IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2020, accepted January 23, 2020, date of publication February 3, 2020, date of current version February 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971391

MaligNet: Semisupervised Learning for Bone Lesion Instance Segmentation Using Bone Scintigraphy

TERAPAP APIPARAKOON[1], NUTTHAPHOL RAKRATCHATAKUL[1], MAYTHINEE CHANTADISAI[2], USANEE VUTRAPONGWATANA[3], KANAUNGNIT KINGPETCH[3], SASITORN SIRISALIPOCH[3], YOTHIN RAKVONGTHAI[3,4], TAWATCHAI CHAIWATANARAT[3,4], AND EKAPOL CHUANGSUWANICH[1]

[1]Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand
[2]Division of Nuclear Medicine, Department of Radiology, Thai Red Cross Society, King Chulalongkorn Memorial Hospital, Bangkok 10330, Thailand
[3]Division of Nuclear Medicine, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[4]Chulalongkorn University Biomedical Imaging Group, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand

Corresponding author: Ekapol Chuangsuwanich (ekapolc@cp.eng.chula.ac.th)

ABSTRACT One challenge in applying deep learning to medical imaging is the lack of labeled data. Although large amounts of clinical data are available, acquiring labeled image data is difficult, especially for bone scintigraphy (i.e., 2D bone imaging) images. Bone scintigraphy images are generally noisy, and ground-truth or gold standard information from surgical or pathological reports may not be available. We propose a novel neural network model that can segment abnormal hotspots and classify bone cancer metastases in the chest area in a semisupervised manner. Our proposed model, called MaligNet, is an instance segmentation model that incorporates ladder networks to harness both labeled and unlabeled data. Unlike deep learning segmentation models that classify each instance independently, MaligNet utilizes global information via an additional connection from the core network. To evaluate the performance of our model, we created a dataset for bone lesion instance segmentation using labeled and unlabeled example data from 544 and 9,280 patients, respectively. Our proposed model achieved mean precision, mean sensitivity, and mean F1-score of 0.852, 0.856, and 0.848, respectively, and outperformed the baseline mask region-based convolutional neural network (Mask R-CNN) by 3.92%. Further analysis showed that incorporating global information also helps the model classify specific instances that require information from other regions. On the metastasis classification task, our model achieves a sensitivity of 0.657 and a specificity of 0.857, demonstrating its great potential for automated diagnosis using bone scintigraphy in clinical practice.

INDEX TERMS Bone scintigraphy, semi-supervised learning, lesion instance segmentation, deep learning.

I. INTRODUCTION

At present, 1.7 million patients are diagnosed with cancer each year [44], and cancer is commonly detected in multiple organs. After cancer has spread to the bones, it can rarely be cured [29]. Therefore, bone cancer detection plays a key role in making treatment decisions [28]. Bone scintigraphy is a nuclear medicine procedure that uses radioactivity to perform bone cancer imaging. Because the spread of cancer often manifests in bones, clinicians usually request bone scintigraphy results before prescribing any type of treatment. The bone scintigraphy results are then used as supporting information for primary decision-making during screening and for identifying the positions of any abnormal regions, called lesions [13], [38].

However, abnormalities found in bone scans include not only cancer but also other bone abnormalities that can be considered benign. A malignant lesion is characterized as a cluster of dangerous tumor cells that can lead to bone cancer metastases [7]. To judge whether a lesion is malignant, the nuclear medicine physician must consider criteria such as pixel intensity (which reflects the level of radioactive uptake),

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Friebe.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

lesion location, number of lesions, etc. In cases where lesion categorization is difficult due to ambiguous characteristics, the time a physician spends during diagnosis to interpret the results may increase by up to an hour per patient. Consequently, using machine learning to support this task could help improve efficiency, resulting in better patient treatment.

The difficulties involved in applying machine learning to medical imaging applications lie in the need for manual labeling. Labeling bone scintigraphy data requires nuclear medicine physicians, making the labeling task expensive and time-consuming. Thus, it is highly likely that only a small portion of the available data will be labeled. Furthermore, when labeling physicians are uncertain about the type of lesion, they may label more than one class per lesion (multilabel data), making the data labeling more complex. The current instance segmentation methods are designed for supervised learning and require large amounts of labeled data for training; they cannot use unlabeled data, and poor results may be obtained when the labeled dataset is small.

Deep learning has become the predominant model for tasks related to medical image interpretation. Convolutional neural networks (CNNs) are usually used in such models due to their ability to handle spatial inputs well [40]. Our work focuses on using both labeled and unlabeled data to improve model accuracy, a method often called semisupervised learning. Specifically, our model uses the feature pyramid network (FPN) architecture [35] as a basis and incorporates the autoencoder structure used in the ladder network [41] to make use of unlabeled data.

The lesion instance segmentation task is responsible for dividing pixels into parts based on lesion characteristics. There are two main approaches to segmentation tasks: semantic segmentation, which aims to group pixels in a semantically meaningful way using a pixelwise classification, and instance segmentation, which not only segments pixels into groups but also identifies the groups as instances. Generally, region-based approaches [16] for object detection are adopted as the first stage in instance segmentation, in which each region is categorized and segmented into a binary mask [22].

