Probabilistic Radiomics: Ambiguous Diagnosis with Controllable Shape Analysis

Jiancheng Yang[1,2,3], Rongyao Fang[1], Bingbing Ni[1,2,3(✉)], Yamin Li[1], Yi Xu[1], and Linguo Li[1]

[1] Shanghai Jiao Tong University, Shanghai, China
nibingbing@sjtu.edu.cn
[2] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Institute for Advanced Communication and Data Science, Shanghai, China

Abstract. Radiomics analysis has achieved great success in recent years. However, conventional Radiomics analysis suffers from insufficiently expressive hand-crafted features. Recently, emerging deep learning techniques, e.g., convolutional neural networks (CNNs), dominate recent research in Computer-Aided Diagnosis (CADx). Unfortunately, as black-box predictors, we argue that CNNs are "diagnosing" voxels (or pixels), rather than lesions; in other words, visual saliency from a trained CNN is not necessarily concentrated on the lesions. On the other hand, classification in clinical applications suffers from inherent ambiguities: radiologists may produce diverse diagnosis on challenging cases. To this end, we propose a controllable and explainable *Probabilistic Radiomics* framework, by combining the Radiomics analysis and probabilistic deep learning. In our framework, 3D CNN feature is extracted upon lesion region only, then encoded into lesion representation, by a controllable Non-local Shape Analysis Module (NSAM) based on self-attention. Inspired from variational auto-encoders (VAEs), an Ambiguity PriorNet is used to approximate the ambiguity distribution over human experts. The final diagnosis is obtained by combining the ambiguity prior sample and lesion representation, and the whole network named *DenseSharp*$^{+}$ is end-to-end trainable. We apply the proposed method on lung nodule diagnosis on LIDC-IDRI database to validate its effectiveness.

Keywords: Radiomics · Deep learning · Attention · Computer-Aided Diagnosis (CADx) · Explainable Artificial Intelligence (XAI)

J. Yang. and R. Fang are contributed equally.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32226-7_73) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 658–666, 2019.
https://doi.org/10.1007/978-3-030-32226-7_73

1 Introduction

Medical images are more than pictures [2]. Mining hidden information using image analysis techniques is referred as *Radiomics* analysis, which raises numerous research attention in clinical decision making. Conventional Radiomics analysis follows the pipeline: (1) manual/automatic delineation of volumes of interest (VOIs); (2) image processing and feature extraction (e.g., SIFT, wavelet); (3) machine learning to associate features and target variables. These hand-craft features are named "Radiomics". Though powerful and successful, emerging deep learning techniques indicate that hand-crafted features could be hardly comparable with end-to-end deep representations given enough data [12].

Deep learning[1] provides a strong alternative to learn representation from raw voxels (or pixels) in an end-to-end fashion. Convolutional neural networks (CNNs) have achieved great success in medical image analysis, though they are classifying **voxels, rather than lesions**. In other words, there is no guarantee that black-box CNNs correctly learn evidence from lesions, especially with limited supervision. We illustrate several failures in Appendix Fig. A.1, by checking the Class Activation Maps (CAMs) [13] from a 3D DenseNet [4,12] on lung nodule malignancy classification. These failures make the predictions given by CNNs unreliable. In contrast, Radiomics analysis is more controllable and transparent for users than black-box deep learning.

On the other hand, classification in clinical applications suffers from inherent ambiguities; on challenging cases, experienced radiologists may produce diverse diagnosis. Though a "ground truth" to eliminate ambiguity could be obtained through a more sophisticated examination (e.g., biopsy) theoretically, this information may be unavailable from imaging only. Discriminative training procedure biases the model towards the mean values rather than ambiguity distribution.

