Physics in Medicine & Biology

Institute of Physics and Engineering in Medicine

PAPER

Automated classification of benign and malignant lesions in ^{18}F-NaF PET/CT images using machine learning

To cite this article: Timothy Perk *et al* 2018 *Phys. Med. Biol.* **63** 225019

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
16 May 2018

REVISED
18 October 2018

ACCEPTED FOR PUBLICATION
26 October 2018

PUBLISHED
20 November 2018

PAPER

Automated classification of benign and malignant lesions in ^{18}F-NaF PET/CT images using machine learning

Timothy Perk[1], Tyler Bradshaw[2], Song Chen[3], Hyung-jun Im[2], Steve Cho[2], Scott Perlman[2], Glenn Liu[4] and Robert Jeraj[1,5,6]

[1] Department of Medical Physics, University of Wisconsin, Madison, WI, United States of America
[2] Department of Radiology, University of Wisconsin, Madison, WI, United States of America
[3] Department of Nuclear Medicine, The First Hospital of China Medical University, Shenyang, Liaoning, People's Republic of China
[4] Department of Medicine, University of Wisconsin, Madison, WI, United States of America
[5] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
[6] Author to whom any correspondence should be addressed.

E-mail: rjeraj@wisc.edu

Keywords: NaF PET/CT, machine learning, lesion classification

Supplementary material for this article is available online

Abstract

Purpose. ^{18}F-NaF PET/CT imaging of bone metastases is confounded by tracer uptake in benign diseases, such as osteoarthritis. The goal of this work was to develop an automated bone lesion classification algorithm to classify lesions in NaF PET/CT images. *Methods.* A nuclear medicine physician manually identified and classified 1751 bone lesions in NaF PET/CT images from 37 subjects with metastatic castrate-resistant prostate cancer, 14 of which (598 lesions) were analyzed by three additional physicians. Lesions were classified on a five-point scale from definite benign to definite metastatic lesions. Classification agreement between physicians was assessed using Fleiss' κ. To perform fully automated lesion classification, three different lesion detection methods based on thresholding were assessed: SUV > 10 g ml^{-1}, SUV > 15 g ml^{-1}, and a statistically optimized regional thresholding (SORT) algorithm. For each ROI in the image, 172 different imaging features were extracted, including PET, CT, and spatial probability features. These imaging features were used as inputs into different machine learning algorithms. The impact of different deterministic factors affecting classification performance was assessed. *Results.* The factors that most impacted classification performance were the machine learning algorithm and the lesion identification method. Random forests (RF) had the highest classification performance. For lesion segmentation, using SORT (AUC $=$ 0.95 [95%CI $=$ 0.94–0.95], sensitivity $=$ 88% [86%–90%], and specificity $=$ 0.89 [0.87–0.90]) resulted in superior classification performance ($p < 0.001$) compared to SUV > 10 g ml^{-1} (AUC $=$ 0.87) and SUV > 15 g ml^{-1} (AUC $=$ 0.86). While there was only moderate agreement between physicians in lesion classification ($\kappa = 0.53$ [95% CI $=$ 0.52–0.53]), classification performance was high using any of the four physicians as ground truth (AUC range: 0.91–0.93). *Conclusion.* We have developed the first whole-body automatic disease classification tool for NaF PET using RF, and demonstrated its ability to replicate different physicians' classification tendencies. This enables fully-automated analysis of whole-body NaF PET/CT images.

Introduction

An important step to allow for the accurate quantitative assessment of ^{18}F-Sodium fluoride (NaF) PET/CT images is to reduce the number of false positive metastases arising from benign diseases. NaF PET/CT is one of the most sensitive imaging modalities for the detection of bone metastases (Even-Sapir *et al* 2006). NaF PET has higher sensitivity and specificity for bone lesion detection than current clinical standard of ^{99m}Tc-methylene diphosphonate (MDP) planar bone scans. NaF PET detects 30% more lesions with 5% fewer false positives than MDP (Even-Sapir *et al* 2006). However, the specificity of NaF PET is still only 70%, due to the non-tumor specific nature of the tracer (Li *et al* 2012). The fluoride accumulation in the skeleton is proportional to blood flow and the rate of bone turnover, and cases of osteoarthritis are associated with increased bone turnover (Burr 1998).

NaF PET SUV metrics have been shown to correlate to clinical outcome, either for progression free survival (Harmon *et al* 2017) or overall survival (Lindgren Belal *et al* 2017). Furthermore, Harmon *et al* showed that the heterogeneity of responses of individual lesions is highly predictive of progression free survival (Harmon *et al* 2016). This suggests that assessment of NaF PET/CT on a per-lesion level may be useful for evaluating treatment efficacy. Misinterpreted false positive lesions may lead to suboptimal treatment decisions. Yet, due to the number of lesions that would need to be analyzed, this type of analysis requires automation, especially in the differentiation of malignant and benign bone disease.

Studies have shown that when using standardized uptake value (SUV) alone it is difficult to determine whether PET uptake is due to metastatic or benign disease (Cook and Fogelman 2000, Muzahir *et al* 2015, Sabbah *et al* 2015). When physicians use CT to classify these regions of ambiguous uptake the specificity of NaF PET/CT can significantly increase (Even-Sapir *et al* 2006). However, due to the large number of lesions in patients with bone metastases (Wang and Shen 2012), it becomes impractical to perform lesion classification manually. Previous studies on automatic detection of osteoarthritis in CT and NaF PET/CT images were limited to only spinal disease in the form of osteophytes, even though osteoarthritis can commonly occur outside of the spine (Punzi *et al* 2004, Rosen *et al* 2006, Munoz *et al* 2013, Yao *et al* 2014, Wang *et al* 2016).

This work aims to develop a method for classification of bone disease on NaF PET/CT scans of metastatic castrate resistant prostate cancer (mCRPC) patients. We compare nine machine learning techniques and assess their performance under different circumstances. We also develop a novel set of imaging features to include in our models that describe spatial probabilities of disease patterns.

Methods and materials

Patients

This study included 37 mCRPC patients with evaluable disease who received PET/CT scans before the start of treatment. All patients received whole body PET/CT scans 60 min post injection of 160.2 ± 9.6 MBq of NaF. Scans at two of the sites were acquired on the Discovery VCT (GE Healthcare, Waukesha, WI) PET/CT scanner, and scans at the third site were acquired on the Gemini (Philips Healthcare, Amsterdam, Netherlands) PET/CT scanner. The acquisition time for a whole-body scan was 3 min per bed position, imaging from the top of the skull to the base of the feet. PET images were attenuation and scatter corrected. Harmonization of image reconstruction parameters was performed to allow quantitative comparisons across these sites as described previously (Jallow and Jeraj 2014, Lin *et al* 2016). Images were quantitatively harmonized using a uniform phantom and the National Electrical Manufacturers Association International Electrotechnical Commission body phantom to measure recovery coefficient (RC) and signal-to-noise ratio (SNR). Reconstruction parameters, such as number of iterations, number of subsets, and post-reconstruction filter, were altered to minimized differences in RC and SNR between phantom images obtained at the different sites.

