Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253574

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Multi-parametric magnetic resonance imaging-based radiomics for differentiation of skull base osteomyelitis from locally advanced nasopharyngeal carcinoma: a multi-center external validation study

Fırat Atak[1]
Hanife Avcı[2]
Yeliz Pekçevik[3]
Ayça Karaosmanoğlu[4]

[1]Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[2]Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye

[3]University of Health Sciences Türkiye, Tepecik Training and Research Hospital, Clinic of Radiology, Division of Neuroradiology and Head and Neck Radiology, İzmir, Türkiye

[4]Hacettepe University Hospital, Department of Radiology, Division of Neuroradiology and Head and Neck Radiology, Ankara, Türkiye

PURPOSE

Skull base osteomyelitis (SBO) and nasopharyngeal carcinoma (NPca) are challenging to differentiate due to overlapping clinical and radiological features. This study aimed to develop and validate a multi-parametric magnetic resonance imaging (MRI)-based radiomics model with high sensitivity, enabling reliable diagnosis of SBO in adult patients presenting with equivocal imaging findings.

METHODS

This was a retrospective, multicenter study using institutional data. The training cohort, comprising 63 adult patients from two classes (31 SBO, 32 NPca) with MRI data, was used for model development and optimization. An external test set (n = 30; 12 SBO, 18 NPca) obtained from two different clinical centers was used for model performance analysis and generalizability. Lesion segmentation was performed using a manual volumetric technique on three axial MRI sequences (pre-contrast T1-weighted, fat-suppressed T2-weighted, and post-contrast fat-suppressed T1-weighted). Hand-crafted radiomic features (n = 2,553) were extracted using the Pyradiomics library. A multi-step process was used to select the final features, including reproducibility analysis using an inter-class correlation coefficient threshold of 0.9, pairwise Spearman correlation analysis with a threshold of 0.8 to reduce redundancy, and least absolute shrinkage and selection operator regression. The final set of five features were used to train six machine learning models. The models were internally validated using 5-fold cross-validation, and performance was confirmed using the unseen external test set. Traditional statistical tests, including the Mann–Whitney U test and chi-squared test, were used to compare baseline characteristics, with a *P* value of <0.05 considered significant.

RESULTS

Among the evaluated classifiers, the random forest model demonstrated the best diagnostic performance, yielding the highest area under the curve (AUC) value in the 5-fold cross-validation analysis. In the external test set, the semantic model demonstrated the best diagnostic performance, achieving an AUC of 0.940 [95% confidence interval (CI): 0.857–1.00], followed by the radiomics model (AUC: 0.903, 95% CI: 0.784–1). The apparent diffusion coefficient (ADC)-based model demonstrated limited discriminative ability (AUC: 0.694, 95% CI: 0.497–0.892). The difference between the semantic and radiomics models did not reach statistical significance ($P = 0.644$), whereas both significantly outperformed the ADC model ($P < 0.05$).

CONCLUSION

Radiomics achieved high and consistent performance in distinguishing SBO from advanced NPca. Although expert-based semantic assessment performed slightly better, radiomics provides an objective alternative. ADC-based methods showed limited generalizability due to inter-center variability.

CLINICAL SIGNIFICANCE

Our study confirms the importance of expert radiologist assessment while demonstrating that radiomics offers a comparably effective and objective decision-support tool. Its ability to provide a consistent, quantitative output is particularly valuable for standardizing the diagnostic approach and empowering less experienced radiologists to make more confident assessments.

KEYWORDS

Radiomics, skull base osteomyelitis, magnetic resonance imaging, nasopharyngeal carcinoma, hand-crafted

The findings of this study were presented as a scientific poster at ECR 2025; however, the full work has not been published elsewhere.

Corresponding author: Ayça Karaosmanoğlu

E-mail: ayca.akgoz@gmail.com

Received 23 July 2025; revision requested 03 September 2025; accepted 24 October 2025.

Epub: 05.12.2025

Publication date:

DOI: 10.4274/dir.2025.253574

You may cite this article as: Atak F, Avcı H, Pekçevik Y, Karaosmanoğlu A. Multi-parametric magnetic resonance imaging-based radiomics for differentiation of skull base osteomyelitis from locally advanced nasopharyngeal carcinoma: a multi-center external validation study. *Diagn Interv Radiol.* 05 December 2025 DOI: 10.4274/dir.2025.253574 [Epub Ahead of Print].

Skull base osteomyelitis (SBO) is defined as an infection of the skull base bones, which often develops as a complication of external otitis. Although it is frequently considered to be synonymous with necrotizing otitis externa, radiologists prefer the broader term SBO because it avoids specific etiological assumptions. SBO can be classified into two subtypes: typical and atypical (central).[1] Typical SBO characteristically arises secondary to external otitis.[2,3] Atypical SBO occurs without typical clinical presentation or a preceding ear infection, posing a diagnostic challenge. It may arise from local infections other than the ear (paranasal sinus or soft tissue) or be idiopathic.[1]

Despite advances in diagnostic techniques and antibiotic regimens, SBO remains a potentially fatal disease. A recent systematic review reported high rates of treatment failure (22%), relapse (7%), and disease-specific mortality (2%) for otogenic SBO.[4] Neurological sequelae are common, with rates reported as high as 48%.[3,5]

Tissue sampling is often required to rule out neoplasms but often yields non-specific results, necessitating repeated biopsies.[1,3] Microbiologic culture of tissue samples plays an important role in establishing a definitive diagnosis of SBO, but an average of 15.8% of cases may yield false-negative culture results.[2] Although computed tomography remains the preferred imaging modality, it has limited sensitivity and may not detect subtle early-stage changes;[6] Tc-99m methylene diphosphonate scintigraphy can overcome this limitation but lacks specificity.[3,7]

Main points

- Magnetic resonance imaging-based radiomics model demonstrated excellent diagnostic accuracy for differentiating skull base osteomyelitis (SBO) from nasopharyngeal carcinoma (NPca), showing statistically superior performance compared with diffusion parameters.
- Semantic assessment demonstrated the best performance; however, given their qualitative and subjective nature, radiomics may provide an important complementary role by offering quantitative and objective information.
- This is the first radiomics study aiming to differentiate SBO from NPca.
- Apparent diffusion coefficient (ADC) threshold values often lack generalizability due to center-specific variability, limiting their use as quantitative biomarkers; however, ADC remains important for qualitative assessment.

Magnetic resonance imaging (MRI) provides superior visualization of soft tissue and intracranial involvement, but differentiation from malignant pathologies can be challenging without typical findings, especially for inexperienced readers.[8-12] Given the absence of a definitive imaging modality for SBO, an often time-consuming multimodal approach is typically necessary to establish an accurate diagnosis. Consequently, the average time from symptom onset to diagnosis is approximately 2 months.[2] Previous studies have demonstrated that early diagnosis and aggressive treatment can improve disease outcomes.[5] Therefore, accurate diagnosis of SBO is essential for appropriate management, with recent studies showing promise in using qualitative and quantitative MRI parameters, including diffusion and dynamic contrast enhancement (DCE) perfusion metrics for distinction.[8-10]

Radiomics, the extraction of quantitative data from medical images, has recently gained significant attention in radiology and oncology research.[13,14] The applications include, but are not limited to, differential diagnosis, prediction of clinical outcomes, and treatment response, as well as characterization of tumor gene expression patterns (radiogenomics).[15]

To date, no prior radiomics studies have investigated the differentiation between SBO and nasopharyngeal carcinoma (NPca). This study evaluates whether MRI radiomics features can distinguish these conditions.

Methods

Study design and ethical approval

This multicenter, retrospective case-control study was conducted and reported in adherence with the CheckList for Evaluation of AI in Radiology guidelines (Supplementary Text).[16] The study received approval from the Hacettepe University Clinical Research Ethics Committee (decision number: GO 23/631, date: 03.10.2023), which waived the need for informed patient consent due to the retrospective nature of the analysis. All patient data were de-identified to ensure data protection.

Patient cohort

Patients who underwent head and neck MRI at one of three institutions between January 2005 and December 2023 were retrospectively screened. The study included adult patients with a final diagnosis of either SBO or NPca. Patients were required to have a definitive diagnosis of SBO or NPca and have the necessary MRI sequences available [axial pre-contrast T1 (T1w), fat-suppressed T2 (FS-T2w), and post-contrast fat-suppressed T1 (T1c)]. Patients were excluded for the following reasons: 1) significant motion artifacts affecting lesion evaluation (n = 3); 2) SBO confined to the temporal bone without further skull base extension (n = 8); 3) prior radiotherapy for NPca (n = 10); or 4) NPca with a tumor stage < T3 (n = 23). Patients with NPca with T-stage < T3 were excluded because skull base involvement is a defining feature of stage T3 disease, and the diagnostic challenge under investigation arises only in this clinical context. The reference standard for NPca was a biopsy-confirmed diagnosis for all patients. The diagnosis of SBO was established based on a combination of the following: 1) compatible clinical presentation; 2) histopathology showing chronic inflammation; 3) microbiological culture isolating non-skin flora pathogens; and 4) clinical and radiological evidence of treatment response on follow-up.

The NPca cohort included 33 (66%) T3 and 17 (34%) T4 tumors; pathologically, 95.5% of analyzed cases (43/45) were of the non-keratinized subtype (the remaining five patients were diagnosed at another center). Within the SBO cohort, seven patients (16.3%) were classified as atypical, presenting with central clivus involvement without evidence of otogenic infection.

Data

The study included 93 patients (43 SBO, 50 NPca). The training set comprised 63 patients (31 SBO, 32 NPca) from a single institution (Center 1). All feature selection, hyperparameter tuning, and model training were performed exclusively on this set. The external test set comprised an independent cohort of 30 patients from two different institutions (Center 2: n = 14; Center 3: n = 16), consisting of 12 SBO and 18 NPca cases. This set was used only once for the final evaluation of the selected models (Figure 1).

Systematic differences between the training and test cohorts, such as variations in MRI scanners and acquisition protocols, were inherent to the multicenter design and served to evaluate the model's generalizability rigorously (Supplementary Text). The classes in the training set (31 SBO, 32 NPca) were well-balanced; therefore, no specific techniques for handling class imbalance were applied.