To address these issues, we propose a controllable and explainable *Probabilistic Radiomics* framework. A *DenseSharp* Network [11] is used as a backbone, which is a multi-task 3D CNN on learning classification and segmentation developed from 3D DenseNet [4,12]. Point clouds, named *feature clouds*, extracted from manual-labeled or predicted VOIs on CNN feature maps are regarded as lesion representations. To enable non-local shape analysis, we further introduce self-attention [8,10] to learn representations from the feature clouds. To capture label ambiguity, an Ambiguity PriorNet is used to approximate the ambiguity distribution over expert labels, inspired by Variational Auto-Encoders (VAEs) [7]. By combining the ambiguity prior sample and lesion representation, the final decision is controllable (by lesion VOI) and probabilistic, which mimics the decision process of human radiologists. Please refer to Appendix Fig. A.2 for comparison among conventional Radiomics analysis, deep learning and Probabilistic Radiomics. On LIDC-IDRI [1] database, we validate the effectiveness of our methodology on lung nodule characterization from CT scans.

[1] We refer to deep learning in a narrow sense, i.e., applying CNNs directly on the medical image analysis problems.

The key contributions of this paper are threefold: (1) We propose a novel viewpoint to regard deep representations from lesions on medical images as point clouds (i.e., feature clouds), and develop a Non-local Shape Analysis Module (NSAM) to end-to-end learn representations from feature clouds (rather than voxels); (2) We explicitly model the diagnosis ambiguity within a probabilistic and controllable approach, which mimics the decision process of human radiologists; (3) The whole network named *DenseSharp*$^+$ is end-to-end trainable.

2 Materials and Methods

2.1 Task and Dataset

Lung cancer is the leading cause of cancer-related mortality worldwide. Early diagnosis of lung cancer with LDCT is an effective way to reduce the related death. In this study, we address the lung nodule malignancy classification problem to explore the performance of the proposed Probabilistic Radiomics method.

We use LIDC-IDRI [1] dataset, one of the largest publicly available databases for lung cancer screening. There are 2,635 nodules from 1,018 CT scans in the dataset, where nodules with diameters $\geq 3\,\text{mm}$ are annotated by at most 4 radiologists. For malignancy classification, rating mode ranges from "1" (highly benign) to "5" (highly malignant), while "3" means undefined/uncertain rating. Besides, each radiologist delineates a VOI for a lesion. Empirically, the malignancy labels and segmentation VOIs are diverse for many instances in the dataset. Prior studies [5,14] define a unique binary label for each instance by voting, we instead treat these labels with **ambiguity**, with all the **5 classes**. We called the whole dataset with 2,635 nodules as *HighAmbig* (high ambiguous) dataset. To fairly compare the model performance, a *LowAmbig* (low ambiguous) dataset is constructed, with a similar nodule inclusion criteria to prior studies [5,14]: (1) the CT slice thickness $\leq 3\,\text{mm}$, (2) annotated by at least 3 radiologists, and (3) the average rating $\neq$ "3". The remaining nodules with average ratings $\leq$ "3" are defined as benign, or malignant otherwise, resulting in 656 benign and 527 malignant.

We pre-process the data as follows: CT are resampled into $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$. The voxel intensity is normalized to $[-1, 1)$ from the Hounsfield unit (HU), by $I = \lfloor \frac{I_{HU}+1024}{400+1024} \times 255 \rfloor / 128 - 1$. Each data sample is a voxel with a size of $32\,\text{mm} \times 32\,\text{mm} \times 32\,\text{mm}$. For simplicity, only single-scale inputs are used.

2.2 Non-local Shape Analysis Module (NSAM)

In our study, we use a CNN (DenseSharp [11] specifically) for extracting representations of nodules. Instead of a typical Global Pooling to derive the final classification, we use the lesion VOIs (manually annotated/automatically predicted) to crop the lesion features into point clouds [10], namely *feature clouds*, for subsequent processing. Inspired by self-attention transformer [8,10], we develop a Non-local Shape Analysis Module (NSAM) to consume the feature clouds.

Define $X \in \mathbb{R}^{N \times c}$ as a feature cloud, X is a permutation-invariant and size-varying set. We figure out that self-attention is well suitable for set; besides, it enables non-local representation learning. We use scaled dot-product attention,

$$Attn(X) = softmax(XX^T/\sqrt{c}) \cdot \sigma(X), \tag{1}$$

where σ is an activation function (e.g., ELU in our study).

