

Supplementary Material

Table S1: Pyradiomics settings for feature extraction

Feature extraction parameters	Parameters
“imageType”	• Original • Wavelet • Gradient • LoG (sigma=2-5 mm)
“featureClass”	• firstorder • glcm • glrlm • glszm • gldm • shape
“settings”	• normalize: True • normalizeScale: 100 • resampledPixelSpacing: [1, 1, 1] • interpolator: 'sitkBSpline' • padDistance: 5 • minimumROIDimensions: 2 • resegmentMode: 'sigma' • binWidth: 25 • removeOutliers: 3

Table S2: Overview of Feature Selection Methods

Feature selection method	Method Name	Description
<i>Filter-based</i>	ANOVA	Compares variances among the means of different groups to determine feature significance.
	Kruskal Wallis	Tests whether two or more classes have an equal median, selecting discriminative features based on p-values.
<i>Wrapper-based</i>	Recursive Feature Elimination (RFE)	Iteratively removes the least significant features using a Support Vector Machine (SVM)-based approach.
	Boruta	Compares original features with randomly shuffled shadow features and applies a p-value correction to select the most relevant ones.
	Sequential Forward Selection (SFS)	Starts with an empty feature set and progressively adds the most impactful feature based on model performance.
	Backward Feature Selection (BFS)	Starts with the full set of features and iteratively removes the least significant feature at each step based on model performance, aiming to retain only the most impactful features.
<i>Embedded</i>	L1-LASSO	Uses L1 regularization to shrink feature coefficients, eliminating less important ones.
	Random Forest variable importance (RF-imp)	Assesses feature significance using tree minimal depth in a Random Forest model.

Table S3: ML classifiers used for the study

ML Classifiers	○ Logistic Regression ○ Random Forest ○ Gradient Boosting ○ SVM ○ AdaBoost ○ Naïve Bayes ○ K-Nearest Neighbors ○ Decision Tree ○ Extra Trees ○ Linear Discriminant Analysis ○ Quadratic Discriminant Analysis ○ LASSO ○ Boosted GLM
-----------------------	---

Table S4: METHodological RadiomICs Score (METRICS)

Categories	No.	Items	Weights	Answers
Study Design	#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.0368	YES
	#2	Eligibility criteria that describe a representative study population	0.0735	YES
	#3	High-quality reference standard with a clear definition	0.0919	YES
Imaging Data	#4	Multi-center	0.0438	YES
	#5	Clinical translatability of the imaging data source for radiomics analysis	0.0292	YES
	#6	Imaging protocol with acquisition parameters	0.0438	YES
	#7	The interval between imaging used and reference standard	0.0292	NO
Segmentation ¹	#8	Transparent description of segmentation methodology	0.0337	NO
	#9	Formal evaluation of fully automated segmentation	0.0225	N/A
	#10	Test set segmentation masks produced by a single reader or automated tool	0.0112	NO
Image Processing and Feature Extraction	#11	Appropriate use of image preprocessing techniques with transparent description	0.0622	YES
	#12	Use of standardized feature extraction software	0.0311	YES
	#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.0415	YES
Feature Processing	#14	Removal of non-robust features	0.0200	YES
	#15	Removal of redundant features	0.0200	YES
	#16	Appropriateness of dimensionality compared to data size ⁴	0.0300	YES
	#17	Robustness assessment of end-to-end deep learning pipelines ⁵	0.0200	N/A
Preparation for Modeling	#18	Proper data partitioning process	0.0599	YES
	#19	Handling of confounding factors	0.0300	YES
Metrics and Comparison	#20	Use of appropriate performance evaluation metrics for task	0.0352	YES
	#21	Consideration of uncertainty	0.0234	YES
	#22	Calibration assessment	0.0176	NO
	#23	Use of uni-parametric imaging or proof of its inferiority	0.0117	YES
	#24	Comparison with a non-radiomic approach or proof of added clinical value	0.0293	NO
	#25	Comparison with simple or classical statistical models	0.0176	YES
Testing	#26	Internal testing	0.0375	YES
	#27	External testing	0.0749	YES
Open Science	#28	Data availability	0.0075	NO
	#29	Code availability	0.0075	NO
	#30	Model availability	0.0075	NO
Total METRICS score				81.4%
Quality category*				EXCELLENT
0≤score<20%, “very low”; 20≤score<40%, “low”; 40≤score<60%, “moderate”; 60≤score<80%, “good”; 80≤score≤100%, “excellent” quality				

