TESTING THE ABILITY OF CONVOLUTIONAL NEURAL NETWORKS TO LEARN RADIOMIC FEATURES

Ivan S. Klyuzhin[1,2,3*] Yixi Xu[2] Anthony Ortiz[2] Juan M. Lavista Ferres[2]
Ghassan Hamarneh[4] Arman Rahmim[1,3]

[1] Department of Integrative Oncology, BC Cancer Research Institute, Vancouver, BC, Canada
[2] AI for Health, Microsoft, Redmond, WA, USA
[3] Department of Radiology, University of British Columbia, Vancouver, BC, Canada
[4] Department of Computing Science, Simon Fraser University, Burnaby, BC, Canada

ABSTRACT

Purpose: To test the ability of convolutional neural networks (CNNs) to effectively capture the intensity, shape, and texture properties of tumors as defined by standardized radiomic features.

Methods: Standard 2D and 3D CNN architectures with an increasing number of convolutional layers (up to 9) were trained to predict the values of 16 standardized radiomic features from synthetic images of tumors, and tested. In addition, several ImageNet-pretrained state-of-the-art networks were tested. The synthetic images replicated the quality of real PET images. A total of 4000 images were used for training, 500 for validation, and 500 for testing.

Results: Radiomic features quantifying tumor size and intensity were predicted with high accuracy, while shape irregularity features had very high prediction errors and generalized poorly between training and test sets. For example, mean normalized prediction error of tumor diameter (mean intensity) with a 5-layer 2D CNN was 4.23 ± 0.25 (1.88 ± 0.07), while the error for tumor sphericity was 15.64 ± 0.93. Similarly-high error values were found with other shape irregularity and heterogeneity features, both with standard and state-of-the-art networks.

Conclusions: Standard CNN architectures and ImageNet-pretrained advanced networks have a significantly lower capacity to capture tumor shape and heterogeneity properties compared to other features. Our findings imply that CNNs trained end-to-end for clinical outcome prediction and other tasks may under-utilize tumor shape and texture information. We hypothesize, that to improve CNN performance, these radiomic features can be computed explicitly and added as auxiliary variables to the dense layers in the networks, or as additional input channels.

***Keywords*:** Deep learning · Radiomics · Cancer · Imaging · Image analysis

[*] Corresponding author: ivan.corr@outlook.com

1 Introduction

Quantitative pattern analysis in radiological images can be used to assess tumor phenotype as well as micro- and macro-environmental conditions [1]. For example, larger and more heterogeneous tumors as measured from PET/CT images have been found to be generally more aggressive and more resilient to treatment [2, 3], while more irregular tumor shapes have been associated with a lower probability of complete response [4]. Given these findings, there have been considerable efforts to develop novel pattern analysis methods for medical imaging. Two distinct approaches have emerged: radiomics and deep learning. Radiomics-based methods utilize hand-crafted features that are intended to capture various properties of the tumor, e.g. its shape and texture [5, 1]. Various radiomic features have been found to be significant predictors of disease-free survival and response to therapy [6, 7, 8]. Deep learning methods in medical imaging typically utilize convolutional neural networks (CNN) trained in an end-to-end fashion, with images serving as inputs and clinical metrics as targets. In the process of training, relevant low- and high-level image features become automatically and implicitly encoded in the layers of the network [9]. Thus, deep learning methods eliminate the need for feature design and selection, and can forego the need for image segmentation [8].

Recent reports of human-level cancer detection performance by CNNs [10, 11, 12, 13] may suggest that emphasis in method development should be placed on deep learning, rather than radiomics. According to the universal approximation theorem [14, 15], hand-crafted radiomic features represent a subset of functions that CNNs can approximate, seemingly obviating the practice of using explicit radiomics for predictive tasks. The problem, however, is that the theorem does not provide any bounds on the required number of neurons to approximate a function: the necessary number of CNN layers or nodes to match the power of a hand-crafted feature may well be impractical. Sample complexity is another concern: the number of samples required to learn a particular feature may be unrealistic, or vary substantially between the features, leading to significant biases in learning of different kinds of information (e.g. texture versus shape) [16].

In the present work, we directly test the ability of CNNs to learn hand-crafted and standardized radiomic features, and measure the sample complexity for different features. To that end, we train standard CNN architectures with a progressively larger number of convolutional layers (up to nine), and several state-of-the-art (SOTA) network architectures, to predict the values of radiomic features for tumor images. A poor prediction accuracy for a particular feature would imply that common CNN architectures may be unable to effectively capture and use the corresponding type of information.

2 Materials and Methods

2.1 Image synthesis

Synthetic 2D and 3D lesion images for CNN training and testing were generated procedurally. We will describe our methodology using the 2D case for brevity; in the 3D case, all aspects of methodology were symmetrically extended into the third dimension. First, a binary region representing the mask of the lesion was generated. To create a variety of mask shapes and sizes, a stochastic region growth algorithm mimicking the typical growth patterns of tumor was used, starting from 1 to 3 seeds that were randomly placed within a binary 48×48-pixel image; using a random number of initial seeds increased the variance of shape features, as confirmed in a post-hoc analysis. A random number (300-550) of region growth iterations was applied (Fig. 1a), and the resulting image was morphologically closed to remove small holes inside the mask.