Reference labeling

A nuclear medicine physician (Physician 1) manually identified and classified each lesion within each patient. Lesions were classified on a five-point scale, including definitely benign (1), likely benign (2), equivocal (3), likely malignant (4), and definitely malignant (5). Benign lesions included spinal osteophytes, disease between joints, inflammation, and dental disease. A subset of 14 patients was analyzed by three additional nuclear medicine physicians (Physician 2–4) working together, but independently from the other nuclear medicine physician. In this subset of patients, one physician identified all of the lesions and then the other three physicians individually classified the lesions, which was followed by the determining of a consensus score. These patients were used to assess agreement between physicians and assess the impact of physician input into machine learning algorithms.

Lesion detection and ROI generation

In order to create a tool that is fully automated, automated detection of NaF PET lesions was performed using different SUV thresholds. Multiple studies assessing use of NaF for metastatic bone cancer imaging use a fixed threshold of SUV $>$ 10 g ml^{-L} for detecting disease (Kurdziel *et al* 2012, Rohren *et al* 2015). However, SUV $>$ 10 g ml^{-1} thresholds often includes uptake in healthy bone, and thus a fixed threshold of SUV $>$ 15 g ml^{-1} has been used as an alternative (Lin *et al* 2016, Harmon *et al* 2017). However, as these methods have either poor sensitivity or specificity for lesion detection, we included optimized bone-specific thresholds based on statistically optimized regional thresholding (SORT) (Perk *et al* 2018), which uses a different threshold in each skeletal region defined from skeletal masks extracted from each patient's CT using atlas-based segmentation (Yip *et al* 2014). In this method receiver operating characteristic (ROC) analysis was performed to determine statistically optimal thresholds for disease detection in each of these regions. For all detection methods, ROIs

Table 1. Class numbers assigned to each ROI.

ROI Classification	Number
Background ROI	0
Definite Benign	1
Likely Benign	2
Equivocal	3
Likely Malignant	4
Definite Malignant	5

with non-specific NaF uptake that did not overlap the skeletal masks, such as the bladder and kidneys, were removed. Automated methods could result in ROIs that overlapped multiple physician identified lesions. If this overlap involved lesions were of multiple classes, manual splitting was performed. Each ROI was assigned a classification number based on whether it was detected by the physician and the classification of the physician described in table 1. A background ROI was defined as an ROI detected by a lesion detection method that was not corroborated by the physician and was labeled as class 0.

Feature extraction

For each individual ROI, 172 imaging-based features were extracted, shown in table 2. This includes 17 histogram/lesion level features extracted on PET or CT images, 41 texture features computed on PET alone (Galavis *et al* 2010), 41 texture features computed on CT alone, a categorical variable indicating ROI location, and 72 spatial distribution features. Global-level features were computed using the whole ROI and include SUV_{max}, volume, and $\text{SUV}_{\text{total}}$. The texture features we extracted include eight histogram-based first-order features, 23 co-occurrence matrix-based features, 11 features based on gray level run length matrices, five neighboring gray level features, and three neighborhood gray tone difference matrix features. To calculate these features, a $5 \times 5 \times 5$ voxel sub volume was extracted around each voxel, and the features were computed on each directional plane (axial, sagittal, and coronal), and then averaged over the three planes to get the feature value for that voxel. The texture feature values for each ROI were calculated as the average feature values of all the voxels within the ROI (Galavis *et al* 2010). The skeletal regions were converted into a categorical variable describing the general location of the lesion (described in table 2). ROIs overlapping multiple regions were assigned to the bone location containing the largest volume of the ROI.

For spatial distribution features, we created population disease distribution models as a way to add *a priori* information about the likelihood of disease in the specific location of the ROI (Perk *et al* 2015). The population disease distribution models were created by using a combination of articulated registration and optical-flow deformable registration to register patient's images and ROIs to a common template (Horn and Schunck 1981, Yip *et al* 2014). The physician labels for each ROI were used to convert the combined template images into population disease distributions for each of the labels listed in table 1: definite metastases, likely metastases, equivocal lesions, likely benign lesions, definite benign lesions, and background ROIs. Two types of distributions were created: probability distribution of disease occurrence (each voxel was normalized by the number of patients) and probability distributions of disease classification (each voxel was normalized by the number of lesions that occurred at that location). Additional distributions were created combining definite benign with likely benign, background ROIs with benign lesions, and definite metastases with likely metastases. When using the spatial distribution features in our models, we adopted a leave-one-out approach where disease distribution models were created using all other patients. For each lesion, four features from each model were extracted: maximum, average, minimum, and standard deviation of probability distributions within the lesion.

ROI classification

The six classes of ROIs were dichotomized into two groups based on physician labels to perform binary classification: 0–2 versus 4–5, 0–3 versus 4–5, 1–2 versus 4–5, 1–3 versus 4–5, 0–1 versus 5, and 1 versus 5. 10-fold cross-validation was used to split lesions into ten distinct training and testing datasets with 90% of the lesions and 10% of the lesions, respectively. Using the testing set in each fold, features with redundant information were removed using correlation analysis across the features. To select features for the model, ROC area under the curve (AUC) for predicting lesion binary classification was used as a surrogate for information gain. The feature with the highest AUC was selected to be included in the model, and if any features were highly correlated with that feature ($R > 0.8$), they were removed. This was done until all features were selected or removed. Then feature values were used as input into nine different machine learning algorithms: random forests (RF), Adaboost decision trees (DT), generalized linear models (GLM), neural network, *k*-nearest neighbors (KNN), linear discriminant analysis (LDA), DT, support vector machines (SVM), and Naive Bayes. Default algorithm hyperparameters (shown in table 3), when available, were used when comparing model performances. Algorithms were trained

Table 2. List of all imaging features used. In the case of location, there is a single feature and below lists the possible categories.