Table S5: CLEAR Checklist

Section	No.	Item	Yes	No	n/a	Page
Title						
	1	Relevant title, specifying the radiomic methodology Indicate the use of radiomics in the title. The following details can also be considered to be specified in the title: radiomic technique (e.g., hand-crafted, engineered, deep, delta, etc.), modality (e.g., computed tomography [CT], magnetic resonance imaging [MRI], ultrasound), important aspects of the scans (e.g., unenhanced, dynamic), use of machine learning (e.g., machine learning-based), external validation, and multi-center design.	☒	☐	☐	1
Abstract						
	2	Structured summary with relevant information Provide a structured summary of the purpose, methods, results, and conclusions, presenting only the most important aspects directly related to the purpose of the study. The abstract should be understandable on its own, without reading the main text. Considering the submission guidelines of the journals, it is recommended to specify the following items: the baseline characteristics (e.g., number of patients, scans, images, classes), data source (e.g., public, institutional), study nature (e.g., prospective, retrospective), segmentation technique (e.g., automated, semi-automated, or manual), feature extraction technique (e.g., hand-crafted, engineered, deep), dimensionality reduction techniques (e.g., feature selection, reproducibility analysis, multi-collinearity), modeling details (e.g., algorithms/models), validation technique (e.g., cross-validation), unseen testing (internal hold-out, external testing), model performance metrics (e.g., the area under the curve) with uncertainty measures (e.g., confidence intervals), number of the final set of features, traditional statistical methods with p-values, and open science status (e.g., public availability of data, code, and/or model).	☒	☐	☐	1
Keywords						
	3	Relevant keywords for radiomics List the primary keywords that indicate (e.g., radiomics, texture analysis) and characterize a radiomic study (e.g., machine learning, deep learning, computed tomography, magnetic resonance imaging, reproducibility), unless the journal requires exclusive use of certain terms (e.g., MeSH terms, which do not yet include radiomics-specific terms).	☒	☐	☐	1
Introduction						
	4	Scientific or clinical background Define the scientific or clinical problem with a summary of the related literature and knowledge gaps, including a short review of the current state of knowledge. Describe why the scientific question is technically or clinically important.	☒	☐	☐	2-3
	5	Rationale for using a radiomic approach Describe why a radiomic approach is considered. Performance and problematic aspects of currently used methods need to be described. Mention what the radiomics approach would offer to solve these problems. Clearly state how radiomics could affect clinical practice considering the study question.	☒	☐	☐	2-3
	6	Study objective(s) Describe the purpose of the study while focusing on the scientific problem. Mention the expected contributions to the current literature.	☒	☐	☐	2-3

Method						
<i>Study Design</i>	7	Adherence to guidelines or checklists (e.g., CLEAR checklist) Indicate that the CLEAR checklist was used for reporting and submit the checklist as supplemental data. Do the same with other checklists or guidelines if used in addition to the CLEAR checklist.	☒	☐	☐	7
	8	Ethical details (e.g., approval, consent, data protection) Describe the ethical questions to ensure that the study was conducted appropriately. Give information about ethical approval, informed consent, and data protection (e.g., de-identification) if the data is from private sources.	☒	☐	☐	22
	9	Sample size calculation Describe how the sample size or power was determined before or after the study (e.g., sample size/power calculation, based on availability).	☒	☐	☐	3-4
	10	Study nature (e.g., retrospective, prospective) Indicate whether the study is prospective or retrospective and case/control or cohort, etc. In the case of prospective studies, provide registration details if available.	☒	☐	☐	3
	11	Eligibility criteria Define the inclusion criteria first. Then, specify the exclusion criteria. Avoid redundancies by using the opposite of the inclusion criteria as exclusion criteria. Specify the selection process (e.g., random, consecutive). Keep the numeric details of eligibility for the results.	☒	☐	☐	4
	12	Flowchart for technical pipeline Provide a flowchart for summarizing the key methodological steps in the study. Due to the complex nature of the radiomic approaches, such flowcharts help readers better understand the methodology.	☒	☐	☐	7
<i>Data</i>	13	Data source (e.g., private, public) State the data source (e.g., private, public, mixed [both private and public]). State clearly which data source is used in different data partitions. Provide web links and references if the source is public. Give the image or patient identifiers as a supplement if public data is used.	☒	☐	☐	3-4
	14	Data overlap State if any part of the dataset was used in a previous publication. Describe the differences between the current study and previous studies in terms of study purpose and methodology.	☐	☒	☐	-
	15	Data split methodology Describe the data split into training, validation, and test sets. Mention that multiple splits are created (e.g., k-fold cross-validation or bootstrapping). Specify how the assignment was done (e.g., random, semi-random, manual, center-wise, chronological order). Indicate the ratio of each partition, with class proportions. Describe at which level the data is split (e.g., patient-wise, image-wise, study-wise, scanner-wise, institution-wise). Clearly state the measures undertaken to avoid information leakage across datasets (e.g., creating the hold-out test set before feature normalization, feature selection, hyperparameter optimization, and model training) [23]. Note that any test data should only be used once for evaluation of the final model to prevent optimistic biases. Declare the systematic differences among the data partitions.	☒	☐	☐	6-7