The lesion texture was created by generating a random Perlin pattern and masking it using the generated mask. The pixel intensities were set to represent PET standardized uptake values (SUV), and were scaled to vary between SUV_{min}, chosen randomly between 2 and 7 for each image, and SUV_{max},

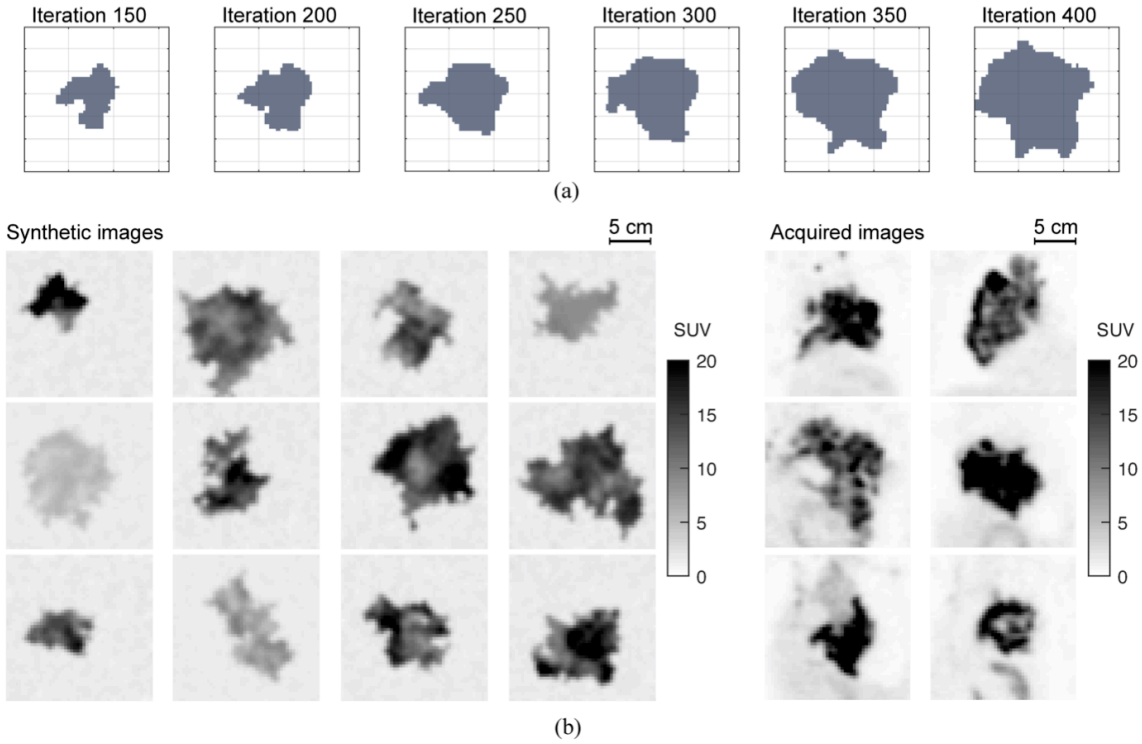

Figure 1: (a) Illustration of the region growth process utilized to generate random lesion shapes. (b) Examples of (left) synthetic lesion images with different lesion intensities, shapes, and textures in comparison to (right) real acquired images of lymphoma tumors. The resolution and noise in synthetic images were matched to those of acquired images. The plotted synthetic and acquired images have the same dimensions (48×48 pixels) and pixel size (3.64 mm).

chosen randomly between 9 and 30; SUV values in the background were set to 1.5. Magnitude-independent Gaussian noise (sigma = 0.15 SUV units) was added everywhere in the image, and spatial Gaussian smoothing (sigma = 0.85 pixels or 3.1 mm) was applied to the entire simulate resolution blurring. The pixel size (3.64 mm), resolution, and noise values were set to match those of the clinical images acquired at our center on a GE Discovery 690 scanner.

The SUV values and lesion sizes in the synthetic images were set to be similar to real PET images of primary mediastinal large B-cell lymphoma (Fig. 1b) obtained with ^{18}F-fludeoxyglucose. In a set of 30 pre-treatment lymphoma images that was available for reference, the SUV_{max} (SUV_{min}) was 23.0 ± 7.2 (4.5 ± 1.8), and the background signal-to-noise ratio was approximately 10%.

A total of 5000 images were generated. Due to the stochastic nature of the algorithm, no two images in the dataset had the same tumor shape or texture.

2.2 Radiomic features

A set of 16 intensity, shape, and texture features, as defined by the Image Biomarker Standardization Initiative (IBSI) [17], was selected for this study (Table 1). The choice of features was based on their simplicity, interpretability and frequency of use in research and clinical practice.

As per IBSI, the 4 intensity features describe first order pixel value statistics within the lesion mask, in SUV units. The coefficient of variation (COV), often used as a measure of lesion heterogeneity [18, 3], was computed as the ratio of the standard deviation to the mean.

The shape features include 4 descriptors of size and 4 descriptors of the shape irregularity. These features do not take into account the pixel intensities or their spatial distributions. Convex area was

Table 1: Radiomic features computed from the synthetic images.

Feature name	Median (Q_1, Q_3)	(min, max)
Intensity features		
Maximum	16.95 (12.51, 21.28)	(6.84, 28.30)
Mean	11.03 (8.56, 13.41)	(4.86, 18.92)
Variance	5.65 (2.68, 9.87)	(0.60, 29.73)
COV	0.21 (0.18, 0.24)	(0.11, 0.43)
Shape features - size		
Area	6.12 (4.21, 8.34) $\times 10^3$	(0.808, 13.23) $\times 10^3$
Convex area	7.28 (5.03, 9.95) $\times 10^3$	(0.833, 16.72) $\times 10^3$
Max diameter	116.54 (97.94, 136.29)	(38.00, 195.34)
Perimeter	381.88 (308.20, 457.69)	(99.42, 690.07)
Shape features - irregularity		
Sphericity	0.69 (0.65, 0.73)	(0.45, 0.90)
Elongation	0.75 (0.66, 0.83)	(0.32, 1.00)
Solidity	0.85 (0.82, 0.87)	(0.62, 0.97)
Extent	0.59 (0.55, 0.63)	(0.39, 0.79)
Texture features		
Contrast	8.59 (7.33, 10.15)	(4.03, 23.46)
Energy	0.8 (0.7, 1.1) $\times 10^{-2}$	(0.4, 4.4) $\times 10^{-2}$
Homogeneity	0.44 (0.41, 0.47)	(0.30, 0.60)
Entropy	7.37 (7.11, 7.59)	(5.52, 8.22)

Statistics computed from 5000 samples. Area and Convex area are given in mm^2, Max diameter and Perimeter are in mm. Intensity features are in SUV units.

defined as the area of the convex envelope of the mask. Solidity is the ratio of lesion's area to the convex area. Extent was defined as the ratio of lesion's area to that of the axis-aligned bounding rectangle.