Multi-head attention [8] is proved to be effective in attention mechanism, where a scaled dot-product attention is applied multiple times on linear transformed input with various weights. The *NSAM* is a variant of multi-head attention, by sharing the linear transformation weights in the K, Q, V-formation [8]. Define g as the number of heads and $c_g = c/g$, the inputs are transformed by the weight $W_g \in \mathbb{R}^{c \times c_g}$ multiple times, before feeding into a scaled dot-product attention module. We further use skip connections [3] to ease the optimization.

$$NSAM(X) = concat\{Attn(X_i)|X_i = XW_i\}_{i=1,..,g}) + X. \tag{2}$$

The whole shape analysis module is a stack of L-layer *NSAM* ($L = 3$, $c = 256$ in this study). The features are subsequently fed into a global average pooling with multi-layer perceptron to obtain a single representation for a lesion VOI.

2.3 Ambiguity PriorNet

To deal with the ambiguous labels, we model the final decision as ambiguity prior distribution over the human experts. Inspired from Variational Auto-Encoders (VAEs) [7], a probabilistic module with a similar structure as 3D DenseNet backbone, named Ambiguity PriorNet (APN), is introduced to model the probabilistic component. APN produces (μ, σ), which controls a Gaussian distribution $N(\mu, \sigma)$ to serve as the ambiguity prior on malignancy labels and segmentation for human experts. To enable the gradient back-propagation, a reparameterization trick [7] is applied to draw a prior sample f_{Ambig} from $N(\mu, \sigma)$.

$$f_{Ambig}(x) = \sigma x + \mu, x \in N(0, 1). \tag{3}$$

In subsequent modules, the prior sample f_{Ambig} is concatenated with lesion representations to produce ambiguous malignancy labels and segmentation.

2.4 *DenseSharp*$^+$ Network Architecture

The proposed *DenseSharp*$^+$ Network (Fig. 1) is based on *DenseSharp* Networks [11], which is a multi-task 3D DenseNet [4,12] with classification and segmentation heads. The *DenseSharp* Network uses a light-weight head for segmentation, which enables a top-down supervision for learning where the lesions are. At each resolution level ($32 \times 32 \times 32$, $16 \times 16 \times 16$ and $8 \times 8 \times 8$), dense blocks with 3D convolution and Batch Normalization [6] are repeated $[3, 8, 4]$ times before each down-sampling. Bottleneck ($B = 4$), compression ($C = 2$) and growth rate $k = 32$ are used following the setting in the *DenseSharp* paper [11].

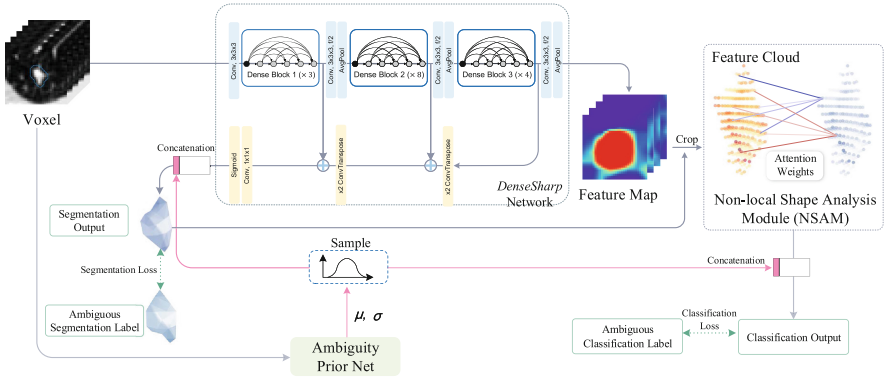

Fig. 1. *DenseSharp*$^+$ **Network Architecture**. A *DenseSharp*$^+$ Network is mainly a *DenseSharp* Network followed by a Non-local Shape Analysis Module (NSAM). *DenseSharp* is a deep 3D CNN based on DenseNet, with a classification head and segmentation head for multi-task learning. We use the feature maps from the classification head, cropped by manual/automatic segmentation, as *feature clouds*, rather than the raw feature maps, for the subsequent NSAM to consume. The NSAM use self-attention to associate non-local spatial information. An Ambiguity PriorNet conditional on the voxel inputs produces prior samples, which is concatenated with the classification and segmentation head to make their outputs probabilistic. Note the whole *DenseSharp*$^+$ Network is end-to-end trainable, with multi-task classification and segmentation loss.

