A radiomics-based machine learning pipeline to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest

Hossein Naseri (hossein.naseri@mail.mcgill.ca)
McGill University

Sonia Skamene
McGill University Health Center (MUHC)

Marwan Tolba
McGill University Health Center (MUHC)

Mame Daro Faye
McGill University Health Center (MUHC)

Paul Ramia
McGill University Health Center (MUHC)

Julia Khriguian
McGill University Health Center (MUHC)

Haley Patrick
McGill University

Aixa X Andrade
University of Texas Southwestern Medical Center

Marc David
McGill University Health Center (MUHC)

John Kildea
McGill University

Article

Keywords:

Posted Date: March 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1446196/v1

A radiomics-based machine learning pipeline to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest

Hossein Naseri[1,*], Sonia Skamene[2], Marwan Tolba[2], Mame Daro Faye[2], Paul Ramia[2], Julia Khriguian[2], Haley Patrick[1], Aixa X. Andrade H.[3], Marc David[2], and John Kildea[1]

[1]Medical Physics Unit, McGill University, Montreal, QC, Canada
[2]Department of Radiation Oncology, McGill University Health Center (MUHC), Montreal, QC, Canada
[3]University of Texas Southwestern Medical Center, Dallas, TX, USA
[*]Corresponding author H.N. (hossein.naseri@mail.mcgill.ca)

ABSTRACT

Radiomics-based machine learning classifiers have shown potential for detecting bone metastases (BM) and for evaluating BM response to radiotherapy (RT). However, current radiomics pipelines require large datasets of images with expert-segmented 3D regions of interest (ROIs). Full ROI segmentation is time consuming and oncologists often outline just RT treatment fields in clinical practice. This presents a challenge for real-world radiomics research. As such, a method that simplifies BM identification but does not compromise the power of radiomics is needed.

The objective of this study was to investigate the feasibility of a radiomics pipeline for BM detection using lesion-center-based geometric ROIs. The simulation-CT images of 170 patients with non-metastatic lung cancer and 189 patients with spinal BM were used. The point locations of 631 BM and 674 healthy bone (HB) regions were identified by experts. ROIs with various geometric shapes were centered and automatically delineated on the identified locations, and 107 radiomics features were extracted. Various re-sampling techniques, feature selection methods, and machine learning classifiers were evaluated.

Our point-based radiomics pipeline was successful in differentiating BM from HB. This approach greatly simplifies the process of preparing images for use in radiomics studies and avoids the bottleneck of full ROI segmentation.

Introduction

In recent years, radiomics-based machine learning (ML) classifiers have shown great potential for use in the early detection of bone metastases (BM) and in assessing response of BM to radiotherapy (RT)[1–20]. However, in order to be clinically acceptable, radiomics models must be trained on large data sets of real-world images. This is challenging as the full 3D segmentation of BM on planning-CT images is time-consuming for radiation oncologists in the clinical context. Often, in the interest of time and given the low doses used in palliative RT, radiation oncologists only delineate treatment field boundaries when treating BM, and they do not fully contour individual BM lesions. As a result, most of the published BM radiomics studies to date were trained and tested with relatively small sample sizes (see Table 1), which diminishes their generalizability and their applicability to clinical RT planning. Motivated by the need for large real-world BM data sets, the objective of this study was to determine if a radiomics model can be trained to distinguish BM from healthy bone (HB) using BM lesions denoted as points on planning-CT images rather than using full 3D segmentation.

Radiomics for BM detection

Radiomics is an automated feature generation method for the extraction of hundreds of quantitative features (radiomics features) from radiology images[21,22]. ML algorithms can be trained to find relationships between radiomics features and cancer outcomes if provided with sufficient and appropriate data. There are three main steps in the training phase of a typical radiomics study. These include: (1) manual or semi-automated segmentation of regions of interest (ROIs) on patients' images, (2) feature extraction from the segmented ROIs, and (3) generation of a statistical or ML model to correlate extracted features to each patient's endpoint data such as their cancer outcome or other clinically-measured biomarkers[8].

In addition to the need for adequate sample sizes, which is the main motivation behind this study, a radiomics model must

overcome two important challenges in order to be reliable in a clinical context. First, it must be clinically reproducible. This is challenging because different radiomics studies use different subsets of radiomics features to achieve optimal models. The variations in published feature selection approaches make radiomics models less clinically reproducible[23,24]. Therefore, to achieve a clinically-reliable radiomics model, it is important to study and account for the effect of the variation in feature selection (FS) methods[25–27]. Second, working with imbalanced data, which is a common property of many real-world clinical data sets, requires a comprehensive study of re-sampling (RS) techniques to adequately account for the effect of sample imbalance when building high-performance radiomics-based ML models[9,28,29].

Depending on the endpoint of interest, various ML classifiers may be used in a radiomics pipeline. Support vector machine (SVM), Bayesian network (BN), multivariate logistic regression (MLR), k-nearest neighbor (kNN), decision trees (DT), random forests (RF), neural network (NNet), and convolutional neural networks (CNN) are among the ML classifiers that are most commonly used in radiomics-based ML pipelines[8–20]. The feasibility of using radiomics-based ML pipelines to distinguish between benign and malignant bone lesions has been reported in previous studies[1–4,6,7]. The main details of these studies are summarized in Table 1.

Author year	Sample size*	Imaging Modality	ROI	labels	Classifier	Performance (AUC, A, P, R)†
Perk et al.[1]	36	PET/CT	Manual	benign and metastatic	RF	0.95, 0.88, 0.88, 0.89
Suhas and Mishra[2]	74	Diagnostic-CT	Semi-automated‡	benign and malignant	RF	0.90, 0.92, 0.92, 0.91
Acar et al.[3]	75	PET/CT	Manual	responded and metastatic	kNN	0.76, 0.74, 0.74, 0.74
Suhas and Kumar[4]	100	Diagnostic-CT	Semi-automated‡	benign and malignant	SVM	0.86, -, 0.85, 0.88
Homayounieh et al.[5]	103	Dual-Energy CT	Semi-automated‡	benign and malignant	RF	0.79, 0.78, 0.72, 0.79
Hong et al.[6]	177	CT	Manual	bone island and metastases	RF	0.96, 0.80, 0.96, 0.86
Sun et al.[7]	206	Diagnostic-CT	Manual	benign and malignant	MLR	0.82, 0.86, 0.93, 0.77

Table 1. Radiomics-based ML pipelines reported in the literature for distinguishing bone lesions. *Sample size is the total number of samples. †A: Accuracy, P: precision (Specificity), R: Recall (sensitivity). ‡In the semi-automated segmentation methods an expert was required to check and modify the computer-segmented ROIs slice-by-slice.

The radiomics-based ML pipelines listed in Table 1 are not readily applicable to our clinical context, palliative RT for BM, for three reasons. First, they have relatively small sample sizes, an inherent problem for generalizability. Second, they require full 3D lesion segmentation, which is challenging to achieve clinically when planning palliative RT for BM. Finally, they were trained on images acquired using diagnostic-CTs or hybrid imaging modalities, whereas palliative RT planning is mostly done on planning-CT (simulation-CT) images alone.

With the above limitations in mind, in this study, we developed a fast and reliable radiomics-based ML pipeline capable of differentiating between BM and HB in RT planning-CT images of cancer patients using just geometric ROIs centered on expert-identified lesion point locations. We investigated the effect of using ROIs with different sizes and geometric shapes. We also examined the performance of different RS and FS methods and ML classifiers in achieving the optimal BM detection pipeline.

Materials and Methods

Ethics declarations

This retrospective study was approved by the Research Ethics Board of the McGill University Health Centre, Montreal, Quebec, Canada, with the waiver of informed consent. We confirmed that all research were performed in accordance with the relevant guidelines and regulations.

Patient selection

The planning-CT images of BM and HB patients used in this study were collected from the Oncology Information System at our institution. Our patient selection procedure is presented in Figure 1.