Normally, classifying the object types in the instance segmentation task first relies on an object detection process to identify the regions of interest (ROIs). Each object is classified independently, which might be appropriate in certain tasks. However, for bone scintigraphy, this method cannot be used because categorizing the type of lesion is reliant on other lesions in the images. For example, if most of the lesions are considered malignant, then the lesions that are not yet classified are also likely to be malignant. We use global features from the core network to support this line of reasoning. The model capitalizes on global features by using the overall composition to help determine lesion types [27].

Comprehensive experimental results show that our model achieves a higher level of accuracy than does the baseline model (Mask R-CNN) on lesion instance segmentation tasks. When used for screening initial diagnosis results our model reaches an accuracy of 74.1% in the bone cancer metastasis classification task, whereas the baseline model has an accuracy of 70.7%. We also studied the effects of including different amounts of labeled and unlabeled data. We find that incorporating 14,786 images of unlabeled data has the same effect as increasing the amount of labeled data by approximately 149 images (20.11% of the available labeled training data).

The remainder of this article is structured as follows. In Section II, we introduce related studies and discuss the differences between those works and our study. In Section III, we present the basic concepts of the model components used in MaligNet. The procedures for chest detection and the lesion instance segmentation model, along with the model architecture, techniques used, and implementation details, are described in Section IV. The dataset and evaluation metrics are presented in Sections V-A and V-B, respectively. The performance comparison results on each subtask are described in Section VI.

II. RELATED WORKS

There is an increasing trend of using deep neural networks for medical image analysis. ChexNet [40] uses DenseNet [25], a CNN variant for detecting pneumonia from chest X-rays. RIANet [46], is an encoder-decoder that efficiently reuses parameters to encode richer representative features for cardiac MRI segmentation. Rather than standard translational convolutions, three-dimensional roto-translation group convolutions have been applied to detect pulmonary nodules in CT scan images to reduce false-positive errors [50]. A combination of three CNNs is used to automatically localize anatomical ROIs of CT scan images [10].

For landmark detection to locate points of interest, a CNN [52] was used to localize geometric landmarks on the femur surface in 3D MRI. The SpatialConfiguration-Net [39] model was used to localize multiple landmarks in hand images using regression heatmaps.

Semantic segmentation has been widely applied in the medical image field to group pixels into semantically meaningful segments. For example, pixels that represent the same tissue or lesion should be grouped into the same segment. Micro-Net [42] and DCNet [33] used CNNs to perform semantic segmentation on microscopy images and multicontrast MRI, respectively. However, semantic segmentation approaches have difficulties separating different instances of the same class, which affects object counting and classification.

To solve the problem of separating instances of the same class, instance segmentation tasks were introduced that identify both objects and their regions and segment the pixels within such regions. However, these methods have some overhead during the object detection phase that occupies time and requires more memory during training.

Compared to instance segmentation, which is rare in medical image analysis [51], the related task of image segmentation is more common. Spine-GAN [21] was developed to perform semantic segmentation on the spinal region

from MRIs. Fully convolutional networks (FCNs) were used for male pelvic organ segmentation from CT scans by [49]. Reference [1] used CNNs for segmentation that had a priori knowledge of the anatomical shapes of knee bones and cartilage. [18] performed gland segmentation using a modified CNN that reintroduced the original image at multiple points within the network to help reduce information loss caused by max pooling. Reference [24] applied a 3D CNN to 3D CT multiorgan medical images. Reference [30] proposed a dual-pathway CNN for brain lesion segmentation. Reference [5] performed four-dimensional segmentation from cardiac 4D flow MRIs.

All the aforementioned works were trained in a supervised manner with labeled data. ChexNet used more than 100,000 chest X-ray images with labels obtained from medical records, which is a high-level classification task for which labels can be acquired relatively easily. However, for complex tasks such as segmentation, the number of training data samples can be as low as a few hundred due to the difficulty of data acquisition and labeling. One way to reduce the effort of data annotation is to use coarser annotation schemes. For example, [32] proposed a constrained-CNN loss for image segmentation on the left ventricle (MRI), vertebral body (MR-T2), and prostate (MR-T2) using segmentation labels that did not cover the entire region.

Another popular approach is to use unlabeled training data to improve the model, a method often called semisupervised learning. Reference [6] provided a comprehensive overview of semisupervised methods applied to medical image analysis. Self-training uses a model previously trained on labeled data to estimate the labels for unlabeled data. Reference [2] proposed a self-training approach for breast lesion segmentation from MRIs. This simple approach is surprisingly effective when the initial model is sufficiently robust.

Another strategy, called graph-based methods in [6], employs unlabeled data to learn the data distribution. Our work falls under this category but is based on a deep learning framework. By modifying the loss function to include an unsupervised loss, the training process is simplified because we treat labeled and unlabeled data almost identically.

For the task of bone scintigraphy, which is the application domain of our work, [3], [9] used CNNs to classify hotspot regions for prostate cancer metastases. Due to the difficulty of acquiring labeled images, this work used only approximately 2,000 images. Reference [4] used the VGG-19 architecture to classify benign and malignant bone lesions. Reference [13] used a sparse autoencoder to automatically learn good features for metastasis classification and then used multiple instance learning (MIL) with a patch-level classifier to perform segmentation. Later [14] proposed EM-MILBoost, which additionally applied expectation-maximization (EM) to MIL to achieve further performance improvements.