The feature maps outputted by the last convolution layer of classification head is upsampled (trilinear interpolation), and then cropped by the lesion segmentation into feature clouds which are consumed by NSAM (Sect. 2.2). Either manual or automatic segmentation by the segmentation head could be used as the lesion segmentation to generate the feature clouds. Although NSAM is able to process size-varying inputs, due to the GPU memory constraint, we sample up to $N^{\max} = 1,024$ points from the feature cloud with sampling strategy Φ. For the manual segmentation, the sampling strategy Φ is random sampling. For the predicted segmentation $\hat{y}_{seg}$, we first estimate the volume by $\hat{v} = \sum \hat{y}_{seg}$. We then sample the $K = \lfloor \hat{v} \rfloor$ points with top-K output scores from the segmentation head. If $N \leq N^{\max}$, all points in the feature cloud are selected.

A DenseNet conditional on the voxel inputs (with a half parameter size of *DenseSharp*) is used as Ambiguity PriorNet (APN), which outputs 6-dimension prior samples to concatenate onto the classification and segmentation heads, to make their outputs probabilistic. Ideally, one prior sample encodes one "human expert", controlling the classification and segmentation results simultaneously.

2.5 Training and Inference

The *DenseSharp*$^+$ Networks is trained with two different schemes individually in order to better evaluate the probabilistic capability of the model. The first scheme trains on the *LowAmbig* dataset (see Sect. 2.1). This scheme denotes

as *LowAmbig* (low ambiguous) training scheme. The second scheme trains the model on the whole labeled dataset, which denotes as *HighAmbig* (high ambiguous) training scheme. In both training schemes, unlike prior studies [5,14] with a unique label on each voxel, we randomly select one of the four experts and the corresponding 5-class malignancy label and segmentation during training.

For training the multi-task neural networks, a cross entropy loss for classification and a dice loss for segmentation are used. The loss weights for classification and segmentation are set as 1 and 0.2, respectively. Online data augmentation is applied on the voxels, including rotation, flipping and shifting within $[-1, 1]$ on a random axis. We use Adam optimizer to train the whole *DenseSharp*$^+$ end-to-end with a batch size of 128 and a learning rate of 0.001 for 150 epochs.

For simplicity, feature maps from the *DenseSharp* are cropped by predicted segmentation to feed into NSAM for training and inference. However, if the prediction segmentation volume is less than 10, the model refuses to use it to classify the nodule. In this case, it is not counted in classification loss during training, and is ignored during the evaluation on classification.

3 Experiments

Our *DenseSharp*$^+$ Network is trained to classify ambiguous labels of 5 malignancy modes from 4 radiologists. N prior samples ($N = 10$ in our experiments) are obtained from the reparameterized conditional Gaussian distribution of the Ambiguous PriorNet. Hence, each tested voxel corresponds to N 5-way outputs. In order to compare with prior studies quantitatively, the corresponding binary classification outputs are computed using Eq. 4.

$$\begin{aligned} (p_1, p_2, p_4, p_5) &= \frac{1}{N} \sum_{i=1}^{N} \text{Softmax}(l_1^i, l_2^i, l_4^i, l_5^i), \\ p_b &= p_1 + p_2, \quad p_m = p_4 + p_5, \end{aligned} \tag{4}$$

where $l_1^i, l_2^i, l_4^i, l_5^i$ denote the i^{th} logit outputs in the N samples of mode 1, 2, 4, and 5 from 5-mode classification. Note mode 3 is ignored in the evaluation since it defines "uncertain" diagnosis.

We evaluate the performance of all models via test AUC and accuracy on *LowAmbig* LIDC-IDRI dataset (see Sect. 2.1) with 5-fold cross validation method. It is worth noting that only *LowAmbig* voxels are evaluated in all our experiments, since the binary labels for data in *HighAmbig* are not trivially defined.