Image feature basis	Features
Location	Skull/mandible, ribs, cervical spine, thoracic spine, lumbar spine, sacrum, shoulders, sternum, humeri, radii/ulnae, hands, pelvis, femora, fibulae/tibiae, feet, patellae
Histogram	Max, mean, integral, standard deviation, skewness, kurtosis, energy, entropy, pet volume
Co-occurrence matrix	Angular moment, contrast-GLCM, correlation, sum of squares variance, inverse difference moment, sum average, sum variance, sum entropy, entropy-GLCM, difference variance, difference entropy, information measure of correlation 1, information measure of correlation 2, maximum probability, diagonal moment, dissimilarity, difference energy, inertia, inverse difference moment, sum energy, cluster shade, cluster prominance
Gray level run length	Small run emphasis, long run emphasis, gray-level nonuniformity, run length nonuniformity, run percentage, low gray-level emphasis, high gray-level emphasis, short run low gray-level emphasis, short run high gray-level emphasis, long run low gray-level emphasis, long run high gray-level emphasis
Neighboring gray level	Small number emphasis, large number emphasis, number nonuniformity, second moment, entropy-NGL
Neighborhood grey tone difference matrix	Coarseness, contrast-NGL, busyness
Disease distribution (occurrence or classification)	Maximum, minimum, mean, standard deviation

and assessed on a single physician's classification of lesions. In the case of neural networks, an additional 20% of the training data was held out to use as a validation set.

Model optimization

Comparisons of classification performance were made to assess the impact of various deterministic factors on the optimal algorithm's performance. We assessed choice of machine learning algorithm, lesion detection method, imaging resolution for extraction of texture, certainty levels of the ground truth by assessing which binary classification task was assessed, different physicians used for ground truth, inclusion of different feature sets, and algorithm hyper-parameters of the optimal model. The impact of image resampling was assessed by either up-sampling the PET or down-sampling the CT prior the texture feature extraction. Algorithm hyper-parameters were optimized using Bayesian optimization. Another comparison was performed to determine the effects of determining the testing and training sets based on the number of patients (90% of patients in training versus 10% testing) or the number of lesions independently of patient (90% of lesions in training versus 10% testing).

Statistics

Agreement between physicians was assessed using Fleiss' κ and Cohen's weighted κ. Fleiss' κ was used to compare the agreement of Physicians 2–4, who had worked together to form a consensus, and Cohen's weighted κ was used to compare the consensus of those three physicians to Physician 1.

The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of each classifier were determined at the classification score threshold that results in the highest summed sensitivity and specificity. A modified McNemar test (Hawass 1997) was used to compare the classification results of the different algorithms and determine if any algorithms had significantly better classification results than the others. Using the classification score of each algorithm, ROC AUC was assessed for each classifier. ROC curves were then tested using DeLong's test (DeLong *et al* 1988), which determines if statistically significant differences exist between ROC curves. Confidence intervals of AUC, sensitivity, and specificity were obtained based on the different folds of the cross-validation, and standard error was used to determine confidence intervals of PPV and NPV.

Machine learning models were implemented in MATLAB 2017a using the machine learning toolbox and statistical comparisons of model outputs was performed in *R*.

Results

In the whole population of 37 patients, 1751 lesions were identified by physician. This dataset was used to train, test and compare various ML algorithms. Additionally, the subgroup of 14 patients, containing 598 lesions, were analyzed by additional three physicians to assess variability of physician classification of lesions. The physicians took in general over an hour to analyze each patient. The agreement between physicians is shown in table 4. Absolute agreement between physicians for all five classes was only 62%. Fleiss' κ between Physician 2–4 was only moderate (0.56). Agreement was higher for lesions that the physicians were more certain of (Class 1 and 5).

Table 3. Hyperparameter settings for machine learning algorithms.

Algorithm	Hyperparameters
RF	Prior: empirical, trees: 1000, maximal number of decision splits (Leaf size): 1, features sampled per node: 9
Adaboost DT	Prior: empirical, learning cycles: 1000, maximal number of decision splits: 1, minimum number of leaf node observations: 1, split criterion: Gini's diversity index
GLM	Distribution: normal
Neural network	Hidden layers = 10, training function: scaled conjugate gradient backpropagation, performance function: Cross entropy
KNN	Prior: empirical, neighbors: 5, distance: standardized euclidean distance, distance weighting function: equal, number of nearest neighbors: 5
LDA	Prior: empirical, discriminant type: pseudoquadratic
Decision tree	Prior: empirical, maximal number of decision splits: 1, minimum number of leaf node observations:1, split criterion: Gini's diversity index
SVM	Prior: empirical
Naive Bayes	Prior: empirical, data distribution: multivariate multinomial distribution, kernel: normal

According to Cohen's weighted κ, there was significant agreement between the Physician 1 and the consensus of Physician 2–4. 2% (12/598) of lesions were classified as definite metastasis by one physician and as definite benign by another physician.

The number of ROIs identified by the physicians and by the automated lesion detection methods for all of the patients are shown in table 5. The spatial disease distributions created from the whole population are shown in figure 1. Table 6 shows the numbers of features used for each model after removing redundant features if feature selection was performed on the whole population. Table 7 lists the features selected for the SORT model, and the features for the models using the other detection methods can be found in supplementary data (stacks.iop.org/PMB/63/225019/mmedia).

Table 8 contains AUC, sensitivity, and specificity for all of the models. The ROC curves showing performance of the different machine learning algorithms when using statically optimized regional thresholding (SORT) for lesion detection are shown in figure 2. RF had the highest overall performance with AUC = 0.95 95%CI [0.93–0.96], sensitivity = 0.88 [0.86–0.90], specificity = 0.89 [0.87–0.90], PPV = 0.83 [0.82–0.84], and NPV = 0.92 [0.92–0.93]. This superior performance was statistically significant ($p < 0.001$ for both DeLong's test and McNemar test) when compared to all algorithms with the exception of Adaboost DT, GLM, and neural networks.

Figure 3 shows a comparison of the classification performance when using RF with three different lesion detection methods. SORT resulted in significantly superior lesion classification ($P < 0.001$) as compared with the global fixed thresholds. Classification using either of the global thresholds was comparable ($P = 0.69$). Table 5 shows the number of ROIs removed by the models for each identification method.

The best model performance was found when only using the classes with the highest physician certainty (class 1 versus 5, $P = 0.04$) and when grouping background ROIs with benign lesions (classes 0–1 versus 5, $P < 0.0001$). The highest performance also comes from models that use all the aforementioned features, followed by models that only exclude spatial distribution features ($P = 0.7$) or texture features ($P = 0.018$). Using the native imaging resolution when extracting texture resulted in similarly performing models as to upsampling the PET or downsampling the CT ($P > 0.5$). Bayesian optimization of RF hyperparameters (using the dataset segmented with SORT) identified the optimal configuration to include 1719 trees, a minimum leaf size of 1, and 17 features sampled at each node, this model did not perform better than models with suboptimal configurations ($P > 0.5$). Dividing up training and testing based on number of patients instead of number of lesions did not impact model performance. Figure 4 shows the ROC curve with confidence intervals for the highest performing model.