Reference [31] performed unsupervised lesion detection on bone scintigraphy images using unsupervised learning on normal images in an autoencoder-like manner instead of using semisupervised learning. Our method uses both supervised and unsupervised data to train our model jointly; this should perform better than a completely unsupervised model in terms of detection capability. Moreover, the model proposed by [31] can detect only lesions; it cannot perform classification or segmentation due to the limitations of unsupervised data.

Our work also differs from previous bone scintigraphy-related works in that our goal is instance segmentation, which means that we both classify the lesion type and segment each lesion. While both semantic segmentation and instance segmentation can identify the locations of lesions, when two lesions overlap or are adjacent to each other, instance segmentation can determine that the two lesions are two separate entities. Moreover, previous works classified lesions only as metastatic or nonmetastatic, whereas our work can classify each lesion into finer classes, i.e., malignant, degenerative change, post-trauma, and inflection/inflammation. This level of classification is closer to the current clinical practice for bone scintigraphy. In some cases, instance segmentation can help the model better differentiate malignant from nonmalignant lesions. For example, if the model understands that lesions on different ribs that form in an orderly manner into a straight line should be classified as post-trauma, then it will be easier for instance segmentation models than for semantic segmentation models to consider this correlation.

III. BACKGROUND

Our overall system consists of two parts: a chest localization model, which localizes the chest area, and an instance segmentation model, which segments and classifies each lesion, as shown in Figure 1. In this paper, we focus on MaligNet, a model for lesion instance segmentation on the chest area in bone scintigraphy that has various internal components. We provide some background for each component and related works.

The **feature pyramid network** (FPN) [35] is chosen as the core component in MaligNet for instance segmentation, as shown in Figure 2. We chose FPN because it was designed to detect objects at different scales, which is the case for lesions in a chest image. An FPN consists of two main parts: bottom-up and top-down pathways. The bottom-up pathway is the feedforward neural network, which can be any object classifier. The top-down pathway, which is connected to the bottom-up pathway through lateral connections, is designed to build semantic feature maps at multiple scales by double upscaling to enhance the feature maps from the bottom-up pathway. Combining high-resolution but semantically weak features with low-resolution but semantically strong features via a lateral connection and top-down pathway imparts rich semantics to all levels of the FPN.

A **region proposal network** (RPN) is a type of fully convolutional network that is used in Faster R-CNN [43]. This model is part of a region-based family that includes R-CNN [16], Fast R-CNN [15] and Mask R-CNN [22]. Region-based object detectors first identify potential regions for objects and then classify each region into object classes.

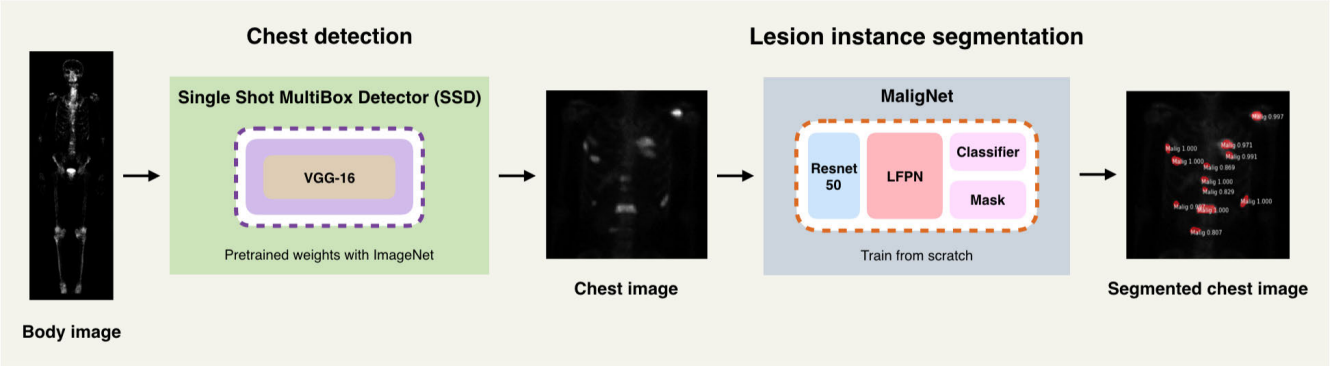

FIGURE 1. An overview of our model workflow. A whole-body bone scintigram (left image) is passed into the single shot multibox detector (SSD) to detect the chest area (middle image) and then sent to the MaligNet model for lesion instance segmentation (right image).

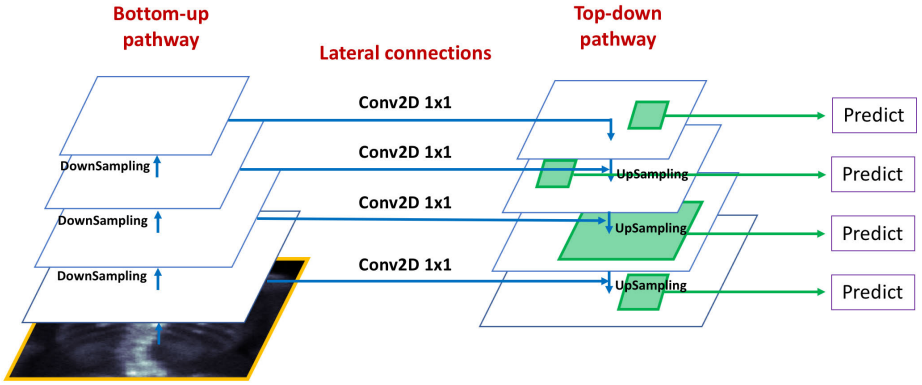

FIGURE 2. Illustration of the feature pyramid network (FPN). The FPN consists of a bottom-up pathway and a top-down pathway. The bottom-up pathway is the feedforward neural network of the core. The top-down pathway is ConvNet, which upsamples the spatially coarser high-level features combined with low-level features through lateral connections.