Texture features are represented by 4 second-order Haralick features computed from the gray level co-occurrence matrix; the pixel intensities were quantized using the constant bin number technique (32 bins). The texture features were computed within the mask.

All features except perimeter and sphericity were computed using the IBSI-compliant SERA radiomics software [19]. For CNN training and testing the features were normalized by subtracting the mean and dividing by one standard deviation.

2.3 Tested neural net architectures

We trained and tested several standard convolution-nonlinearity-pooling (CNP) architectures with an increasing number of convolutional layers (Table 2). The rectified linear unit (ReLU) nonlinearity was used throughout each CNP network, and max-pooling was used as the downsampling operation. After the flattening layer, all networks included one dense layer with 16 nodes, followed by a regression output layer. All parameters were trained. Since the number of max-pooling layers, flattened features, and dense layer nodes were fixed, the number of trainable parameters depended solely on the number of convolutional layers. In all convolutional filters, isotropic kernels of the size 5×5 were used; the kernel size of max-pooling layers was 2×2.

Additionally, several SOTA CNN architectures were tested that included non-standard computation blocks and connections (Table 2). The ImageNet pre-trained models were downloaded from

Table 2: Parameters of the tested CNP and SOTA CNNs.

Standard CNP Networks				
Network	Num. layers (Conv. layers)	Layer structure	Trainable parameters	Num. features
CN-2D-3	9 (3)	c-m-c-m-c-m-f-d-r	11,649	512
CN-2D-5	11 (5)	cc-m-cc-m-c-m-f-d-r	14,865	512
CN-2D-7	13 (7)	ccc-m-cc-m-cc-m-f-d-r	18,081	512
CN-2D-9	15 (9)	ccc-m-ccc-m-ccc-m-f-d-r	21,297	512

SOTA Networks				
Network	Num. layers	Total parameters	Trainable parameters	Num. features
MobileNetV2	155	2,259,265	1,281	1,280
NASNetMobile	769	4,270,773	1,057	1,056
DenseNet201	707	18,323,905	1,921	1,920
Xception	132	20,863,529	2,049	2,048
InceptionV3	311	21,804,833	2,049	2,048
InceptionResNetV2	780	54,338,273	1,537	1,536

The CN-2D-*X* abbreviations denote different CNP networks, where *X* is the number of convolutional layers; each convolutional layer consisted of 8 filters. The "Layer structure" column shows the layer sequence for each network: c = convolution, m = max pooling, f = flattening, d = dense, r = regression layer. The "Num. features" column contains the number of flattened features entering the dense layer.

the TensorFlow model repository. The total number of parameters ranged from ~2M to ~54M. The pre-trained head (dense layers) of each network was removed, and a single new regression layer was added and trained, with the rest of the network frozen. The resulting number of trainable parameters was between 1057 and 2049, depending on the number of flattened features after the last convolutional layer.

The inputs to CNP and SOTA networks were the synthetic SUV images, and the target variables were the normalized values of the radiomic features. One feature was tested at a time, i.e. each network only had one regression output. The networks were implemented in Python using the Keras module within TensorFlow v.2.2.

2.4 Network training and testing

The networks were trained in end-to-end using the AdaGrad algorithm, with the base learning rate set to 0.01. The minimized loss function was the mean absolute error between predicted and ground truth values of radiomic features. Training was performed for 200 epochs, in mini-batches of 32 images. Out of 5000 generated images, 4000 were used for training, 500 for validation, and 500 for testing. The differences in feature distributions between the sets were insignificant, as they were generated using the same process.

The test set was used to assess the efficacy of CNP and SOTA networks in learning radiomic features. Two metrics were used to quantify prediction error: 1) the normalized mean absolute error (nMAE):

$$nMAE = \frac{100\%}{N} \frac{\sum_{i=1}^{N} |y_i^{\text{pred}} - y_i|}{\text{pRange}(y_i)} \tag{1}$$