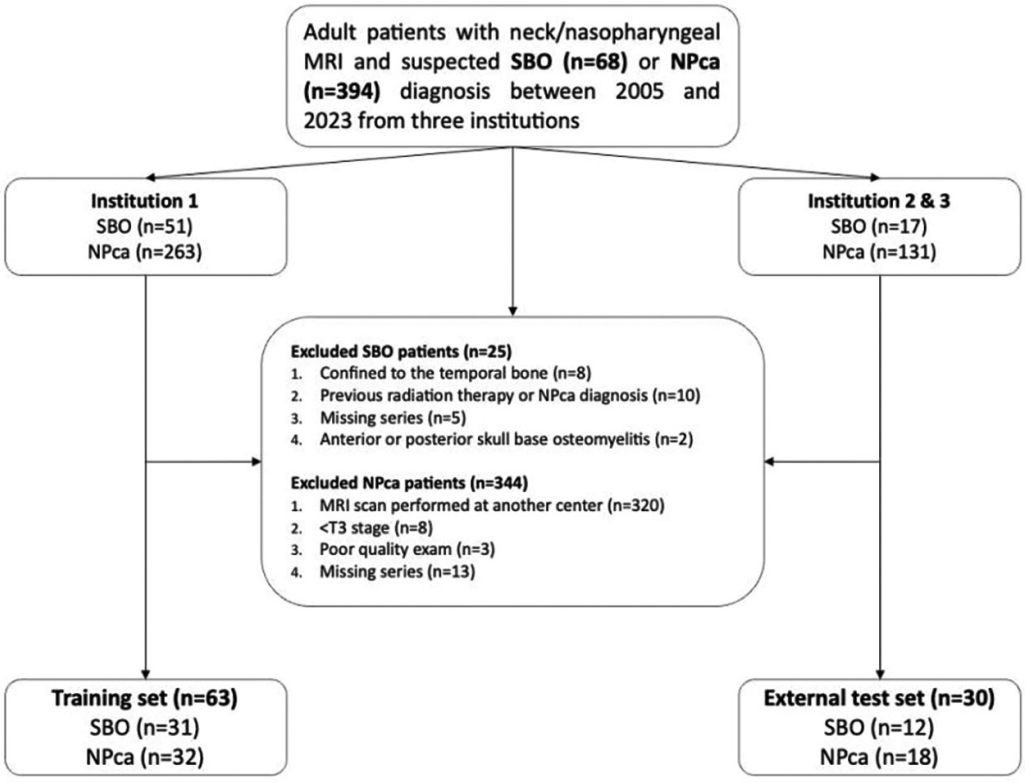


Figure 1. Flowchart of the study. SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; MRI; magnetic resonance imaging.

Non-radiomics data collection

Two radiologists (F.A., A.K.), blinded to diagnoses, evaluated nine predefined semantic MRI features, including the presence of lateral extension, abscess or fluid collection, architectural distortion, lymphadenopathy, increased T2 signal, marked contrast enhancement, dural venous thrombosis, dural thickening, and otologic inflammation (Supplementary Figure 1).

Lateral extension was considered positive if an abnormal signal was detected lateral to the temporomandibular joint.[10] An abscess was considered positive only if diffusion restriction was present. Lesions exhibiting T2 hypointensity, ring-like contrast enhancement, facilitated diffusion, and T2 hyper- or hypointensity were classified as fluid collections. Lymphadenopathy was defined as retropharyngeal lymph nodes measuring >0.8 cm or cervical nodes measuring >1.1 cm in the shortest diameter.[10] Increased T2 signal intensity and marked contrast enhancement were assessed by comparing the lesion signal characteristics with nasal mucosal tissue.[10] Particular attention was given to the predominant signal features exhibited by the majority of the lesion volume rather than focal variations. Otologic inflammation was considered positive if contrast enhancement was observed in the ear cavities on post-contrast images and negative only in the presence of effusion.

A radiologist (F.A.), blinded to diagnoses, measured apparent diffusion coefficient (ADC) values by placing three 0.2 cm^2 regions of interest (ROIs) on ADC maps, excluding cystic or necrotic areas, and targeting the visually lowest signal regions of the lesion. The arithmetic mean of the ADC values (ADCmean) was recorded. Reference ADC (rADC) values were calculated by averaging measurements from ROIs placed over the spinal cord across two subsequent slices. Normalized ADC (nADC) values were derived by dividing ADCmean by rADC (Supplementary Figure 2). Diffusion-weighted imaging was unavailable in six patients (6.45%).

Radiomics workflow

The complete radiomics pipeline is illustrated in Figure 2.

Segmentation

A radiologist (F.A., with 7 years of experience) performed manual volumetric segmentation of the entire lesion on T1w, FS-T2w, and T1c sequences using Slicer 3D 5.2.2 (open source, https://www.slicer.org) (Figure 2).[17] The segmentation volumes were comparable between the two groups, with a mean volume of 29,057.97 mm^3 for SBO and 17,282.68 mm^3 for NPca.

Image preprocessing

The images were resampled to an isotropic voxel size of $2 \times 2 \times 2$ mm^3 using BSpline interpolation. Voxel signal intensities were normalized using z-score normalization (3σ) with a scaling factor of 200. To improve the reliability of texture feature calculations, voxels with intensities outside the range of $\mu \pm 3\sigma$ within the segmentaw tion mask were excluded from the radiomics analysis.[18] Gray-level discretization was performed using a fixed bin width (FBW) approach (Supplementary Figure 3). The bin width was set to 5, which in our data typically resulted in approximately 30–130 bins per volume of interest. This range has been suggested in the Pyradiomics documentation as a practical guideline to ensure that discretized features remain informative while retaining reproducibility.[19] Although there are currently no universally accepted standards for MRI, FBW was considered the most appropriate choice following intensity normalization.[19-21] Moreover, several reports have emphasized that selecting lower bin width values may enhance feature reproducibility.[21]

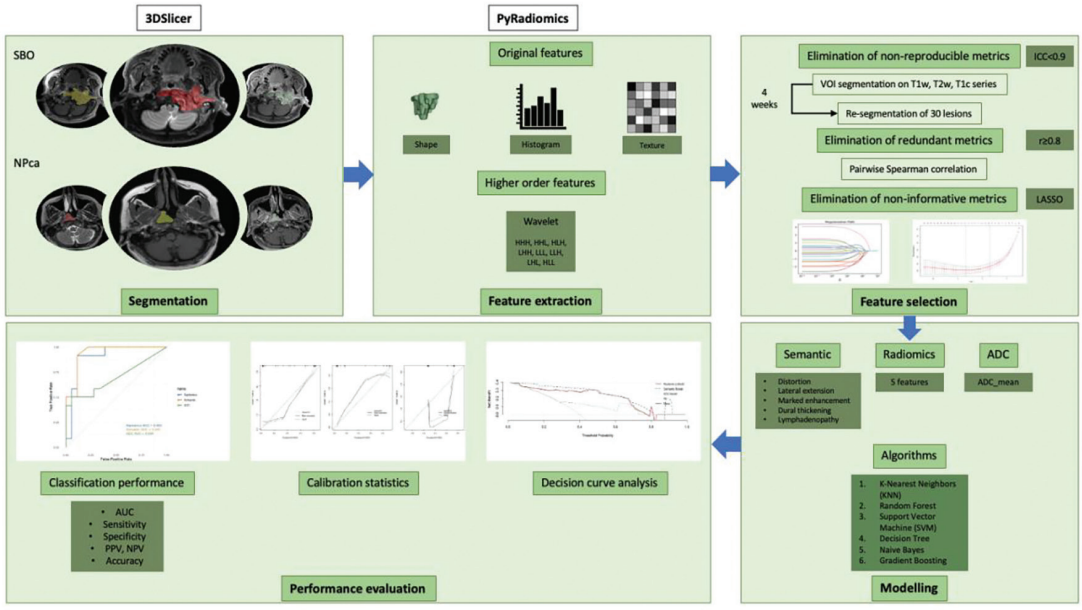


Figure 2. Radiomics workflow. SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; ICC, interclass correlation coefficient; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; ADC, apparent diffusion coefficient.

Feature extraction

From each of the three MRI sequences, 851 radiomic features were extracted, resulting in a total of 2,553 hand-crafted features per patient. These included shape descriptors, first-order statistics, texture features, and wavelet-transformed features. Feature extraction was performed using Pyradiomics (v3.0.1),[19] an open-source Python package. All Pyradiomics settings not explicitly mentioned remained at their default configuration.

Feature selection and modeling

In order to reduce the risk of overfitting and select robust features, a multi-step feature selection was performed. To ensure reproducibility, two radiologists (F.A., A.K.) independently segmented a random subset of 30 cases. Radiomic features with poor inter-rater reliability [interclass correlation coefficient (ICC) < 0.90] were excluded, yielding 1,064 reproducible features. Subsequently, pairwise Spearman correlation analysis was conducted to address multicollinearity; among highly correlated feature pairs ($|\rho| > 0.80$), the feature with the lower univariate discriminative power was discarded, resulting in 79 non-redundant features. The remaining feature set was then subjected to the binomial least absolute shrinkage and selection operator (LASSO) ($\alpha = 1$), with the regularization parameter (λ) optimized via leave-one-out cross-validation using the minimum cross-validated deviance criterion. To avoid overfitting, the 1-standard error rule was applied, resulting in the selection of 5 features (Supplementary Table 1).

Six machine learning (ML) classifiers [random forest (RF), support vector machine (SVM), XGBoost, Decision Tree (DT), Naïve Bayes, k-nearest neighbors (kNN)] were trained. Model training and hyperparameter tuning were performed on the training set using 5-fold cross-validation. The classifier with the highest area under the curve (AUC) value during cross-validation was chosen as the final model for each feature set. During the entire model development process, the test dataset remained completely apart from the other data and was used only for final performance evaluation, thereby preventing the possibility of data leakage.

Univariate logistic regression analysis was performed to identify significant imaging predictors for the non-radiomic semantic and ADC features. Subsequently, semantic and ADC models were developed based on these selected predictors using the best ML algorithm.

Performance evaluation

The final models (semantic, ADC, and radiomics) were evaluated on the external test set using sensitivity, specificity, accuracy, balanced accuracy, positive predictive value, negative predictive value, and F1-score. Discriminative ability was assessed by generating receiver operating characteristic (ROC) curves and calculating the AUC. Pairwise comparisons of AUCs were conducted using the DeLong test to determine statistical significance. Model calibration was evaluated with calibration curves, and clinical utility was assessed using decision curve analysis (DCA). Finally, radiomics model interpretability was enhanced by applying SHapley Additive exPlanations (SHAP), which allows quantification of the contribution of each radiomic feature to model predictions.

Statistical analysis

All statistical analyses were performed using R (v4.3.2; R Foundation for Statistical Computing, Vienna, Austria) and SPSS (v23.0; IBM Corp., Armonk, NY, USA) under the supervision of an experienced data analyst. Normality of continuous variables was assessed using the Shapiro–Wilk test, and homogeneity of variance was tested using Levene's test. Descriptive statistics were reported as median (25^{th} percentile–75^{th} percentile) for continuous variables and as frequencies (percentages) for categorical variables. Differences between groups were compared using the Student's t-test or Mann–Whitney U test for continuous variables, and the chi-squared or Fisher's exact test for categorical variables. Univariate logistic regression was used to identify significant predictors. A P value of <0.05 was considered statistically significant.