The RPN is a module that generates the initial region proposals. It is designed to detect objects based on the convolutional feature maps from the core network. The module predicts region proposals at various scales and aspect ratios (see Table 9) using multiple anchor boxes. Specifically, for each location and scaling factor on a regular grid, the RPN outputs object region boundaries and their associated objectless scores, which specify the likelihood that each proposed region contains an object of interest, as shown in Figure 3. The cost function for classifying each region proposal (R_c) is categorical cross-entropy loss, which was defined as follows:

$$R_c = -\frac{1}{N}\sum_{n=1}^{N}\sum_{a=1}^{A}\log P(\tilde{\mathbf{y}} = \mathbf{y}_{n,a} \mid \mathbf{x}_{n,a}), \tag{1}$$

where $\mathbf{x}$ is the feature map from the FPN, $\mathbf{y}$ is the ground truth of the anchor box, $\tilde{\mathbf{y}}$ is the anchor box prediction, N is the minibatch size, and A is the number of anchors per image.

For the region bounding boxes, the smoothed-L1 loss was used as the cost function for the bounding box prediction, as shown below:

$$d_{n,a} = \left\|\mathbf{b}^*_{n,a} - \mathbf{b}_{n,a}\right\|_1 \tag{2}$$

$$s(x) = \begin{cases} x - 0.5 & if\ |x| > 1 \\ 0.5x^2 & otherwise, \end{cases} \tag{3}$$

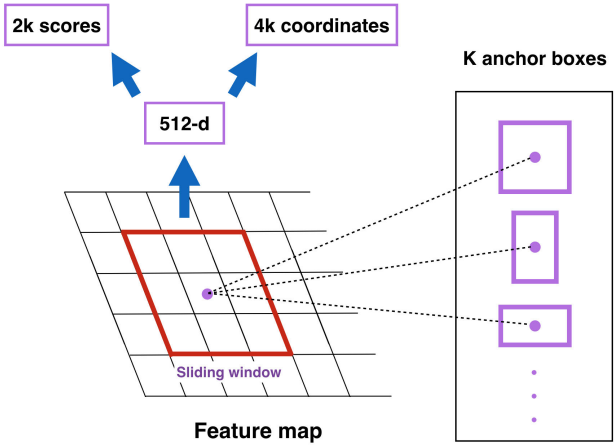

FIGURE 3. Illustration of the region proposal network (RPN). The input is a feature map. The RPN produces $2k$ anchor scores and $4k$ bounding box coordinates per pixel in the feature map, where k is the number of anchor boxes.

$$R_b = -\frac{1}{N}\sum_{n=1}^{N}\sum_{a=1}^{A} s(d_{n,a}), \tag{4}$$

where $\|\ \|_1$ denotes the L1 norm; $\mathbf{b}^*_{n,a}$ and $\mathbf{b}_{n,a}$ are vectors containing the coordinates of the predicted bounding box and labeled bounding box, respectively; $s(d_{n,a})$ is the smoothed

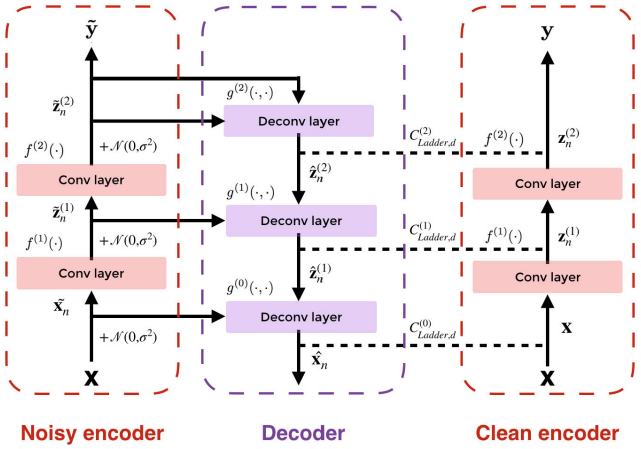

FIGURE 4. The structure of the ladder network, which is a convolutional neural network, consists of two neural network parts: an encoder and a decoder. The encoder includes a clean encoder ($x \rightarrow z^{(m)} \rightarrow y$) and a noisy encoder ($\tilde{x} \rightarrow \tilde{z}^{(m)} \rightarrow \tilde{y}$), which share the same mapping function f. The decoder (g) ($\tilde{z}^{(m)} \rightarrow \hat{z}^{(m)} \rightarrow \hat{x}$) recovers the lost information by making comparisons with the clean encoder in the lateral connections. When labels are present, i.e., supervised learning, the loss signal will be top-down information in the noisy encoder, while for unsupervised learning, the loss signal will be the lateral information between the decoder and the clean encoder.

L1 loss; and R_b is the sum of the region proposal bounding box loss at all anchors.