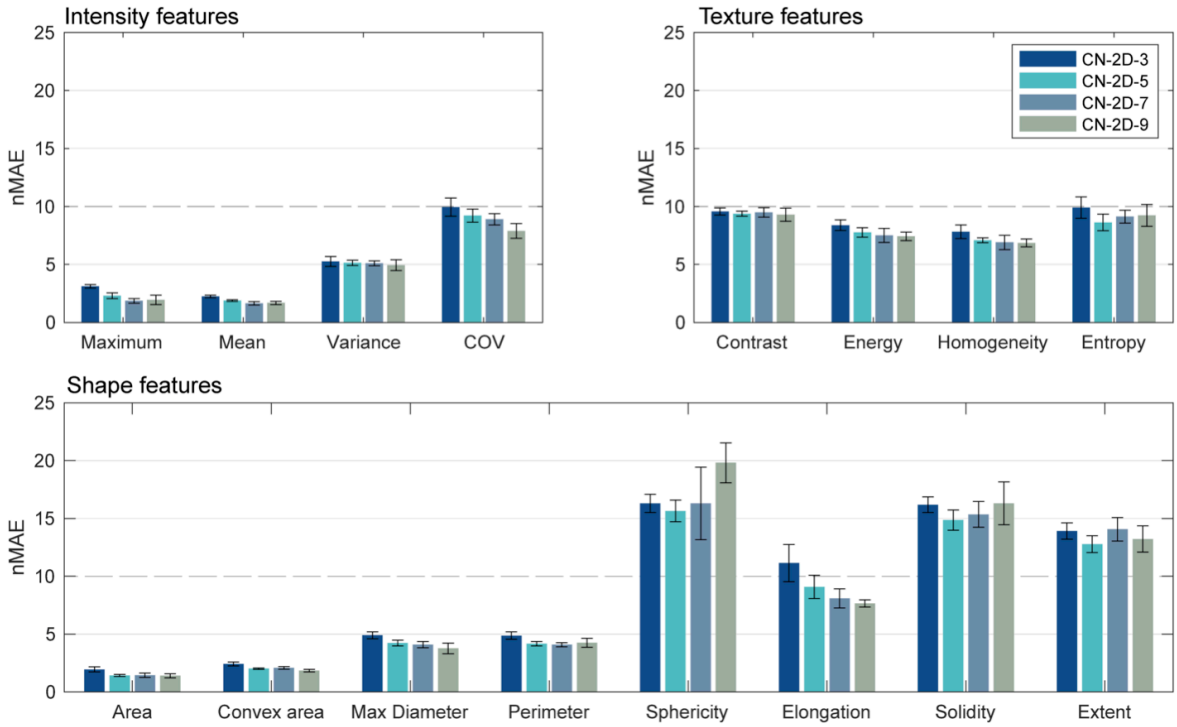

Figure 2: Radiomic feature prediction errors (nMAE) for the CNP networks. The mean values and standard deviations were measured using 5 independent training trials.

where N is the number of test samples, y_i is the true feature value, y_i^{pred} is the predicted feature value, and pRange is the percentile range (2.5%–97.5%); 2) Spearman's correlation coefficient ρ between the ground truth and network-predicted feature values. The values of nMAE and ρ were computed on the test set, for all tested features and networks, as reported below.

To analyze the sample complexity for different features, we trained the CNP networks using different numbers of samples, ranging from 100 to 4500. The test loss and the difference between the training and test loss (i.e. generalization) were measured as functions of the number of samples.
Two additional tests were performed to aid the interpretation of results: 1) binary masks were tested as CNN inputs to investigate the effect of contrast on learning of shape features; 2) a dataset with fixed lesion size was tested to prevent networks from using size as a proxy for shape. We briefly report the results of these auxiliary tests.

3 Results

3.1 Standard CNP networks

Radiomic feature prediction errors on the test set are plotted in Figure 2 for all 2D CNP networks. Scatter plots of true versus predicted values obtained with the CN-2D-5 network are plotted in Figure 3. The lowest values of nMAE, approximately in the range from 1 to 5, were measured with size features (Area, Convex area, Max diameter, Perimeter), and with the Mean and Maximum intensity features. The texture prediction errors were intermediate, in the range from 7 to 10. The highest prediction errors (nMAE range from 13 to 20) were measured for features that quantified shape irregularity – Sphericity, Solidity, and Extent. Notably, while Area and Convex area were predicted with low errors, their ratio defined as Solidity was predicted with high errors. The improvement in prediction performance with added

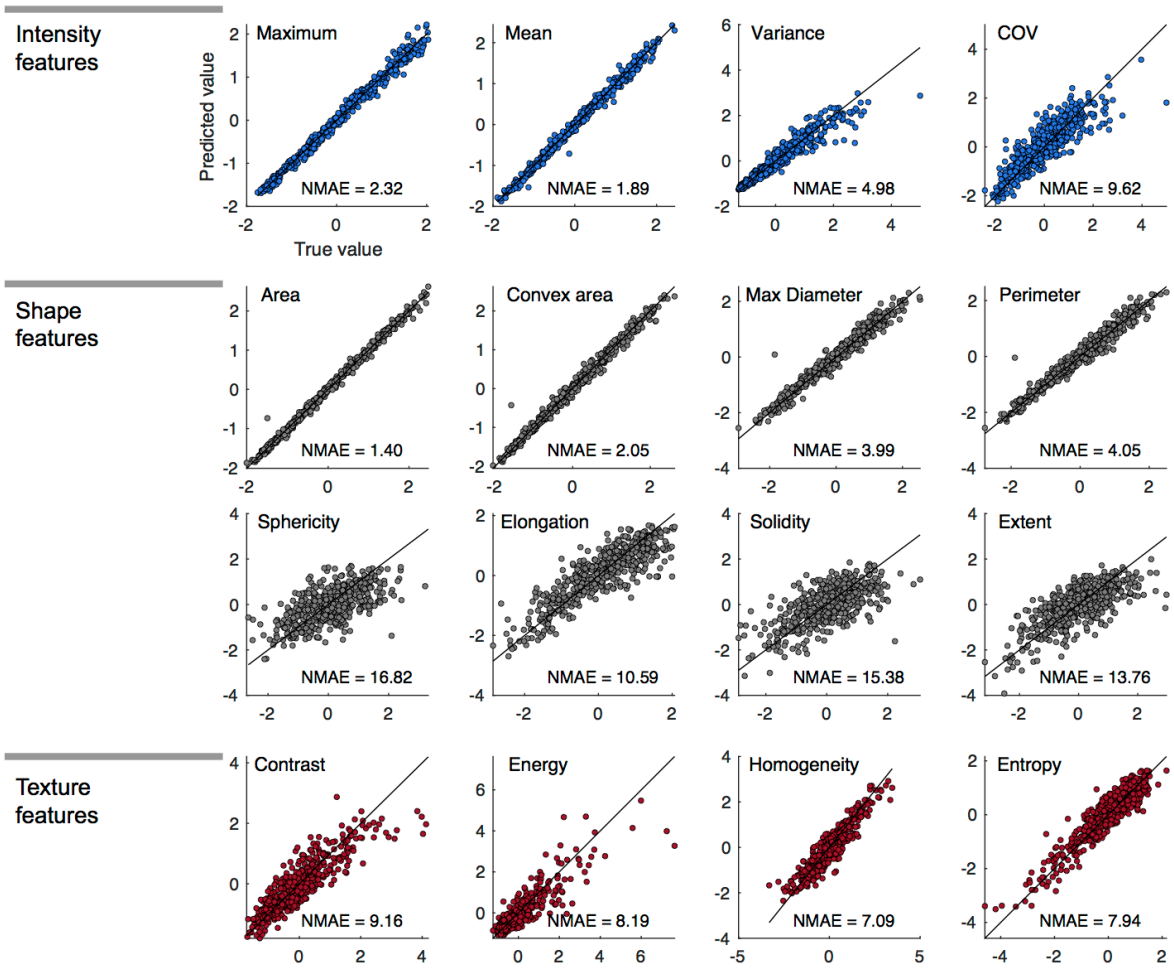

Figure 3: Predicted feature values (normalized, y-axes) plotted against true feature values (normalized, x-axes) in the test set of 500 samples, for the CN-2D-5 network. The identity line is plotted in solid black color.

convolutional layers was either small or insignificant. The only 2 features with significant improvement were COV and Elongation. Prediction error for Sphericity was highest with the CN-2D-9 network due to increased overfitting.