Figure 5 shows classification performance of the RF model when each physician is independently used for training and testing. Models trained and testing using each physician's labels had high classification performance. In particular, models trained and tested on Physicians 2–4 had very similar algorithm performance ($P > 0.7$). The model predicting the labels from Physician 1 had higher sensitivity for benign diseases, but had a lower specificity; however, this was not a significant difference ($P = 0.4$). Additionally, the use of one physician for the training set and predicting a different physician's labels resulted in high AUCs, ranging from 0.90–0.94.

Discussion

The primary challenge for developing automated NaF PET lesion classification is subjective nature of the classification, different physicians may not classify lesions the same way. This is noticeable in our analysis, where

Table 4. Assessment of the agreement of lesion classification between the different physicians.

Assessment metric	Value [95% CI]
Percent agreement between all 4	62%
Fleiss overall κ between cooperating physicians	0.56 [0.54–0.56]
Fleiss κ for definite metastases	0.74
Fleiss κ for likely metastases	0.26
Fleiss κ for equivocal lesions	0.21
Fleiss κ for likely benign lesions	0.41
Fleiss κ for definite benign lesions	0.59
Cohen's weighted κ between independent physicians	0.76 [0.69–0.83]

Table 5. Number of ROIs coming from the different lesion identification method. Additionally for each method the numbers of lesions removed by the classification are shown. Classification performed by physician was manual removal of class 1 lesions. Classification for the other identified lesions was performed using random forest (RF) models.

Number of ROIs	Total		Definite metastases		Definite benign		Background ROI	
Classificaition status	Before	After	Before	After	Before	After	Before	After
Number of lesions identified by physician	1751	1432	925	925	319	0	—	—
Number of ROIs identified by SORT thresholds	2989	1358	938	818	381	65	1031	76
Number of ROIs identified by SUV $> 10\,\mathrm{g\,ml^{-1}}$	3508	1278	1061	802	308	93	1536	108
Number of ROIs identified by SUV $> 15\,\mathrm{g\,ml^{-1}}$	1278	748	791	625	131	24	29	4

there was only moderate agreement between physicians. While in our work lesion detection was performed by a single physician, different physicians would have different sensitivities or specificities for lesion detection, and this would further reduce the agreement between physicians. The large numbers of lesions and moderate inter-observer variability in lesion classification demonstrates the need for automation in lesion classification in NaF PET/CT imaging. Ideally, biopsies or extended follow-up would be used to corroborate the physician findings, but performing biopsies on such a large number of lesions is impractical. Extended follow-up was not available for these subjects, and using extended follow-up as ground truth has its own limitations. Thus, we chose to use the combined experience of four nuclear medicine physicians as our reference.

We tested nine different machine learning algorithms for lesion classification. RF were found to more accurately replicate physician classifications of benign and malignant lesions (AUC $=$ 0.95) than other machine learning algorithms. RF had significant improvement over most models except for Adaboost DT, GLM and neural networks. This implies that there is a complex relationship between benign and malignant lesions, in which there is no feature hyperplane that can be used to cleanly separate benign lesions from malignant lesions. An alternative approach would be to use deep learning through convolutional neural networks. This would reduce the need for assessment of different features and further investigation into the use of deep learning for benign disease classification is merited.

It should be noted that hyperparameters were not investigated for all machine learning models. With default parameters RF outperformed the other models; however, while RF hyperparameters did not improve testing performance of the algorithm it is possible that hyperparameters of the other models could significantly improve the performance of those algorithms. Further investigation might be unnecessary, as it has been shown that RF models often outperform other machine learning algorithms in various tasks (Fernandez-Delgado *et al* 2014).

As can be noted by our training results, RF almost perfectly fit the training data in each fold of cross-validation, which would lead to a concern of overfitting. However, the models maintained high classification performance on the testing data. This is a strength of RF, which have been shown to avoid decreased testing performance, even when perfect training is achieved (Svetnik *et al* 2003). We also assessed the impact of other hyperparameters as well as the number of features necessary. Each of these factors indicate that overfitting was not a concern.

Of the automatic lesion detection methods, the statically optimized regional thresholding (SORT) method resulted in a superior classification model than using the global thresholds. This is likely due to SORT's improved sensitivity and specificity for detecting disease compared to the global thresholds; it detected more disease than either global threshold while outputting fewer background ROIs than SUV $> 10\,\mathrm{g\,ml^{-1}}$ thresholds.

We focused on binary groupings for classification due to the limited numbers of equivocal, likely benign, or likely metastatic lesions (i.e. classes 2–4). Grouping background ROIs with definite benign lesions and performing binary classification of these against definite metastases (classes 0–1 versus 5) resulted in the best algorithm performance. However, the algorithm still had good performance in differentiating lesions of which physicians (class 0–2 versus 4–5) were less certain (AUC $=$ 0.92). Likely the reduced model performance on the uncertain

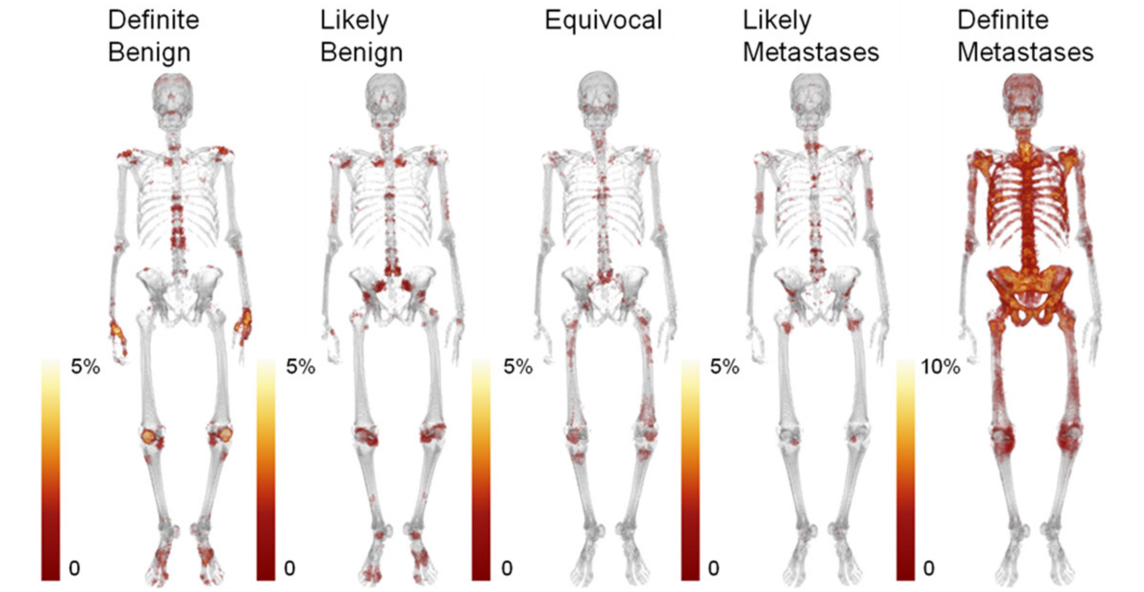

Figure 1. Spatial probability distributions of disease occurrence across the patient population projected onto skeleton renderings. Each voxel represents the percent of patients that had that disease throughout the population.

