Characterization of Pulmonary Nodules in Computed Tomography Images Based on Pseudo-Labeling Using Radiology Reports

Yohei Momoki, Akimichi Ichinose, Yutaro Shigeto, Ukyo Honda, Keigo Nakamura, and Yuji Matsumoto

***Abstract*— A computer-aided diagnosis (CAD) system that characterizes nodules in medical images can help radiologists determine its malignancy. Preparing large volumes of labeled data for CAD systems, however, requires advanced medical knowledge. This makes it extremely difficult to develop such systems, despite their growing demand. In this paper, we propose a new training method to build an image classifier for characterization of nodules utilizing pseudo-labels, i.e., image labels automatically retrieved from radiology reports. A radiology report is a type of record in which radiologists present a summary of lesion characteristics and diagnosis. Labeling radiology reports is much easier than labeling radiology images, and can be done without high expertise. Using several thousand labeled reports, we constructed a hierarchical attention network-based text classifier to assign pseudo-labels of the characteristics of pulmonary nodules with high accuracy (macro F1-score of 0.941). Experimental results show that the image classifier trained with the pseudo-labels can achieve almost the same performance as the one trained with the labels annotated by radiologists: AUC 0.848 for the model trained with the pseudo-labels on 3,000 computed tomography (CT) images and 0.847 for the model trained with the manual labels on 800 CT images.**

***Index Terms*— Computer-aided diagnosis, lung nodule, nodule characterization, pseudo-labeling, radiology report.**

I. Introduction

LUNG cancer is the most common cancer and the leading cause of cancer-related deaths in the world [1]. Early detection and diagnosis are therefore crucial for an improved prognosis. Radiologists spend countless hours detecting nodules in computed tomography (CT) images. In addition, for each detected nodule, it takes a considerable amount of time and effort to confirm the detailed characteristics such as spiculation and pleural indentation to determine the disease name and malignancy.

Based on this background, many researchers have tackled pulmonary nodule detection from CT images [2]–[4] and characterization [5]–[8]. An issue of these tasks is the intensive annotation cost to build training data. Although the medical institutions have a large number of nodule images, these images cannot be used for training as they are unlabeled in most cases. It requires the advanced medical knowledge to manually annotate the detailed characteristics of nodules on images. Thus, the manual annotation is cost intensive and not scalable. Since it is difficult to construct a large training dataset which has many kinds of labels solely by manual annotation, researchers focus on nodule characterization of a single category, such as malignancy [7], [8] and opacity [6]. Although a previous study [5] that can classify multiple characteristics of lung nodules exists, it does not cover the characteristics required for qualitative assessment of lung nodules.

To reduce the cost of the training data construction, we propose a pseudo-labeling approach for automatic characterization of pulmonary nodules. Different from the standard approach to annotate labels directly on images, our approach utilizes radiology reports for labeling.

A radiology report is a record in which radiologists present a summary of the image findings and the corresponding diagnosis in order to communicate them to the physicians, who use these reports as reference. All the information regarding the characteristics identified by the radiologists is included in the reports. Therefore, we build a pseudo-labeler based on these reports, and the pseudo-labels generated as a result can be used for training the image classifier.

Our pseudo-labeler is a text classifier trained solely on radiology reports, meaning that our approach does not require labeled images. Although building such a pseudo-labeler requires a certain manual annotations on the reports, the cost of the annotation is much lower than that of the annotation on medical images: the reports are much easier to understand than the medical images, thus non-specialists can correctly annotate labels on the reports. Fig 1 shows an overview of our approach.

The main contributions of this study are as follows:

- We propose a pseudo-labeling approach for building an image classifier to mitigate the lack of training data. In our approach, the classifier is trained with pseudo labels. The advantage of our method is that it does not require any manual image annotations. This advantage is useful for low or zero resource setting as with our case.
- In our experiments (section IV), we show that the classifier trained using our pseudo-labeling approach achieves

Manuscript received October 14, 2020; revised January 21, 2021 and March 5, 2021; accepted March 29, 2021. Date of publication April 13, 2021; date of current version May 5, 2022. This article was recommended by Associate Editor B. Yao. *(Corresponding author: Yohei Momoki.)*

Yohei Momoki, Akimichi Ichinose, and Keigo Nakamura are with the Imaging Technology Center, Fujifilm Corporation, Tokyo 107-0052, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yohei.momoki@fujifilm.com; akimichi.ichinose@fujifilm.com; keigo.nakamura@fujifilm.com).

Yutaro Shigeto, Ukyo Honda, and Yuji Matsumoto are with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: yutaro.shigeto@riken.jp; ukyo.honda@riken.jp; yuji.matsumoto@riken.jp).

Color versions of one or more figures in this article are available at https:doi.org/10.1109/TCSVT.2021.3073021.

Digital Object Identifier 10.1109/TCSVT.2021.3073021

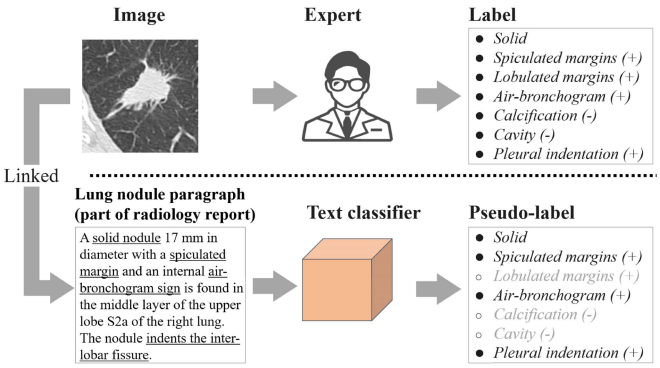

Fig. 1. The overview of our method. The top figure shows the procedure of manual annotation: expert (e.g., radiologist) annotates labels (the detail of characteristics) of a nodule image. The bottom figure shows our proposed pseudo-labeling. In our method, a text classifier predicts pseudo-labels from a given nodule paragraph (corresponding to a given nodule image) which is a part of a radiology report. The text classifier operates by inputting Japanese text, but in the figure, it is translated into English for explanation.

almost the same performance as the one trained on ground truth labels on images manually annotated by radiologists.

This study was approved by the institutional review boards of all participating institutions, which waived the requirement for patient consent.

II. Method

A. Problem Setup

Nodule characterization is a multi-label image classification problem, i.e., given an input image (nodule image), a classifier may predict multiple classes (characterization, such as well-defined margin, solid opacity, and air-bronchogram sign inside). In general, image classifiers are trained using labeled images. However, this conventional manner is hardly applicable in our task owing to the difficulty of creating labeled images. Instead of labeled images, we consider the use of two datasets; one is a set containing unlabeled image-text pairs. The other is a set of labeled texts.