The predicted-vs-true value scatter plots for Sphericity, Solidity and Extent (Fig. 3) demonstrate that the high error values did not originate from outliers or biases. Indeed, the data points for these features are substantially more scattered around the identity line compared to features like the Maximum, Mean, Area, and Homogeneity. The high values of Contrast and Energy were predicted with relatively high errors, likely due to a low representation of such values in the training set.

Spearman's rank correlation coefficients ρ between predicted and true feature values for 2D networks are listed in Table 3; greater values correspond to better prediction performance. The relative standing of features in terms of ρ was similar to that of nMAE: Sphericity, Solidity, and Extent had distinctly and significantly lower ρ values compared to other features.

Results obtained with 3D CNP networks were very similar to 2D networks (Fig. E1 and Table E1 [supplement]). The intensity and size features were predicted with relatively low errors, while shape features were predicted with high errors.

To further investigate the high prediction errors for the shape irregularity features, training and test losses were inspected as functions of the training epoch. For illustration, training and test losses for the CN-2D-5 network are plotted in Fig. 4. After 200 epochs, the test loss had converged for all tested

Table 3: Spearman's rank correlation coefficients (ρ) between predicted and true feature values for the CNP networks.

Feature name	CN-2D-3	CN-2D-5	CN-2D-7	CN-2D-9
Intensity features				
Maximum	0.99	1.00	1.00	1.00
Mean	0.99	1.00	1.00	1.00
Variance	0.98	0.98	0.98	0.98
COV	0.86 ± 0.03	0.88 ± 0.01	0.89 ± 0.01	0.91 ± 0.02
Shape features				
Area	1.00	1.00	1.00	1.00
Convex area	0.99	0.99	0.99	1.00
Max diameter	0.97	0.98	0.98	0.98
Perimeter	0.97	0.98	0.98	0.98
Sphericity	**0.65 ± 0.03**	**0.68 ± 0.04**	**0.66 ± 0.13**	**0.54 ± 0.07**
Elongation	0.82 ± 0.05	0.89 ± 0.03	0.91 ± 0.02	0.92 ± 0.01
Solidity	**0.58 ± 0.04**	**0.67 ± 0.04**	**0.65 ± 0.04**	**0.62 ± 0.09**
Extent	**0.69 ± 0.04**	**0.75 ± 0.03**	**0.69 ± 0.04**	**0.73 ± 0.04**
Texture features				
Contrast	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.02
Energy	0.83 ± 0.02	0.85 ± 0.01	0.86 ± 0.02	0.86 ± 0.01
Homogeneity	0.90 ± 0.01	0.91 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
Entropy	0.84 ± 0.03	0.88 ± 0.02	0.85 ± 0.01	0.86 ± 0.03

The three lowest (worst) values in each column are highlighted in bold. The mean values and standard deviations were measured using 5 independent CNN training trials. Where omitted, the standard deviation was less than 0.01.

features. The loss convergence rates varied between features: the fastest convergence (i.e. achieved in a fewest number of epochs) was observed with Maximum/Mean intensity, Variance, and size-related features such as Area and Volume. In comparison, texture features such as Energy and Entropy required more epochs to converge. Features quantifying shape irregularity and COV had the slowest convergence.

The values of the training and test losses after 200 epochs were lowest for intensity (except COV) and size-related features, followed by texture features; Solidity, Solidity end Extent had the highest loss values. With the latter 3 features, there was also a marked difference between the training and test losses. The relatively high training loss for these features indicates that the networks were less effective at approximating the respective functions. On the other hand, the even higher test loss indicates that the networks did not generalize well when predicting these features for new images.

3.2 SOTA networks

Radiomic feature prediction errors obtained with the SOTA networks are plotted in Fig. 5. The intensity (Maximum, Mean, Variance) and size features were predicted with higher errors compared to the CNP networks. SOTA networks with a greater number of trainable parameters or layers did not produce lower prediction errors. On the contrary, the intensity features were predicted best with the MobileNetV2 network, which had the fewest number of parameters. Inspection of the true-vs-predicted value scatter plots (not shown) confirmed that the high prediction errors were distributed uniformly among the test samples and not originate from a few outliers.

