan, "MRI of spinal bone marrow: Part I, techniques and normal age-related appearances," *AJR. Amer. J. Roentgenol.*, vol. 197, pp. 1298–1308, Dec. 2011.

[3] C. Ricci, M. Cova, Y. S. Kang, A. Yang, A. Rahmouni, W. W. Scott, and E. A. Zerhouni, "Normal age-related patterns of cellular and fatty bone marrow distribution in the axial skeleton: MR imaging study," *Radiology*, vol. 177, pp. 8–83, Oct. 1990.

[4] S. M. Navarro, G. R. Matcuk, D. B. Patel, M. Skalski, E. A. White, A. Tomasian, and A. J. Schein, "Musculoskeletal imaging findings of hematologic malignancies," *Radiographics, Rev. Publication Radiol. Soc. North Amer.*, vol. 37, pp. 881–900, May/Jun. 2017.

[5] S. A. Mattonen, G. A. Davidzon, J. Benson, A. N. C. Leung, M. Vasanawala, G. Horng, J. B. Shrager, S. Napel, and V. S. Nair, "Bone marrow and tumor radiomics at 18F-FDG PET/CT: Impact on outcome prediction in non-small cell lung cancer," *Radiology*, vol. 293, no. 2, pp. 451–459, Nov. 2019.

[6] H. Li, C. Xu, B. Xin, C. Zheng, Y. Zhao, K. Hao, Q. Wang, R. L. Wahl, X. Wang, and Y. Zhou, "18F-FDG PET/CT radiomic analysis with machine learning for identifying bone marrow involvement in the patients with suspected relapsed acute leukemia," *Theranostics*, vol. 9, no. 16, pp. 4730–4739, 2019.

[7] M. E. Mayerhoefer, C. C. Riedl, A. Kumar, A. Dogan, P. Gibbs, M. Weber, P. B. Staber, S. Huicochea Castellanos, and H. Schöder, "[18F]FDG-PET/CT radiomics for prediction of bone marrow involvement in mantle cell lymphoma: A retrospective study in 97 patients," *Cancers*, vol. 12, no. 5, p. 1138, May 2020.

[8] S. Gitto, R. Cuocolo, D. Albano, V. Chianca, C. Messina, A. Gambino, L. Ugga, M. C. Cortese, A. Lazzara, D. Ricci, R. Spairani, E. Zanchetta, A. Luzzati, A. Brunetti, A. Parafioriti, and L. M. Sconfienza, "MRI radiomics-based machine-learning classification of bone chondrosarcoma," *Eur. J. Radiol.*, vol. 128, Jul. 2020, Art. no. 109043.

[9] L. Filograna, J. Lenkowicz, F. Cellini, N. Dinapoli, S. Manfrida, N. Magarelli, A. Leone, C. Colosimo, and V. Valentini, "Identification of the most significant magnetic resonance imaging (MRI) radiomic features in oncological patients with vertebral bone marrow metastatic disease: A feasibility study," *La Radiologia Medica*, vol. 124, no. 1, pp. 50–57, Jan. 2019.

[10] L. He, Z. Liu, C. Liu, Z. Gao, Q. Ren, L. Lei, and J. Ren, "Radiomics based on lumbar spine magnetic resonance imaging to detect osteoporosis," *Acad. Radiol.*, to be published, doi: 10.1016/j.acra.2020.03.046.

[11] H. Song, W. Kang, Q. Zhang, and S. Wang, "Kidney segmentation in CT sequences using SKFCM and improved GrowCut algorithm," *BMC Syst. Biol.*, vol. 9, no. 5, p. S5, 2015.
[12] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017.
[13] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Stat. Soc., B Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, Jun. 2011.
[14] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers Neuroinform.*, vol. 8, p. 14, 2014.
[15] Y. Jiang, C. Chen, J. Xie, W. Wang, X. Zha, W. Lv, H. Chen, Y. Hu, T. Li, J. Yu, Z. Zhou, Y. Xu, and G. Li, "Radiomics signature of computed tomography imaging for prediction of survival and chemotherapeutic benefits in gastric cancer," *EBioMedicine*, vol. 36, pp. 171–182, Oct. 2018.
[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[17] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202.
[18] J. C. Stoltzfus, "Logistic regression: A brief primer," *Acad. Emergency Med.*, vol. 18, no. 10, pp. 1099–1104, Oct. 2011.
[19] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Making, Int. J. Soc. Med. Decis. Making*, vol. 26, pp. 565–574, Nov./Dec. 2006.
[20] E.-J. Hwang, J.-Y. Jung, S. K. Lee, S.-E. Lee, and W.-H. Jee, "Machine learning for diagnosis of hematologic diseases in magnetic resonance imaging of lumbar spines," *Sci. Rep.*, vol. 9, no. 1, p. 6046, Apr. 2019.
[21] G. P. Liney, C. P. Bernard, D. J. Manton, L. W. Turnbull, and C. M. Langton, "Age, gender, and skeletal variation in bone marrow composition: A preliminary study at 3.0 Tesla," *J. Magn. Reson. Imag.*, vol. 26, pp. 787–793, Sep. 2007.
[22] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: The facts and the challenges of image analysis," *Eur. Radiol. Experim.*, vol. 2, no. 1, Nov. 2018.
[23] J. E. Park, S. Y. Park, H. J. Kim, and H. S. Kim, "Reproducibility and generalizability in radiomics modeling: Possible strategies in radiologic and statistical perspectives," *Korean J. Radiol.*, vol. 20, pp. 1124–1137, Jul. 2019.
[24] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and reproducibility of radiomic features: A systematic review," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018.

EO-JIN HWANG received the B.S. degree in biomedical engineering from the Boston University College of Engineering, Boston, MA, USA, in 2006, and the M.A. degree in bio imaging from the Boston University School of Medicine, Boston, in 2008.

From 2009 to 2011, she was a Research Assistant with Sungkyunkwan University. From 2011 to 2014, she was a Research Assistant with the Kyung Hee University Hospital, Gangdong. Since 2014, she has been a Research Scientist with the Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea. Her research interest includes development of machine learning models for disease prediction using medical images.

SANGHEE KIM received the M.D. degree from The Catholic University of Korea, in 2015.

She is currently on Fellowship Training of musculoskeletal radiology with the Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea. Her research interests include image analysis, image processing in musculoskeletal radiology, and application of AI in clinical practice.

JOON-YONG JUNG graduated from the Medical College. He received the M.D. degree and the Ph.D. degree in medical science from The Catholic University of Korea, Seoul, in 2002 and 2015, respectively.

He completed the Residency in radiology with the Catholic Medical Center. He finished his Fellowship in musculoskeletal radiology with the Seoul St. Mary's Hospital, where he was a Faculty Member of radiology, in 2012. He is currently an Associate Professor of radiology with The Catholic University of Korea. His research interests include quantitative MR imaging, standardization of quantitative imaging, and AI development for musculoskeletal disease.