	16	Imaging protocol (i.e., image acquisition and processing) Provide the imaging protocol and acquisition parameters with post-processing details. Define physical pixel and voxel dimensions. Clearly state whether single or multiple or various scanners are used, with the number of instances for each protocol. Define the timing of the phase if a contrast medium was used. State the patient preparation (drug administration, blood sugar control before the scans, etc.) if performed.	☐	☒	☐	-
	17	Definition of non-radiomic predictor variables Describe the data elements appearing as non-radiomic predictors. Non-radiomic variables might be demographic characteristics (e.g., age, gender, ethnicity), widely used traditional laboratory biomarkers (e.g., carcinoembryonic antigen), or traditional approaches used in daily clinical practice (e.g., radiologist's qualitative reading, Hounsfield Unit evaluation, Response Evaluation Criteria in Solid Tumors [RECIST], Response Assessment in Neuro-Oncology [RANO] criteria). It would be helpful to know how these predictors were identified (e.g., based on a literature review). If applicable, describe any transformation of predictors (e.g., binarization of continuous predictors, the grouping of levels of categorical variables).	☒	☐	☐	2
	18	Definition of the reference standard (i.e., outcome variable) Describe the reference standard or outcome measure that the radiomic approach will predict (e.g., pathological grade, histopathological subtypes, genomic markers, local-regional control, survival, etc.). Provide the rationale for the choice of the reference standard (e.g., higher reproducibility rates). Clearly state the reproducibility concerns, potential biases, and limitations of the reference standard.	☒	☐	☐	3
<i>Segmentation</i>	19	Segmentation strategy Indicate which software programs or tools are used for segmentation or annotation. Specify the version of the software and the exact configuration parameters. Provide reference and web link to the software. Describe the segmentation method (e.g., automatic, semi-automatic, manual). Provide the rules of the segmentation (e.g., margin shrinkage or expansion from the visible contour, included/excluded regions). Provide figures to show the segmentation style. Provide image registration details (e.g., software, version, link, parameters) if segmentation is propagated for multi-modal (e.g., CT and MR), multi-phase (e.g., unenhanced, arterial, venous phase CT), or multi-sequence (e.g., T2-weighted, post-contrast T1-weighted, diffusion-weighted imaging) analyses. If radiomic features are extracted from 2D images on a single slice, please explain with which criteria the slice is chosen. In the case of several lesions, explain if all the lesions are segmented and describe how the feature values are aggregated. If only one lesion is chosen, describe the criteria (e.g., the primitive or the most voluminous).	☒	☐	☐	4
	20	Details of operators performing segmentation State how many readers performed the segmentation, as well as their experience. In the case of multiple readers, describe how the final form of segmentation is achieved (e.g., the consensus of readers, intersection of segmentations, independent segmentation for further reproducibility analysis, sequential refinements from numerous expert raters until convergence), which is particularly important for the training data because the segmentation process on the test data should be as close to the clinical practice as possible, that is, the segmentation of a single reader.	☒	☐	☐	4

<i>Pre-processing</i>	21	Image pre-processing details Indicate which software programs or tools are used for pre-processing. Specify the version of the software and the exact configuration parameters. Provide reference and web link to the software, if available. Describe all pre-processing techniques and associated parameters applied to the image including the normalization (e.g., minimum-maximum normalization, standardization, logarithmic transformation, bias field correction), de-noising, skull stripping (also known as brain extraction), interpolation to create uniform images (e.g., in terms of slice thickness), standardized uptake value conversion, and registration. Also, state if an image or feature-based harmonization technique was used.	☒	☐	☐	4-5
	22	Resampling method and its parameters Specify the resampling technique (e.g., linear, cubic b-spline) applied to the pixels or voxels. Provide the physical pixel and voxel dimensions after resampling.	☒	☐	☐	4
	23	Discretization method and its parameters Specify the discretization method (e.g., fixed bin width, fixed bin count method, or histogram equalization) used for hand-crafted radiomic feature extraction. Report the rationale for using a particular discretization technique. Indicate the number of grey levels for the fixed bin count method or the bin width as well as the value of the first level (or minimum and maximum bounds) for the fixed bin width method. Any experimental detail with different discretization methods and values is important to declare.	☒	☐	☐	4
	24	Image types (e.g., original, filtered, transformed) Provide the image types from which the radiomic features are extracted, e.g., original or images with convolutional filters (e.g., Laplacian of Gaussian edge enhancement, wavelet decomposition) [24]. Also, give nuances about the parameters of transformed image types (e.g., sigma values of Laplacian of Gaussian filtering).	☒	☐	☐	4-5
<i>Feature extraction</i>	25	Feature extraction method Indicate which software programs or tools are used for radiomic feature extraction. Specify the version of the software and the exact configuration parameters (also see Item#55). Provide reference and web link to the software. Indicate if the software adheres to the benchmarks/certification of IBSI [25]. Specify the general feature types, such as deep features, hand-crafted features, engineered features, or a combination. Refer to the mathematical formulas of the hand-crafted and engineered features. Provide formulas and code if new hand-crafted features are introduced. Present the architectural details for deep feature extraction. Provide details of any feature engineering performed. Specify whether radiomic features are extracted in a two-dimensional (2D) plane, 2D tri-planar, or three-dimensional (3D) space. If 2D features are extracted from 3D segmentation, provide reasons (e.g., large slice thickness) as to why such an approach is followed.	☒	☐	☐	4-5
	26	Feature classes Provide the radiomic feature classes (e.g., shape, first-order, grey-level co-occurrence matrix). Use IBSI terminology for feature classes [25]. Specify the number of features per feature class. Mention if any feature class is excluded with reason.	☒	☐	☐	4-5
	27	Number of features Indicate the total number of features per instance. If applicable, provide the number of features per imaging modality and its components (e.g., phase for CT, sequence for MRI, etc.).	☒	☐	☐	5