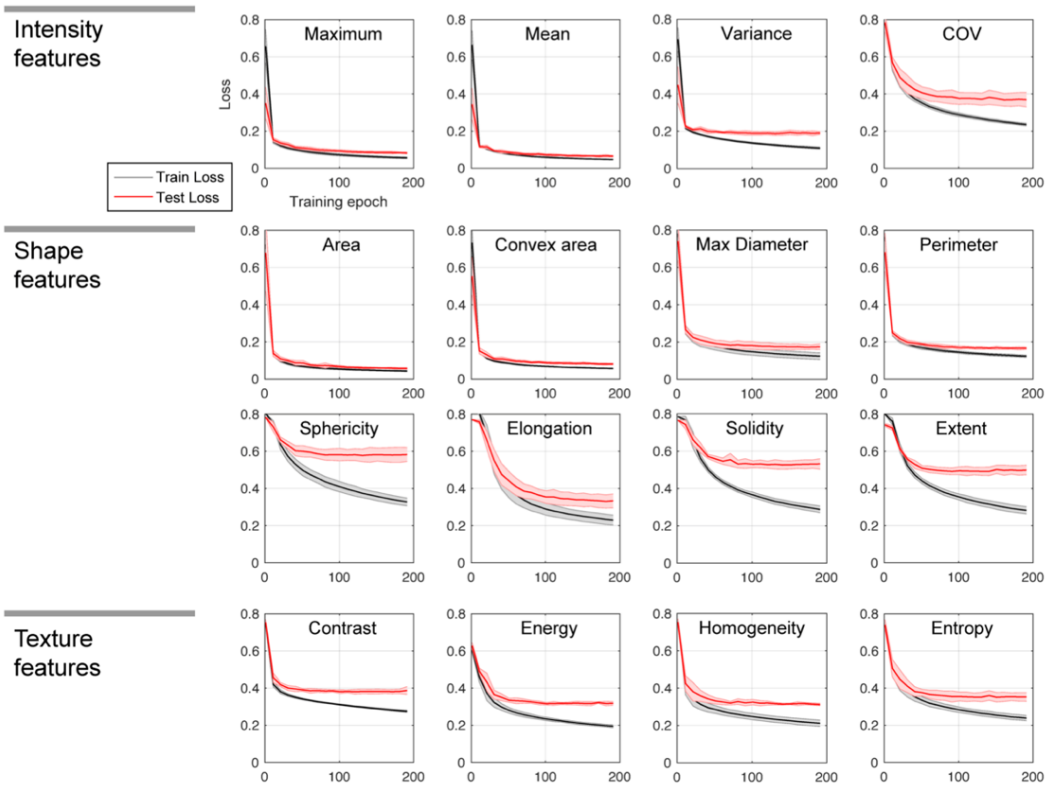

Figure 4: Training and test losses plotted against the training epoch for the CN-2D-5 network. The mean and standard deviation of loss values from 5 independent training trials are plotted.

Across all networks, high prediction errors were measured with Sphericity, Solidity, and Extent, similarly to the CNP networks. However, the highest prediction errors were obtained with Elongation, which is something that was not observed with CNP networks.

The standard deviations of errors with the SOTA networks were markedly lower compared to those of the CNP networks, like due to the frozen parameters being constant between different trials. Likewise, the training and test loss of SOTA networks converged on average within the first 20 epochs, much faster compared to the CNP networks. We found that the test loss closely followed the training loss with most features, although the difference between the training and test losses was again greater with the shape irregularity features.

Spearman's rank correlation coefficients ρ between predicted and true feature values for the SOTA networks are given in Table 4. The ranking of features and networks in terms of ρ was similar to that obtained with nMAE.

3.3 Sample complexity analysis

The generalization capacity of a trained CNN can be assessed from the difference between the training and test loss; sample complexity represents the number of training samples required to achieve good generalization. We measured the train-test loss difference with the CN-2D-3 network (simplest network tested) for a representative group of features, using various numbers of training samples (Fig. 6). The graphs demonstrate the significantly different sample complexities for different features. In the extreme case, to achieve the same level of generalization, Area required ~100 samples, and Sphericity required 3900 samples. Note that the corresponding test loss values for Area and Sphericity were 0.3 and 0.6, respectively, i.e. a similar generalization capacity does not imply a similar prediction error.

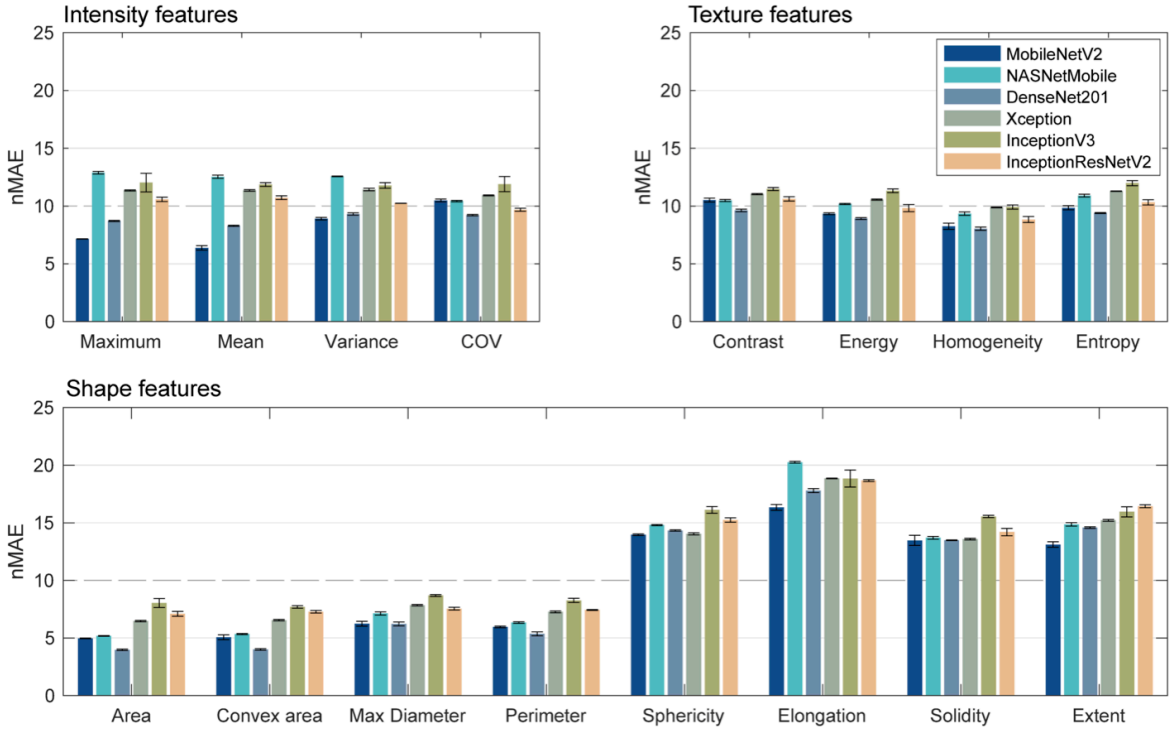

Figure 5: Feature prediction errors (nMAE) for the SOTA networks. The mean values and standard deviations were measured using 3 independent trials of the regression layer training.