The objective of this study is to find a function f, which can be used to assign a set of classes Y to an input image x, without the need for labeled images. In other words, our learning problem is to obtain an image classifier trained on the following two datasets:

$$\mathcal{D}_{\text{pair}} = \{(x_i, z_i) \mid i = 1, \cdots, n\}, \tag{1}$$

$$\mathcal{D}_{\text{text}} = \{(z'_i, Y'_i) \mid i = 1, \cdots, m\}, \tag{2}$$

where x_i is an image, z_i and z'_i are texts, and Y'_i is a set of labels. Note that there are no common texts in the two datasets.

B. Proposed Approach

To obtain an image classifier without the need for manual image annotation, we propose an approach of pseudo-labeling. In this approach, labels are provided to each image in $\mathcal{D}_{\text{pair}}$ through text classification.

Concretely speaking, we first train a text classifier g using $\mathcal{D}_{\text{text}}$, and subsequently predict pseudo-labels Y for each image in $\mathcal{D}_{\text{pair}}$ by the predictions of the texts:

$$Y_i = g(z_i). \tag{3}$$

After the pseudo-labels are obtained, we construct an image classifier f using ordinary supervised training with the pseudo-labels:

$$\mathcal{D}_{\text{pseudo}} = \{(x, Y) \mid (x, z) \in \mathcal{D}_{\text{pair}}\} \tag{4}$$

$$= \{(x, g(z)) \mid (x, z) \in \mathcal{D}_{\text{pair}}\}. \tag{5}$$

C. Classifiers

In our approach, any text and image classifiers can be used, and the choice depends on the datasets in use. In this paper, we use an attention-based text classifier and a CNN based image classifier.

1) Text Classifier: Our attention-based classifier is a modified version of the hierarchical attention networks (HAN) [9]. HAN first builds the document representation from the word representations and then predicts the labels of the document. To capture the hierarchical structure of the documents (a document is comprised of sentences, and a sentence is comprised of words), it uses a hierarchical attention mechanism: a document representation is computed by the weighted average of the sentence representations; similarly, the sentence representation is computed by the weighted average of the word representations.

The weights, or attention scores, for the computation of the weighted average are obtained using the attention mechanism [10]. Let s be a sentence, and $\mathbf{s}$ be its representation vector. The attention score for the i-th sentence in the document (i.e., text z in our case) is therefore computed as follows:

$$\alpha_i = \frac{\mathbf{q}^\top \mathbf{s}_i}{\sum_{s_j \in z} \mathbf{q}^\top \mathbf{s}_j} \tag{6}$$

where $\mathbf{q}$ is a query vector (trainable parameter) to compute the attention scores. The attention scores for words, which is used to take the weighted average of word embeddings to form $\mathbf{s}$, are computed in the same manner except that a different query vector is applied.

A problem with the attention computation is that the attention parameter vectors are common for all the labels. Consequently, the same attention scores are obtained, regardless of the labels for which the likelihood was estimated. In general, the important sentences in a document are different for each label. Then, the attention score for each label is calculated as follows:

$$\alpha_{c,i} = \frac{\mathbf{q}_c^\top \mathbf{s}_{c,i}}{\sum_{s_{c,j} \in z} \mathbf{q}_c^\top \mathbf{s}_{c,j}} \tag{7}$$

where $\mathbf{q}_c$ is a query vector associated with class c and $\mathbf{s}_{c,i}$ indicates the representation of the i-th sentence for the class c. This technique is almost the same as the selective attention [11] used in the relation extraction task. Selective attention calculates attention scores using queries associated with the relation. The difference from the selective attention model is that our model has a sentence level encoder. As shown in our experiments (Sec. IV), this modification has a beneficial effect on the results of classification.

We compute the document representations $\mathbf{z}_c$ as follows:

$$\mathbf{z}_c = \sum_i \alpha_{c,i} \mathbf{s}_{c,i} \tag{8}$$

After document representations are constructed, HAN assigns labels to the given document:

$$p_c = \text{sigmoid}(\mathbf{w}_c^{\top}\mathbf{z}_c + \mathrm{b}_c) \tag{9}$$

where p_c indicates the predicted confidence score for the class c and $\mathbf{w}_c$ and b_c are trainable parameters.

Given the training data $\mathcal{D}_{\text{text}}$, we optimize the model with binary cross-entropy loss as follows:

$$L_{\text{text}} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{c\in C}\left(y_c^{(i)}\log(p_c^{(i)}) + (1-y_c^{(i)})\log(1-p_c^{(i)})\right) \tag{10}$$

where C is a set of classes, $p_c^{(i)}$ is the confidence score of z_i' for class c, and $y_c^{(i)}$ is an indicator function such that $y_c^{(i)} = 1$ if $c \in Y_i'$ and 0 otherwise.

2) Image Classifier: We use a CNN based model, which is a simpler version of VGG [12]. VGG is a well-known CNN classifier, which secured the first and second places in the localization and classification tasks, respectively, in the ImageNet Challenge 2014. Although it works well on other datasets, training a VGG may require a vast amount of training data. In our preliminary experiment, VGG did not work well on our dataset because our training data was limited. Thus, we use a more shallow model. Each convolution layer consists of $3\times3\times3$ sized kernel followed by batch normalization [13]. In order to avoid overfitting, we reduce the number of trainable parameters by applying the global average pooling (GAP) layer. In our classifier, we used ReLU as the activation function in all layers except the final layer, like original VGG. According to the prior works [14], [15], ReLU function offers better performance in deep learning compared to the sigmoid or tanh functions. Recently, new activation functions such as Mish [16] and Swish [17] have been proposed, but they did not offer any performance improvement for our tasks. The configuration of our model is detailed in Table I. In this table, the batch normalization and ReLU activation in each convolution layer is not shown for brevity.

Some classes such as well/ill-margin class have different criteria depending on the radiologists to read CT images. Therefore, it is extremely difficult to produce consistent ground truth data for such classes. In order to handle these inconsistent and noisy training data, we adopt bootstrapping cross-entropy loss [18]. Given the training data $\mathcal{D}_{\text{pseudo}}$, the object function is as follows:

$$\begin{aligned} L_{\text{image}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c\in C}\Big[&\left(\beta y_c^{(i)} + (1-\beta)t_c^{(i)}\right)\log(p_c^{(i)}) \\ &+\left(\beta(1-y_c^{(i)}) + (1-\beta)(1-t_c^{(i)})\right)\log(1-p_c^{(i)})\Big] \end{aligned} \tag{11}$$

where $p_c^{(i)}$ is the confidence score of the image x_i for class c, $y_c^{(i)}$ is an indicator function such that $y_c^{(i)} = 1$ if $c \in Y_i$ and 0 otherwise, $t_c^{(i)}$ is a threshold function such that $t_c^{(i)} = 1$ if $p_c^{(i)} > 0.5$ and 0 otherwise, and β is the scale factors to balance the terms, which is set to 0.8 in our experiments.