	28	Default configuration statement for remaining parameters After providing all modified parameters of pre-processing and radiomic feature extraction, state clearly that all other parameters remained as a default configuration.	☒	☐	☐	5
<i>Data preparation</i>	29	Handling of missing data State if, and how much, missing data are present in the study. If so, provide details as to how it was addressed (e.g., deletion, substitution, or imputation).	☒	☐	☐	4
	30	Details of class imbalance Indicate the balance status of the classes according to the reference standard. Provide details about how class imbalance is handled. Specify the techniques (e.g., synthetic minority over-sampling, simple over-sampling through replication, under-sampling) used to achieve the class balance. Clearly state these data augmentation and under-sampling strategies are applied only in the training set.	☒	☐	☐	4, 7
	31	Details of segmentation reliability analysis Describe the reliability analysis done to assess the influence of segmentation differences. An intra- and inter-rater reproducibility analysis must be considered in manual and semi-automatic methods. Provide details about the statistical tests used for the reliability analysis (e.g., intraclass correlation coefficient along with types) [26]. Mention the independence of assessment. Clearly state the reliability analysis is performed using the training set only.	☐	☒	☐	-
	32	Feature scaling details (e.g., normalization, standardization) If applicable, describe the normalization technique applied to the radiomic feature data (e.g., minimum-maximum normalization, standardization, logarithmic transformation, ComBat normalization [choice of the batch, parametric or not, with or without empirical Bayes]). Specify the normalization scale. It is important to emphasize that this procedure is applied to the numeric radiomic feature data, not the images, in the training set and independently applied to the validation and test sets.	☒	☐	☐	5
	33	Dimension reduction details Specify the dimension reduction methods used, if applicable (e.g., collinearity analysis, reproducibility analysis, algorithm-based feature selection). Provide details about the statistical methods used. For example, provide the relevant statistical cut-off values for each step (e.g., features with intraclass correlation coefficient ≤ 0.9 are excluded). Clearly state the dimension reduction that is performed using the training set. Specify how the final number of features is achieved, for instance, the “rule of thumb” of ten features maximum for each instance.	☒	☐	☐	5-6
<i>Modeling</i>	34	Algorithm details Provide the name and version of software programs or packages used for modeling. Refer to the related publication of the software if available. Specify the algorithms used to create models with architectural details including inputs, outputs, and all intermediate components. The description of the architecture should be complete to allow for exact replication by other investigators (also see Item#55 and Item#56). When a previously described architecture is used, refer to the previous work and specify any modification. If the final model involved an ensemble of algorithms, specify the type of ensemble (e.g., stacking, majority voting, averaging, etc.).	☒	☐	☐	4, 6

	35	Training and tuning details Describe the training process with adequate detail. Specify the augmentation technique, stopping criteria for training, hyperparameter tuning strategy (e.g., random, grid-search, Bayesian), range of hyperparameter values used in tuning, optimization techniques, regularization parameters, and initialization of model parameters (e.g., random, transfer learning). If transfer learning is applied, clearly state which layers or parameters are frozen or affected.	☒	☐	☐	6
	36	Handling of confounders Describe the method (e.g., directed acyclic graphs) for the detection of potential confounders (e.g., differences in tumor size between cohorts, different image acquisition parameters such as slice thickness, and differences in patient populations between primary and secondary hospitals) [27, 28]. Please describe how confounding was addressed (e.g., covariate adjustment).	☐	☒	☐	-
	37	Model selection strategy Describe how the final model was selected. Two broad categories for these are probabilistic (e.g., Akaike information criterion, Bayesian information criterion) and resampling methods (e.g., random train-test split, cross-validation, bootstrap validation) [12, 29]. Clearly state that only the training and validation sets are used for model selection. State if the model complexity is considered in selection, for instance, the “one standard error rule” [30]. Specify which performance metrics were used to select the final model.	☒	☐	☐	6
<i>Evaluation</i>	38	Testing technique (e.g., internal, external) Clearly state whether the model was internally or externally tested. The term “external testing” should only be used for the process that involves data usage from different institutions. In the case of external testing, specify the number of sites providing data and further details about whether they are used for multiple testing or in a single test. Describe the data characteristics and state if there are any differences among training, validation, internal testing, and external testing datasets (e.g., different scanners, different readers for segmentation, different ethnicity). Again, note that any test data should only be used once for evaluation to prevent biased performance metric estimates.	☒	☐	☐	6
	39	Performance metrics and rationale for choosing Specify the performance metrics to evaluate the predictive ability of the models. Justify the selected metrics according to the characteristics of the data (e.g., class imbalance). Beware of the potential pitfalls and follow recommendations when selecting the appropriate performance metrics [7, 31].	☒	☐	☐	6
	40	Uncertainty evaluation and measures (e.g., confidence intervals) Describe the uncertainty evaluation (e.g., robustness, sensitivity analysis, calibration analysis if applicable) and measures of uncertainty quantification (e.g., confidence intervals, standard deviation).	☒	☐	☐	7
	41	Statistical performance comparison (e.g., DeLong’s test) Specify the statistical software and version used. Indicate which method was used for the comparison of the model performance such as the DeLong’s test [32, 33], McNemar’s test [34], or Bayesian approaches [35]. Provide a statistical threshold for the comparison (e.g., $p < 0.05$) along with confidence intervals if applicable to the method or metric. Also, state if multiplicity is considered and corrected when comparing multiple models (e.g., p-value adjustment, Bonferroni correction, false-discovery rate). Report threshold values to stratify data into groups for statistical testing (e.g., the operating point on the receiver operating characteristic [ROC] curve to define the confusion matrix, and cut-off values for defining strata in survival analysis).	☒	☐	☐	7