A **ladder network** is a semisupervised learning method proposed by [41] that can utilize labeled and unlabeled data simultaneously. A ladder network is similar in concept to a denoising autoencoder (DAE). A DAE is an autoencoder that receives a corrupted data point as input and is trained to reconstruct the uncorrupted data point [17]. A ladder network takes this a step further by introducing noise at every layer, not just the input. Figure 4 illustrates a simple ladder network. The network consists of three parts: the noisy encoder (the leftmost stack), the denoising decoder (the middle stack), and the clean encoder (the rightmost stack), which is the original network. Let us first consider the clean encoder. Let $\mathbf{z}^{(m)}$ be the output of the m-th layer of the clean encoder. The function $f^{(m+1)}(\cdot)$ represents the $(m+1)$-th layer, which in our case is a convolutional layer. The relationship between each layer is given by the following:

$$\mathbf{z}^{(m+1)} = f^{(m+1)}(\mathbf{z}^{(m)}). \tag{5}$$

Note that $\mathbf{z}^{(0)}$ refers to the network input, $\mathbf{x}$. The encoder yields a final prediction, $\mathbf{y}$.

The noisy encoder uses the same weights and layers as the original network. However, we corrupt the inputs preceding each layer by adding Gaussian noise:

$$\tilde{\mathbf{z}}^{(m+1)} = f^{(m+1)}(\tilde{\mathbf{z}}^{(m)} + \mathbf{h}^{(m)}) \tag{6}$$

where $\mathbf{h}^{(m)}$ is random noise sampled from a Gaussian distribution with zero mean and variance σ^2. Here, we set $\sigma^2 = 1$ (unit variance) for all layers. The output of the m-th layer of this corrupted network is $\tilde{\mathbf{z}}^{(m)}$.

Finally, the denoising decoder aims to recover the original output at each layer by using only the features from the noisy encoder and the higher-level layer. Let $g^{(m)}$ be the inverse mapping decoder function for the m-th layer, which in this case is a transposed convolutional layer [53] that outputs $\hat{\mathbf{z}}^{(m)}$ and takes $\tilde{\mathbf{z}}^{(m)}$ and $\hat{\mathbf{z}}^{(m+1)}$ as inputs:

$$\hat{\mathbf{z}}^{(m)} = g^{(m)}(\tilde{\mathbf{z}}^{(m)}, \hat{\mathbf{z}}^{(m+1)}). \tag{7}$$

The supervised cost, $C_{Ladder,c}$, is the average negative log probability of the noisy output $\tilde{y}$ matching the ground-truth target $\mathbf{y}$ given input $\mathbf{x}$:

$$C_{Ladder,c} = -\frac{1}{N}\sum_{n=1}^{N} logP(\tilde{\mathbf{y}} = \mathbf{y}_n|\mathbf{x}_n), \tag{8}$$

where N is the minibatch size, and n is the index of the training data.

The goal of the denoising decoder is to output a $\hat{\mathbf{z}}^{(m)}$ that matches $\mathbf{z}^{(m)}$. This is accomplished by adding the unsupervised cost function, $C_{Ladder,d}$, which attempts to minimize the mean square error as shown below:

$$\begin{aligned} C_{Ladder,d} &= \sum_{m=0}^{M} C_{Ladder,d}^{(m)} \\ &= \sum_{m=0}^{M} \frac{1}{Nw_l} \sum_{n=1}^{N} \left\| \mathbf{z}_n^{(m)} - \hat{\mathbf{z}}_n^{(m)} \right\|^2, \end{aligned} \tag{9}$$

where w_l is the layer width and M is the number of layers in the ladder network.

The ladder network uses this unsupervised loss to learn the important features even without any labeled data. The two losses can be combined and used jointly for training when labels are present.

IV. PROPOSED METHOD

Our proposed model is a neural network to perform lesion instance segmentation, named MaligNet. Although lesions can occur anywhere throughout the body, they are often found in the chest area. This area is often the hardest to diagnose in the chest area, which consists of small bones, due to the complexity and overlap of the ribs. Therefore, we focused only on finding lesions in the chest area. Figure 1 shows an overview of our system. We start by locating the chest area using the single shot multibox detector (SSD) [37] in both anterior and posterior whole-body views in the bone scintigram. The resulting identified chest area is subsequently used in the lesion instance segmentation process.

A. CHEST DETECTION

The chest detector is the first part of our pipeline and is used to detect the chest area. Because it is relatively straightforward to detect the chest area, we use a standard SSD [37] to detect the chest areas from both the anterior and posterior views. We chose the SSD because it is a one-stage detection model that can be trained and makes inferences at high speed. Moreover, it maintains good accuracy compared to other object