TABLE I
STRUCTURE OF OUR IMAGE CLASSIFIER

layer name	output size	layers*
conv-1	$96\times128\times128$	conv: $\left[\ 3\times3\times3, 8, 1\ \right]\times2$
pool-1	$48\times64\times64$	max pool: $\left[\ 2\times2\times2, 8, 2\ \right]$
conv-2	$48\times64\times64$	conv: $\left[\ 3\times3\times3, 32, 1\ \right]\times2$
pool-2	$24\times32\times32$	max pool: $\left[\ 2\times2\times2, 32, 2\ \right]$
conv-3	$24\times32\times32$	conv: $\left[\ 3\times3\times3, 64, 1\ \right]\times2$
conv-4	$24\times32\times32$	conv: $\left[\begin{array}{c}3\times3\times3, 128, 1\\ 3\times3\times3, 256, 1\end{array}\right]$
avg_pool	256	global average pool: 256-d
fc1	512	fully connected: 512-d, ReLU, DropOut
fc2	11	fully connected: 11-d, sigmoid

*For convolution and max pool layers, filter configurations are described in the following manner: [kernel size, output channel, stride]. Each convolution layer is followed by batch normalization and ReLU.

III. RELATED WORK

Image Analysis for Lung Nodules: Lung cancer is an important disease and attracts a great deal of interest in the field of radiology. Many applications for detection [2]–[4], [19], segmentation [20], and characterization [5]–[8] of lung nodules, which are candidate lesions for lung cancer, have been developed so far. Most of the studies on characterization of lung nodules are on classification of their malignancy [7], [8], and there are still few studies on the prediction of imaging features such as morphology and internal characteristics of lung nodules [5]. Our study belongs to the latter. These studies are similar to our problem setting in that they used ROI (Region of Interest) images of a lung nodule as input and predicted its characteristics. However, these studies differ from ours in that they trained their image classifier on manually-annotationed data by radiologists, and the number of nodules used for training is limited (less than 1,500). In addition, the number of characteristics to be predicted was also smaller than that of ours. The difficulty in preparing a large number of annotated data may have led to a lack of the comprehensiveness of the characteristics required for qualitative assessment of lung nodules.

Medical Image Analysis Using Radiology Report: There are some studies on the use of radiology reports for tasks of analyzing radiological images to reduce the cost of the training data construction [19]. For example, a lesion annotation network, which is trained using descriptions of target lesions extracted from radiology reports, has been proposed [21]. This framework does not require manually annotation to images like our approach. In this study, labels are automatically obtained by text mining based on words in the reports and lesion ontology. However, at least for the Japanese language, ontologist of radiology terms such as RadLex [22] have not been developed, and thus, this method cannot be used in this study.

Pseudo Labeling Approach: Our approach is related to self-training, which is a well-known pseudo-labeling approach, but

differs from it in terms of how the pseudo-labels are created. In the self-training approach, the classifier is first trained using a small number of labeled samples and the labels of unlabeled examples are subsequently predicted. Next, these examples with pseudo-labels are included in the training set and the classifier is re-trained. Thus, this method requires image supervision. In our setting, however, a considerable effort is required to create even a small amount of manually labeled data. Therefore, self-training is not applicable, considering our settings.

Due to the difficulty of data annotation, semi-supervised learning is a popular approach in medical domain [2], [23]–[26]. In this context, pseudo-labeling has been discussed [24], [25]. Although they have produced promising results, the methods assume that there exist (small) labeled images. Thus, their methods are not applicable to our settings as with self-training.

Multi-view learning [27] is similar to our approach in that it uses multiple information sources such as image and text information. The objective of multi-view learning is to train better classifiers by analyzing multiple information sources simultaneously. In our setting, text information is not given in inference phase, i.e., the goal of our task is to build an image classifier (not multi-view classifiers). Our approach uses two distinct views (the text and image information), however, it does not use them at the same time. The text information is used to construct the training data for image classifiers, and image information is used in inference phase (nodule characterization). Once we learn an image classifier using the pseudo-labeled data, and if we have an additional set of image and report pairs, we can apply multi-view learning to obtain better pseudo-labels for such a additional set. We regard this as a future work.

IV. Experiments

This section describes two experiments performed in this study. The first experiment (Sec. IV-A) is pseudo-label recognition. This task is to predict categories of a given text. Through this experiment, we investigate the quality of pseudo-labeling for images. The experiment described in Sec. IV-B shows the proposed image classifier works well in pulmonary nodule characterization without manual labeling on images.

Notably, our target is pulmonary nodule and mass. We defined the classes based on the terms described in [28]. We selected 13 key terms from the guideline and used them for class definition. Specifically, six types of marginal characteristics (well-defined/ill-defined/ smooth/irregular/spicula/lobulation), three types of opacities (ground glass type/solid type/part solid type), three types of internal characteristics (air-bronchogram/cavity/calcification) and one type of external characteristic (pleural indentation) were adopted.

A. Pseudo-Label Recognition

1) Dataset: We prepared the following datasets for the pseudo-label recognition.

TABLE II
Label Definition and Volume for the Pseudo-Label Recognition

		the count of label appearances		
category	name	training	validation	test
margin	ill-defined	95	19	96
	well-defined	521	27	278
	irregular	166	17	150
	smooth	162	18	77
	spiculated	208	25	272
	lobulated	293	25	260
opacity	solid type	787	44	933
	part solid type	175	16	279
	ground glass type	690	20	637
internal	air-bronchogram (+)	205	23	347
	air-bronchogram (-)	27	16	95
	cavity (+)	216	16	253
	cavity (-)	301	24	419
	calcification (+)	241	18	203
	calcification (-)	438	34	577
external	pleural identation	438	34	577

- **dataset 1** is a set of 200,000 radiology reports including those of the chest collected from a university hospital in Japan. This dataset contains data spanning over a period of 10 years.
- **dataset 2** is a set of 6,540 manually labeled texts. We used this dataset for training and validation sets. The validation set was randomly extracted such that each label contained at least 15 instances, and the rest were used to form the training set. This dataset is equivalent to $\mathcal{D}_{\text{text}}$ (mentioned in Sec. II-A).
- **dataset 3** is a set of 3,144 nodules in which CT images and manually labeled texts exist in pairs. We used this dataset as a test set. It is equivalent to $\mathcal{D}_{\text{pair}}$ (mentioned in Sec. II-A)