For radiomics feature selection, ICC was calculated using the "irr" and "lpSolve" packages, and redundant features were removed with pairwise Spearman correlation. Binomial LASSO regression was performed with the "glmnet" package. The RF classifier was im-

plemented using the "randomForest" package with 1,000 trees and mtry tuned across values of 2–4, with the best performance identified at mtry = 3. In addition, five other supervised ML algorithms were developed within the "caret" framework: SVM ("e1071"), Naïve Bayes ("e1071"), XGBoost ("xgboost"), kNN ("class"), and DT ("rpart").

Model performance was assessed using 5-fold cross-validation in the training set and confirmed in the external test set. Discrimination was evaluated by ROC curves generated with the "pROC" package, and AUCs were compared using DeLong's test. Calibration was assessed using calibration plots ("rms" package), and clinical utility was evaluated by DCA ("rmda" package). Model interpretability was enhanced with SHAP implemented via the "fastshap," "shapviz," and "ggplot2" packages.

Results

Patient characteristics

The baseline radiological characteristics of the 93 patients are summarized in Table 1. Patients with SBO had a mean age of 65 ± 11 years, whereas those with NPca had a mean age of 50 ± 14 years. Both groups were predominantly male (SBO ratio: 4:1; NPca ratio: 3.5:1).

A comparison of the training and test cohorts revealed no significant differences in the distribution of semantic MRI findings ($P > 0.05$). However, a statistically significant difference was observed in the ADC metrics (ADCmean and nADC) between the two cohorts ($P < 0.05$)

Table 1. Comparison of semantic MRI findings and ADC values between training and test sets

	Training set (n = 63)			Test set (n = 30)			
	SBO (n = 31)	NPca (n = 32)	*P* value	SBO (n = 12)	NPca (n = 18)	*P* value	*P* value
ADC variables							
ADCmean (10^{-3} mm²/s)	1.078 (0.988–1.183)	0.699 (0.605–0.799)	< 0.001[b]	0.967 ± 0.309	0.726 ± 0.118	0.023[a]	0.037[b]
nADC	1.33 (1.17–1.55)	0.92 (0.81–1.03)	< 0.001[b]	1.14 ± 0.30	0.73 ± 0.15	< 0.001[a]	< 0.001[b]
Semantic MRI variables							
Architectural distortion							
No	27 (87.1%)	7 (21.9%)	< 0.001[c]	10 (83.3%)	3 (16.7%)	< 0.001[c]	0.461[c]
Yes	4 (12.9%)	25 (78.1%)		2 (16.7%)	15 (83.3%)		
Involvement of lateral structures							
No	13 (41.9%)	30 (93.8%)	< 0.001[c]	4 (33.3%)	18 (100%)	< 0.001[c]	0.797[c]
Yes	18 (58.1%)	2 (6.3%)		8 (66.7%)	0		
T2-signal							
Low	23 (74.2%)	32 (100%)	0.002[c]	9 (75%)	18 (100%)	0.054[c]	> 0.05[c]
High	8 (25.8%)	0		3 (25%)	0		
Enhancement							
Low	11 (35.5%)	28 (87.5%)	< 0.001[c]	2 (16.7%)	17 (94.4%)	< 0.001[c]	> 0.05[c]
High	20 (64.5%)	4 (12.5%)		10 (83.3%)	1 (5.6%)		
Dural thickening							
No	13 (41.9%)	19 (59.4%)	0.258[c]	4 (33.3%)	16 (88.9%)	0.004[c]	0.223[c]
Yes	18 (58.1%)	13 (40.6%)		8 (66.7%)	2 (20%)		
Venous thrombosis							
No	23 (74.2%)	32 (100%)	0.002[c]	9 (75%)	18 (100%)	0.054[c]	0.973[c]
Yes	8 (25.8%)	0		3 (25%)	0		
Abscess/fluid collection							
No	17 (54.8%)	31 (96.9%)	< 0.001[c]	7 (58.3%)	18 (100%)	0.006[c]	0.607[c]
Yes	14 (45.2%)	1 (3.1%)		5 (41.7%)	0		
Lymphadenopathy							
No	27 (87.1%)	4 (12.5%)	< 0.001[c]	8 (66.7%)	4 (22.2%)	0.024[c]	0.542[c]
Yes	4 (12.9%)	28 (87.5%)		4 (33.3%)	14 (77.8%)		
Otologic inflammation							
No	5 (16.1%)	31 (96.9%)	< 0.001[c]	4 (33.3%)	18 (100%)	<0.001[c]	0.201[c]
Yes	26 (83.9%)	1 (3.1%)		8 (66.7%)	0		

Data are presented mean ± SD or median (25^{th} percentile–75^{th} percentile), as appropriate. Categorical variables reported as frequency (percent). Differences between groups were analyzed using: [a], Independent t-test; [b], Mann–Whitney U test; [c], Yates continuity correction chi-squared test; [d], Fisher's exact test. MRI; magnetic resonance imaging; ADC, apparent diffusion coefficient; SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma; ADCmean, mean of the ADC values; nADC, normalized apparent diffusion coefficient.

Non-radiomic data analysis

All nine semantic MRI findings showed statistically significant differences between the groups ($P < 0.05$, Table 1). Lateral extension, high T2 signal, marked enhancement, dural venous thrombosis, dural thickening, fluid loculation/abscess, and ear inflammation were more common in SBO, whereas architectural distortion and lymphadenopathy were more frequent in NPca. The most common MRI findings in SBO included absence of architectural distortion (n = 27, 87.1%), presence of ear inflammation (n = 26, 83.9%), and presence of marked enhancement (n = 20, 64.5%). Based on univariate logistic regression, five significant predictors of SBO were identified: presence of lateral extension [odds ratio (OR): 36.71], marked enhancement (OR: 20.77), dural thickening (OR: 3.57), absence of architectural distortion (OR: 0.04), and absence of cervical lymphadenopathy (OR: 0.044) (Supplementary Table 2).

In the training set, both the ADCmean and nADC values were significantly higher in the SBO group than in the NPca group ($P < 0.001$). The ADCmean (OR: 1.012) value was identified as a significant predictor of SBO based on univariate logistic regression ($P < 0.001$) (Table 2).

Cut-off values for ADCmean and nADC, determined using the Youden index, were 0.954 10^{-3} mm^2/s and 1.133 10^{-3} mm^2/s, respectively. These predefined thresholds demonstrated good diagnostic performance in the training set [ADCmean: AUC = 0.849, 95% confidence interval (CI): 0.65–1.00; nADC: AUC: 0.897, 95% CI: 0.70–1.00], but performance declined in the external test set (ADCmean: AUC: 0.688, 95% CI: 0.49–0.88; nADC: AUC: 0.705, 95% CI: 0.51–0.90) (Table 2).

Radiomics model performance

The performance of six candidate classifiers was assessed using 5-fold cross-validation. Among these, the RF achieved the highest mean AUC (0.985, 95% CI: 0.956–1.00) and was selected as a final model (Table 3).

Comparative analysis of final models

The top-performing radiomics RF model was compared with two other models developed using the same RF classifier: a semantic model (using the five selected qualitative features) and an ADC model (using ADCmean). The performance of all three models on the external test set is detailed in Table 4. The semantic model achieved the highest diagnostic performance (AUC: 0.940, 95% CI: 0.857–1.00), followed by the radiomics model (AUC: 0.903, 95% CI: 0.784–1.00) and ADC model (AUC: 0.694, 95% CI: 0.497–0.892). The difference between the semantic and radiomics models did not reach statistical significance ($P = 0.644$), whereas both significantly outperformed the ADC model ($P < 0.05$) (Figure 3). The confusion matrices corresponding to the three evaluated models are presented in Supplementary Figure 4.

Clinical utility and model interpretability

The calibration curves indicated that the radiomics model was well-calibrated, with its bias-corrected curve closely tracking the ideal line for predicted probabilities between approximately 0.20 and 0.85 (Figure 4). The semantic model also demonstrated good overall calibration, though with a slight tendency to overestimate risk at very high probabilities (> 0.8) (Figure 5). In contrast, the ADC model was poorly calibrated, showing considerable underestimation in the mid-range and erratic overestimation at the upper end of the probability scale.

DCA demonstrated that both the radiomics and semantic models provided substantial clinical utility over default strategies within a 10%–70% threshold range. The semantic model was slightly superior at lower thresholds, whereas the radiomics model showed comparable or greater benefit in the mid-range. Conversely, the ADC model offered minimal net benefit, indicating a general lack of clinical utility. The clinical relevance of all models decreased markedly beyond a 0.70 threshold.

The SHAP analysis revealed that t2_wavelet_LLL_firstorder_Minimum was the most influential feature in model predictions, exerting a strong negative contribution to the probability of SBO (i.e., lower values were associated with a higher likelihood of SBO). In contrast, t1_original_firstorder_Skewness demonstrated a positive contribution, indicating that greater skewness increased the predicted probability of SBO (Supplementary Figure 5). The models' outputs and reference standard were presented through the use of sample cases to enhance the transparency of the models (Figures 6, 7).

Table 2. Performance analysis of ADC parameters in training and external test cohorts

Parameter	Cut-off	Dataset	Accuracy	Sensitivity	Specificity	AUC (95% CI)
ADCmean	0.954 × 10^{-3} mm^2/s	Training	0.91 (0.84–0.98)	0.84 (0.69–0.98)	1.00 (0.89–1.00)	0.85 (0.65–1.00)
		Test	0.77 (0.61–0.92)	0.50 (0.22–0.78)	0.94 (0.84–1.00)	0.69 (0.49–0.88)
nADC	1.133	Training	0.90 (0.82–0.97)	0.87 (0.74–1.00)	0.92 (0.84–1.00)	0.89 (0.70–1.00)
		Test	0.80 (0.66–0.94)	0.50 (0.22–0.78)	1.00 (0.81–1.00)	0.70 (0.51–0.90)

AUC, area under the curve; CI, confidence interval; ADCmean, mean apparent diffusion coefficient; nADC, normalized apparent diffusion coefficient.