Table 6. Number of features from each type of feature that could be included in the models after features selection, the location feature was included in all models unless otherwise stated.

Detection method	Location	PET histogram	CT histogram	PET texture	CT texture	Distribution	Total number of features
SORT	1	4	4	10	14	39	73
SUV $>$ 10 g ml^{-1}	1	3	5	14	13	39	75
SUV $>$ 15 g ml^{-1}	1	4	5	12	14	42	78

Table 7. List of all features selected by for the SORT model.

Image feature basis	Features
Location	Location
PET histogram	Standard deviation(PET), skewness(PET), kurtosis(PET), energy(PET)
CT histrogram	Max(CT), mean(CT), integral(CT), standard deviation(CT), kurtosis(CT)
PET texture	Sum mean(PET), sum variance(PET), maximum probability(PET), dissimilarity(PET), sum energy(PET), run percentage(PET), short run low gray-level emphasis(PET), Coarseness(PET), contrast-NGL(PET), busyness(PET)
CT texture	Correlation(CT), inverse difference moment(CT), sum mean(CT), sum variance(CT), information measure of correlation 1(CT), dissimilarity(CT), sum energy(CT), cluster shade(CT), small run emphasis(CT), number nonuniformity(CT), second moment(CT), coarseness(CT), contrast-NGL(CT), busyness(CT)
Disease distribution	Maximum(background occurrence), minimum(background occurrence), mean(background occurrence), minimum(definite metastases occurrence), mean(definite metastases classification), minimum(likely metastases occurrence), mean(likely metastases occurrence), standard deviation(likely metastases occurrence), maximum(likely metastases classification), minimum(likely metastases classification), mean(likely metastases classification), minimum(equivocal occurrence), mean(equivocal occurrence), standard deviation(equivocal occurrence), minimum(equivocal classification), mean(equivocal classification), standard deviation(equivocal classification), maximum(likely benign occurrence), minimum(likely benign occurrence), mean(likely benign occurrence), minimum(likely benign classification), mean(likely benign classification), standard deviation(likely benign classification), mean(definite benign occurrence), standard deviation(definite benign occurrence), maximum(definite benign classification), minimum(definite benign classification), mean(likely/definite metastases occurrence), standard deviation(likely/definite metastases occurrence), minimum(likely/definite metastases classification), standard deviation(likely/definite metastases classification), minimum(likely/definite benign occurrence), maximum(background/likely/definite benign occurrence), minimum(background/likely/definite benign occurrence), mean(background/likely/definite benign occurrence), standard deviation(background/likely/definite benign occurrence), minimum(background/likely/definite benign classification), mean(background/likely/definite benign classification), standard deviation(background/likely/definite benign classification)

Table 8. Summary of model performances on the training and testing sets under different conditions, with 95% CI. In the top section machine learning models were compared and in the remaining sections, the RF model was altered by one parameter.

		Training set			Testing set		
Comparison	Model	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Machine Learning Algorithm	RF	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.95 [0.93–0.96]	0.88 [0.86–0.90]	0.89 [0.87–0.90]
	Adaboost DT	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.95]	0.85 [0.83–0.86]	0.88 [0.85–0.91]
	GLM	0.95 [0.95–0.95]	0.87 [0.87–0.87]	0.90 [0.90–0.90]	0.94 [0.92–0.95]	0.87 [0.86–0.89]	0.87 [0.85–0.90]
	Neural network	0.94 [0.93–0.94]	0.87 [0.86–0.88]	0.87 [0.87–0.88]	0.93 [0.91–0.94]	0.89 [0.87–0.91]	0.84 [0.81–0.87]
	KNN	0.96 [0.96–0.96]	0.93 [0.93–0.93]	0.84 [0.84–0.85]	0.90 [0.89–0.92]	0.88 [0.86–0.90]	0.80 [0.77–0.83]
	LDA	0.91 [0.91–0.91]	0.87 [0.84–0.89]	0.82 [0.80–0.84]	0.87 [0.85–0.90]	0.87 [0.84–0.90]	0.77 [0.73–0.80]
	Decision tree	1.00 [1.00–1.00]	0.98 [0.97–0.98]	0.97 [0.97–0.98]	0.86 [0.84–0.87]	0.82 [0.81–0.84]	0.85 [0.82–0.87]
	SVM	0.82 [0.81–0.82]	0.73 [0.71–0.75]	0.79 [0.76–0.82]	0.82 [0.80–0.84]	0.69 [0.65–0.74]	0.83 [0.80–0.85]
	Naive bayes	0.97 [0.97–0.97]	0.89 [0.88–0.90]	0.89 [0.88–0.90]	0.77 [0.74–0.80]	0.80 [0.78–0.82]	0.67 [0.62–0.71]
Lesion Detection Method	SORT	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.95 [0.93–0.96]	0.88 [0.86–0.90]	0.89 [0.87–0.90]
	SUV > 10	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.89 [0.87–0.90]	0.93 [0.92–0.94]	0.71 [0.68–0.74]
	SUV > 15	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.88 [0.87–0.90]	0.76 [0.69–0.82]	0.87 [0.85–0.89]
Physician Certainty	0–2 versus 4–5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.92 [0.91–0.93]	0.83 [0.82–0.85]	0.86 [0.84–0.89]
	0–3 versus 4–5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.92 [0.91–0.92]	0.84 [0.83–0.86]	0.83 [0.81–0.84]
	1–2 versus 4–5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.89 [0.88–0.90]	0.85 [0.83–0.87]	0.76 [0.73–0.79]
	1–3 versus 4–5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.88 [0.87–0.89]	0.84 [0.83–0.85]	0.76 [0.74–0.78]
	0–1 versus 5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.95 [0.93–0.96]	0.88 [0.86–0.90]	0.89 [0.87–0.90]
	1 versus 5	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.93 [0.92–0.94]	0.84 [0.83–0.86]	0.85 [0.84–0.86]
Features Included	CT	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.71 [0.70–0.73]	0.71 [0.69–0.73]	0.64 [0.61–0.66]
	PET	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.89 [0.88–0.90]	0.87 [0.85–0.89]	0.80 [0.77–0.82]
	Spatial distribution	0.99 [0.99–0.99]	0.95 [0.95–0.95]	0.93 [0.93–0.94]	0.85 [0.84–0.86]	0.68 [0.67–0.70]	0.86 [0.84–0.87]
	PET/CT	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.90 [0.89–0.91]	0.89 [0.87–0.90]	0.76 [0.73–0.79]
	Simple features[a]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.93 [0.92–0.94]	0.92 [0.90–0.94]	0.76 [0.74–0.78]
	No texture features	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.93 [0.92–0.94]	0.89 [0.88–0.90]	0.82 [0.81–0.84]
	PET/CT and location	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.95]	0.87 [0.85–0.88]	0.87 [0.85–0.90]
	All	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	**0.95 [0.93–0.96]**	**0.88 [0.86–0.90]**	0.89 [0.87–0.90]
RF Hyperparameters	$T = 1729$, LS = 1, $F = 17$[b]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.96]	0.87 [0.85–0.89]	0.88 [0.85–0.90]
	$T = 1000$, LS = 1, $F = 9$	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.95 [0.93–0.96]	0.88 [0.86–0.90]	0.89 [0.87–0.90]
	$T = 115$, LS = 2, $F = 2$[c]	1.00 [1.00–1.00]	0.98 [0.98–0.99]	0.99 [0.99–0.99]	0.94 [0.93–0.95]	0.88 [0.87–0.90]	0.86 [0.84–0.88]
	$T = 408$, LS = 11, $F = 61$	0.99 [0.99–0.99]	0.94 [0.94–0.94]	0.94 [0.94–0.95]	0.94 [0.94–0.95]	0.87 [0.85–0.89]	0.88 [0.86–0.90]
	$T = 1000$, LS = 1, $F =$ All	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.95]	0.92 [0.91–0.93]	0.83 [0.80–0.85]
Cross-validation Method	Equal patients per fold	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.95]	0.86 [0.82–0.90]	0.87 [0.85–0.89]
	Equal lesions per fold	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.95 [0.93–0.96]	0.88 [0.86–0.90]	0.89 [0.87–0.90]
	1	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.93 [0.91–0.95]	0.93 [0.91–0.94]	0.78 [0.73–0.82]