	42	Comparison with non-radiomic and combined methods Indicate whether comparisons with non-radiomic approaches (e.g., clinical parameters, laboratory parameters, traditional radiological evaluations) are performed. Non-radiomic approaches can be combined with radiomic data as well (e.g., clinical-radiomic evaluation). Explain how the clinical utility is assessed, such as with decision curve analysis [36].	☒	☐	☐	8-10, 15-19
	43	Interpretability and explainability methods Describe the techniques used to increase the interpretability and explainability of the models created, if applicable [37]. Figures (e.g., class activation maps, feature maps, SHapley Additive exPlanations, accumulated local effects, partial dependence plots, etc.) related to the interpretability and explainability of the proposed radiomic model can be provided.	☒	☐	☐	13
Results						
	44	Baseline demographic and clinical characteristics Provide the baseline demographic, clinical, and imaging characteristics in text and/or tables. Report the information separately for training, validation (i.e., cross-validation), and test datasets, along with grouping based on the reference standard or non-radiomic variables. Associated statistical tests should also be provided to identify if the sets are identical or not. Provide whether any confounder is detected and handled appropriately.	☐	☒	☐	-
	45	Flowchart for eligibility criteria Provide a flowchart for summarizing eligibility criteria with the number of included and excluded instances. If more than one data source is involved, please give details for each source separately.	☒	☐	☐	7
	46	Feature statistics (e.g., reproducibility, feature selection) Give statistical information (e.g., distribution of features based on outcome variables) of the selected features for inclusion into the model. Provide the name and number of reproducible features (e.g., for segmentation reproducibility, for reproducibility against image perturbations). Create a table for the selected features with details of feature name, class, and image type. Also, provide results of reproducibility statistics. Reproducibility metrics of selected features can be presented in tables or supplementary files. Figures (e.g., boxplots, correlation matrix, feature importance plots) and tables of descriptive summaries of features can be provided.	☒	☐	☐	13
	47	Model performance evaluation Provide the performance metrics for training, validation (e.g., multiple splits like cross-validation, bootstrapping, etc.), and unseen test data, separately. A summary of the most important findings should be given in the text. Provide the 'no information rate' as well. Details can be provided in figures (e.g., ROC curves, precision-recall curves) and tables. It is a good practice to provide figures for calibration statistics to show the robustness of model performance. Present additional figures to showcase examples of true and false predictions to help readers better understand the strengths and limitations of the proposed strategy.	☒	☐	☐	8-10, 14

	48	Comparison with non-radiomic and combined approaches Give the results about the comparison of radiomic approaches with non-radiomic (e.g., visual analysis, clinical only parameters) or combined approaches in the text and preferably on a table. Present the results for training, validation, and test data, separately. Provide uncertainty measures (e.g., confidence intervals, standard deviation, etc.) and statistical comparison results with p-values for each. Confusion matrices must also be provided. Aside from the predictive performance, specify which model is superior to others in terms of clinical utility. The clinical utility can be presented with a decision curve analysis. For the decision curve analysis, quantify the net benefit according to optimal probability thresholds, with multiple cut-points associated with different clinical views. Also, provide the rationale for why a specific threshold could be appropriate and clearly state what is meant by all and none strategies.	☒	☐	☐	8-9
Discussion						
	49	Overview of important findings Provide a summary of the work and an overview of the most important findings. No statistical information is needed. Try to position the study into one of the following categories: proof of concept evaluation, technical task-specific evaluation, clinical evaluation, and post-deployment evaluation [38]. Summarize the contribution to the literature.	☒	☐	☐	20-21
	50	Previous works with differences from the current study Provide the most important and relevant previous works. Mention the most prominent differences between the current study and the previous works.	☒	☐	☐	20-21
	51	Practical implications Summarize the practical implications of the results. Describe the key impact of the work on the field. Highlight the potential clinical value and role of the study. Discuss any issues that may hamper the successful translation of the study into real-world clinical practice. Also, provide future expectations and possible next steps that others might build upon the current work.	☒	☐	☐	21
	52	Strengths and limitations (e.g., bias and generalizability issues) Clearly state the strengths and the limitations of the current work. Any issue that may lead to potential bias, uncertainty, reproducibility, robustness, and generalizability problems should be declared.	☒	☐	☐	21
Open Science						
<i>Data availability</i>	53	Sharing images along with segmentation data [Please note this item is “not essential” but “recommended”] Provide relevant raw or processed image data considering the regulatory constraints of the institutions involved. Segmentation data can also be shared unless the segmentation is done as part of the workflow. In situations where sharing of the entire dataset is not possible, an end-to-end analysis workflow applied to a representative sample, or a public dataset with similar characteristics can facilitate the ability of the readers in reproducing key components of the analysis [39]. Also, specify the reason if the data is not available.	☐	☐	☒	Data privacy constraints
	54	Sharing radiomic feature data Share selected radiomic feature data along with clinical variables or labels with the public, if possible (i.e., in accordance with the regulatory constraints of the institute). Specify the reason if the radiomic feature data is not available.	☐	☐	☒	Data privacy constraints