Table 3. Five-fold cross-validation classification performance of six different modeling strategies

Model	Performance measures							
	Accuracy (95% CI)	Sensitivity	Specificity	PPV	NPV	AUC (95% CI)	BA	F1
Support vector machine	0.921 (0.824–0.974)	0.936	0.906	0.906	0.936	0.968 (0.920–1.00)	0.921	0.921
Naive bayes	0.936 (0.845–0.982)	0.968	0.906	0.909	0.967	0.962 (0.917–1.00)	0.937	0.937
Random forest	0.921 (0.824–0.974)	0.968	0.875	0.882	0.966	0.985 (0.956–1.00)	0.921	0.923
K-nearest neighbors	0.952 (0.867–0.990)	0.967	0.936	0.938	0.968	0.979 (0.952–1.00)	0.953	0.952
XGBoost	0.905 (0.804–0.964)	0.936	0.875	0.879	0.933	0.968 (0.928–1.00)	0.905	0.906
Decision tree	0.889 (0.784–0.954)	0.903	0.875	0.875	0.903	0.914 (0.840–0.989)	0.889	0.889

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; BA, balanced accuracy.

Table 4. Diagnostic performances of semantic, ADC, and radiomics (random forest) models in the cross-validation and external test set for differentiating skull base osteomyelitis from nasopharyngeal carcinoma

Model	Dataset	Accuracy (95% CI)	Sens.	Spec.	PPV	NPV	AUC (95% CI)	BA	F1
Semantic	Training (CV)	0.888 (0.784–0.954)	0.871	0.906	0.900	0.878	0.953 (0.900–1.00)	0.888	0.885
	Test	0.900 (0.735–0.978)	0.916	0.888	0.846	0.941	0.940 (0.857–1.00)	0.903	0.880
Radiomics	Training (CV)	0.921 (0.824–0.974)	0.968	0.875	0.882	0.966	0.985 (0.956–1.00)	0.921	0.923
	Test	0.800 (0.614–0.923)	0.917	0.722	0.688	0.929	0.903 (0.784–1.00)	0.819	0.786
ADC-only	Training (CV)	0.929 (0.830–0.980)	0.903	0.961	0.965	0.892	0.981 (0.955–1.00)	0.932	0.933
	Test	0.700 (0.506–0.852)	0.500	0.833	0.667	0.714	0.694 (0.497–0.892)	0.667	0.571

ADC, apparent diffusion coefficient; CI, confidence interval; Sens., sensitivity; Spec., specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; BA, balanced accuracy; CV, cross-validation.

Discussion

In this study, we developed and externally validated an MRI-based radiomics model to differentiate SBO from locally advanced NPca. The model demonstrated robust and generalizable performance on the external test set (AUC: 0.903), comparable with prior literature using quantitative DCE and ADC parameters.[8,9] Importantly, only modest performance degradation was observed in the independent cohort, highlighting the generalizability and robustness of our radiomics model. Notably, although a semantic model based on expert interpretation achieved a slightly higher performance (AUC: 0.940), the principal value of our radiomics model lies in its objectivity. By providing a reproducible, quantitative analysis, it can complement the radiologist's assessment, reduce inter-reader variability, and serve as a valuable decision-support tool, particularly for less experienced readers.

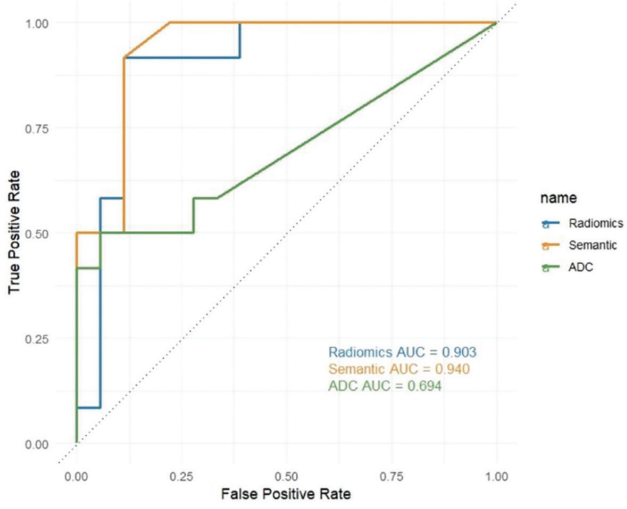


Figure 3. Receiver operating characteristic curves (ROC) and area under the curve (AUC) values of apparent diffusion coefficient (ADC), semantic, and radiomics models in the external test set.

The diagnostic challenge of distinguishing SBO from NPca is well-documented, as both can present with similar features, such as asymmetric soft tissue thickening and cranial nerve palsy.[8,10-12] Our findings on semantic MRI features align with and expand upon previous work. We reproduced many of the key indicators of SBO identified by Goh et al.[10] However, we noted some discrepancies. For instance, lateral extension was seen in 60.5% of our SBO cases, lower than the 92% reported by Goh et al.,[10] which may be due to our inclusion of more non-otogenic

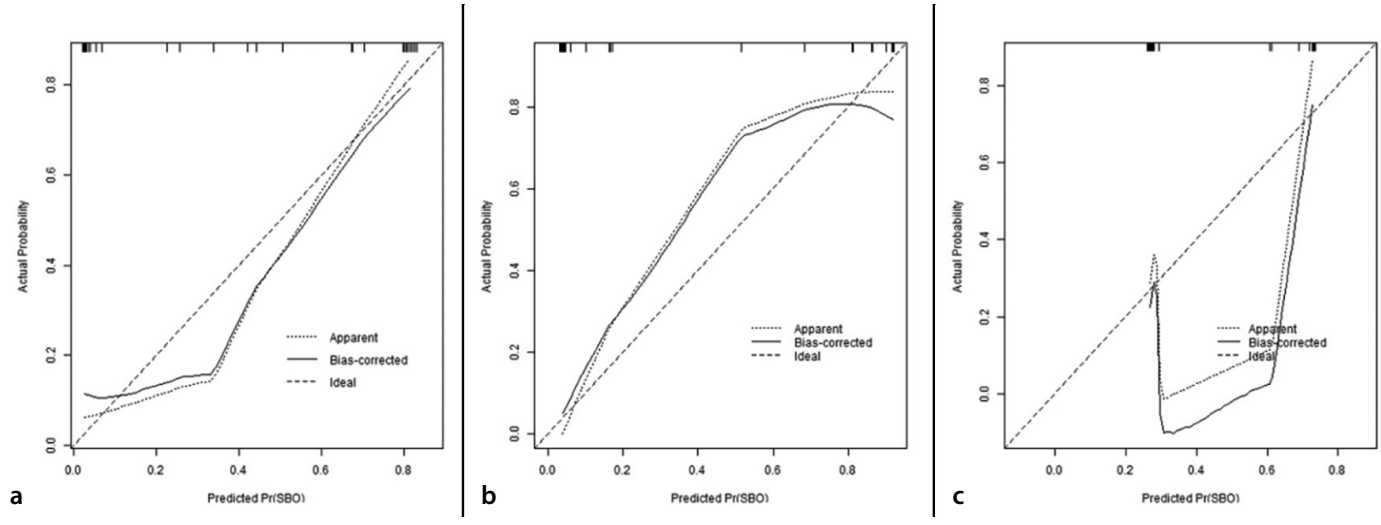


Figure 4. Calibration plots for the random forest models using radiomics features **(a)**, semantic features **(b)**, and apparent diffusion coefficient measurements **(c)**. Each plot shows the calibration of predicted probabilities for the class "SBO" against the observed actual probabilities. The "Apparent" curve represents the raw calibration, and the "Bias-corrected" curve accounts for overfitting. The "Ideal" line indicates perfect calibration, where predicted probabilities exactly match observed outcomes. SBO, skull base osteomyelitis.

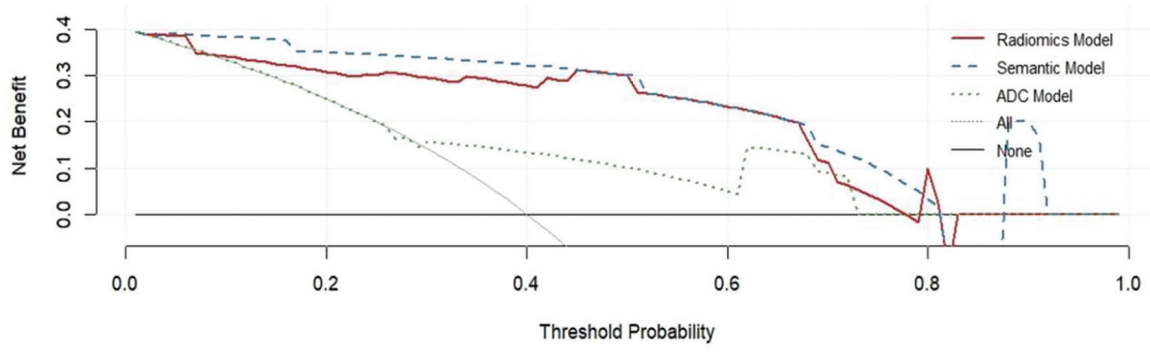


Figure 5. Decision curve analysis of the three predictive models. The plot compares the net benefit of the radiomics model (red line), the semantic model (blue dashed line), and the apparent diffusion coefficient-only model (green dotted line) across a range of threshold probabilities. The horizontal black line ("None") represents the net benefit of treating no patients, and the thin gray line ("All") represents the net benefit of treating all patients. The model with the highest curve for a given threshold probability range offers the greatest clinical utility. ADC, apparent diffusion coefficient.