3.4 Additional tests

Using binary masks as inputs to the CNP networks (instead of SUV images) to predict the shape features resulted in the reduction of nMAE by approximately 20% for the Sphericity, Solidity and Extent features. The relative prediction errors for these features remained to be the highest.

On a different synthetic image set with fixed lesion size, the errors of shape feature predictions were similar to those plotted in Fig. 2. This may indicate that, when predicting shape features, the networks did not utilize any possible correlations between the lesion shape and size.

4 Discussion

We have directly quantified the relative expressive power of standard CNN architectures with respect to standardized intensity, shape, and texture features commonly used in oncological imaging. We found that features quantifying lesion size as well as maximum and mean intensities exhibited lowest prediction errors. On the other hand, features quantifying shape irregularity had highest prediction errors, and generalized poorly from the training to test sets. Given that tumor shape has been found to be a significant predictor of clinical outcomes [4, 20], this finding may bear significant implications for the use of CNNs in clinical prediction tasks. For example, CNNs that are trained to predict progression-free survival from tumor images, may preferentially learn to leverage the intensity and size information, while the shape-irregularity information may be under-utilized.

In addition to standard CNNs trained end-to-end, we tested several ImageNet pre-trained SOTA networks that were fine-tuned on our data.. We found that all radiomic features predicted by SOTA networks had high errors, higher than those obtained with standard CNNs, implying that radiomics-related information is poorly represented in the high-level feature output layers of ImageNet pre-trained

Table 4: Spearman's rank correlation coefficients (ρ) between predicted and true feature values for the SOTA networks.

Feature name	MobileNetV2	NASNetMobile	DenseNet201	Xception	InceptionV3	InceptionResNetV2
			Intensity features			
Maximum	0.95	0.84	0.94	0.88	0.88 ± 0.01	0.90
Mean	0.96	0.84	0.94	0.87	0.86	0.89
Variance	0.93	0.83	0.93	0.87	0.86	0.89
COV	0.85	0.83	0.88	0.83	0.80 ± 0.01	0.86
			Shape features			
Area	0.98	0.97	0.99	0.96	0.94	0.95
Convex area	0.98	0.97	0.98	0.96	0.94	0.95
Max diameter	0.96	0.94	0.95	0.93	0.91	0.93
Perimeter	0.96	0.95	0.97	0.95	0.92	0.94
Sphericity	**0.72**	0.70	0.72	0.74	0.66 ± 0.01	0.70 ± 0.01
Elongation	**0.62 ± 0.01**	**0.31 ± 0.01**	**0.51 ± 0.01**	**0.45**	**0.48 ± 0.01**	**0.50**
Solidity	**0.71 ± 0.01**	**0.70**	**0.71**	**0.70**	**0.63**	**0.68 ± 0.01**
Extent	0.74 ± 0.01	**0.65 ± 0.01**	**0.68 ± 0.01**	**0.64**	**0.61 ± 0.01**	**0.55 ± 0.01**
			Texture features			
Contrast	0.84	0.83 ± 0.01	0.86	0.80	0.80 ± 0.01	0.83
Energy	0.80	0.79	0.81	0.77	0.73 ± 0.01	0.80 ± 0.01
Homogeneity	0.88 ± 0.01	0.86	0.89	0.84	0.84 ± 0.01	0.88 ± 0.01
Entropy	0.84	0.81	0.85	0.80	0.77	0.83 ± 0.01

The three lowest values in each column are highlighted in bold. The mean values and standard deviations were measured using 3 independent trials of the regression layer training. Where omitted, the standard deviation was less than 0.01.

networks. The errors were highest for the shape features, mirroring findings with the standard CNNs. Based on these observations, we conclude that simple CNNs trained end-to-end on domain-specific images should capture radiomic features better than advanced networks pre-trained on large image sets like ImageNet.

Sample complexity analysis showed that intensity and size metrics required around 100-500 training samples to achieve good train-test generalization. On the other hand, shape irregularity features required around 2000-4000 training samples. Hence, a relatively large number of examples is required for CNNs to capture the shape-related information from the images. In contrast, medical imaging studies that use CNNs often have far fewer than 1000 training samples — typically the number of samples is on the order of 100 or less, particularly for PET studies (according to the Cancer Imaging Archive, https://www.cancerimagingarchive.net) [21]. In studies with tens or hundreds of samples, CNNs may only be able to implicitly learn "easier" features related to the intensity and size (image augmentation may help to alleviate this issue).

We hypothesize that high prediction errors for some features may be attributed to two factors. First, the tested networks lacked direct ability to capture global context, which may be important for capturing global shape properties. Designing and using CNNs that can capture the global context and have larger perceptive fields may lead to a better implicit learning of shape properties. Second, the high prediction errors could have originated from the limited ability of CNNs to approximate ratio-type features or functions (such as COV, sphericity, solidity and extent). For example, solidity is a ratio of area and convex area, both of which were predicted with a much lower error compared to solidity. Including a non-standard division operation in the network graph, or adding the reciprocal image as an input, may improve prediction performance. Alternatively, features with high prediction errors can be added