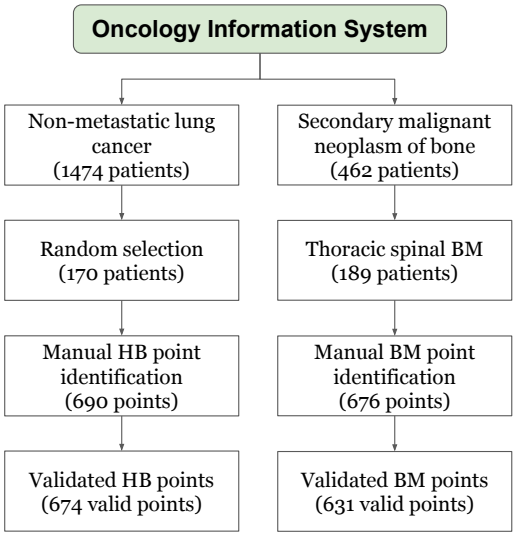

Figure 1. Flow chart of patient selection.

BM samples were from patients who received palliative RT for a secondary malignant neoplasm of bone in the thoracic spine between January 2016 and September 2019. HB samples were from individuals who received curative RT for non-metastatic lung cancer (as their CT images covered the same anatomy) during this period.

In total, we found 189 BM patients (96 male and 93 female; mean age 69 $\pm$ 13 y) and 1474 HB patients in our database. To reduce the large imbalance between the number of BM and HB patients, we randomly shuffled the HB sample (in a Microsoft Excel file) and selected the first 170 patients (86 male and 84 female; mean age 71 $\pm$ 12 y) to include in our study (see Figure 1). We purposely keep a slight imbalance to investigate the effect of various re-sampling (RS) techniques, as described below.

Planning-CT images

All planning-CT images were generated using one of three Philips' Brilliance Big Bore RT CT scanners at our institution with the acquisition parameters provided in Table 2. Planning-CT DICOM files were manually de-identified and exported to a secured hard drive from the Eclipse radiation therapy treatment planning software (Varian Medical Systems, Palo Alto, California), into which they had been previously imported for RT planning.

Tube voltage (kV)	Tube current (mA)	exposure (mm)	Field of view (pixel)	Matrix size (mm)	Slice thickness (mm)	Pixel spacing
120	165-366	240-450	600	512×512	3.0	0.77-1.37

Table 2. Planning-CT image acquisition parameters.

Lesion identification

The planning-CT images of the BM patients were randomly divided into five sets using the Python `random.shuffle` module and were loaded into our custom-written 3D DICOM visualization web application (diCOMBINE[30]) for lesion identifying. diCOMBINE is an open-source software developed by our group using the Python Flask[31] framework for DICOM 3D visualization and lesion point location labeling. The center points of BM lesions were labeled by an expert team comprising one staff radiation oncologist and four radiation oncology fellows. Each expert was asked to label BM center points in one of the five data sets, and a peer expert was tasked with reviewing them and validating the labels. A total of 631 validated BM center points were thus identified in the BM data set. Similarly, the planning-CT images of the HB patients were randomly divided into three sets and were loaded into diCOMBINE for HB labeling. One staff medical physicist and two medical physics graduate students were asked to identify HB points in one of the data sets each. When identifying HB points, the medical physicists were instructed to avoid non-metastatic skeletal complications (such as surgically-treated bone lesions). An average of four HB points were identified in each planning-CT image. Then, we asked each physicist to independently review and confirm the HB points that one of their peers had labeled. A total of 674 validated HB points were identified in this way. Screenshots of our diCOMBINE 3D lesion labeling web application are presented in Figure 2. These BM and HB points were used as center points for our automated ROI delineation.

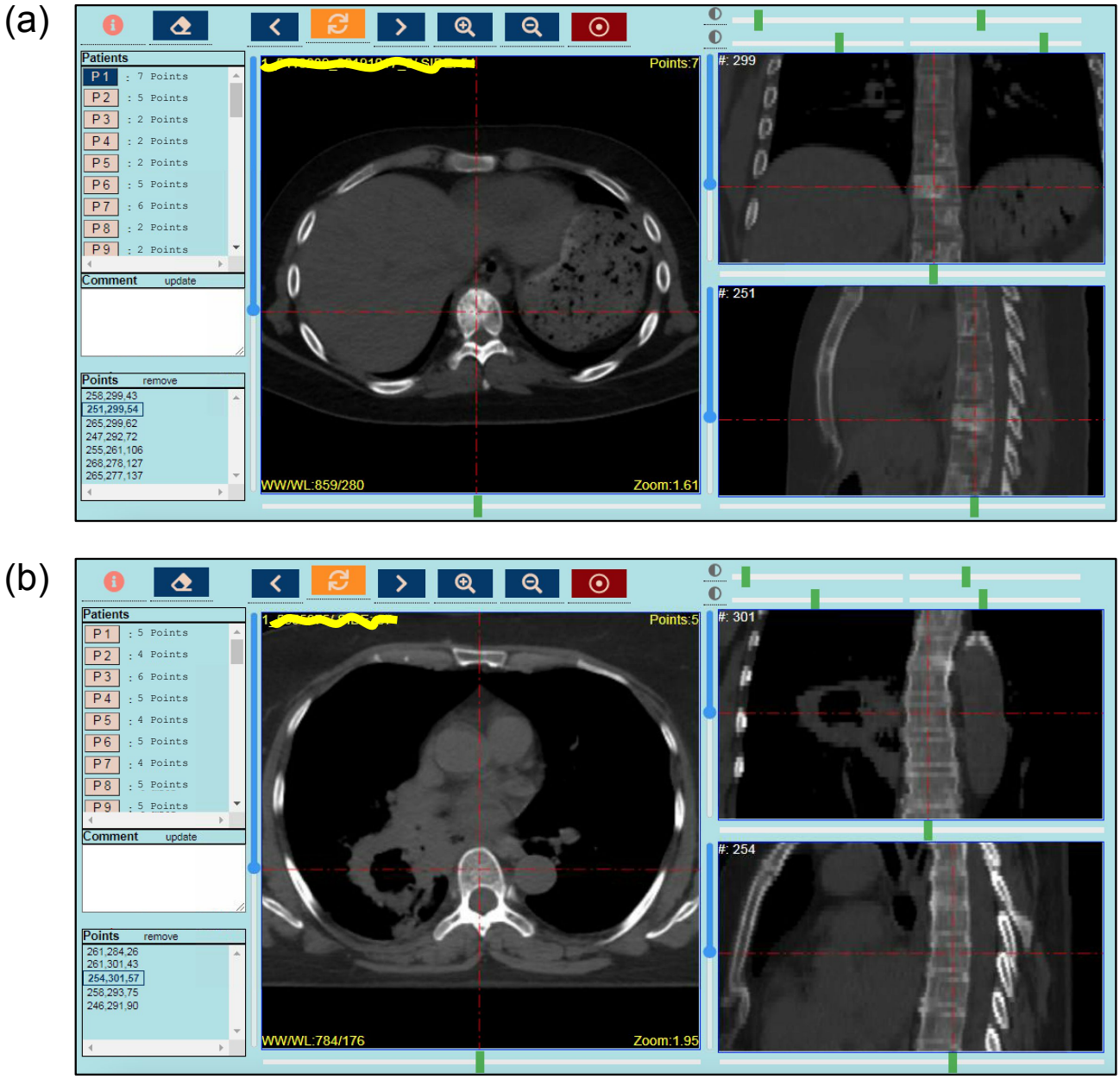

Figure 2. Screenshots of our diCOMBINE 3D lesion labeling web application showing expert-labeled points. (a) A BM lesion, and (b) a HB point.

ROI Delineation

ROIs were automatically delineated in the planning-CT images using geometric shapes centered on the expert-identified point locations. We used four spherical (SP) and five cylindrical along the z-axis (CY) ROIs of various sizes. The characteristics of the ROIs used are specified in Table 3. The size ranges were defined to extend from the size of a large bone lesion (∼15 mm)[32] to the maximum size of a spinal vertebra (∼50 mm)[33,34].