<i>Code availability</i>	55	Sharing pre-processing scripts or settings Share the pre-processing and feature extraction parameter scripts or settings (e.g., YAML file in PyRadiomics or complete textual description). If it is not available in a script format, then the parameter configuration as appeared in the software program can be shared as a screenshot.	☒	☐	☐	Supp. Table S1
	56	Sharing source code for modeling Share the modeling scripts [40]. Code scripts should include sufficient information to replicate the presented analysis (e.g., to train and test pipeline), with all dependencies and relevant comments to easily understand and build upon the method. Even if the actual input dataset used cannot be shared, in situations where a similar dataset is available publicly, it should be used to share an example workflow with all pre- and post-processing steps included. Specify the reason in case the source code is not available.	☐	☒	☐	Available upon request
<i>Model availability</i>	57	Sharing final model files Share the final model files for internal or external testing [40]. Describe how inputs should be prepared to use the model. Also, include the source code that was used for pre-processing the input data. Specify the reason in case the final model data is not available.	☐	☒	☐	Available upon request
	58	Sharing a ready-to-use system [Please note this item is “not essential” but “recommended”] An easy-to-use tool (e.g., standalone executable applications, notebooks, websites, virtual machines, etc.) can be created and shared with or without source code that is based on the model created [40]. The main aim is to be able to test or validate the model by other research groups. With this approach, users even without experience in machine learning or coding can also test the proposed models.	☐	☒	☐	-

Yes, details provided; No, details not provided; n/a, not applicable

Table S6: TRIPOD Checklist

Section/Topic	Item Page	Checklist Item	
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	2-3
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	3
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	3-4
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	3-4
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	3-4
	5b	Describe eligibility criteria for participants.	3
	5c	Give details of treatments received, if relevant.	-

Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	3
	6b	Report any actions to blind assessment of the outcome to be predicted.	3
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	4-5
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	-
Sample size	8	Explain how the study size was arrived at.	3
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	-
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	5-7
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	6-7
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	7
Risk groups	11	Provide details on how risk groups were created, if done.	5
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	-
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	8
Model development	14a	Specify the number of participants and outcome events in each analysis.	3
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	-
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	9-10
	15b	Explain how to use the prediction model.	9
Model performance	16	Report performance measures (with CIs) for the prediction model.	9-10
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	22-23
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	21-23
Implications	20	Discuss the potential clinical use of the model and implications for future research.	21
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	23
Funding	22	Give the source of funding and the role of the funders for the present study.	24

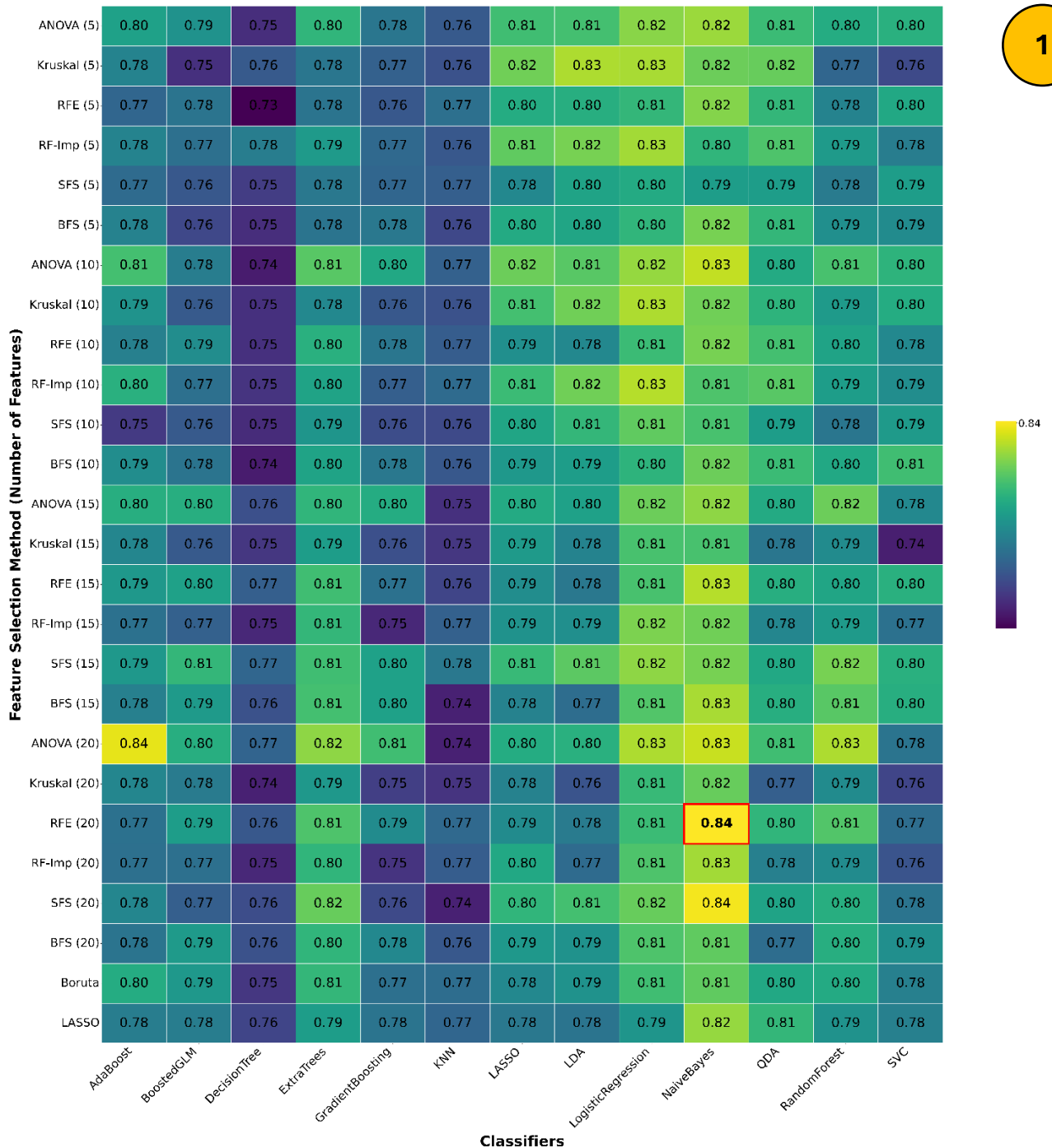
Table S7: Optimal hyperparameters for the best-performing models in Scenario 1