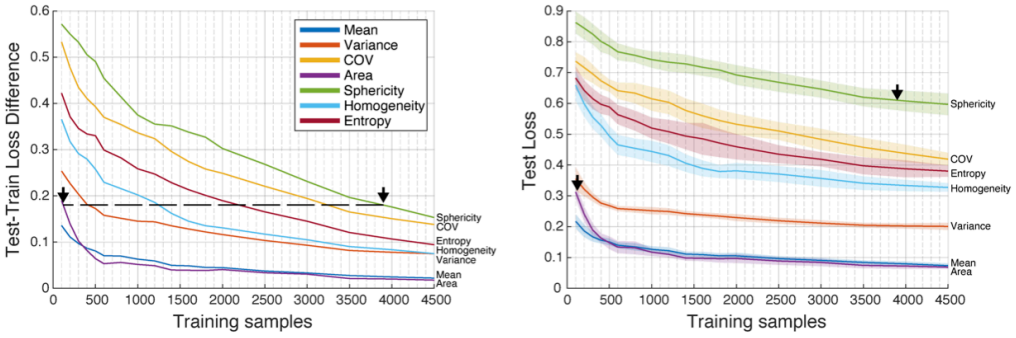

Figure 6: Left: difference between the train and test loss plotted against the number of training samples; arrows indicate the same value of the difference for Area and Sphericity. Right: test loss plotted against the number of training samples. Arrows indicate the values of test loss for Area and Sphericity achieved at the same level of generalization. The mean and standard deviations of loss values from 10 independent trials are plotted.

explicitly as auxiliary variables to the dense layers in the "heads" of the networks, or as additional input channels. We propose that making these modifications to existing and previously published models for image-based diagnosis may improve the performance of the models, which is an area of future research.

Among other findings, there was an unexpectedly small improvement in the prediction error with added network depth. It is of interest to explore how the width of the network, i.e. the number of filters or channels in the convolutional layers, affects prediction errors: shallower and wider CNNs may perform as well or better than deeper networks in medical imaging applications. Recent theoretical studies suggest that the expressive power of neural networks grows faster with added depth than with added width [22, 23]. However, this may or may not apply to functions that represent low-level image features.

A limitation of our study is that the tests were performed on synthetic PET images and for a specific range of radiomic features given in Table 1; using synthetic images allowed us to produce a large number of samples. We made efforts to match the image properties of synthetic lesions to those of real tumors, including size, shape, resolution, and signal-to-noise ratios. We believe that our findings should generalize to other modalities, and it is of particular interest to reproduce our experiments on CT and MRI images, where large datasets are available. We also note that there is a chance for a CNN not to properly capture certain "important" radiomic features, but to instead implicitly discover other patterns, not captured by existing radiomic features, that lead to equally high end-task performances.

In conclusion, our work shows that conventional CNNs architectures readily learn first-order intensity and size-related radiomic features from less than 500 samples. On the other hand, features describing tumor heterogeneity (e.g. COV) and shape irregularity are difficult to learn, and require an order of magnitude more samples; the capacity of CNNs to learn texture features is intermediate. Therefore, CNNs may not be as effective as explicit radiomic features at capturing certain tumor properties. This is in fact more strongly the case for CNNs pretrained on image sets like ImageNet. In our view, the use of explicit radiomics and traditional machine learning techniques may not be properly discarded in favor of existing CNNs when it comes to medical image analysis, as the strengths of these two approaches appear to be complementary: a combination of the two approaches or appropriate next-generation deep networks are likely to produce improved results.

Acknowledgment

This work was supported by the National Institutes of Health (NIH) / Canadian Institutes of Health Research (CIHR) Quantitative Imaging Network (QIN) Grant number 137993, and in part through computational resources and services provided by Microsoft and the Vice President Research and Innovation at the University of British Columbia.

References

[1] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563–577.

[2] Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson ARA. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. Cancer Res. 2015;75(8):1567–1579.

[3] Ceriani L, Milan L, Martelli M, Ferreri AJM, Cascione L, Zinzani PL, et al. Metabolic heterogeneity on baseline 18FDG-PET/CT scan is a predictor of outcome in primary mediastinal B-cell lymphoma. Blood. 2018;132(2):179–186.

[4] Hsu CY, Wang CW, Kuo CC, Chen YH, Lan KH, Cheng AL, et al. Tumor compactness improves the preoperative volumetry-based prediction of the pathological complete response of rectal cancer after preoperative concurrent chemoradiotherapy. Oncotarget. 2017 jan;8(5):7921–7934.

[5] Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: The process and the challenges. Magn Reson Imaging. 2012;30(9):1234–1248.

[6] Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging. 2017 jan;44(1):151–165.

[7] Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Rev Precis Med Drug Dev. 2016 mar;1(2):207–226.

[8] Bodalal Z, Trebeschi S, Nguyen-Kim TDL, Schats W, Beets-Tan R. Radiogenomics: bridging imaging and genomics. Abdom Radiol. 2019;44(6):1960–1984.

[9] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–833.

[10] Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836–1842.

[11] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–961.

[12] Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists ’Performance in Breast Cancer Screening. IEEE Trans Med Imaging. 2019;p. 1–1.

[13] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 jan;577(7788):89–94.

[14] Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals, Syst. 1989;.

[15] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989 jan;2(5):359–366.

[16] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. 7th Int Conf Learn Represent ICLR 2019. 2018 nov;(c):1–22.

[17] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 may;295(2):328–338.

[18] Watabe T, Tatsumi M, Watabe H, Isohashi K, Kato H, Yanagawa M, et al. Intratumoral heterogeneity of F-18 FDG uptake differentiates between gastrointestinal stromal tumors and abdominal malignant lymphomas on PET/CT. Ann Nucl Med. 2012 apr;26(3):222–227.

[19] Ashrafinia S. Quantitative nuclear medicine imaging using advanced image reconstruction and robotics [PhD Dissertation]. Johns Hopkins University; 2019.

[20] Wang G, Wu F, Wang J, Yang C, Zhou C, Niu W, et al. Volumetric imaging parameters are significant for predicting the pathological complete response of preoperative concurrent chemoradiotherapy in local advanced rectal cancer. J Radiat Res. 2019;60(5):666–676.

[21] Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G. Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev. 2020 jun.

[22] Telgarsky M. Benefits of depth in neural networks. In: 29th Annual Conference on Learning Theory. vol. 49 of Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR; 2016. p. 1517–1539.

[23] Liang S, Srikant R. Why deep neural networks for function approximation? In: International Conference on Learning Representations; 2017.