Radiomics feature extraction

The open-source PyRadiomics package (version 3.0.1)[35] was used to calculate the 3D quantitative radiomics features. For each of the nine ROIs listed in Table 3, we extracted 107 radiomics features from each of the planning-CT images. We did not apply

	Spherical				
Abbreviation	SP50	SP30	SP20	SP15	
Diameter (mm)	50	30	20	15	
	Cylindrical along the z-axis				
Abbreviation	CY50	CY30	CY20	CY15	CY5030
Width (mm) × Height (mm)	50×50	30×30	20×20	15×15	50×30
	Ensemble				
Abbreviation	E4SP	E4CY	E5CY	E9SC	
ROIs	SP50 +SP30 +SP20 +SP15	CY50 +CY30 +CY20 +CY15	CY50 +CY30 +CY20 +CY15 +CY5030	SP50+SP30 +SP20+SP15 +CY50+CY30 +CY20+CY15 +CY5030	

Table 3. The characteristics of the ROIs used in this study. ROIs from the planning-CT images were segmented using cylindrical and spherical ROIs with various sizes around the expert-labeled BM and HB points.

any filters prior to feature extraction. These 107 features include 18 First Order, 14 Shape, 24 Gray Level Co-occurrence Matrix (GLCM), 16 Gray Level Size Zone Matrix (GLSZM), 16 Gray Level Run Length Matrix (GLRLM), 14 Gray Level Dependence Matrix (GLDM), and five Neighbouring Gray Tone Difference Matrix (NGTDM) features[36,37]. We also aggregated radiomics features from multiple ROIs to define four ensemble ROIs, including; 1) E4SP: 428 features extracted from all four spherical ROIs, 2) E4CY: 428 features extracted from the first four cylindrical ROIs, 3) E5CY: 535 features extracted from all five cylindrical ROIs, and 4) E9SC: 963 features extracted from all nine ROIs combined. Our rationale for this approach was that by aggregating features extracted from ROIs with various sizes around the BM centers, we could extract sufficient information about the BMs' shape, size, and other characteristics and distinguish them from HBs using ML classifiers. Similar feature aggregation approaches were used in other studies[38,39].

Pipeline development

Our complete radiomics-based ML pipeline is presented in Figure 3. After extracting radiomics features for each ROI, we normalized the feature space using z-score normalization[40]. Then, we randomly divided the data set into 70% and 30% stratified training and testing sets, respectively. Each stratified set contained approximately the same BM/HB samples ratio as the initial data set. The training set was used for RS, FS, and ML pipeline development using 5-fold cross-validation[41]. The test set was used for the final performance evaluation. No RS was performed on the test set. In the present study, we examined the performance of four RS techniques, 13 FS methods, and 12 ML classifiers as shown in Figure 3 and described in the following sections.

RS techniques

Most of the ML classifiers are designed to work with balanced data sets with the same number of samples in each class[42,43]. However, having an imbalanced sample ratio is common in many real-world radiation oncology outcome data sets[29,44]. One tactic to tackle an imbalanced data set issue is to implement RS techniques[45]. In this study, we tested our pipeline with and without using RS techniques. For RS, we examined four techniques, including the random oversampling (ROS), random undersampling (RUS), Synthetic Minority Oversampling Technique[46] (SMOTE), and Tomek Links[47,48] (TL) algorithms.

FS methods

Radiomics calculates hundreds of features from images and some of them are redundant or are not useful for detecting BM[49]. To identify the most useful radiomics features for differentiating BM and HB, we investigated several supervised and unsupervised FS methods, including principal component analysis[50] (PCA), fast independent component analysis[51] (Fast ICA), zero variance threshold[52] (VT_0), near-zero variance threshold[52], least absolute shrinkage and selection operator[53] (LASSO) logistic regression algorithm, recursive feature elimination with cross-validation[54] (RFECV), and decision-tree-based feature selection[55] (TREE). For the PCA, motivated by Zack et al.[9] we used 20, 24, and 30 features. For the LASSO method, we examined least-squares penalty (α) values of 0.1, 0.5 and 1.0. α controls the stability of the selected features. A LASSO method with a larger α keeps fewer features (the most stable ones)[53]. For near-zero variance, we selected the variance threshold

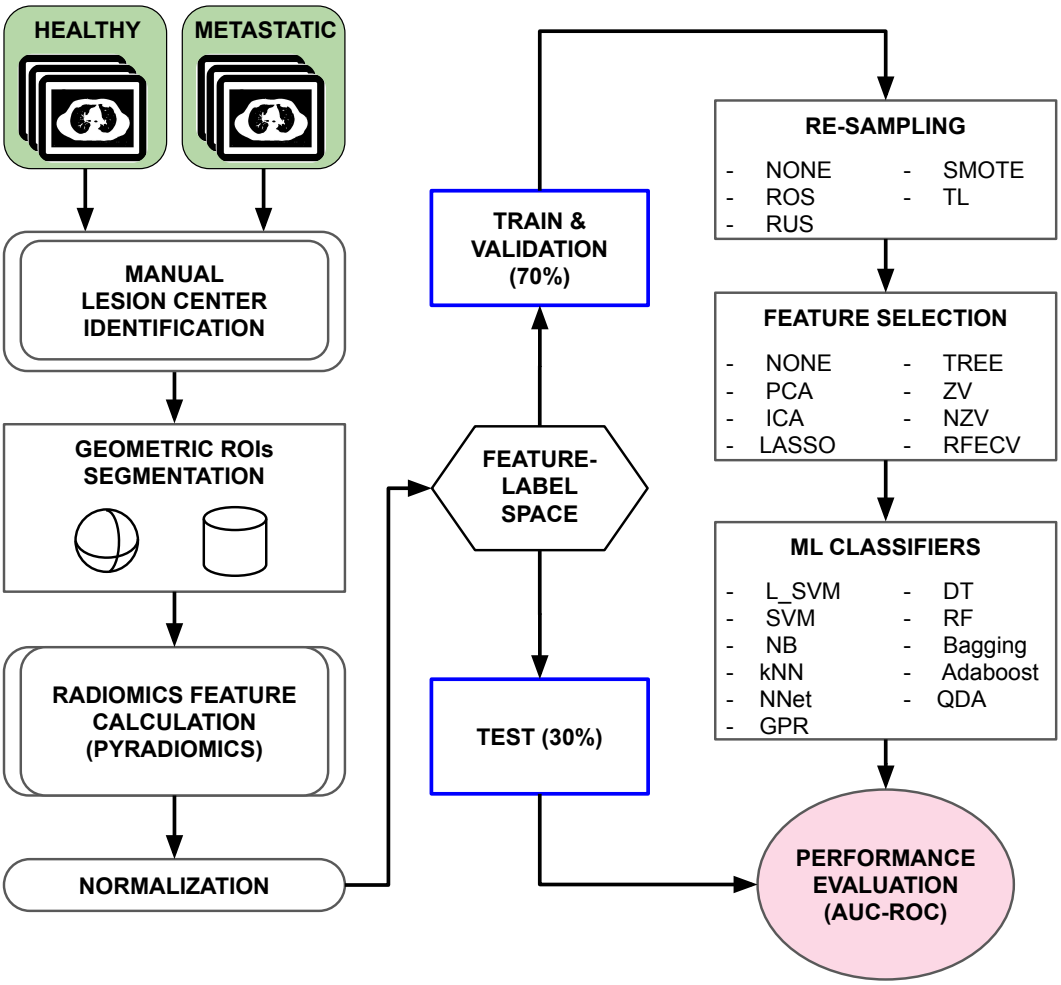

Figure 3. Our radiomics-based ML pipeline for classifying metastatic (BM) and healthy (HB) spinal bones.

of 0.8 (VT_0.8) as used by Zack et al.[9]. The performance of these FS methods, along with no FS, was then evaluated using 12 supervised ML classifiers.