SBO cases. Similarly, the prevalence of high T2 signal intensity in our SBO cohort (25.6%) was lower than in other studies.[10,22] This likely reflects the assessment methodology and pathological heterogeneity of SBO, which can include not only edema but also fibrosis and necrosis, both of which can result in a low T2 signal.[7,23,24] Furthermore, the time interval between symptom onset and imaging may have influenced the T2 signal intensity. Patients imaged at more advanced stages may exhibit a lower T2 signal due to chronic inflammation and fibrosis. Architectural distortion was noted in 14% of our SBO patients, which is higher than the previous reports in

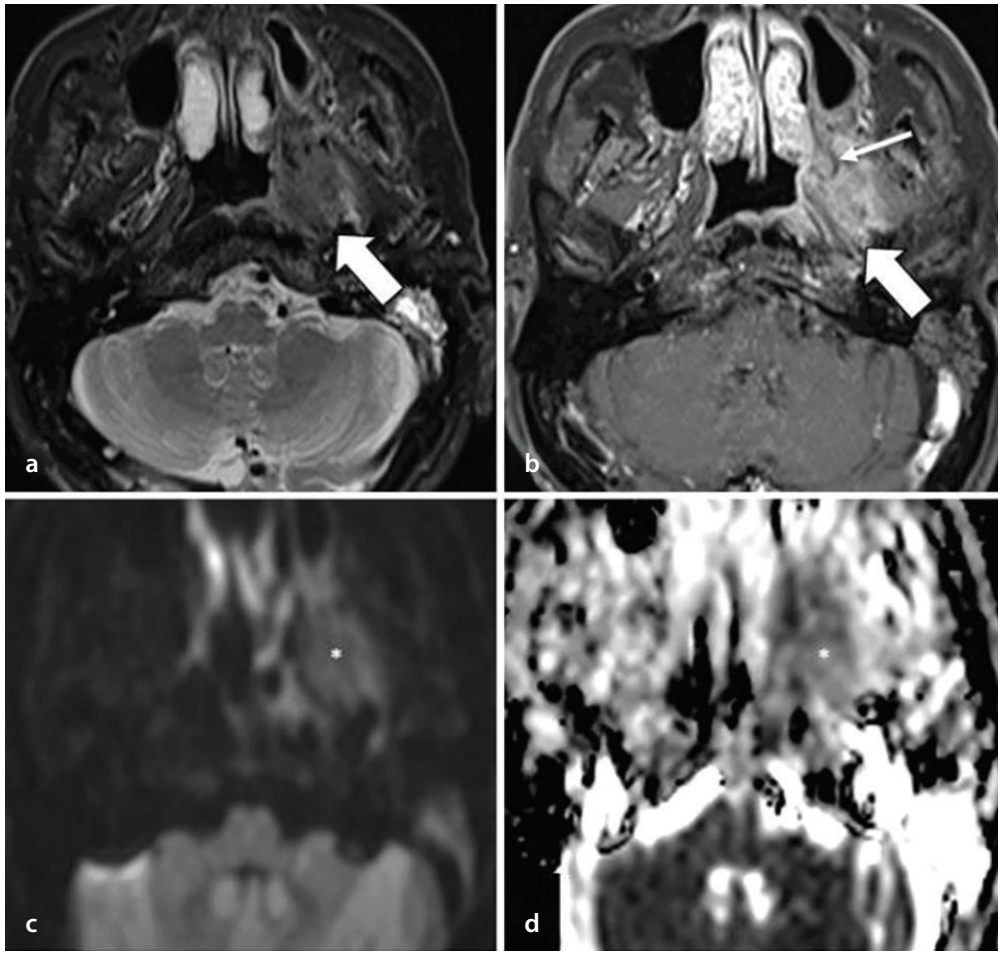


Figure 6. A patient with histologically confirmed nasopharyngeal carcinoma misclassified as skull base osteomyelitis by the semantic model but correctly identified by the radiomics model. **(a)** Axial fat-suppressed T2-weighted image showing a homogeneous, hypointense lesion relative to nasal mucosa with mild architectural distortion (arrow). Left mastoid cells show effusion without inflammatory enhancement (not shown). **(b)** Post-contrast T1-weighted image demonstrating an infiltrative, enhancing lesion in the left parapharyngeal space (thick arrow) extending into the pterygoid body (thin arrow), pterygopalatine fossa, and floor of the middle cranial fossa (not shown). The left Rosenmüller fossa is compressed without mucosal irregularity. Lateral structures, including the temporomandibular joint and parotid space, are unremarkable. The lesion enhancement pattern is similar to adjacent mucosa. **(c, d)** Diffusion-weighted imaging reveals relatively restricted diffusion (asterisks) with an apparent diffusion coefficient (ADC) of 0.86×10^{-3} mm^2/s and a normalized ADC of 0.97.

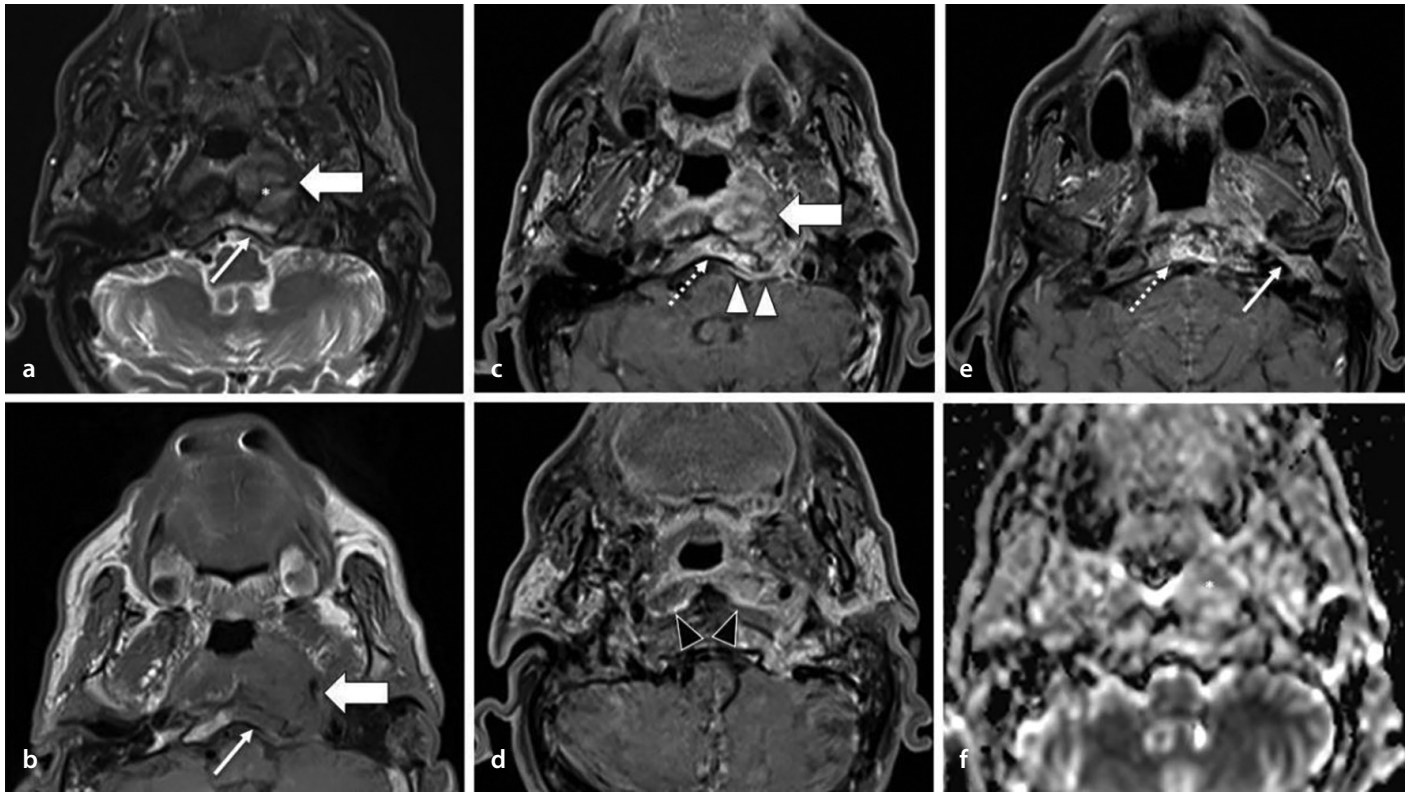


Figure 7. A patient with a confirmed diagnosis of skull base osteomyelitis (SBO) was misclassified as nasopharyngeal carcinoma by the semantic model but correctly identified as SBO by the radiomics model. **(a)** Axial fat-suppressed T2-weighted image demonstrates mild hyperintensity of the lesion (thick arrow) and bone marrow edema (thin arrow) in the left occipital condyle. Mild architectural distortion and asymmetric expansion of the left prevertebral muscles are also visible (asterisk). **(b)** Axial T1-weighted image shows deep hypointense bone marrow signal in the occipital condyle (thin arrow). **(c, e)** Axial post-contrast T1-weighted images show a markedly enhancing lesion involving the left prevertebral muscles, occipital condyle, and clivus (thick arrows), extending along the left Eustachian tube **(e)**, (thin arrow). Note the linear enhancement along the retropharyngeal fascia **(d),** (black arrowheads), and compressed left Rosenmüller fossa without mucosal irregularities **(c)**. Lateral structures, including the temporomandibular joint and parotid space, are preserved. Lesion enhancement is comparable with adjacent mucosa, with multifocal regions of higher enhancement within the bone **(a, e)**; (dashed arrows). Dural thickening and enhancement are also present (**d**), (white arrowheads). **(f)** Apparent diffusion coefficient (ADC) map demonstrates facilitated diffusion (asterisk) relative to brain parenchyma (ADC: 1.24×10^{-3} mm^2/s; normalized ADC: 1.38).

the literature.[10,22] This discrepancy may be explained by differences in inclusion criteria. Unlike previous studies, our cohort excluded SBO cases confined to the temporal bone, instead focusing on advanced-stage patients with prominent soft tissue involvement, which can mimic NPca.

Although prior studies by Ozgen et al.[8] and Baba et al.[9] reported high efficacy for ADC parameters, with 96% accuracy and an AUC of 0.96, respectively, our ADC cut-off and model performance were notably lower. This discrepancy likely arises from two factors: lower mean ADC values observed in our advanced SBO cases[25] and variations in the time interval between symptom onset and imaging. Unlike previous studies, we applied established cohorts to an independent, unseen patient population and observed a significant decline in the diagnostic performance of ADC values. It should be acknowledged that significant differences in ADC distributions between the training and test cohorts likely contributed to the reduced performance of the ADC-based model on external validation, rendering the comparison with radiomics not entirely equivalent. Notably, despite the application of normalization procedures in line with commonly adopted practices in the literature, ADC metrics still exhibited limited generalizability. This observation underscores the particular sensitivity of ADC values to inter-center variability and domain shifts, which may constrain their utility as stand-alone biomarkers in multicenter studies. This may be attributed to various factors, such as the selected b-values, patient-related factors, and differences between scanners. This highlights the challenge of applying a single quantitative ADC threshold across different institutions without standardized protocols.[26,27]

Radiomics has been extensively studied in NPca for applications such as differential diagnosis, risk stratification, and the prediction of patient prognosis and treatment response.[28] However, to our knowledge, no radiomics studies have specifically addressed SBO.

Our SHAP analysis revealed that the most influential radiomic predictors were t2_wavelet_LLL_firstorder_Minimum and t1_original_firstorder_Skewness. The minimum intensity feature extracted from wavelet–filtered T2-weighted images was inversely associated with SBO probability, suggesting that lower signal minima were more indicative of NPca. This finding is biologically plausible when considered alongside qualitative assessment, as SBO often demonstrates a uniformly elevated T2 signal, similar to the mucosal tissue, due to edema and inflammatory exudation in the early disease phase. In contrast, NPca typically contains lower intensity foci related to dense cellularity, consistent with the observed association of lower minimum intensity values with malignancy. Positive skewness indicates a right-tailed distribution, where most voxels exhibit relatively low-to-intermediate signal intensities. In our analysis, increased skewness was positively associated with SBO, which is biologically plausible given that inflammatory lesions are typically fluid-rich and therefore

demonstrate lower baseline T1 intensities than the denser, more solid cellular architecture of NPca.