Model	Feature Selection	Classifier	Opt. Hyperparameters
Radiomics (No ComBat)	RFE_20	Naïve Bayes	No hyperparameter tuning applied
Radiomics–ADC ratio (No ComBat)	LASSO	Naïve Bayes	No hyperparameter tuning applied
Radiomics (ComBat)	BFS_20	Random Forest	max_depth=10, n_estimators=200
Radiomics–ADC ratio (ComBat)	KW_20	Ada Boost	n_estimators=100

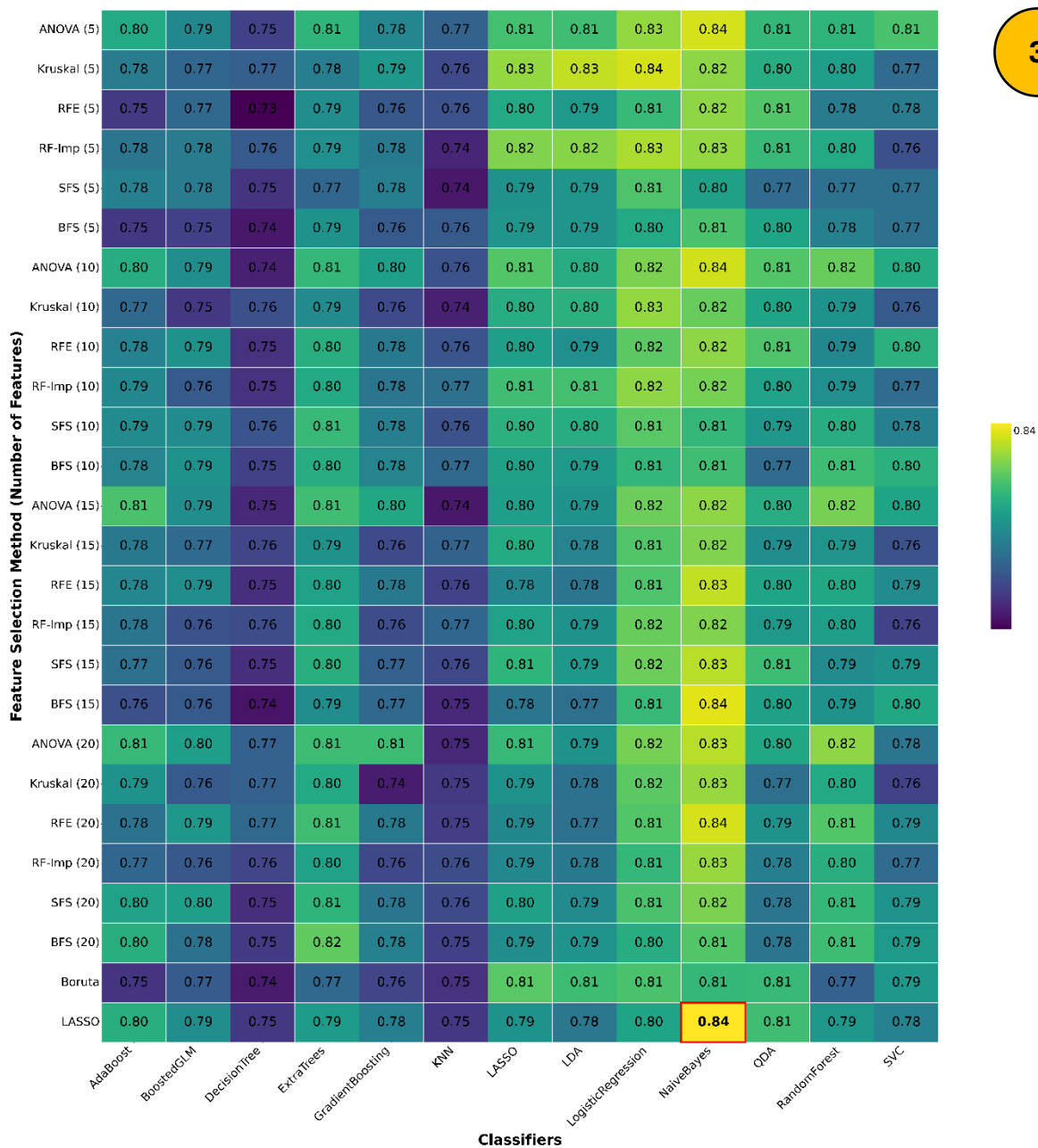
Table S8: Optimal hyperparameters for the best-performing models in Scenario 2