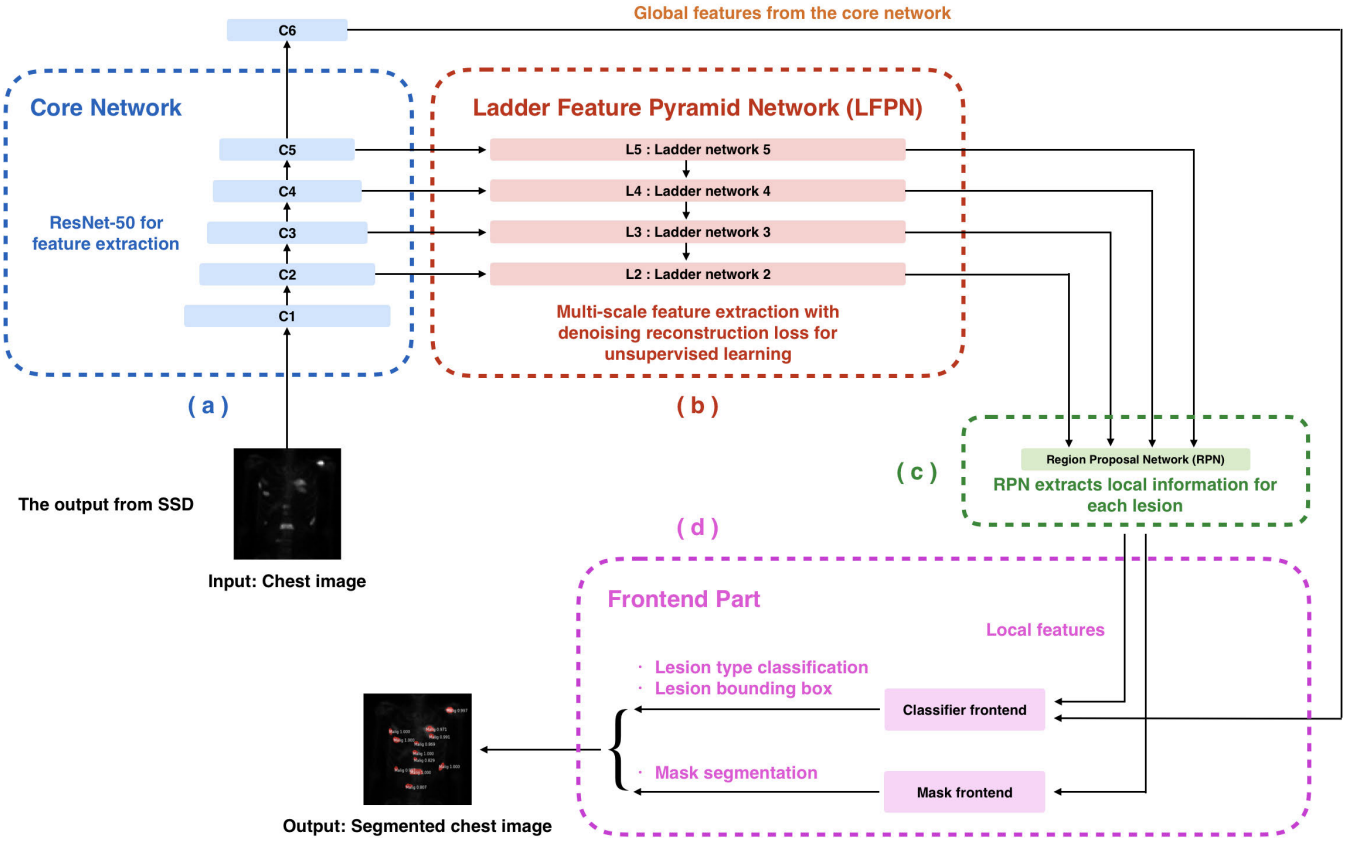

FIGURE 5. An overview of MaligNet. Figure 5a (in blue) shows the core network of MaligNet, a ResNet-50 architecture that extracts features at different scales for input to the ladder feature pyramid network (LFPN). Figure 5b (in red) shows the LFPN, which consists of a feature pyramid network combined with a ladder network to facilitate semisupervised learning. Figure 5c (ingreen) depicts the region proposal network (RPN), which selects regions of interest for object classification and regression. Figure 5d (in pink) shows the front-end part, which consists of two parts: a classifier front-end (Figure 5d) and a mask front-end (Figure 8). The classifier front-end performs lesion classification and refines the bounding box. The mask front-end outputs the segmentation masks. Unlike Mask R-CNN, the classifier front-end also receives global features from the core network.

detection models and was easy to adapt and apply to our task. We used VGG-16 [45] as the core network in the SSD. The VGG-16 was pretrained on ImageNet [11]; then, all the layers were fine-tuned based on the source model parameters with our data. The hyperparameters used for retraining the SSD are shown in Table 8.

B. LESION INSTANCE SEGMENTATION

MaligNet is a CNN model based on an FPN with some modifications. Specifically, we added the ladder feature pyramid network to the top-down pathway to allow semisupervised training. We also added an additional layer that extracts global features from the core network to the classifier head. As shown in Figure 5, MaligNet consists of four parts. Similar to the FPN, the first part is an image classification core model used to extract features. We tested several standard architectures (see Table 9) to choose the core model and ultimately settled on ResNet-50 [23]. ResNets have the advantageous property of using a stride of two for every scale reduction, which makes incorporating ResNet-50 into the FPN straightforward when we have to upscale the feature maps in the top-down pathway. Moreover, ResNet-50 is a relatively small network and is based on modern standards; thus, it is appropriate for our limited labeled data.

The second part is the LFPN, which corresponds to the top-down pathway of the FPN but includes additional denoising decoder components inspired by the ladder network [41]. This allows MaligNet to utilize the training data from both labeled and unlabeled data simultaneously. The features from the top-down pathway are used by the third part, which detects and localizes lesions using an RPN. The front-end part is the final part of our model and is designed to perform lesion instance segmentation and provides two output results. The classifier front-end adjusts the bounding boxes and categorizes each lesion into different classes. The mask front-end is used for mask prediction. However, unlike the traditional FPN, which focuses on local information in each region, the classifier front-end also exploits the global information taken from the topmost level of the core network, which condenses all of the image information.