Our semantic model, built with only five features, demonstrated strong performance. A semantic model developed on a larger dataset incorporating a broader range of features may yield even stronger results. Notably, three of the five selected radiomic predictors were first-order features, and SHAP analysis identified two of these first-order metrics as the most influential contributors to model predictions. This finding is biologically plausible, as these quantitative features are the mathematical equivalent of qualitative assessments—such as T2 signal intensity—that radiologists routinely use. Crucially, our radiomic approach achieved performance comparable with the human-led semantic evaluation but with the key advantages of being fully objective and reproducible. This suggests that with larger datasets, radiomics holds the potential to surpass purely visual assessment by quantifying subtle signal variations that are imperceptible to the human eye.

Future research should validate the robustness and generalizability of this radiomics approach in larger, prospective, multicenter cohorts. Additionally, comparative studies assessing the diagnostic performance of the model against that of expert neuroradiologists are necessary to formally establish its added clinical value. Furthermore, integration of relevant clinical risk factors, such as diabetes, inflammatory markers, and immune status, into the predictive model could enhance its diagnostic accuracy and clinical applicability. Due to the retrospective data collection, clinical history or acute phase reactants, such as erythrocyte sedimentation rate and C-reactive protein, were unavailable for a significant portion of the cohort, which prevented their inclusion in our models.

There were several limitations to our study. First, the retrospective design may have introduced selection bias. Second, the small sample size, particularly for certain subgroups, may have limited the performance of the semantic model and necessitates validation in larger cohorts. Moreover, future studies could compare these models with qualitative assessments by expert radiologists. Third, our narrow focus on skull base pathologies may limit the broader clinical applicability of the model. Finally, the significant differences in ADC parameters between the training and test sets may have contributed to the suboptimal performance of the ADC model, likely owing to the small sample size. Although data processing may address these discrepancies, it was not applied to avoid introducing further uncertainties.

In conclusion, the radiomics model demonstrated high and robust performance in differentiating SBO from advanced NPca. Although the semantic model based on expert assessment achieved slightly higher performance, radiomics offers a complementary, quantitative, and objective approach. In contrast, the ADC-based strategy proved highly sensitive to inter-center variability, suggesting that site-specific threshold determination may be more appropriate.

Acknowledgements

The authors gratefully acknowledge Ahmet Husrev Kösebalaban for his invaluable assistance with technical support, image preprocessing, and radiomic feature extraction.

Footnotes

Conflict of Interest

The authors declared that they have no conflict of interest.

References

1. Chapman PR, Choudhary G, Singhal A. Skull base osteomyelitis: a comprehensive imaging review. *AJNR Am J Neuroradiol*. 2021;42(3):404-413. [Crossref]
2. Mahdyoun P, Pulcini C, Gahide I, et al. Necrotizing otitis externa: a systematic review. *Otol Neurotol*. 2013;34(4):620-629. [Crossref]
3. Álvarez Jáñez F, Barriga LQ, Iñigo TR, Roldán Lora F. Diagnosis of skull base osteomyelitis. *Radiographics*. 2021;41(1):156-174. [Crossref]
4. Takata J, Hopkins M, Alexander V, et al. Systematic review of the diagnosis and management of necrotising otitis externa: highlighting the need for high-quality research. *Clin Otolaryngol*. 2023;48(3):381-394. [Crossref]
5. Başaran S, Evlice O, Benli A, et al. Skull base osteomyelitis and long term outcome. *Klimik Derg*. 2021;34(2):129-137. [Crossref]
6. Chawdhary G, Pankhania M, Douglas S, Bottrill I. Current management of necrotising otitis externa in the UK: survey of 221 UK otolaryngologists. *Acta Otolaryngol*. 2017;137(8):818-822. [Crossref]
7. van Kroonenburgh AMJL, van der Meer WL, Bothof RJP, van Tilburg M, van Tongeren J, Postma AA. Advanced imaging techniques in skull base osteomyelitis due to malignant otitis externa. *Curr Radiol Rep*. 2018;6(1):3. [Crossref]
8. Ozgen B, Oguz KK, Cila A. Diffusion MR imaging features of skull base osteomyelitis compared with skull base malignancy. *AJNR Am J Neuroradiol*. 2011;32(1):179-184. [Crossref]
9. Baba A, Kurokawa R, Kurokawa M, Ota Y, Srinivasan A. Dynamic contrast-enhanced MRI parameters and normalized ADC values could aid differentiation of skull base osteomyelitis from nasopharyngeal cancer. *AJNR Am J Neuroradiol*. 2023;44(1):74-78. [Crossref]
10. Goh JPN, Karandikar A, Loke SC, Tan TY. Skull base osteomyelitis secondary to malignant otitis externa mimicking advanced nasopharyngeal cancer: MR imaging features at initial presentation. *Am J Otolaryngol*. 2017;38(4):466-471. [Crossref]
11. See A, Tan TY, Gan EC. Atypical culture-negative skull base osteomyelitis masquerading as advanced nasopharyngeal carcinoma. *Am J Otolaryngol*. 2016;37(3):236-239. [Crossref]
12. Lee B, Bae YJ, Choi BS, et al. Radiologic differentiation between granulomatosis with polyangiitis and its mimics involving the skull base in humans using high-resolution magnetic resonance imaging. *Diagnostics (Basel)*. 2021;11(11):2162. [Crossref]
13. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91. [Crossref]
14. Shur JD, Doran SJ, Kumar S, et al. Radiomics in oncology: a practical guide. *Radiographics*. 2021;41(6):1717-1732. [Crossref]
15. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. *Eur J Radiol*. 2020;127:108991. [Crossref]
16. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics Research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [Crossref].
17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30(9):1323-1341. [Crossref]
18. 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. [Crossref]
19. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. [Crossref]
20. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One*. 2019;14(3):e0213459. [Crossref]
21. Koçak B, Yüzkan S, Mutlu S, et al. Influence of image preprocessing on the segmentation-

based reproducibility of radiomic features: *in vivo* experiments on discretization and resampling parameters. *Diagn Interv Radiol*. 2024;30(3):152-162. [Crossref]

22. Brenner A, Cavel O, Shendler G, et al. CT findings in temporal bone sites in skull base osteomyelitis from malignant otitis externa. *Eur Arch Otorhinolaryngol*. 2023;280(6):2687-2694. [Crossref]
23. Urbančič J, Vozel D, Battelino S, et al. Atypical skull-base osteomyelitis: comprehensive review and multidisciplinary management viewpoints. *Trop Med Infect Dis*. 2023;8(5):254. [Crossref]
24. Adams A, Offiah C. Central skull base osteomyelitis as a complication of necrotizing otitis externa: imaging findings, complications, and challenges of diagnosis. *Clin Radiol*. 2012;67(10):e7-e16. [Crossref]
25. Cherko M, Nash R, Singh A, Lingam RK. Diffusion-weighted magnetic resonance imaging as a novel imaging modality in assessing treatment response in necrotizing otitis externa. *Otol Neurotol*. 2016 Jul;37(6):704-707. [Crossref]
26. Ogura A, Hatano I, Osakabe K, Yamaguchi N, Koyama D, Watanabe H. Importance of fractional b value for calculating apparent diffusion coefficient in DWI. *AJR Am J Roentgenol*. 2016;207(6):1239-1243. [Crossref]
27. Kim SY, Lee SS, Park B, et al. Reproducibility of measurement of apparent diffusion coefficients of malignant hepatic tumors: effect of DWI techniques and calculation methods. *J Magn Reson Imaging*. 2012;36(5):1131-1138. [Crossref]
28. Duan W, Xiong B, Tian T, et al. Radiomics in nasopharyngeal carcinoma. *Clin Med Insights Oncol*. 2022;16:11795549221079186. [Crossref]

Supplementary Text[16]

CLEAR Checklist v1.0

Section	No.	Item	Yes	No	n/a	Page
Title						
	1	Relevant title, specifying the radiomic methodology	☑	☐	☐	
Abstract						
	2	Structured summary with relevant information	☑	☐	☐	
Keywords						
	3	Relevant keywords for radiomics	☑	☐	☐	
Introduction						
	4	Scientific or clinical background	☑	☐	☐	
	5	Rationale for using a radiomic approach	☐	☐	☐	
	6	Study objective(s)	☑	☐	☐	
Method						
Study design	7	Adherence to guidelines or checklists (e.g., CLEAR checklist)	☑	☐	☐	
	8	Ethical details (e.g., approval, consent, data protection)	☑	☐	☐	
	9	Sample size calculation	☐	☑	☐	
	10	Study nature (e.g., retrospective, prospective)	☑	☐	☐	
	11	Eligibility criteria	☑	☐	☐	
	12	Flowchart for technical pipeline	☑	☐	☐	
Data	13	Data source (e.g., private, public)	☑	☐	☐	
	14	Data overlap	☐	☑	☐	
	15	Data split methodology	☑	☐	☐	
	16	Imaging protocol (i.e., image acquisition and processing)	☑	☐	☐	
	17	Definition of non-radiomic predictor variables	☑	☐	☐	
	18	Definition of the reference standard (i.e., outcome variable)	☑	☐	☐	
Segmentation	19	Segmentation strategy	☑	☐	☐	
	20	Details of operators performing segmentation	☑	☐	☐	