(*Continued*)

Table 8. (*Continued*)

		Training set			Testing set		
Comparison	Model	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Physician Ground Truth	2	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.92–0.95]	0.81 [0.79–0.83]	0.94 [0.91–0.97]
	3	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.93 [0.92–0.95]	0.80 [0.77–0.83]	0.94 [0.92–0.96]
	4	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.92–0.95]	0.79 [0.76–0.81]	0.96 [0.95–0.97]
	Consensus	1.00 [1.00–1.00]	1.00 [1.00–1.00]	1.00 [1.00–1.00]	0.94 [0.93–0.95]	0.83 [0.81–0.85]	0.91 [0.88–0.93]

[a] Simple features are the global level features from PET or CT and the location feature.

[b] Optimal model identified by Bayesian optimization,T represents number of trees, LS represents minumum leaf size and F represents the number of features sampled per node.

[c] Worst model identified by Bayesian optimization.

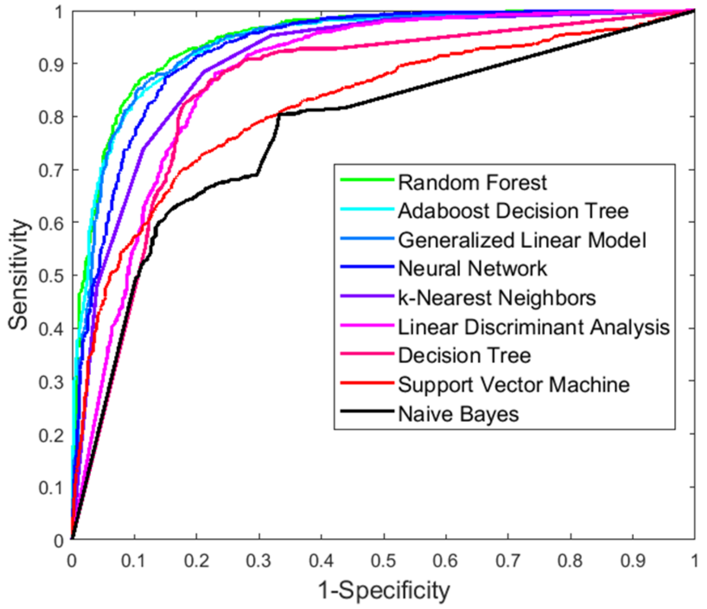

Figure 2. ROC curves of the different machine learning algorithms with lesions identified using statically optimized regional thresholding (SORT). The RF model had the highest AUC.

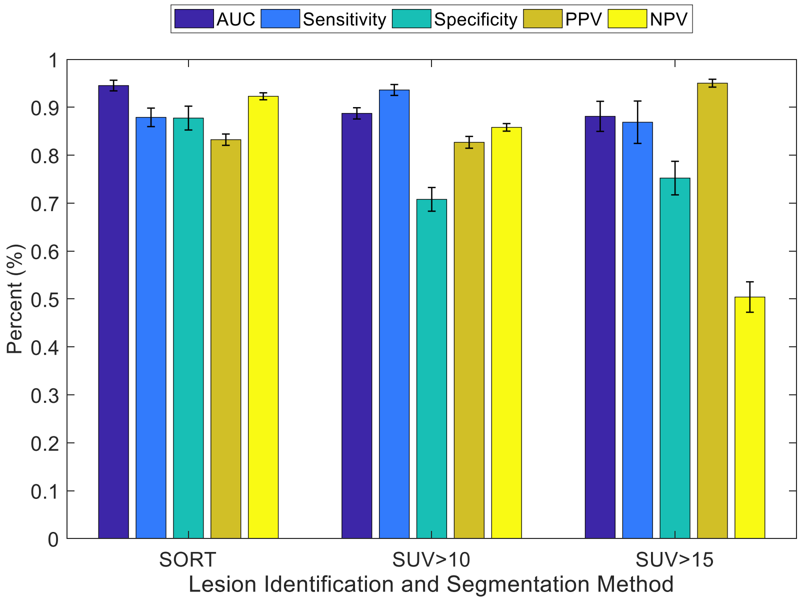

Figure 3. Performance statistics of random forest (RF) models trained using the different lesion identification and segmentation methods. Data resulting from using SORT thresholds resulted in a superior classification model ($P < 0.05$) than data resulting from the other thresholding methods.