Dataset 2 and dataset 3 were created from dataset 1. The annotation work was completed by two annotators (non-doctors). For classification, we defined 16 types of labels (see Table II). These labels are based on the 13 terms mentioned in [28] and their modality usage in a report. We defined positive and negative labels separately because some terms are written in a positive or negative context. Only those labels concerned with internal characteristics have both positive (+) and negative (-) modalities. Labels without brackets only have positive modality. The annotators assigned a value of 1 if a corresponding description to the label was found in the text, and a 0, otherwise. Here, 0 means underspecified. Fig 2 shows the flowchart of our text labeling. The annotators carried out text labeling as follows: 1) They read the "Findings" section of a radiology report and extracted only those paragraphs that were related to solitary pulmonary nodules; 2) They provided labels to a paragraph based on the facts presented in the text without guesswork. The reports are written in Japanese and involved five types of characters (Alphabet / Number / Kanji / Hiragana / Katakana). Note that more than 100 radiologists were involved for the creation of the reports, including primary and secondary readers. As such, in a report, a label appears

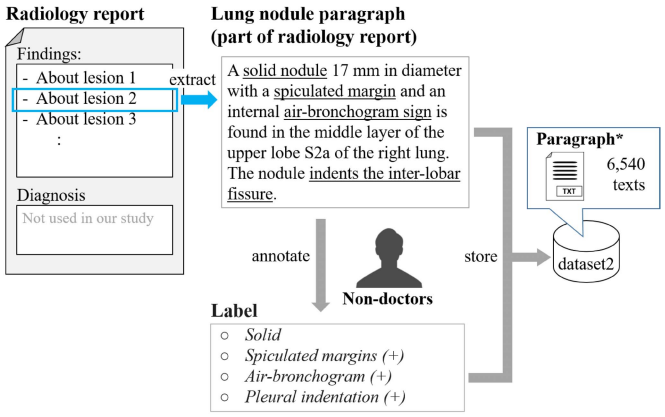

Fig. 2. Flowchart of our text labeling. Annotators (non-doctors) extract texts which is a paragraph refering to a solitary lung nodule in the 'Findings' section of a radiology report. These labeled texts are used to train our pseudo-labeler.

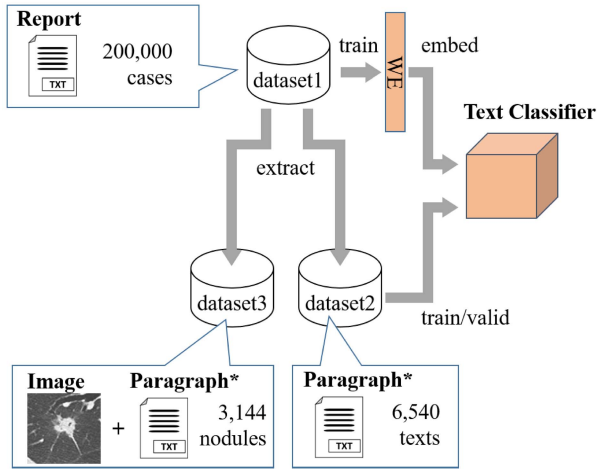

Fig. 3. Flowchart of the experiment of pseudo-label recognition. Here, '∗' means manually labeled datas. In this task, dataset 3 is regarded as a test set. Dataset 2 and 3 were created from dataset 1 and are mutually disjoint.

in the form of its various synonyms or in different wordings. During labelling, they did not refer to any nodule images.

The dataset statistics are summarized in Table II. We show the clinical information (age and sex) for the dataset 1, 2 and 3 in Table VI. We summarize the flowchart of the experiment of pseudo-label recognition in Fig 3.

For preprocessing, all the alphabets with uppercase were changed to lowercase.

2) Input Representation: We created the input representation in two ways:

- *BoW* : Bag of words. We used the count of each word in the paragraph as the input representation. The words are created by tokenizing dataset 2 by MeCab [29], which is commonly used approach for tokenizing Japanese sentences. The number of unique words required to create a BoW was 2,073.
- *WE* : Word embeddings. We used the Sentencepiece model in unigram mode [30] to tokenize dataset 1. Next, we trained the Skip-gram [31] model to obtain the word embeddings. We set the maximum vocabulary size for Sentencepiece to 32,000, and the dimension for word embeddings to 256. The numbers of words and sentences were fixed to be the maximum length amongst all the inputs, and padding was done for the shorter sequences.

TABLE III
THE OPTIMAL HYPERPARAMETERS FOR EACH NEURAL NETWORK-BASED METHOD

		CNN	HAN	HAN+SA
WE		rand	static	static
learning rate		1e−4	1e−3	1e−2
batch size		8	8	64
CNN	units	512	-	-
	kernel	3×256		
	stride	1		
	padding	VALID		
	activation	ReLU		
GRU	units	-	128	64
Attention	units	-	256	128

3) Text Classifiers: We evaluated the following methods.

- **SVM** is a support vector machine model with an RBF kernel. We adopted binary relevance to build multi-label classifier. The model receives a BoW as an input and performs classification. To obtain better accuracy, we performed Grid search to identify the best (C, γ) pair in the range of $C = [0.001, 0.01, 0.1, 1, 10, 100]$, $\gamma = [0.001, 0.01, 0.1, 1, 10, 100]$.
- **CNN** is a model similar to [32]. We used Max pooling with a single-layered CNN.
- **HAN** is a model similar to [9]. We used a single-layered Gated recurrent unit (GRU) in both word and sentence level encoders. The word and sentence level attention scores were calculated according to eq.6.
- **HAN+SA** is a modified version of **HAN**, in which Selective Attention [11] was applied. The word and sentence level attention scores were calculated according to eq.7.

Neural network-based methods (CNN, HAN, and HAN+SA) use a Sigmoid activation function in the last layer. The models take the WE as input and output a probability vector corresponding to the label. We optimized the cross-entropy loss using the Adam optimizer with the optimal hyperparameters for each method (see Table III). Here, 'static' means that the pre-trained WE are kept static during training and 'rand' means that the WE are randomely initialized and modified during training.

4) Evaluation: We computed the F1-score which is often adopted in multilabel image classification tasks [27]. To change the confidence score into a label decision, we adjusted the threshold for each label that produced the best F1-score in the validation set and applied it to the test set. We used the macro-averaged F1-score obtained by averaging the F1-scores over the entire label to compare model performances.