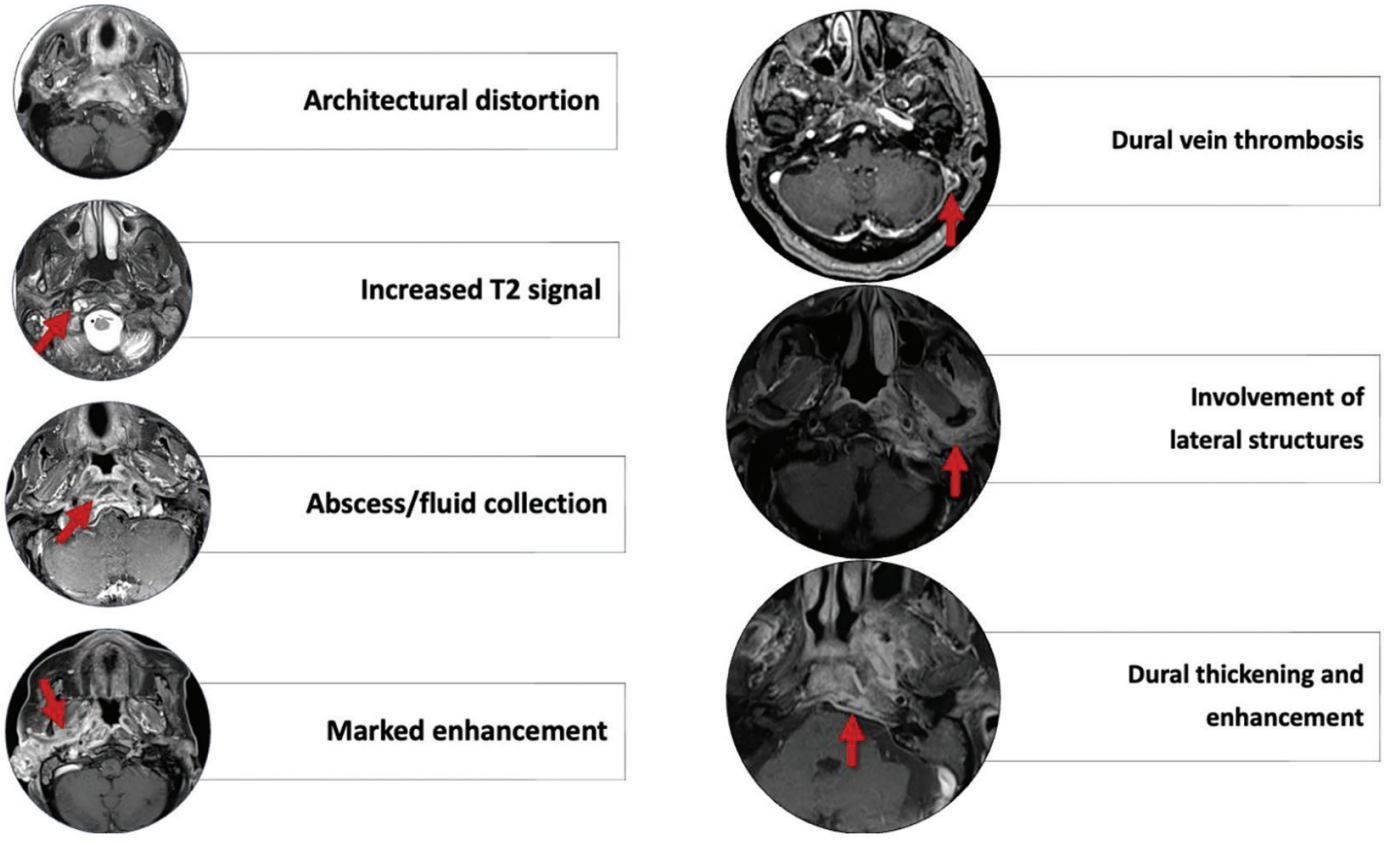
Section	No.	Item	Yes	No	n/a	Page
Pre-processing	21	Image pre-processing details	☑	☐	☐	
	22	Resampling method and its parameters	☑	☐	☐	
	23	Discretization method and its parameters	☑	☐	☐	
	24	Image types (e.g., original, filtered, transformed)	☑	☐	☐	
Feature extraction	25	Feature extraction method	☑	☐	☐	
	26	Feature classes	☑	☐	☐	
	27	Number of features	☑	☐	☐	
	28	Default configuration statement for remaining parameters	☑	☐	☐	
Data preparation	29	Handling of missing data	☐	☐	☑	
	30	Details of class imbalance	☑	☐	☐	
	31	Details of segmentation reliability analysis	☑	☐	☐	
	32	Feature scaling details (e.g., normalization, standardization)	☑	☐	☐	
	33	Dimension reduction details	☑	☐	☐	
Modeling	34	Algorithm details	☑	☐	☐	
	35	Training and tuning details	☑	☐	☐	
	36	Handling of confounders	☐	☑	☐	
	37	Model selection strategy	☑	☐	☐	
Evaluation	38	Testing technique (e.g., internal, external)	☑	☐	☐	
	39	Performance metrics and rationale for choosing	☑	☐	☐	
	40	Uncertainty evaluation and measures (e.g., confidence intervals)	☑	☐	☐	
	41	Statistical performance comparison (e.g., DeLong's test)	☑	☐	☐	
	42	Comparison with non-radiomic and combined methods	☑	☐	☐	
	43	Interpretability and explainability methods	☑	☐	☐	
Results						
	44	Baseline demographic and clinical characteristics	☑	☐	☐	
	45	Flowchart for eligibility criteria	☑	☐	☐	
	46	Feature statistics (e.g., reproducibility, feature selection)	☑	☐	☐	
	47	Model performance evaluation	☑	☐	☐	
	48	Comparison with non-radiomic and combined approaches	☑	☐	☐	

Section	No.	Item	Yes	No	n/a	Page
Discussion						
	49	Overview of important findings	☑	☐	☐	
	50	Previous works with differences from the current study	☑	☐	☐	
	51	Practical implications	☑	☐	☐	
	52	Strengths and limitations (e.g., bias and generalizability issues)	☑	☐	☐	
Data availability	53	Sharing images along with segmentation data [n/e]	☐	☑	☐	
	54	Sharing radiomic feature data	☑	☐	☐	
Code availability	55	Sharing pre-processing scripts or settings	☑	☐	☐	
	56	Sharing source code for modeling	☐	☑	☐	
Model availability	57	Sharing final model files	☐	☑	☐	
	58	Sharing a ready-to-use system [n/e]	☐	☑	☐	

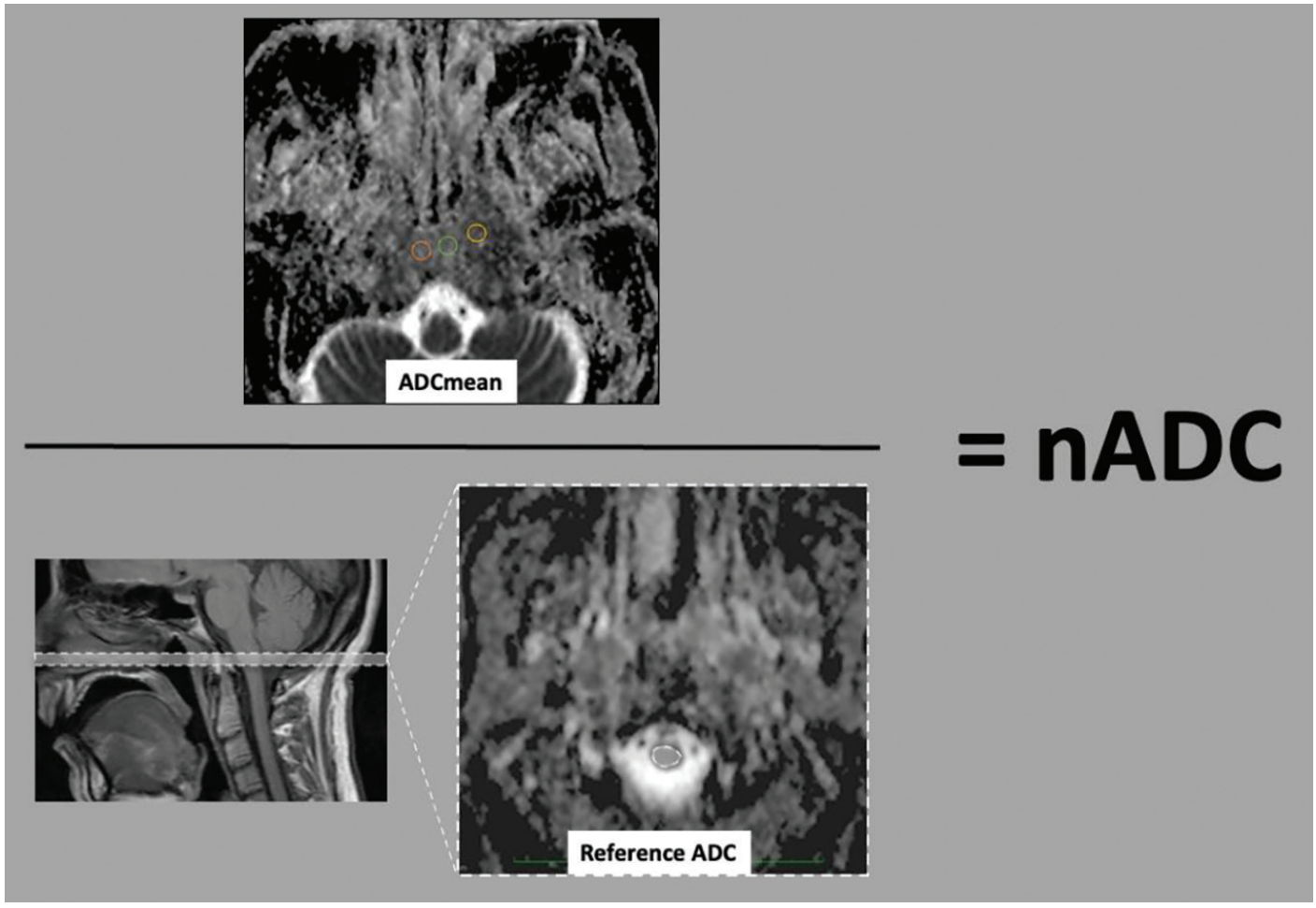
MRI acquisition

Imaging protocols varied due to studies being performed across different centers, timeframes, and scanners. Diffusion-weighted imaging was not available in 6 patients (6.45%). The distribution of magnetic resonance imaging (MRI) vendors was as follows: Siemens (54.8%), Philips (28%), and GE (17.2%). Axial contrast-enhanced T1-weighted images were obtained after the injection of 0.1 mmoL/kg of gadolinium-based contrast agents. All but three patients (3.2%) underwent imaging in a 1.5T MRI unit.

The median in-plane resolutions of the T1w, FS-T2w, and T1c series were 0.625 x 0.625 mm^2, 0.69 x 0.69 mm^2, and 0.68 x 0.68 mm^2, respectively (25^{th}-75^{th} percentile: 0.45-0.765 mm, 0.46-0.75 mm, 0.51-0.765 mm pixel width); the median slice thickness for all three series was 4 mm (range: 3-5 mm, 3-5 mm, 1-5 mm). Additionally, the T2 series were obtained using different fat suppression techniques [chemical fat suppression, short tau inversion recovery (STIR), Dixon]. Diffusion-weighted imaging was performed using two b-values, with the highest b-value set at 1000 on Philips scanners and 800 on other scanners.



Supplementary Figure 1. Illustration of the evaluated semantic features. The presence of architectural distortion, involvement of lateral structures, marked enhancement, dural thickening, and lymphadenopathy were used for semantic model.