ML classifiers

The Python scikit-learn ML package (version 0.20.4)[56] was used to implement our ML classifiers. We used 12 supervised classification models, including the linear support vector machine[57] (L-SVM), SVM with Radial-basis function kernel[57] (SVM), Gaussian Naive Bayes[58] (NB), K-Nearest Neighbors[59] (kNN), Quadratic Discriminant Analysis[60] (QDA), Gaussian Process Regression[61] (GPR), Decision Tree[62] (DT), Random Forest[63] (RF), Bagging[63], AdaBoost[63], Neural network with stochastic gradient-based solver[64,65] (NNet) and NNet with Limited-memory Broyden–Fletcher–Goldfarb–Shanno solver[66] (NNet-LBFGS). For both NNet classifiers, we used the rectified linear unit activation function[67] (RELU).

Performance evaluation

The performance of our radiomics-based ML pipeline was measured using the test data set. The standard error of calculations was reported using 5-fold cross-validation on the training data set. We used the area under the receiver operating characteristic curve[68] (AUC) for performance evaluation. Also, we reported precision and recall for our best-performing pipeline.

Results

Radiomics feature space

A JSON file of the metadata of extracted radiomics features is available in the supplementary dataset in our public repository[69]. The predictive performance of the different RS techniques, FS methods, and ML classifiers was evaluated for each ROI on the test set using the AUC, precision, recall, and F-1 scores. Examples of receiver operating characteristic (ROC) curves are presented in Figure 4 for the a) NB (a poor performance), b) RF (a good performance), and c) GPR (the best performance) ML classifiers on the test data set (red squares) and 5-fold validation set (pink lines) using 20 mm spherical ROI (SP20) with no FS and no RS. Note that 20 mm SP ROI was selected for visualization purposes throughout this paper for no particular reason. The effect of using the various geometric ROIs will be presented later in this paper. Raw data values, including confusion matrices, ROC graphs, and performance tables (precision, recall, F-1, ROC-AUC values on training, validation, and test sets) for all ML classifiers on all ROIs are provided in the output data folder in our public repository[69].

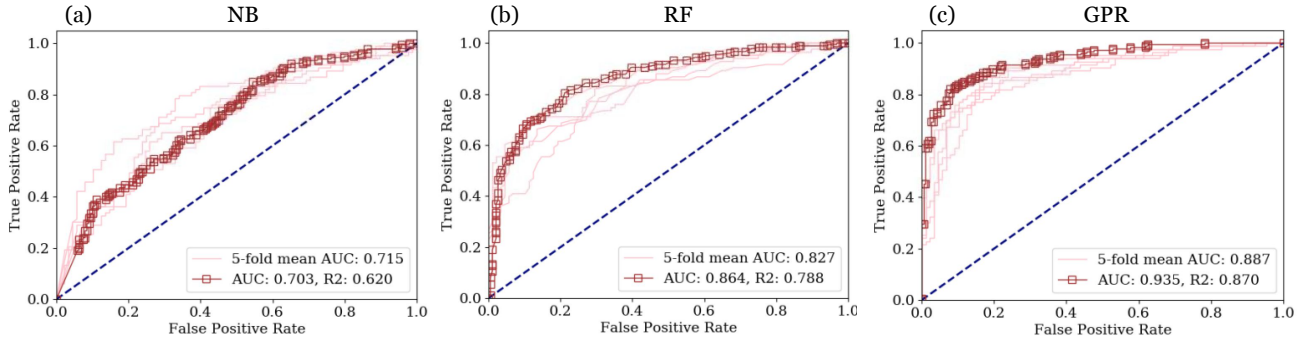

Figure 4. Example ROC curves for our radiomics-based ML pipeline with the a) NB, b) RF, and c) GPR ML classifiers. We used 20 mm spherical ROI (SP20) with no FS and no RS for this example. AUC values (presented in the legends) were calculated for the test set (red squares) and 5-fold validation set (pink lines). The 20 mm SP ROI was used for visualization purposes. Full data is available in the supplementary dataset[69].

Effect of the RS

An example of an AUC grid for different combinations of ML classifiers and RS techniques is presented in Figure 5 for the 20 mm spherical ROI. No FS method was used in this example. As can be seen, RS did not have much of an effect on the performance of the ML classifiers. This is because our data set was nearly balanced with 631 BM lesions and 674 HBs. The GPR classifier resulted in the best performance (AUC= 93% ± 0.5%).

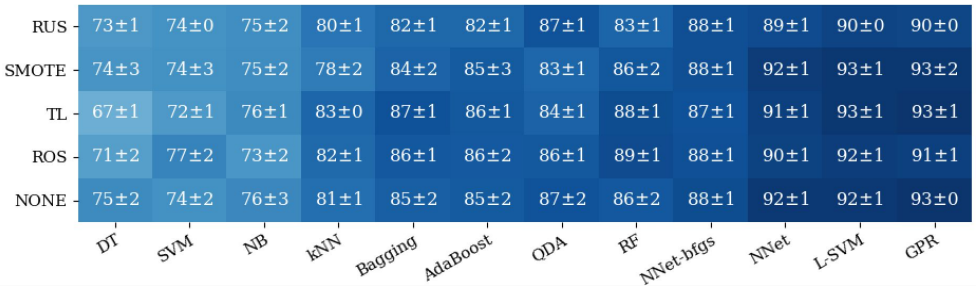

Figure 5. The AUC grid for different combinations of ML classifiers (x-axis) and RS techniques (y-axis).

Effect of the FS

An example of an AUC grid for different combinations of ML classifiers and FS methods is presented in Figure 6 for the 20 mm SP ROI. SMOTE RS technique was used in this example. As can be seen in Figure 6, the best results were achieved by the GPR and NNet classifiers with LASSO FS methods. The RFECV, VT, LASSO, and TREE FS methods outperformed PCA and ICA FS methods. Overall, FS did not have much effect on the performance of the pipeline for the 20 mm ROI. For example, for the GPR ML classifier, the performance of our pipeline increased only 2% (from 93% to 95%) with the LASSO method compared to with no FS (NONE).

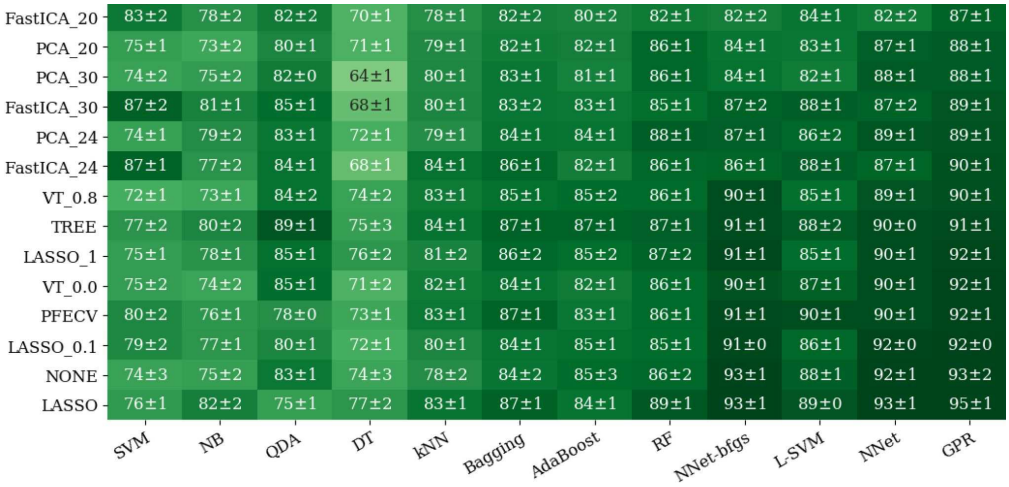

Figure 6. The AUC grid for different ML (x-axis) classifiers and FS methods (y-axis) combinations. The number in front of each PCA or FastICA method is the number of selected features used. The number in front of each LASSO method corresponds to the α penalty value (the default value is 0.5). The number in front of each VT method is its variance threshold value.