5) Results: As shown in Table IV, Fig 4 and Fig 5, HAN+SA model performed the best with the macro-averaged F1-score. In class analysis, the HAN+SA generally showed a better F1-score than the other models. Compared to other deep learning-based methods, HAN+SA has the advantage

TABLE IV
PER-LABEL AND MACRO AVERAGED F1-SCORES IN THE TEST SET OF PSEUDO-LABEL RECOGNITION

category	name	SVM	CNN	HAN	HAN+SA
margin	ill-defined	0.73	0.55	0.81	**0.88**
	well-defined	**0.99**	0.96	0.91	0.95
	irregular	0.00	0.87	0.92	**0.92**
	smooth	**0.97**	0.76	0.85	0.87
	spiculated	0.00	0.95	0.95	**0.98**
	lobulated	**0.99**	0.93	0.96	0.94
opacity	solid type	0.96	0.95	0.96	**0.98**
	part solid type	0.00	0.79	**0.89**	**0.89**
	ground glass type	0.95	0.93	**0.96**	0.95
internal	air-bronchogram (+)	0.00	0.67	0.88	**0.95**
	air-bronchogram (-)	0.00	0.68	0.89	**0.91**
	cavity (+)	0.00	0.88	0.93	**0.96**
	cavity (-)	0.00	0.93	0.96	**0.99**
	calcification (+)	0.01	0.64	0.91	**0.93**
	calcification (-)	0.91	0.98	0.98	**0.99**
external	pleural identation	0.05	0.94	0.95	**0.96**
(macro averaged F1-score)		0.410	0.838	0.919	**0.941**

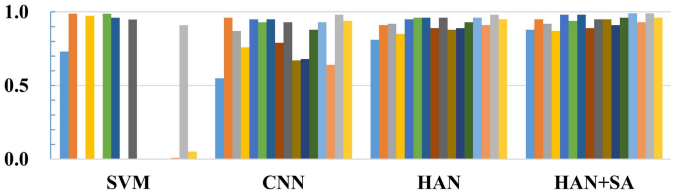

Fig. 4. Per-label F1-scores in the test set of the pseudo-label recognition. The bars are in the row order of Table IV.

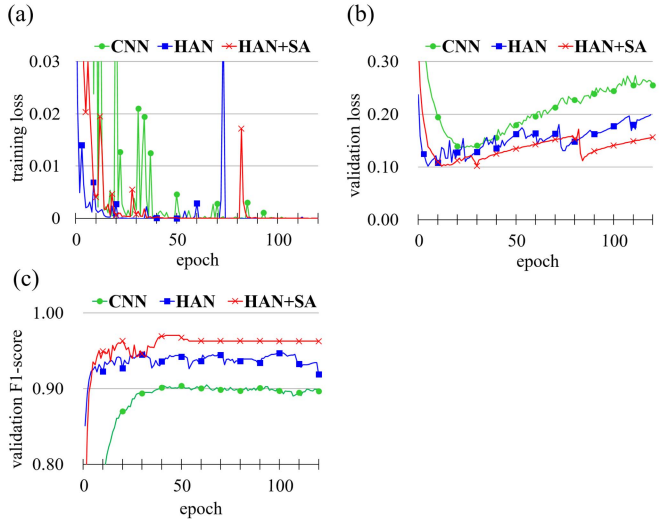

Fig. 5. Training/Validation loss and F1-score plot of the neural network-based methods. (a) training loss, (b) validation loss, (c) validation F1-score.

in that it can pay attention to the words for each class and was therefore effective in this task. On the contrary, classification of the "well-defined" and "smooth" class showed much inferior results than SVM. Not only these classes are formed with one word as features, but they also have only a few variants of synonyms. As such, it is easier to classify than other classes that are formed with words as features or have many variants of synonyms. A close examination on these labels showed that they have high precision but low recall and are not tend to be labeled. These labels seem to be overlooked by accidentally focusing on unintended words, leading to lower F1-scores compared to SVM.

B. Nodule Characterization

1) Setups: For each nodule, we cropped a three-dimensional nodule patch from the CT images. The intensity of the extracted image I is then scaled to the Hounsfield unit $I_{HU} \in [-1800, 600]$ and linearly normalized to $I_{norm} \in [0, 1]$ range using the transformation $I_{norm} = (I_{HU} + 1800)/2400$ before being used as input to the network. The spacings (mm/pixel) of CT scans between patients and machines were different. Therefore, all the images were rescaled to 1 mm isotropic voxels in preprocessing steps. In addition, the size of pulmonary nodules differs greatly. Therefore, we fixed the size of a nodule patch to $98 \times 128 \times 128$ with zero padding.

We optimized the bootstrapping cross entropy loss using Adam optimizer with a base learning rate of 0.0001. During training, we used a batch size of 16 and set a dropout ratio to 0.5. The following data augmentation methods were used to ensure robustness of the proposed framework: random flipping in three directions, resizing, rotation by any angle in 3D, addition of Gaussian noise, smoothing and sharpening using Gaussian filter, changing slice thickness by thinning out slices.

In most cases, only those characteristics that form the basis for diagnosis are mentioned in the radiology reports. That is, some characteristics are not included in the reports. Consequently, the pseudo-labels obtained from the pseudo-label recognition task are not always available for all classes. Therefore, during training, we do not calculate the loss for classes for which a pseudo-label are not obtained from the pseudo-label recognition task.

2) Dataset: We prepared the dataset 3 mentioned in Sec. IV-A and the following dataset 4 for training.

- **dataset 4** is the subset of LIDC-IDRI dataset [33], which is a public and comprehensive dataset of pulmonary nodules. This dataset does not contain any reports. From the LIDC-IDRI dataset, we selected only 821 nodules that are greater than 1 cm in diameter.

The detailes of the dataset used for training are shown in Table V. We compared the images of datasets 3 and 4 in detail and found that the domain difference between the two datasets is small.

We also prepared a validation set for hyperparameter tuning, which is a set of 100 CT images containing at least one nodule. It is a subset of dataset 1 and there is no overlap with dataset 2 and 3. One radiologist annotated against this dataset.

3) Evaluation: We select AUC as an evaluation metric, which is the area under the receiver operating characteristic (ROC) curve and a popular metric in image classification tasks. We calculate the average AUC simply by averaging the AUC for each class as a measure of the performance. For the test set, we selected 300 nodule images from dataset 1. The test set has no overlap with datasets 2,3. We show the clinical information (age and sex) for the test set in Table VI.

TABLE V
INFORMATION ABOUT CT SCANS

characteristics	dataset 3	dataset 4	test set
No. of cases	2,905	486	298
No. of series	3,128	486	298
(contrast series)	598	250	71
No. of nodules	3,144	821	300
slice thickness [mm]			
< 1.0	1,079	16	96
< 2.0	1,233	176	127
< 3.0	0	208	0
< 4.0	0	83	0
< 5.0	0	1	0
≥ 5.0	826	2	75
radiation dose [mAs]			
< 50	64	99	9
< 100	613	64	70
< 150	605	84	61
< 200	1,159	47	98
< 250	509	79	32
< 300	27	44	5
≥ 300	151	69	23

TABLE VI
CLINICAL INFORMATION

characteristics	dataset 1	dataset 2	dataset 3	test set
No. of cases	200,000	6,540	2,905	298
age				
< 20	0	0	0	0
< 30	6,692	52	24	1
< 40	10,584	142	62	5
< 50	18,117	317	173	11
< 60	27,222	685	291	26
< 70	57,740	2,156	898	99
< 80	56,569	2,301	1,094	116
< 90	21,357	823	344	37
< 90	1,668	64	18	3
≥ 100	51	0	1	0
sex				
male	117,374	4,093	1,775	182
female	82,618	2,447	1,130	116
unkown	8	0	0	0

For each nodule, one radiologist with an experience of more than 10 years manually built the ground truth data of the 11 categories mentioned above. These 11 categories are the same as the 16 labels (shown in Table II); Internal characteristics defined separately for positive and negative are grouped into one category, and ill-defined/well-defined, irregular/smooth labels are grouped into two categories.