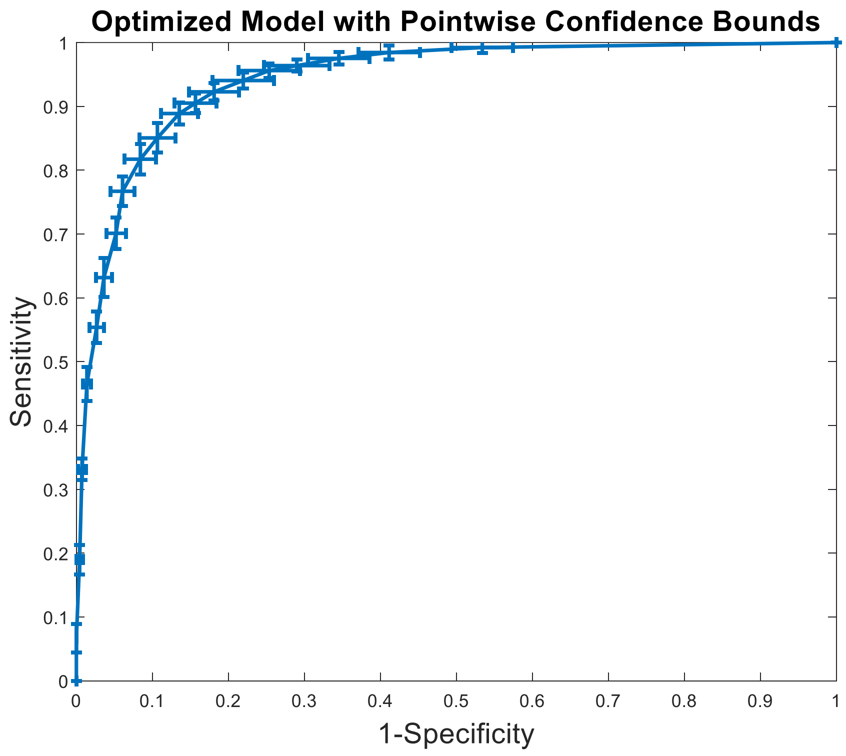

Figure 4. ROC curve of the random forest (RF) trained using SORT to identify lesions, on class 0–1 versus 5, including all features, and using the hyperparameters: 1719 trees, a minimum leaf size of 1 and 17 features sampled at each node. Confidence bounds come from ten-fold cross-validation.

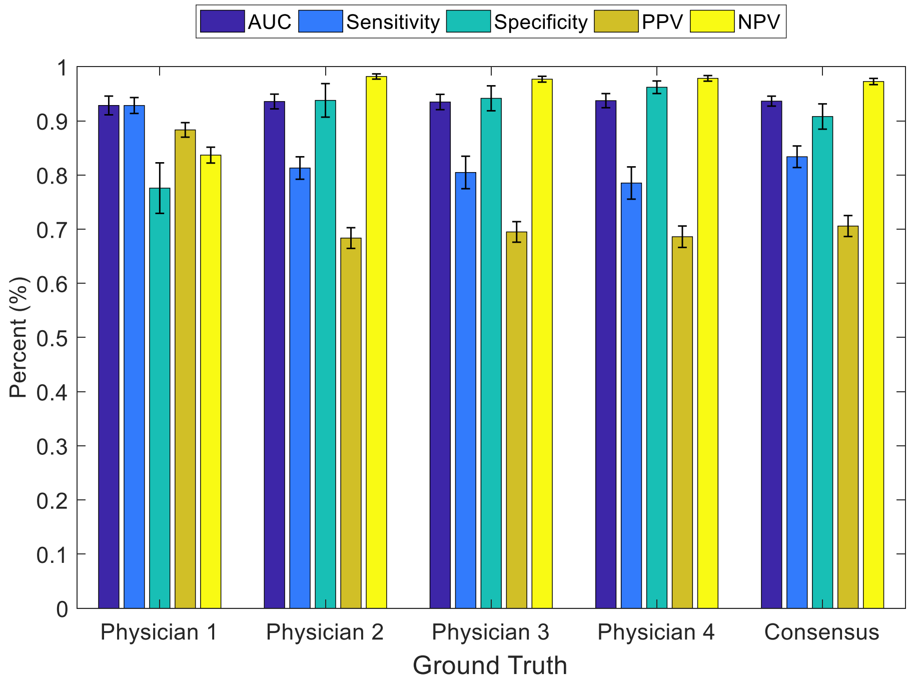

Figure 5. Performance characteristics of random forest models trained using different physician labels for training and testing of the model. The consensus category represents the consensus of Physicians 2–4. All models had AUC $>$ 0.93.

lesions is a result of physician inconsistency. An additional benefit of grouping background ROIs with metastases is the reduction of the false positive rate of the automated lesion detection methods.

Using only PET features or using only CT features resulted in worse classification performance than when using both sets of features. The highest performance was found using all of PET, CT, spatial distribution, and texture features. Interestingly, this fully inclusive model was not statistically superior to the model that included all features except the spatial distribution features. This suggests that the spatial distribution features may not be necessary to achieve a top performing classification model and that in cases where the model speed would be optimized these more complicated features could be excluded. Texture features appear to be another set of features that provided minimal gain to the model, even though excluding texture resulted in a significant decrease in performance. Additionally, as the impact of voxel size for extraction of image texture was minimal on the classification performance of RF ($P > 0.5$), the CT can be downsampled to reduce computation time of extracting texture.

An interesting trend was found when we performed cross-validation to select equal numbers of patients in each training and testing dataset rather than equal numbers of lesions. We found that the underlying patient level correlation of lesions from the same patient did not provide a significant improvement in model performance. This likely implies that the RF model did not use any patient specific information in the classification.

With the different physician classifications available to us, we were able to assess if the model performance is dependent on the physician that performed the classification. Our model had high performance on predicting how each physician classified lesions. This implies that the model is able to replicate tendencies in each physician's classification performance.

This study had several limitations. As discussed, the primary limitation is the lack of ground truth classification labels. However, as we were able to create models that could replicate the scoring tendencies of each physician, if a ground truth became available in the future, we expect the model could be trained to learn those classifications as well. Another limitation is the use of a simple lesion segmentation method. With a method that accurately determines lesion boundaries, the splitting of lesions that were merged together during lesion detection would not need to be performed and the quantification of each ROI would be more accurate. We hypothesize that this would result in superior classification performance of machine learning models. However, even the current level of performance of our algorithm exceeds any available alternative.

Conclusion

We have developed the first automated lesion classification tool for NaF PET/CT images. This tool, when combined with the automated lesion detection tool, SORT, can allow physicians to quickly and accurately classify and analyze lesions in NaF PET/CT images. We found that RF outperformed other classification algorithms when classifying lesions in NaF PET/CT images. When the model was trained and tested using classification labels from four different physicians, it maintained high classification performance for predicting each physician's labels.