Table 1 shows the performance of our models and baselines[2]. It is noticeable that 3D DenseNet reveals a comparable performance with 3D DPN [14]. The *DenseSharp*$^+$ network with *HighAmbig* training scheme outperforms the one with *LowAmbig* training scheme. The *HighAmbig DenseSharp*$^+$ is trained

[2] Note that all counterparts use (sightly) different evaluation protocols.

Table 1. AUC and accuracy of DenseNet, *DenseSharp*, *DenseSharp*$^+$, and prior studies. The performance of our models is evaluated on *LowAmbig* LIDC-IDRI [1] dataset (see Sect. 2.1) with 5-fold cross validation.

Method	AUC	Accuracy (%)
3D DPN [14]	–	88.28
3D DPN ensemble [14]	–	90.44
3D CNN w. MTL [5]	–	80.08
3D CNN w. sparse MTL [5]	–	91.26
3D DenseNet (our implementation)	0.9218	87.82
DenseSharp [11] (our implementation)	0.9393	89.26
DenseSharp$^+$ (LowAmbig)	0.9480	90.87
DenseSharp$^+$ (HighAmbig)	**0.9566**	**91.52**

on an ambiguous dataset with a larger scale, resulting in a better performance than that of *LowAmbig DenseSharp*$^+$, which shows an excellent ability to learn from the ambiguous data distribution. The performance of *HighAmbig* trained *DenseSharp*$^+$ is also better than 3D DPN ensemble [14] and 3D CNN w. sparse MTL [5]. Notably, compared with other methods, we adopt a coarser dataset preprocessing strategy and a simpler evaluation setting. For instance, both counterparts [5,14] use 10-fold cross validation, with more training samples than 5-fold in our study. The 3D DPN [14] only evaluates its performance on the overlapping nodules with LUNA16 dataset, which are easier to classify. The sparse MTL [5] resamples voxels at a higher resolution (spacing of 0.5 mm), besides the CNN is pre-trained on large-scale video dataset, rather than randomly initialized.

As for the segmentation output of *DenseSharp*$^+$, the average segmentation dice coefficient is 0.7625 on *LowAmbig* LIDC-IDRI with 5-fold cross validation. The segmentation output is of good quality with such a light-weight segmentation head. Due to the probabilistic segmentation output, *DenseSharp*$^+$ with automatic segmentation refuses to classify the nodules whose predicted volume is less than 10; 73 nodules are refused by *HighAmbig*-trained *DenseSharp*$^+$.

For further evaluation of probabilistic property of *DenseSharp*$^+$ model, we compute the mean standard deviation of softmax outputs as a diversity metric, derived from the softmax outputs of all the tested voxels (Eq. 5),

$$\mathrm{DIV} = \frac{1}{5}\sum_{i=1}^{5} \mathrm{Std}_{j=1\ldots N}(p_{ij}), \tag{5}$$

in which p_{ij} is the softmax output of malignancy mode i and j^{th} sample of Gaussian distribution from one voxel. Std$(\cdot)$ is the standard deviation operation. The distribution of *DIV* from all the tested voxels reflects the probabilistic output variance of *DenseSharp*$^+$ Networks. Figure 2 shows the *DIV* distribution of all the tested voxels. The two highlight samples show that the classification predictions from the model mimic the ambiguous labels from different experts.

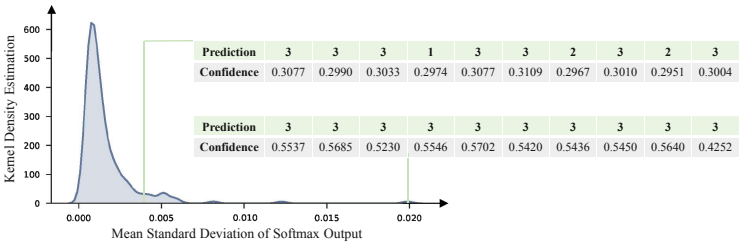

Fig. 2. The diversity metric (DIV) distribution of all tested voxels. The two highlight examples show that the output of $DenseSharp^{+}$ model varies as the prior sample varies, thanks to its probabilistic property.