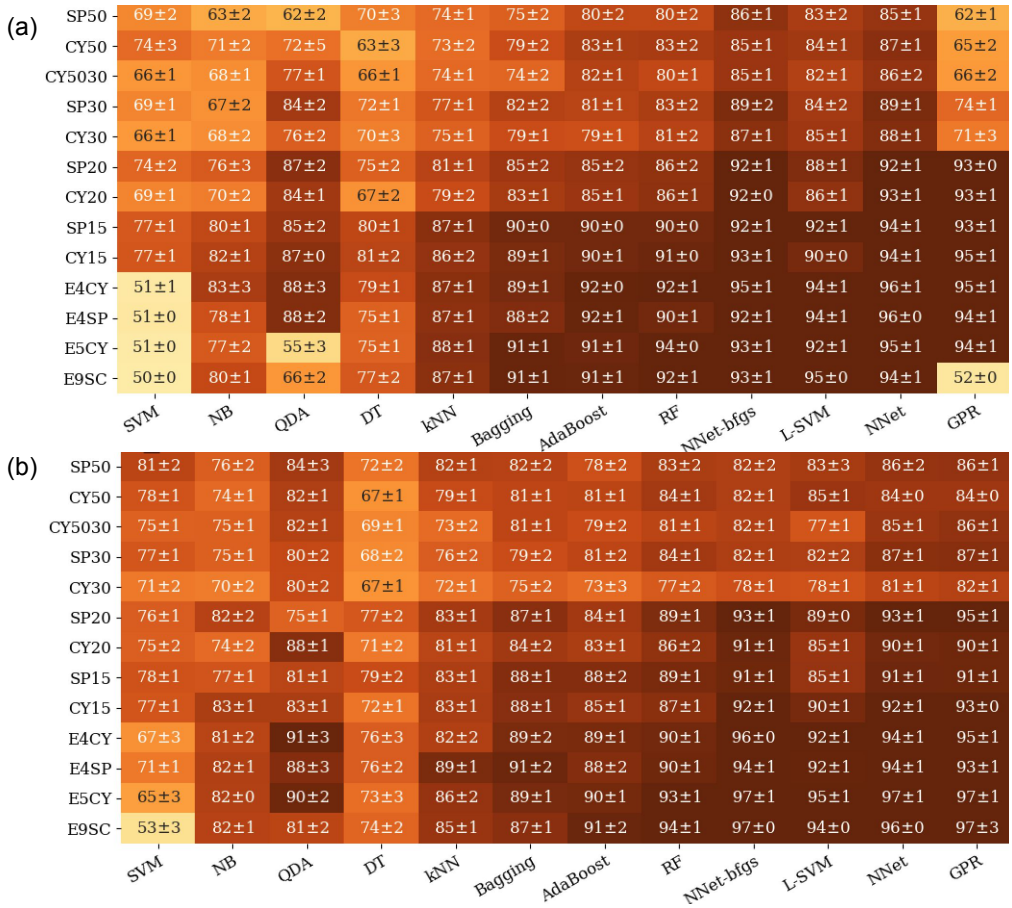

Figure 7. The AUC grid for different combinations of ML classifiers (x-axis) and geometric ROIs (y-axis) with various sizes and shapes (a) with no FS method and no RS techniques, and (b) with LASSO as the FS method and SMOTE as the RS technique.

Effect of ROI

Two examples of the effect of using geometric ROIs with different sizes and shapes are presented in Figures 7. For plot (a), we used no FS and no RS techniques. For plot (b), we used the best performing FS method (LASSO) and the best performing

RS technique (SMOTE). As can be seen in Figure 7, the size of the ROI had a significant effect on the performance of our radiomics-based ML pipeline. In general, a smaller ROI resulted in superior performance of the pipeline. For example, for the NNet classifier with no FS and no RS (the rightmost column of Figure 7-a), the AUC was improved from 85% to 94% when we moved from the SP50 to the CY15 ROI. CY15 resulted in the best overall performance when no FS was used. When we employed FS methods, the ensemble ROIs (like E4SP and E9SC) out-performed the single-size ROIs. This was most pronounced for the LASSO method, which is presented in Figure 7-b. The AUC grids for other FS methods are provided in the output data folder in our public repository[69].

Comparing Figure 7-a and 7-b revealed that some ML classifiers (like SVM or GPR) were more sensitive to the use of FS than others (like NNet or RF). Also, we noticed that FS was more important when using large ROIs (such as SP50 or CY50) or ensemble ROIs (such as E4SP or E9SC).

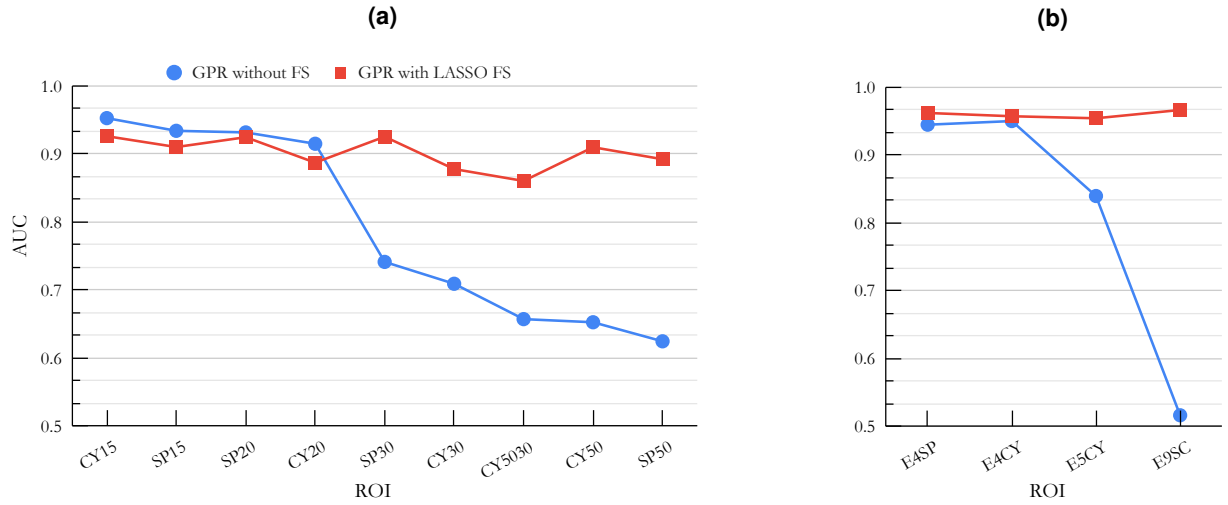

Figure 8. An example of AUCs versus the volume of the ROI for (a) single geometric ROIs and (b) for ensemble ROIs (Ref. Table 3). For this graph, we used our best performing ML classifier (GPR), with our best performing FS method (LASSO), and without FS method.

To visualize the effect of the size of ROI on the performance of our pipeline, in Figure 8 we show the AUCs of our best performing ML classifier (GPR) for (a) single geometric ROIs (sorted by volume), and (b) for ensemble ROIs (sorted by total volume). To show the effect of the use of FS, we plotted the results without FS (blue circles), and with our best-performing FS method (LASSO) (red squares). It can be seen that a smaller ROI resulted in a better performance. Also, FS was more important for larger ROIs (like SP50 and CY50) and ensemble ROIs (like E9SC).

The grid of the F-1 scores for the best performing FS method (LASSO) and RS technique (SMOTE) is presented if Figure 9. The GPR, NNet, and L-SVM classifiers achieved higher than 0.9 F-1 score in detecting BM using the ensemble ROIs. The AUC, precision, recall, and F1 score of our best performing pipeline, corresponding to the E9SC ROI, SMOTE RS technique, LASSO FS method, and GPR ML classifier, were 97%, 92%, 91%, and 0.9, respectively. The performance of our pipeline for all combinations of RS techniques, FS methods, ML classifiers, and ROIs are provided in the supplementary dataset[69].