1) LADDER FEATURE PYRAMID NETWORK (LFPN)

To allow semisupervised learning, we incorporated ladder-network-like structures into each level of the FPN. The new

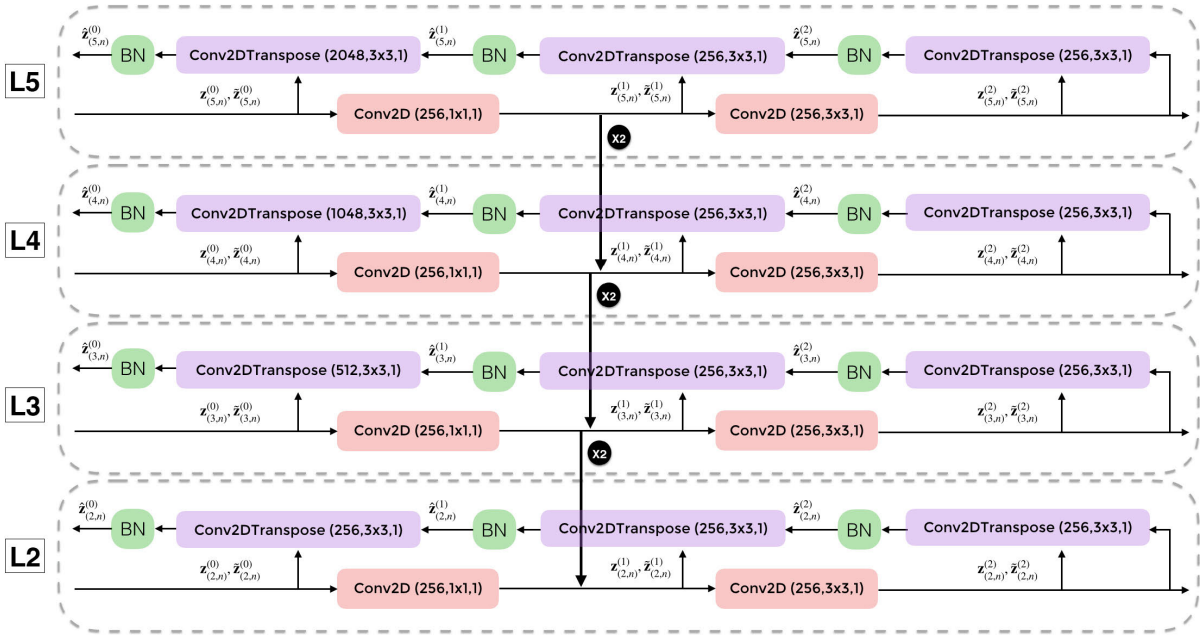

FIGURE 6. Illustration of the ladder feature pyramid network (LFPN), the feature pyramid network combined with the ladder network to enhance the features by unsupervised learning. In the figure, we have four lines of convolutional neural networks (red blocks) in the feature pyramid network, and each line represents an encoder. We also add a decoder to invert the mapping at each encoder layer(purple blocks); thus, we have a total of four ladder networks, denoted as L2, L3, L4, and L5. The upsampling layer (the x2 symbol) in the LFPN is a bicubic algorithm for scaling up the feature map.

structure is referred to as the LFPN. As shown in Figure 6, each lateral connection is similar to an encoder part in the ladder network. The lateral connections that do not add noise into the features are considered to be the clean encoder $\mathbf{z}_{(l,n)}^{(m)}$, which is defined as follows::

$$\mathbf{z}_{(l,n)}^{(m+1)} = \begin{cases} g(f_l^{(m+1)}(\mathbf{z}_{(l,n)}^{(m)}), \mathbf{z}_{(l+1,n)}^{(m+1)}) & if\ m = 0 \\ f_l^{(m+1)}(\mathbf{z}_{(l,n)}^{(m)}) & if\ m = 1. \end{cases} \tag{10}$$

The main difference between our model and a regular ladder network is the additional connection from the upsampling layer (denoted as x2 in the figure).

For the noisy encoder, the features in the LFPN are as follows:

$$\tilde{\mathbf{z}}_{(l,n)}^{(m+1)} = \begin{cases} g(f_l^{(m+1)}(\tilde{\mathbf{z}}_{(l,n)}^{(m)}), \tilde{\mathbf{z}}_{(l+1,n)}^{(m+1)}) + \mathbf{h}_{(l,n)}^{(m+1)} & if\ m = 0 \\ f_l^{(m+1)}(\tilde{\mathbf{z}}_{(l,n)}^{(m)}) + \mathbf{h}_{(l,n)}^{(m+1)} & if\ m = 1, \end{cases} \tag{11}$$

where f is the convolution function, and g is the feature combination function. Note that $\mathbf{z}_{(l,n)}^{(0)}$ refers to $\mathbf{x}_{(l,n)}^{(0)}$, and $\tilde{\mathbf{z}}_{(l,n)}^{(0)}$ refers to $\mathbf{x}_{(l,n)}^{(0)} + \mathbf{h}_{(l,n)}^{(m+1)}$.