Moreover, thanks to the explicit modeling, only voxels in lesions are counted, the visual saliency maps produced by the $DenseSharp^{+}$ is highly calibrated with the nodules. Please refer to Appendix Fig. A.3 for illustration.

4 Conclusion and Further Work

In this study, a Probabilistic Radiomics framework is proposed, which is well-performing, controllable and explainable in Computer-Aided Diagnosis (CADx). The proposed method is more expressive than conventional Radiomics analysis, more controllable and explainable than conventional deep learning approaches. Moreover, we explicitly model the ambiguity of the classification with a probabilistic approach. However, there are still limitations to make the Probabilistic Radiomics an *omics*-level approach (e.g., genomics, proteomics, immunomics).

Compared to other "omics" approaches, Radiomics is generally less reproducible [2]. Perturbations (e.g., rotations, different imaging parameters, adversarial attacks) on the images/point clouds [9] could introduce large variances to the outputs. Besides, the data-hungriness issue makes current MIC research a Sisyphean challenge; model learning on a certain task is non-trivial to transfer to another task. A more generalizable representation learning is the key to this problem, (probably) following a route of self-supervised learning and meta-learning. We will explore the robustness, transferability, and reproducibility of Probabilistic Radiomics in the future study.

Acknowledgment. This work was supported by National Science Foundation of China (U1611461, 61502301, 61521062). This work was supported by SJTU-UCLA Joint Center for Machine Perception and Inference, China's Thousand Youth Talents Plan, STCSM 17511105401, 18DZ2270700 and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China. This work was also jointly supported by SJTU-Minivision joint research grant.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
2. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2015)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
5. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 249–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_20
6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
7. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
8. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)
9. Yang, J., Zhang, Q., Fang, R., Ni, B., Liu, J., Tian, Q.: Adversarial attack and defense on point sets. arXiv preprint arXiv:1902.10899 (2019)
10. Yang, J., Zhang, Q., Ni, B., et al.: Modeling point clouds with self-attention and gumbel subset sampling. In: CVPR, pp. 3323–3332 (2019)
11. Zhao, W., Yang, J., et al.: 3D deep learning from ct scans predicts tumorinvasiveness of subcentimeter pulmonary adenocarcinomas. Cancer Res. **78**(24), 6881–6889 (2018)
12. Zhao, W., Yang, J., et al.: Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3d deep learning. Cancer Med. (2019)
13. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016)
14. Zhu, W., Liu, C., Fan, W., Xie, X.: Deeplung: 3D deep convolutional nets for automated pulmonary nodule detection and classification. In: WACV (2017)

A Appendix from *Probabilistic Radiomics: Ambiguous Diagnosis with Controllable Shape Analysis*

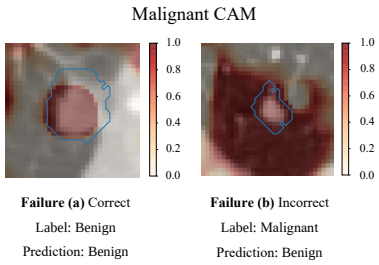

Fig. A.1. Two types of failures from a well-trained 3D DenseNet, visualized by CAM techniques. Only malignant CAMs on the central slices are depicted. The blue contours on each plot are manual segmentation of lesions by radiologists. The voxels with higher intensity are more malignant, and those with intensity ≤ 0.5 are benign. For failure (a), the model predicts "benign" on a benign nodule correctly. However, this "correct" prediction comes from the prediction apart from lesions on voxels, which means the model uses incorrect evidences. For failure (b), the model outputs "benign" on a malignant nodule incorrectly. Whereas, within the lesion voxels it is indeed predicted as malignant, indicating that the model performance could be boosted further if it uses correct evidences.