Discussion

In this study, we investigated the feasibility of using a single-point-based geometric ROI to develop a radiomics pipeline to distinguish BM and HB locations in planning-CT images of cancer patients with BM. We investigated various RS techniques, FS methods, and ML classifiers using point-based geometric ROIs with various shapes and sizes.

The time and effort needed for manual 3D segmentation of ROI are significant limitations to achieving large real-world image data sets. This, in turn, hinders the generation of generalizable radiomics-based prognostic ML models[70] for use in the clinic. Another limitation of manual lesion segmentation is inter-observer variability, which has been shown to have a significant impact on the performance and reproducibility of radiomics-based pipelines[71]. Furthermore, manual segmentation tools, designed for radiation therapy treatment planning, intend to load one patient at a time. Therefore, switching between patients is another time-consuming process that slows down the lesion delineation for multiple patients in the research context[72].

Figure 9. The F-1 score grid for different combinations of ML classifiers (x-axis) and ROIs (y-axis) with different sizes and shapes with the SMOTE RS technique and LASSO FS method.

Our in-house-developed open-source 3D DICOM visualization and lesion identification tool (diCOMBINE[30]) allowed our collaborating radiation oncologists to quickly review planning-CT images of several hundred patients and efficiently identify 676 BM centers. They found diCOMBINE fast and easy to use, allowing each expert to label around 150 lesions per hour. Based on our experts' anecdotal experience, single-point-based geometric ROI delineation was 10 to 15 times faster than full manual 3D segmentation. These lesion centers were used to generate ROIs automatically. Defining point-based geometric ROIs, instead of full 3D manual segmentation of the ROIs, allowed us to rapidly generate a large sample set, minimize expert imposed uncertainties, and investigate the effect of the size and shape of the ROIs in the performance of our radiomics pipeline. Besides, our point-based radiomics pipeline will allow us to study the feasibility of building an automated BM-identifying pipeline. To the best of our knowledge, no studies on automated BM delineation have been published previously.

Radiomics extracts hundreds of features from an ROI. However, these features are generally highly correlated and contain much noise. Therefore, it is essential to apply proper FS methods to achieve a robust radiomics-based ML pipeline. Among the seven FS methods we examined in this study, we found that PCA and ICA resulted in lower AUC values than the VT, LASSO, and TREE FS methods. One reason for this difference was that VT, LASSO, and TREE methods automatically defined the optimal number of features, while in PCA and ICA, the number of features was predefined. For highly-correlated features, the optimal number of features (f) is roughly proportional to the square root of the sample size (n)[73]. Accordingly, 30 features would appear to be less than the optimal number of features for our sample size ($f = \sqrt{n} = \sqrt{1305} = 36$). For studies with small sample sizes, such as Zhang et al.[9], that used 112 samples, PCA with 10 features seems to be a suitable FS method. We also noticed that the effect of the FS method depends on the selected ML classifier. For ML classifiers that had built-in FS methods (i.e., RF and NNet), applying FS methods in some cases worsened the overall performance of the pipeline. Inversely, for ML classifiers that did not have built-in FS methods (i.e., GPR), adding FS had a significant effect on the performance of the ML classifier. The effect of the FS method was more significant when working with ensemble ROIs that had many more features. For example, the AUC value for the GPR ML classifier using the E9SC ROI (963 features) improved from 0.52 to 0.97 when the LASSO FS method was used, as shown in Figure 7.

Among the ML methods we examined in this study, we found that GPR, NNet, SVM, and RF resulted in the highest AUC values and F-1 scores. We showed that the GPR classifier outperformed the NNet classifier for most ROIs. However, for the ensemble ROIs (in which the number of features was large), GPR required a proper FS method (i.e., LASSO). The dimensionality issue of GPR classifiers and their requirement for FS was also discussed in the literature[74,75].

We found that our radiomics-based ML pipeline performed slightly better on spherical ROIs compared to cylindrical ROIs of similar volumes. More significantly, we found that the smaller ROIs (15 and 20 mm) resulted in better performance compared to the larger ROIs (30 and 50 mm) (Figure 7). This might be due to the fact that in larger ROIs there are probably more outlier features captured from bone or organs/tissue surrounding the lesion of interest. The performance of our pipeline did not improve considerably by decreasing the size of ROI below 20 mm, which is roughly the size of a large BM lesion[32]. As can be seen in Figure 8, our pipeline performed better on the ensemble ROIs compared to the single ROI when used with FS methods. This could be due to having many features in the ensemble ROIs. For example, the E9SC ROI contains $9 \times 107 = 963$ features. For such a prominent feature space, FS methods become very important.

Although various radiomics pipelines have been previously developed and reported to classify bone lesions, our radiomics-

based ML pipeline, reported here, offers several advantages compared to preceding efforts, mainly in the context of palliative radiotherapy planning. First, we, pragmatically, used planning-CT images of cancer patients for extracting radiomics features, whereas prior studies used hybrid modalities or diagnostic-CT images (as listed in Table 1). Hybrid modalities allow the development of high-quality prognostic pipelines. However, these pipelines are less clinically applicable in palliative radiotherapy treatment planning for BM, which is often primarily based on a patient's planning-CT scan. Second, all ML classifiers presented in the prior studies were restricted to full 3D segmentation of the lesion volumes. In the real-world clinical workflow for palliative radiotherapy of BM, it is common to use single-slice or lesion-center-based treatment planning with radiation oncologists often defining treatment field limits rather than lesion contours. Therefore, pipelines that require full 3D segmentation of ROI have limited application in real-world palliative radiotherapy[72]. Moreover, 3D segmentation of the ROI is a time-consuming bottleneck that likely compelled all the prior studies to train and test their radiomics pipelines with limited sample sizes. Training on a small sample size diminishes the generalizability and clinical applicability of a radiomics pipeline. In comparison, our point-based pipeline allowed us to avoid the labor-intensive manual segmentation step and train and test our pipeline on a large data set. Finally, in this study, we investigated the effects of different RS techniques, FS methods, and ML classifiers in achieving the optimal prognostic model using geometric ROIs. To the best of our knowledge, no prior study performed such a comprehensive optimization.

Our study had some limitations. First, we selected BM and HB from two sets of separate patients. This selection might drive the risk of potential susceptibility to bias if there is a systematic difference between the two sets of images. However, our rationale for using non-metastatic cancer patients to select HBs was to eliminate the possibility of error in labeling HBs by our medical physicists. Second, our collaborating medical physicists could not identify non-metastatic skeletal complications from metastatic bone lesions. Therefore, the non-metastatic skeletal complications (i.e., surgically-removed lesions or bone islands) were ignored when labeling HB points. A solution for this problem would be using pathology data to identify non-metastatic and metastatic lesions but this would significantly increase the required effort. Third, we used single-center planning-CT images from 359 patients in this retrospective study. A multi-center study with a more extensive data set is required to test the generalizability of our radiomics pipeline. Such a big data set would allow us to try more robust deep learning ML classifiers[76,77] to build an AI tool to scan patients' planning-CT images and identify BM lesions automatically. The present work provides strong motivation to pursue such a multi-center study.

Conclusion

We demonstrated that our radiomics-based ML pipeline can successfully distinguish between metastatic and healthy bones in planning-CT images using lesion-center-based geometric ROIs. Our results suggest that the GPR classifier with ensemble ROIs is particularly promising for the differentiation of BM and HB. Optimum pipeline performance was obtained using elimination-based FS methods such as LASSO. Our lesion-center-based ROI delineation methodology demonstrates that full 3D lesion segmentation is not necessary to train radiomics-based ML classifiers to distinguish between bone lesions. This opens the door to big data artificial intelligence research for cancer patients with BM.