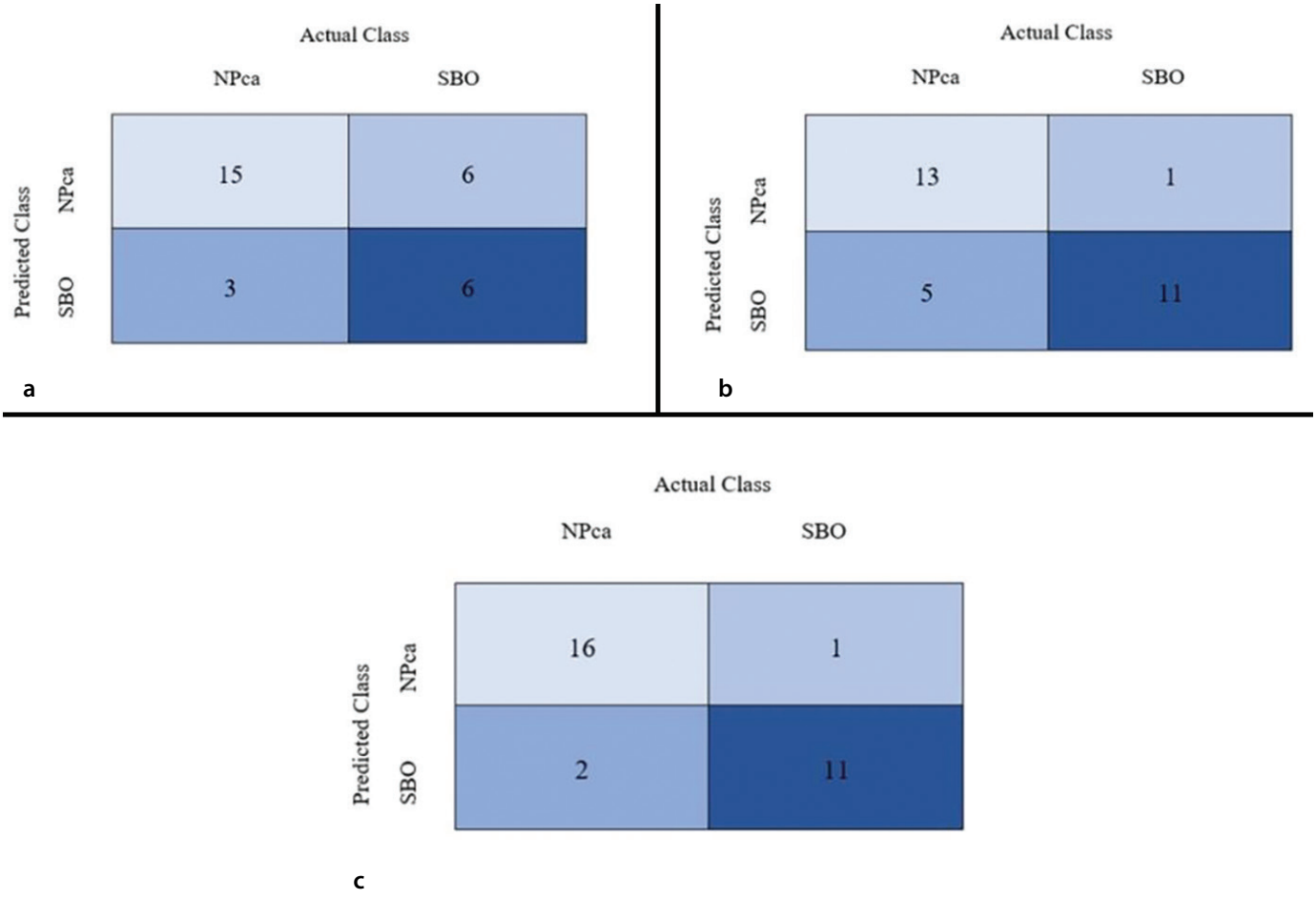


Supplementary Figure 2. Apparent diffusion coefficient (ADC) parameters. For the mean of the ADC values (ADCmean) calculation, three regions of interest (ROIs) of equal size were drawn on the lesion at the level of the skull base selecting areas of higher restricted diffusion visually. Subsequently, the arithmetic mean of these ADC values was calculated. The reference ADC value was calculated as the arithmetic mean of two ADC values which were obtained by the largest possible ROI drawn to cover the majority of the spinal cord on the slice nearest to the obex level and the slice immediately below it. Subsequently, the normalized ADC (nADC) was derived as the ratio of the ADCmean to the reference ADC. This approach was intended to mitigate the impact of center- or device-related factors.

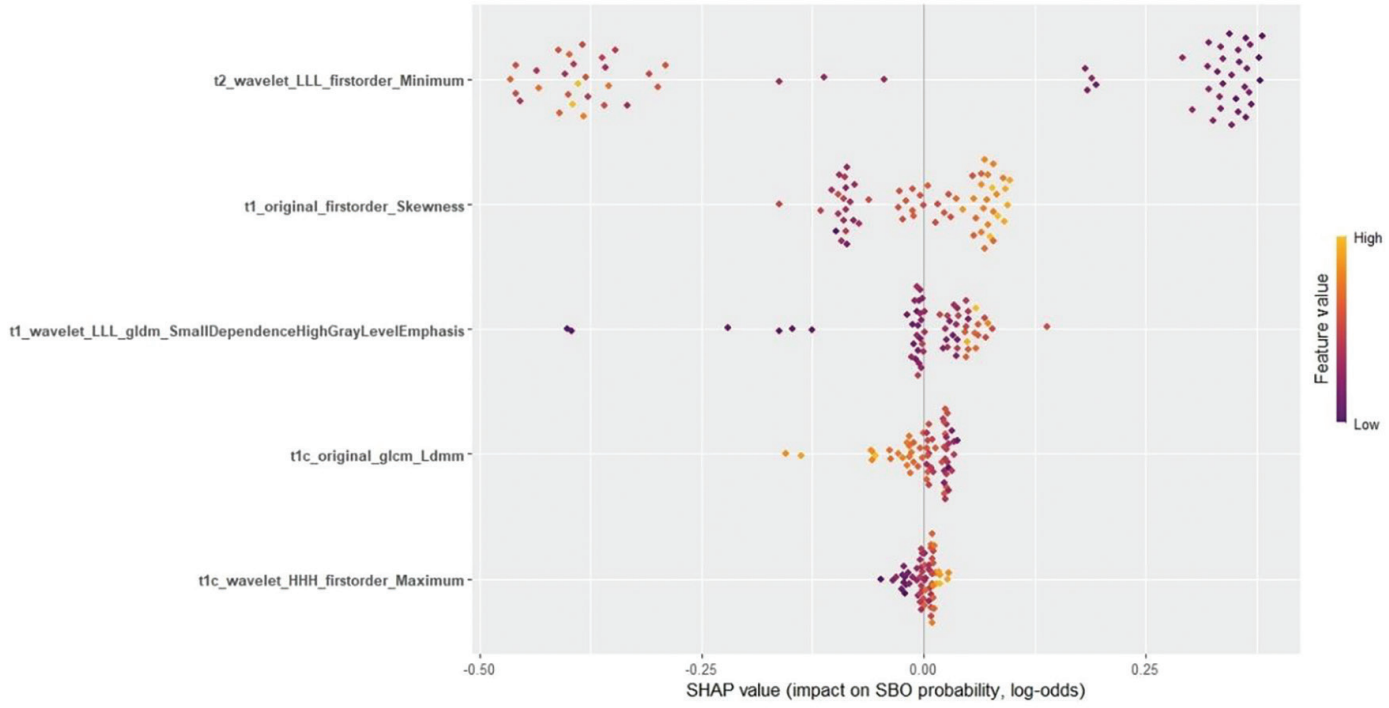
```yaml
minimumROIDimensions: 2
minimumROISize: null
normalize: true
normalizeScale: 200
removeOutliers: 3
resampledPixelSpacing:
  - 2.0
  - 2.0
  - 2.0
interpolator: "sitkBSpline"
preCrop: false
padDistance: 5
distances:
  - 1
force2D: false
force2Ddimension: 0
resegmentRange:
  - -3
  - 3
label: 1
additionalInfo: true
binWidth: 5
symmetricalGLCM: true
correctMask: true
resegmentMode: "sigma"

```

Supplementary Figure 3. The YAML configuration file used for radiomic feature extraction with PyRadiomics (v3.0.1). This file defines key settings for image preprocessing and feature calculation. Key settings for standardization include isotropic resampling to 2 x 2 x 2 mm voxels, intensity normalization, and discretization with a fixed bin width of 5. The full software citation can be found in the Methods section.



Supplementary Figure 4. Confusion matrices of the **(a)** ADC model; **(b)** radiomics model; and **(c)** semantic model. ADC, apparent diffusion coefficient; SBO, skull base osteomyelitis; NPca, nasopharyngeal carcinoma.



Supplementary Figure 5. SHapley Additive exPlanations (SHAP) summary plot for radiomics model. This plot illustrates the impact of individual feature values on the model's output (SBO probability, expressed in log-odds). Each point represents a single instance (a patient), showing its SHAP value for a given feature. The x-axis indicates the SHAP value, where positive values suggest an increase in the predicted probability of SBO, and negative values suggest a decrease. The y-axis lists the top contributing radiomics features. The color of each point represents the actual feature value for that instance, ranging from "Low" (purple) to "High" (yellow), providing insight into how the magnitude of a feature's value correlates with its impact on the prediction. For instance, high values of "t2_wavelet_LLL_firstorder_Minimum" appear to be associated with a positive impact on SBO probability, while low values are associated with a negative impact. SBO, skull base osteomyelitis.

Supplementary Table 1. List of 5 selected features and corresponding inter-class correlation coefficient (ICC) values

MRI sequence	Image type	Feature class	Feature name	ICC
T1c (2 features)	Original	GLCM	Idmn	0.928
	Wavelet_HHH	First order	Maximum	0.932
T1w (2 features)	Original	First order	Skewness	0.903
	Wavelet_LLL	GLDM	SmallDependenceHighGrayLevelEmphasis	0.905
FS – T2w (1 feature)	Wavelet_LLL	First order	Minimum	0.925

Supplementary Table 2. Univariate analysis results of semantic MRI findings and ADC parameters

Variables		OR (95% CI)	*P* value
Semantic MRI findings	**Architectural distortion**	0.04 (0.013-0.122)	< 0.001
	Lateral extension	36.71 (7.86-171.38)	< 0.001
	Marked enhancement	20.77 (6.71-64.30)	< 0.001
	Dural thickening	3.569 (1.510-8.433)	0.004
	Lymphadenopathy	0.044 (0.015-0.128)	< 0.001
ADC parameters	**ADCmean**	1.012 (1.007-1.016)	< 0.001

MRI; magnetic resonance imaging; ADC, apparent diffusion coefficient; ADCmean, mean of the ADC values; OR, odds ratio; CI, confidence interval.