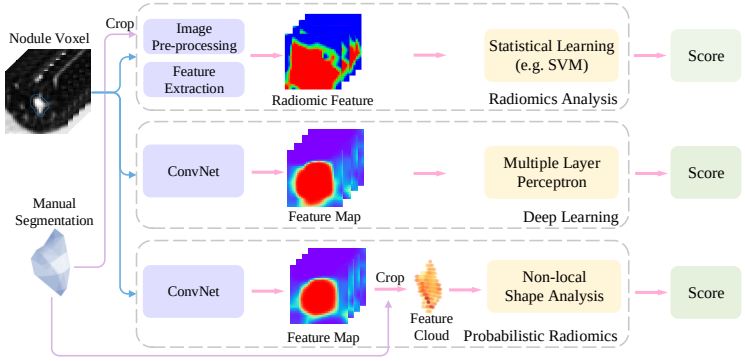

Fig. A.2. Comparison of conventional Radiomics analysis, deep learning, and our proposed Probabilistic Radiomics framework. Radiomics analysis (top) only responds to the user-delineated VOIs, while the hand-crafted features are pre-defined and not learnable. Conventional Deep learning (middle) learns expressive representations end-to-end from voxels of CT scans, however, it could possibly learn "evidences" outside lesions, making its prediction unreliable and unexplainable. The proposed Probabilistic Radiomics framework (bottom) uses *feature clouds* (instead of voxels) for a final decision, which are CNN feature maps cropped by the automatic segmentation of lesions. The feature clouds are then consumed by a Non-local Shape Analysis Module (NSAM) based on self-attention for deeper representation. The proposed framework takes advantage of the expressiveness of deep learning and the controllability of Radiomics analysis, thus defining a *Probabilistic Radiomics*.

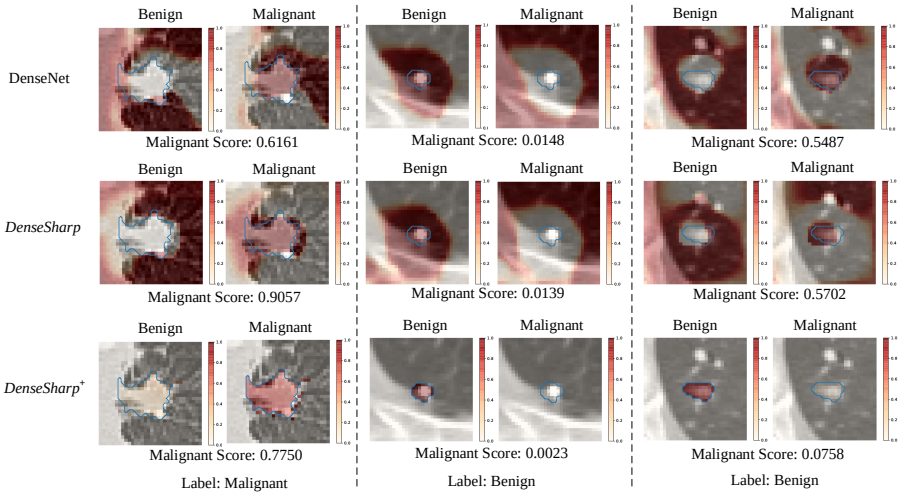

Fig. A.3. Three nodule samples classified by DenseNet, *DenseSharp*, and *DenseSharp*$^+$, visualized by CAM techniques. As the benign and malignant CAMs have gone through a *softmax*, the sum of benign CAM and malignant CAM in a corresponding voxel equals to 1. The blue contours on each plot are manual segmentation of lesions. The "labels" are the classification by radiologists. The malignant scores are the possibilities of malignancy (predicted by models). The threshold of output score is 0.5 (larger than 0.5 classify as malignant and vice versa). As illustrated, the segmentation head of *DenseSharp* helps the model better locate the lesions than DenseNet, making the CAM of *DenseSharp* appears a more precise activation than that of DenseNet to the manual segmentation. In most cases, DenseNet and *DenseSharp* models not only activate the features in lesions' locations, but also activate the locations in the background, which not precisely utilizes the features of lesions themselves (the "correct evidences"). In some other cases, the two models face the two failures described in Fig. A.1, making their classification incorrect or lack of interpretability. *DenseSharp*$^+$ model only adapts the features upon lesions to classify the nodule, with better controllability and interpretability.