Data availability

The supporting dataset is provided as a figshare repository[69]. This repository contains three files: 1) "featurespace_metadata.json.zip" file that includes radiomics features extracted from 1273 spinal lesions (healthy or metastatic) from radiotherapy planning-ct images using geometrical regions of interest (ROIs). 2) "output.zip" folder that contains the results of our radiomics-based machine learning pipeline that validated and tested using several RS, FS, and ML on single-point-based geometric ROIs with various shapes and sizes. 3) A README.md file that is provided to explain the information about the data structure and file naming patterns.

References

1. Perk, T. *et al.* Automated classification of benign and malignant lesions in 18 F-NaF PET/CT images using machine learning. *Phys. medicine biology* **63**, DOI: 10.1088/1361-6560/AAEBD0 (2018).

2. Suhas, M. V. & Mishra, A. Classification of benign and malignant bone lesions on CT images using random forest. *2016 IEEE Int. Conf. on Recent Trends Electron. Inf. Commun. Technol. RTEICT 2016 - Proc.* 1807–1810, DOI: 10.1109/RTEICT.2016.7808146 (2017).

3. Acar, E., Leblebici, A., Ellidokuz, B. E., Başbinar, Y. & Kaya, G. C. Machine learning for differentiating metastatic and completely responded sclerotic bone lesion in prostate cancer: A retrospective radiomics study. *Br. J. Radiol.* **92**, DOI: 10.1259/bjr.20190286 (2019).

4. Suhas, M. V. & Kumar, R. Classification of benign and malignant bone lesions on CT imagesusing support vector machine: A comparison of kernel functions. *2016 IEEE Int. Conf. on Recent Trends Electron. Inf. Commun. Technol. RTEICT 2016 - Proc.* 821–824, DOI: 10.1109/RTEICT.2016.7807941 (2017).

5. Homayounieh, F. *et al.* Semiautomatic Segmentation and Radiomics for Dual-Energy CT: A Pilot Study to Differentiate Benign and Malignant Hepatic Lesions. *AJR. Am. journal roentgenology* **215**, DOI: 10.2214/AJR.19.22164 (2020).

6. Hong, J. H. *et al.* Development and validation of a radiomics model for differentiating bone islands and osteoblastic bone metastases at abdominal CT. *Radiology* **299**, 626–632, DOI: 10.1148/RADIOL.2021203783/ASSET/IMAGES/LARGE/RADIOL.2021203783.VA.JPEG (2021).

7. Sun, W. *et al.* A CT-based radiomics nomogram for distinguishing between benign and malignant bone tumours. *Cancer Imaging* **21**, 1–10, DOI: 10.1186/S40644-021-00387-6/FIGURES/4 (2021).

8. Vial, A. *et al.* The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. *Transl. Cancer Res.* **7**, DOI: 10.21037/21823 (2018).

9. Zhang, Y., Oikonomou, A., Wong, A., Haider, M. A. & Khalvati, F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Sci. Reports 2017 7:1* **7**, 1–8, DOI: 10.1038/srep46349 (2017).

10. Baessler, B. *et al.* Radiomics allows for detection of benign and malignant histopathology in patients with metastatic testicular germ cell tumors prior to post-chemotherapy retroperitoneal lymph node dissection. *Eur. radiology* **30**, 2334–2345, DOI: 10.1007/S00330-019-06495-Z (2020).

11. Duron, L. *et al.* A Magnetic Resonance Imaging Radiomics Signature to Distinguish Benign From Malignant Orbital Lesions. *Investig. radiology* **56**, 173–180, DOI: 10.1097/RLI.0000000000000722 (2021).

12. Laderian, B. *et al.* Role of radiomics to differentiate benign from malignant pheochromocytomas and paragangliomas on contrast enhanced CT scans. *J. Clin. Oncol.* **37**, e14596–e14596, DOI: 10.1200/JCO.2019.37.15{_}SUPPL.E14596 (2019).

13. Li, S. *et al.* A radiomics approach for automated diagnosis of ovarian neoplasm malignancy in computed tomography. *Sci. Reports 2021 11:1* **11**, 1–9, DOI: 10.1038/s41598-021-87775-x (2021).

14. Yin, P. *et al.* Machine and Deep Learning Based Radiomics Models for Preoperative Prediction of Benign and Malignant Sacral Tumors. *Front. Oncol.* **10**, 2235, DOI: 10.3389/FONC.2020.564725/BIBTEX (2020).

15. Wang, H. *et al.* Radiomics nomogram for differentiating between benign and malignant soft-tissue masses of the extremities. *J. magnetic resonance imaging : JMRI* **51**, 155–163, DOI: 10.1002/JMRI.26818 (2020).

16. Wang, J. *et al.* Prediction of malignant and benign of lung tumor using a quantitative radiomic method. *Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc. IEEE Eng. Medicine Biol. Soc. Annu. Int. Conf.* **2016**, 1272–1275, DOI: 10.1109/EMBC.2016.7590938 (2016).

17. Zhou, L. *et al.* A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors. *Transl. oncology* **12**, 292–300, DOI: 10.1016/J.TRANON.2018.10.012 (2019).

18. Guo, B. J. *et al.* Benign and malignant thyroid classification using computed tomography radiomics. *https://doi.org/10.1117/12.2549087* **11314**, 954–961, DOI: 10.1117/12.2549087 (2020).

19. Paul, R. *et al.* Deep radiomics: deep learning on radiomics texture images. *https://doi.org/10.1117/12.2582102* **11597**, 8–17, DOI: 10.1117/12.2582102 (2021).

20. Chen, A. *et al.* CT-based radiomics model for predicting brain metastasis in category T1 lung adenocarcinoma. *Am. J. Roentgenol.* **213**, 134–139, DOI: 10.2214/AJR.18.20591 (2019).

21. Mayerhoefer, M. E. *et al.* Introduction to Radiomics. *J. nuclear medicine : official publication, Soc. Nucl. Medicine* **61**, 488–495, DOI: 10.2967/JNUMED.118.222893 (2020).

22. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. journal cancer (Oxford, Engl. : 1990)* **48**, 441–446, DOI: 10.1016/J.EJCA.2011.11.036 (2012).

23. Sugai, Y. *et al.* Impact of feature selection methods and subgroup factors on prognostic analysis with CT-based radiomics in non-small cell lung cancer patients. *Radiat. Oncol.* **16**, 1–12, DOI: 10.1186/S13014-021-01810-9/FIGURES/2 (2021).

24. Demircioğlu, A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights into Imaging 2021 12:1* **12**, 1–10, DOI: 10.1186/S13244-021-01115-1 (2021).

25. Yin, P. *et al.* Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. *Eur. radiology* **29**, 1841–1847, DOI: 10.1007/S00330-018-5730-6 (2019).

26. Delzell, D. A., Magnuson, S., Peter, T., Smith, M. & Smith, B. J. Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data. *Front. oncology* **9**, DOI: 10.3389/FONC.2019.01393 (2019).

27. Ligero, M. *et al.* Selection of Radiomics Features based on their Reproducibility. *Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc. IEEE Eng. Medicine Biol. Soc. Annu. Int. Conf.* **2019**, 403–408, DOI: 10.1109/EMBC.2019.8857879 (2019).

28. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. resonance imaging* **30**, 1234–1248, DOI: 10.1016/J.MRI.2012.06.010 (2012).

29. Xie, C. *et al.* Effect of machine learning re-sampling techniques for imbalanced datasets in 18 F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *Eur. journal nuclear medicine molecular imaging* **47**, 2826–2835, DOI: 10.1007/S00259-020-04756-4 (2020).

30. Naseri, H. diCOMBINE: 3D-DICOM Visualization and Lesion Identification Web Application, DOI: 10.5281/ZENODO.5218743 (2021).

31. Flask Web Development, 2nd Edition [Book].

32. Hall, G. & Wright, J. Bone Lesions. *Gnepp's Diagn. Surg. Pathol. Head Neck* 689–742, DOI: 10.1016/B978-0-323-53114-6.00008-0 (2021).

33. Zhou, S. H., McCarthy, I. D., McGregor, A. H., Coombs, R. R. & Hughes, S. P. Geometrical dimensions of the lower lumbar vertebrae – analysis of data from digitised CT images. *Eur. Spine J. 2000 9:3* **9**, 242–248, DOI: 10.1007/S005860000140 (2000).