4) Compared Methods: To verify the effect of our proposed method, we compared the characterization results trained on the following datasets:

(i) dataset 3 with pseudo-labels obtained using the HAN+SA model, which is equivalent to $\mathcal{D}_{\text{pseudo}}$ (**proposed method**),
(ii) dataset 4 with labels manually annotated by radiologists (**supervised**),
(iii) dataset 3 and dataset 4 with pseudo-labels obtained using the model trained on (ii) (**self-training**).

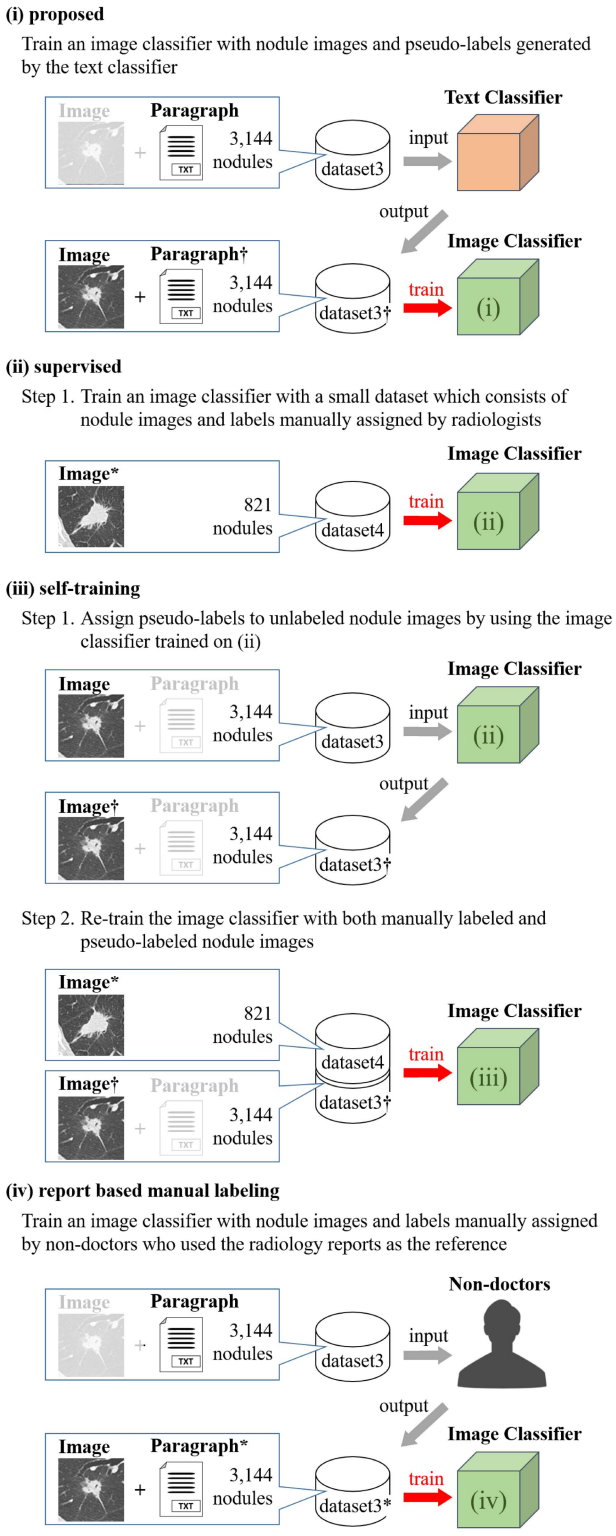

Fig. 6. Flowchart of the experiment of nodule characterization. Here, '$*$' means manually labeled datas and '†' means pseudo-labeled datas.

In addition, to demonstrate that the accuracy of pseudo-labels obtained by pseudo-labeler is sufficient to perform nodule characterization, we compared the results obtained as a result of training on (i):

(iv) dataset 3 with labels assigned by two non-doctors who used the reports as the reference (**manual labeling**).

TABLE VII
COMPARISON OF THE PERFORMANCE FOR VARIOUS METHODS

method	average AUC 8 Classes	11 Classes
(i) proposed	0.848 ± 0.012	-
(ii) supervised	0.847 ± 0.006 †	0.842 ± 0.005
(iii) self-training	0.856 ± 0.008 †	0.850 ± 0.006
(iv) manual labeling	0.855 ± 0.008 †	-
(i) + (ii)	0.877 ± 0.007 **	0.873 ± 0.006

** There was significant difference compared to (i) ($p < 0.001$)
† There was no significant difference compared to (i) ($p > 0.05$)

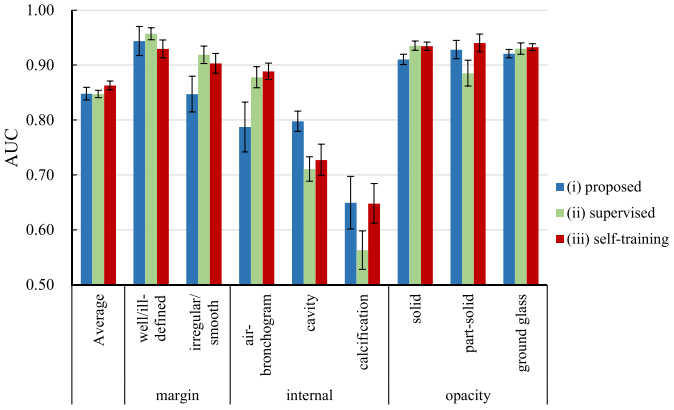

Fig. 7. Performance comparison among (i)–(iii). For each model trained on (i)–(iii), the AUC for 8 aforementioned classes are shown.

We summarize the flowchart of the experiment of nodule characterization in Fig 6.

If the pseudo-label recognition model performs effectively, the resulting pseudo-labels should be identical to the ones annotated in (iv). However, as shown in Sec. IV-A, our pseudo-label recognition model cannot always predict labels correctly. In other words, the pseudo-labels are noisier than the labels annotated in (iv). By comparing the results obtained by training on (i) and (iv), we reveal the effect of that noise on the nodule characterization task.