Model	Feature Selection	Classifier	Opt. Hyperparameters
Radiomics (No ComBat)	RFE_20	Random Forest	max_depth=10, n_estimators=200
Radiomics–ADC ratio (No ComBat)	LASSO	Boosted GLM	learning_rate=0.1, n_estimators=100, base_estimator_penalty=l2
Radiomics (ComBat)	BFS_20	Extra Trees	n_estimators=200
Radiomics–ADC ratio (ComBat)	KW_20	Boosted GLM	learning_rate=0.1, n_estimators=100, base_estimator_penalty=l2

Figure S1: Mean cross-validation performance in Scenario 1 across feature selection methods, classifiers, feature counts, and harmonization settings (1: radiomics-model without ComBat, 2: radiomics-model with ComBat, 3: radiomics-ADC ratio without ComBat, 4: radiomics-ADC ratio with ComBat). Red boxes denote the top-performing feature selection–classifier combinations per metric.







3

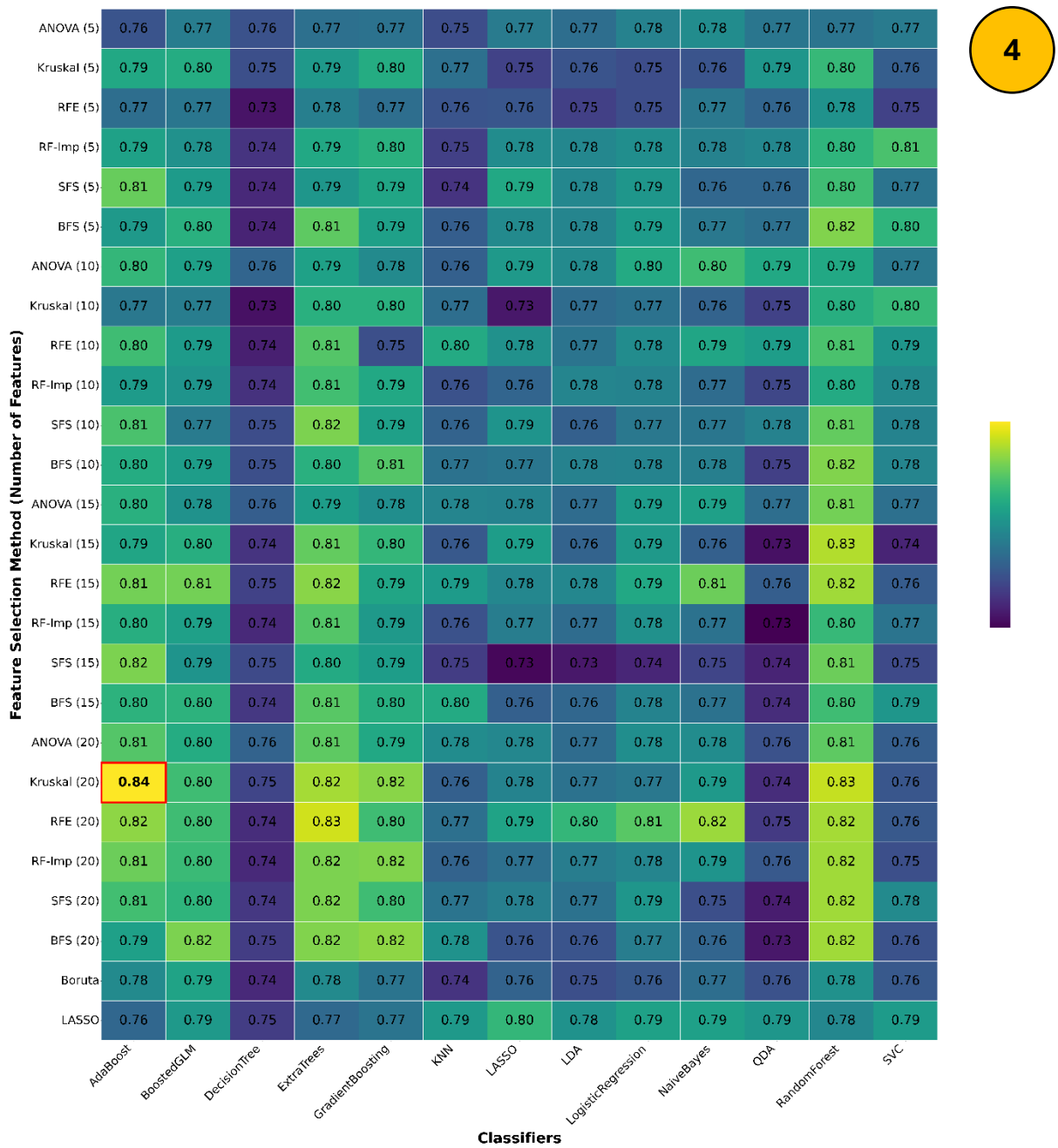


Figure S2: Mean performance in Scenario 2 across feature selection methods, classifiers, feature counts, and harmonization settings (1: radiomics-model without ComBat, 2: radiomics-model with ComBat, 3: radiomics-ADC ratio without ComBat, 4: radiomics-ADC ratio with ComBat). Red boxes denote the top-performing feature selection–classifier combinations per metric.