34. Busscher, I., Ploegmakers, J. J., Verkerke, G. J. & Veldhuizen, A. G. Comparative anatomical dimensions of the complete human and porcine spine. *Eur. Spine J.* **19**, 1104–1114, DOI: 10.1007/S00586-010-1326-9/FIGURES/8 (2010).

35. Van Griethuysen, J. J. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research* **77**, e104–e107, DOI: 10.1158/0008-5472.CAN-17-0339 (2017).

36. Radiomic Features — pyradiomics v3.0.1.post9+gdfe2c14 documentation.

37. Zwanenburg, A. *et al.* The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* **295**, 328–338, DOI: 10.1148/RADIOL.2020191145 (2020).

38. Fontaine, P. *et al.* The importance of feature aggregation in radiomics: a head and neck cancer study. *Sci. Reports 2020 10:1* **10**, 1–11, DOI: 10.1038/s41598-020-76310-z (2020).

39. Wakabayashi, K. *et al.* A predictive model for pain response following radiotherapy for treatment of spinal metastases. *Sci. Reports* | **11**, 12908, DOI: 10.1038/s41598-021-92363-0 (123).

40. Kochendörffer, R. Kreyszig, E.: Advanced Engineering Mathematics. J. Wiley & Sons, Inc., New York, London 1962. IX + 856 S. 402 Abb. Preis s. 79.—. *Biom. Zeitschrift* **7**, 129–130, DOI: 10.1002/BIMJ.19650070232 (1965).

41. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. Royal Stat. Soc. Ser. B (Methodological)* **36**, 111–147, DOI: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x (1974).

42. Sun, Y., Wong, A. K. & Kamel, M. S. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *http://dx.doi.org/10.1142/S0218001409007326* **23**, 687–719, DOI: 10.1142/S0218001409007326 (2011).

43. He, H. & Ma, Y. Imbalanced learning: Foundations, algorithms, and applications. *Imbalanced Learn. Foundations, Algorithms, Appl.* 1–210, DOI: 10.1002/9781118646106 (2013).

44. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232, DOI: 10.1007/S13748-016-0094-0/TABLES/1 (2016).

45. Fernández, A. *et al.* Learning from Imbalanced Data Sets. *Learn. from Imbalanced Data Sets* DOI: 10.1007/978-3-319-98074-4 (2018).

46. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Of Artif. Intell. Res.* **16**, 321–357, DOI: 10.1613/jair.953 (2002).

47. TomekLinks — Version 0.9.0.dev0.

48. Tomek, I. TWO MODIFICATIONS OF CNN. *IEEE Transactions on Syst. Man Cybern.* **SMC-6**, 769–772, DOI: 10.1109/TSMC.1976.4309452 (1976).

49. Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. *Eur. radiology experimental* **2**, DOI: 10.1186/S41747-018-0068-Z (2018).

50. 2.5. Decomposing signals in components (matrix factorization problems) — scikit-learn 1.0.1 documentation.

51. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430, DOI: 10.1016/S0893-6080(00)00026-5 (2000).

52. 1.13. Feature selection — scikit-learn 1.0.1 documentation.

53. Kim, S.-J., Koh, K., Lustig, M., Boyd, S. & Gorinevsky, D. An Interior-Point Method for Large-ScalèScalè 1-Regularized Least Squares. *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING* **1**, DOI: 10.1109/JSTSP.2007.910971 (2007).

54. 1.13. Feature selection — scikit-learn 1.0.1 documentation.

55. Breiman, L. Random Forests. *Mach. Learn. 2001 45:1* **45**, 5–32, DOI: 10.1023/A:1010933404324 (2001).

56. 1. Supervised learning — scikit-learn 0.20.4 documentation.

57. 1.4. Support Vector Machines — scikit-learn 1.0.1 documentation.

58. 1.9. Naive Bayes — scikit-learn 1.0.1 documentation.

59. 1.6. Nearest Neighbors — scikit-learn 1.0.1 documentation.

60. 1.2. Linear and Quadratic Discriminant Analysis — scikit-learn 1.0.1 documentation.

61. 1.7. Gaussian Processes — scikit-learn 1.0.1 documentation.

62. 1.10. Decision Trees — scikit-learn 1.0.1 documentation.

63. 1.11. Ensemble methods — scikit-learn 1.0.1 documentation.

64. 1.17. Neural network models (supervised) — scikit-learn 1.0.1 documentation.

65. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *3rd Int. Conf. on Learn. Represent. ICLR 2015 - Conf. Track Proc.* (2015).

66. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program. 1989 45:1* **45**, 503–528, DOI: 10.1007/BF01589116 (1989).

67. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning - whole book. *Nature* **521**, 800 (2016).

68. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874, DOI: 10.1016/J.PATREC.2005.10.010 (2006).

69. Hossein Naseri *et al.* A radiomics-based machine learning pipeline to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest; dataset, DOI: https://doi.org/10.6084/m9.figshare.19224615.v1 (2022).

70. Kocak, B., Durmaz, E. S., Ates, E. & Kilickesmez, O. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn. interventional radiology (Ankara, Turkey)* **25**, 485–495, DOI: 10.5152/DIR.2019.19321 (2019).

71. Haarburger, C. *et al.* Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. *Sci. Reports 2020 10:1* **10**, 1–10, DOI: 10.1038/s41598-020-69534-6 (2020).

72. Kocak, B., Durmaz, E. S., Kaya, O. K., Ates, E. & Kilickesmez, O. Reliability of Single-Slice-Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. *AJR. Am. journal roentgenology* **213**, 377–383, DOI: 10.2214/AJR.19.21212 (2019).

73. Hua, J., Xiong, Z., Lowey, J., Suh, E. & Dougherty, E. R. Optimal number of features as a function of sample size for various classification rules. *Bioinforma. (Oxford, England)* **21**, 1509–1515, DOI: 10.1093/BIOINFORMATICS/BTI171 (2005).

74. Tripathy, R., Bilionis, I. & Gonzalez, M. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *J. Comput. Phys.* **321**, 191–223, DOI: 10.1016/J.JCP.2016.05.039 (2016).

75. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning. *Gaussian Process. for Mach. Learn.* DOI: 10.7551/MITPRESS/3206.001.0001 (2005).

76. Bibault, J. E. *et al.* Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci. Reports 2018 8:1* **8**, 1–8, DOI: 10.1038/s41598-018-30657-6 (2018).

77. He, Y. *et al.* Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. *eBioMedicine* **62**, 103121, DOI: 10.1016/J.EBIOM.2020.103121 (2020).

Acknowledgements

This research has been supported by the startup grant of Dr. John Kildea at Research Institute of the McGill University Health Centre (RI-MUHC), Ruth and Alex Dworkin scholarship award from the McGill University, Faculty of Medicine, RI-MUHC studentship award, Grad Excellence Award-00293 from the McGill University, Department of Physics, and CREATE Responsible Health and Healthcare Data Science (SDRDS) grant of the Natural Sciences and Engineering Research Council. The authors would like to thank Dr. Luc Galarneau for his help with statistical analysis.

Competing interests

The authors declare no competing interests.

Author contributions statement

H.N. contributed to the methodology, literature review, software, formal analysis, investigation, visualization, and writing the original draft. S.S. participated in data collection, interpretation, and validation. M.T. participated in data collection, interpretation, and validation. M.F. participated in data collection, interpretation, and validation. P.R. participated in data collection, interpretation, and validation. J.Kh. participated in data collection, interpretation, and validation. H.P. participated in data collection, interpretation, and validation. A.X.A.H. participated in data collection, interpretation, and validation. M.D. participated in conceptualization and methodology. J.Ki. participated in data collection and contributed to the conceptualization, investigation, supervision, funding acquisition, and editing of the original draft. All authors contributed to the review of the paper and approved the final manuscript.