Amongst all the classes, the following three (spicular, lobular, and pleural indentation classes) were included in the reports, in most cases, only if these characteristics are present and have been determined to be important for making diagnosis. Therefore, negative labels cannot be retrieved by the pseudo-label recognition for these 3 classes, and consequently, we compare the results of the experiments for the remaining 8 classes.

Finally, we show that the combined use of a small dataset with manually annotated labels and a large dataset with pseudo-labels generated by the pseudo-labeler leads to performance improvement. For all the 11 classes, we compare the results trained on (ii) and (i)+(ii).

We conduct the experiments 9 times for each method and evaluate the difference in performance between the methods by two-tailed Student's t-test.

5) Results: The results are shown in Table VII and Fig 7, and training/validation accuracy and loss plot of the proposed method is shown in Fig 8. As shown in Table VII, there was no significant difference between the results from the proposed

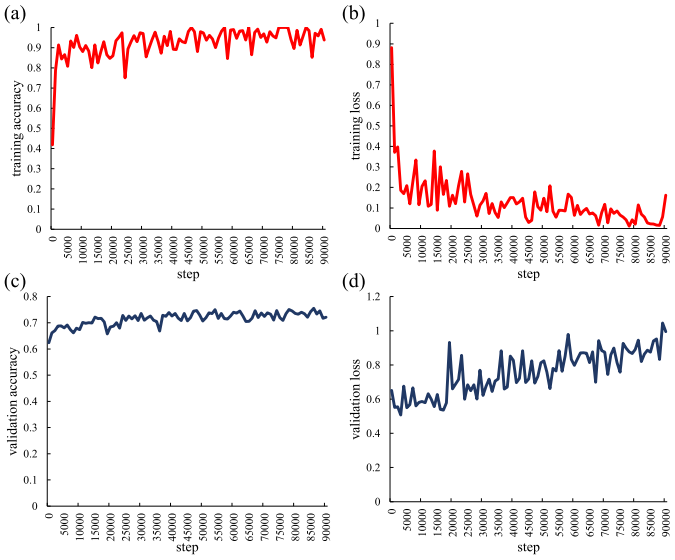

Fig. 8. Training/Validation accuracy and loss plot of the proposed method. (a) training accuracy, (b) training loss, (c) validation accuracy, (d) validation loss.

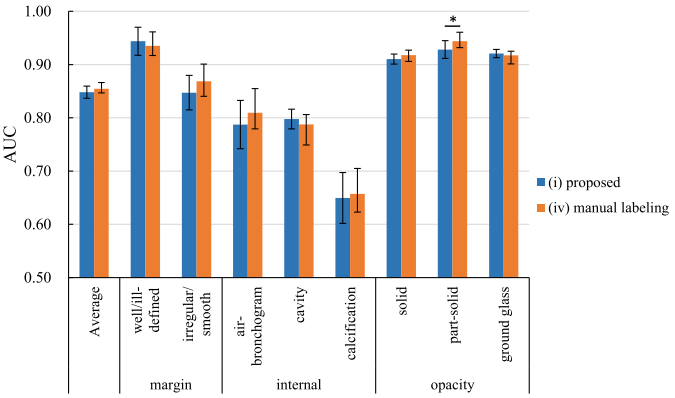

Fig. 9. Performance comparison between (i) proposed method versus (iv) manual labeling. '*' indicates there was significant difference ($p < 0.05$).

method and those from supervised training (method (i) vs (ii); $p = 0.86$). From the results, we conclude that the proposed method can achieve almost the same performance compared to the method using image supervision. The examples of the model output trained using the proposed method are shown in Fig 10. However, the differences in performance are relatively small considering the differences in the sizes of training datasets ((i) 3,144 vs (ii) 821). As mentioned above, all the characteristics are not included in a radiology report. Furthermore, pseudo-labels obtained from the reports are not always available for all the classes, whereas the labels that radiologists manually annotated to the image are always available for all the classes. Therefore, as presented in Table II, the number of labels is relatively small compared to the number of nodule patches used for training (For example, air-bronchogram class is labeled only for 442/3,144 nodules). However, since there are far more reports and images in hospitals than collected this time, the better characterizer can be achieved by utilizing them, without further manual image annotation.

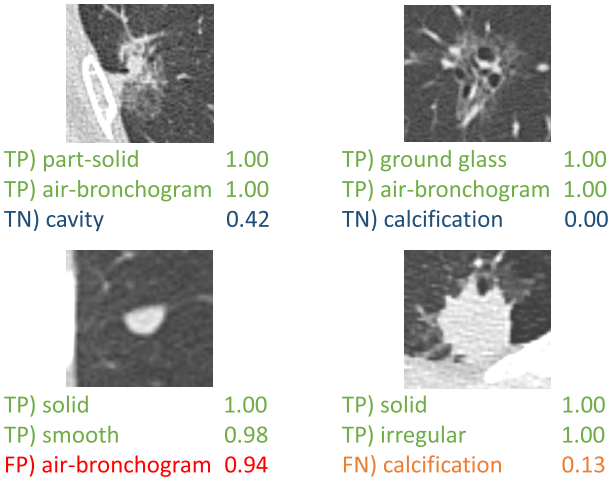

Fig. 10. Examples of the model output trained on only pseudo-labels obtained from the reports.

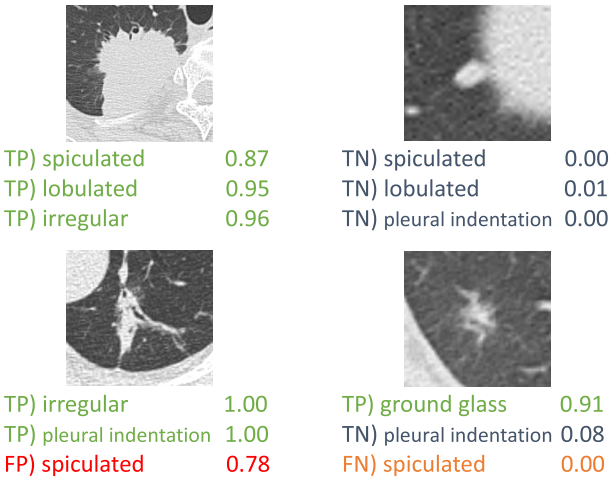

Fig. 11. Examples of the model output trained on supervised-labels by radiologists and pseudo-labels obtained from the reports.