Acknowledgments

We would like to thank the patients who volunteered their time and imaging technologists for data acquisition. We would like to thank Erin Meyer and Saima Muzahir for their contributions to this work during its initial stages. We would like to thank the UW Carbone Cancer Center, Madison, WI for providing support to this work (Support Grant P30 CA014520). This study was supported by the Prostate Cancer Foundation (PCF) through the PCF Creativity Award (Liu, Jeraj) and PCF Mazzone Challenge Award (Jeraj, Liu), and conducted within the Prostate Cancer Clinical Trials Consortium (PCCTC). Additional support provided by the United States Department of Defense Prostate Cancer Research Program (W81XWH-17-0020).

ORCID iDs

Timothy Perk https://orcid.org/0000-0002-9906-5087
Tyler Bradshaw https://orcid.org/0000-0001-9549-7002
Song Chen https://orcid.org/0000-0002-2639-6301

References

Burr D B 1998 The importance of subchondral bone in osteoarthrosis *Curr. Opin. Rheumatol.* **10** 256–62

Cook G and Fogelman I 2000 The role of positron emission tomography in the management of bone metastases *Cancer* **88** 2927–33

Delong E R, Delong D M and Clarke-Pearson D L 1988 Comparing the areas under two more more correlated receiver operating characteristic curves: a nonparametric apporach *Biometrics* **44** 837–45

Even-Sapir E, Metser U, Mishani E, Lievshitz G, Lerman H and Leibovitch I 2006 The detection of bone metastases in patients with high-risk prostate cancer: ^{99m}Tc-MDP planar bone scintigraphy, single- and multi-field-of-view SPECT, ^{18}F-fluoride PET, and ^{18}F-fluoride PET/CT *J. Nucl. Med.* **47** 287–97

Fernandez-Delgado M, Cernadas E, Barro S and Amorim D 2014 Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–81

Galavis P E, Hollensen C, Jallow N, Paliwal B and Jeraj R 2010 Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters *Acta Oncol.* **49** 1012–6

Harmon S A *et al* 2017 Quantitative assessment of early [^{18}F] sodium fluoride positron emission tomography/computed tomography response to treatment in men with metastatic prostate cancer to bone *J. Clin. Oncol.* **35** 2829–37

Harmon S *et al* 2016 Mo-Ab-Bra-05: [^{18}F] NaF PET/CT imaging biomarkers in metastatic prostate cancer *Med. Phys.* **43** 3691

Hawass N E 1997 Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients *Br. J. Radiol.* **70** 360–66

Horn B K P and Schunck B G 1981 Determining optical-flow *Artif. Intell.* **17** 185–203

Jallow N 2014 Comprehensive assessment of uncertainties and errors in [^{18}F]NaF PET imaging *PhD Thesis* University of Wisconsin-Madison

Kurdziel K A *et al* 2012 The kinetics and reproducibility of ^{18}F-sodium fluoride for oncology using current PET camera technology *J. Nucl. Med.* **53** 1175–84

Li Y, Schiepers C, Lake R, Dadoarvar S and Berenji G 2012 Clinical utility of ^{18}F-fluoride PET/CT in benign and malignant bone diseases *Bone* **50** 128–39

Lin C *et al* 2016 Repeatability of quantitative ^{18}F-NaF PET: a multicenter study *J. Nucl. Med.* **57** 1872–9

Lindgren Belal S *et al* 2017 3D skeletal uptake of ^{18}F sodium fluoride in PET/CT images is associated with overall survival in patients with prostate cancer *EJNMMI. Res.* **7** 15

Munoz H E, Yao J, Burns J E and Summers R M 2013 Detection of vertebral degenerative disc disease based on cortical shell unwrapping *SPIE Proc.* **8670** 86700C

Muzahir S, Jeraj R, Liu G, Hall L T, Rio A M, Perk T, Jaskowiak C and Perlman S B 2015 Differentiation of metastatic versus degenerative joint disease using semi-quantitative analysis with ^{18}F-NaF PET/CT in castrate resistant prostate cancer patients *Am. J. Nucl. Med. Mol. Imaging* **5** 162–8

Perk T, Bradshaw T, Harmon S, Perlman S, Liu G and Jeraj R 2015 Su-D-303-01: spatial distribution of bone metastases in metastatic castrate-resistant prostate cancer *Med. Phys.* **42** 3214–5

Perk T, Chen S, Harmon S, Lin C, Bradshaw T, Perlman S, Liu G and Jeraj R A statistically optimized regional thresholding method (SORT) for bone lesion detection in ^{18}F-NaF PET/CT imaging *Phys. Med. Biol.* **63** 225018

Punzi L, Ramonda R and Sfriso P 2004 Erosive osteoarthritis *Best Pract. Res. Clin. Rheumatol.* **18** 739–58

Rohren E M, Etchebehere E C, Araujo J C, Hobbs B P, Swanston N M, Everding M, Moody T and Macapinlac H A 2015 Determination of skeletal tumor burden on ^{18}F-fluoride PET/CT *J. Nucl. Med.* **56** 1507–12

Rosen R S, Fayad L and Wahl R L 2006 Increased ^{18}F-FDG uptake in degenerative disease of the spine: characterization with ^{18}F-FDG PET/CT *J. Nucl. Med.* **47** 1274–80

Sabbah N, Jackson T, Mosci C, Jamali M, Minamimoto R, Quon A, Mittra E S and Iagaru A 2015 ^{18}F-sodium fluoride PET/CT in oncology: an atlas of SUVs *Clin. Nucl. Med.* **40** E228–31

Svetnik V, Liaw A, Tong C, Culberson J C, Sheridan R P and Feuston B P 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling *J. Chem. Inf. Model.* **43** 1947–58

Wang C Y and Shen Y 2012 Study on the distribution features of bone metastases in prostate cancer *Nucl. Med. Commun.* **33** 379–83

Wang Y, Yao J, Burns J E, Liu J and Summers R M 2016 Detection of degenerative osteophytes of the spine on PET/CT using region-based convolutional neural networks *Computational Methods and Clinical Applications for Spine Imaging: 4th Int. Workshop and Challenge, Csi 2016, Held in Conjunction with Miccai 2016* (*Athens, Greece, 17 October 2016*) ed J Yao *et al* (Cham: Springer) (Revised Selected Papers) pp 116–24

Yao J, Munoz H, Burns J E, Lu L and Summers R M 2014 Computer aided detection of spinal degenerative osteophytes on sodium fluoride PET/CT *Computational Methods And Clinical Applications For Spine Imaging* (New York: Springer) pp 51–60

Yip S, Perk T and Jeraj R 2014 Development and evaluation of an articulated registration algorithm for human skeleton registration *Phys. Med. Biol.* **59** 1485–99