The results also show that there was no significant difference in performance when compared to those trained with labels given by two non-doctors who used the reports as the reference (method (i) vs (iv); $p = 0.20$). However, the AUC was significantly lower for the proposed method in determining part solid type ($p = 0.046$). In addition, as shown in Fig 9, the performance of the proposed method was lower for irregular/smooth class and air-bronchogram class. For these classes, the performance of pseudo-label recognition was slightly lower, and thus, the generated labels may have been noisy. Although there was a slight difference in performance when comparing self-training and the proposed method ($p = 0.123$), there was almost no difference between the performance of self-training and those trained with manually labeled dataset ($p = 0.701$). Thus it is highly likely that the proposed method can achieve the same performance as the self-training by improving the performance of the label recognition in the future.

It is also shown by (i)+(ii) that training using both of the datasets, i.e. a small dataset with labels manually annotated and a large dataset with pseudo-labels, led to significant improvement in the performance (method (i)+(ii) vs (ii); $p < 0.001$). Thus, our pseudo-labeling approach can also be used to effectively augment the data, with much less cost than directly annotating the images. The examples of the model output trained using both of the datasets are shown in Fig 11.

V. Conclusion

In this paper, we tackled the characterization of pulmonary nodules. A challenge of this task is how to handle the lack of training data (labeled images). To solve this issue, we proposed a pseudo-labeling approach. The pseudo-labeler was trained solely on radiology reports, and thus, our approach does not require manual image annotation. Our approach achieved almost the same performance as the model trained on the data manually labeled by radiologists.

As future work, we plan to employ this method to build a CAD system that performs more detailed characterization of pulmonary nodules, other lung diseases such as pneumonia, and organ diseases such as a liver tumor or brain stroke by only annotating a moderate size of radiology reports.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] D. Wang, Y. Zhang, K. Zhang, and L. Wang, "FocalMix: Semi-supervised learning for 3D medical image detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3951–3960.

[3] J. Ding, A. Li, Z. Hu, and L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 559–567.

[4] A. Masood *et al.*, "Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images," *J. Biomed. Informat.*, vol. 79, pp. 117–128, Mar. 2018.

[5] F. Ciompi *et al.*, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Sci. Rep.*, vol. 7, no. 1, Jun. 2017, Art. no. 46479.

[6] X. Tu *et al.*, "Automatic categorization and scoring of solid, part-solid and non-solid pulmonary nodules in CT images with convolutional neural network," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Dec. 2017.

[7] S. Hussein, R. Gillies, K. Cao, Q. Song, and U. Bagci, "TumorNet: Lung nodule characterization using multi-view convolutional neural network with Gaussian process," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 1007–1010.

[8] B. Veasey *et al.*, "Lung nodule malignancy classification based ON NLSTx data," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1870–1874.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[11] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 2124–2133.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[14] F. Ertam and G. Aydin, "Data classification with deep learning using tensorflow," in *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, Oct. 2017, pp. 755–758.

[15] M. D. Zeiler *et al.*, "On rectified linear units for speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3517–3521.
[16] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*. [Online]. Available: http://arxiv.org/abs/1908.08681
[17] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*. [Online]. Available: http://arxiv.org/abs/1710.05941
[18] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," 2014, *arXiv:1412.6596*. [Online]. Available: http://arxiv.org/abs/1412.6596
[19] E. Pesce, S. J. Withey, P.-P. Ypsilantis, R. Bakewell, V. Goh, and G. Montana, "Learning to detect chest radiographs containing pulmonary lesions using visual attention networks," *Med. Image Anal.*, vol. 53, pp. 26–38, Apr. 2019.
[20] H. Tang, C. Zhang, and X. Xie, "Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 266–274.
[21] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8523–8532.
[22] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," *RadioGraphics*, vol. 26, no. 6, pp. 1595–1597, Nov. 2006.
[23] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 605–613.
[24] W. Bai *et al.*, "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 253–260.
[25] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2020, pp. 614–623.
[26] G. Xu *et al.*, "CAMEL: A weakly supervised learning framework for histopathology image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10682–10691.
[27] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," 2013, *arXiv:1304.5634*. [Online]. Available: http://arxiv.org/abs/1304.5634
[28] *General Rule for Clinical and Pathological Record of Lung Cancer*, 8th ed., KANEHARA Co., LTD, Japan Lung Cancer Soc., Tokyo, Japan, 2017.
[29] T. Kudo. (2005). *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. [Online]. Available: https://taku910.github.io/mecab/
[30] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 66–75.
[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[32] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
[33] S. G. Armato *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.*, vol. 38, no. 2, pp. 915–931, Jan. 2011.

Yohei Momoki was born in Tokyo, Japan, in 1984. He received the B.S. degree in information science from the Tokyo University of Science in 2006, and the M.S. degree in information science from the Tokyo Institute of Technology, in 2009, respectively.

He is currently a Researcher with the Imaging Technology Center, Fujifilm Corporation, Tokyo, Japan. His main research interests include natural language processing and data-to-text generation.

Akimichi Ichinose was born in Nagano, Japan, in 1993. He received the B.S. and M.S. degrees in mechano-informatics from the University of Tokyo, Tokyo, Japan, in 2015 and 2017, respectively.

Since 2017, he has been a Researcher with the Imaging Technology Center, Fujifilm Corporation, Tokyo. His main research interest includes medical image recognition.

Yutaro Shigeto was born in Kumamoto, Japan, in 1989. He received the Ph.D. degree from the Nara Institute of Science and Technology, Japan, in 2017.

He has been a Research Scientist with the Software Technology and Artificial Intelligence Research Laboratory, Chiba Institute of Technology, since 2017. He is also a Visiting Scientist at the RIKEN Center for Advanced Intelligence Project (AIP). His research interests include data mining and natural language processing.

Ukyo Honda was born in Miyagi, Japan, in 1991. He received the B.A. degree in law from Keio University, Japan, in 2016, and the M.S. degree in engineering from the Nara Institute of Science and Technology, Japan, in 2019, where he is currently pursuing the Ph.D. degree.

His research interests include natural language processing and multimodal language processing.

Keigo Nakamura was born in Tokyo, Japan, in 1977. He received the B.S. and M.S. degrees in informatics from the University of Chiba, Chiba, Japan, in 2000 and 2003, respectively.

Since 2003, he has been a Researcher with the Imaging Technology Center, Fujifilm Corporation, Tokyo. His main research interest includes medical image recognition.

Yuji Matsumoto received the B.S., M.S., and Ph.D. degrees in information science from Kyoto University, Japan, in 1977, 1979, and 1990, respectively.

He joined the Machine Inference Section, Electrotechnical Laboratory, in 1979. He has been an Academic Visitor with the Imperial College of Science and Technology, London, U.K., the Deputy Chief of the First Laboratory with the Institute for New Generation Computer Technology (ICOT), an Associate Professor with Kyoto University, and a Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, until 2020. He is currently a Team Leader with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan. His main research interests include natural language understanding and machine learning.

Dr. Matsumoto is a fellow of ACL and the Information Processing Society of Japan.