Here, the noise for all layers is sampled from a Gaussian distribution with zero mean at a fixed variance level, which is a tunable hyperparameter (see Table 9). Even though bone scintigraphy raw images have a Poisson noise distribution [47], the purpose of adding noise in the LFPN is to augment the feature space, not the raw image, and the weight distribution in the network is Gaussian. Thus, chose to inject a Gaussian distribution instead of a Poisson distribution. We also tested injecting Poisson noise into the LFPN, but the resulting model performed worse than did the baseline FPN model.

To denoise the noisy features, a transposed convolution layer is used in the denoising decoder because it uses the inverse function of the convolutional layer [41], [53] used in the CNN-Ladder network. The reconstruction $\hat{\mathbf{z}}_{(l,n)}^{(m)}$ is the output of its upper layer $\hat{\mathbf{z}}_{(l,n)}^{(m+1)}$ and the noisy lateral layer $\tilde{\mathbf{z}}_{(l,n)}^{(m)}$ via the Conv2DTranspose layer and batch normalization. We add a denoising decoder (the purple blocks in the figure) to each lateral connection of the FPN such that we can incorporate the unsupervised loss. For each lateral connection (L2 to L5 in Figure 6), there are three targets for performing denoising; these correspond to the outputs of different layers on that level. Thus, the unsupervised loss from Equation 9 becomes the following:

$$\begin{aligned} C_{LFPN,d} &= \sum_{l=2}^{L=5} \sum_{m=0}^{m=2} C_{LFPN,d}^{(m)} \\ &= \frac{1}{NLM} \sum_{l=2}^{L=5} \sum_{m=0}^{M=2} \sum_{n=1}^{N} \left\| \mathbf{z}_{l,n}^{(m)} - \hat{\mathbf{z}}_{l,n}^{(m)} \right\|^2, \end{aligned} \tag{12}$$

where N is the minibatch size, L is the number of levels in the LFPN, and M is the number of layers in each lateral connection.

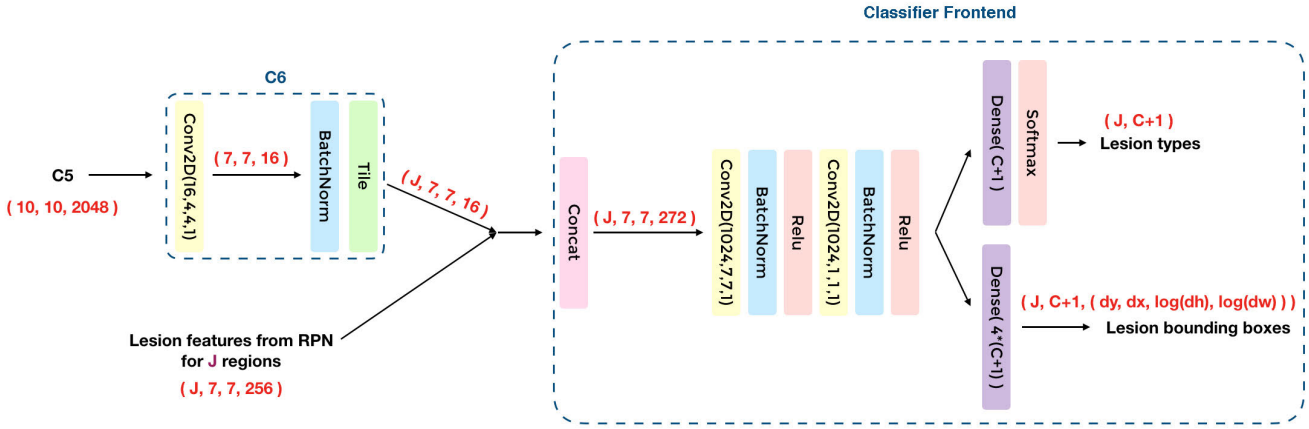

FIGURE 7. Illustration of the classifier front-end. Global features are applied with lesion features by concatenation. The classifier front-end separates into two branches that perform lesion-type classification and bounding box prediction. Each lesion prediction in both sub-front ends has $C + 1$ outputs.

During the training process, the RPN accepts the noisy output rather than the clean output. Because noise is added to the features, the model will capture the important information from these features as augmented information, which makes the model more generalizable and helps avoid overfitting.

Finally, regarding the choice of location for adding the ladder network structure to the model. We add the ladder network to the FPN rather than to the core network because the noise introduced in the ladder network can accumulate, as there are more noising layers. Adding the ladder network to the core network, which has many layers, greatly reduces the model performance. However, adding the noise to the FPN only adds three noise terms per lateral connection.

2) APPLYING GLOBAL FEATURES FOR LESION CLASSIFICATION

Typically, object detection for natural images is performed by detecting each object independently—regardless of the other objects in the same image. However, physicians usually use both other lesions and additional cues beyond the lesion itself into account when determining the lesion types. For example, if a lesion is isolated, without other nearby lesions, it is more difficult to assert that the isolated lesion is malignant. However, when multiple lesions occur in the same region, they are usually malignant. Thus, we incorporate global features (the features from layer C6 in Figure 5) that summarize the overall image information to enhance the prediction of each individual lesion.

The architecture of the classifier front-end is shown in Figure 7. The output from the core network (C5) is embedded into a lower-dimensional space using a convolutional layer and then tiled and replicated (using the tile layer in TensorFlow) J times, allowing it to be concatenated with the features from each region proposal. The concatenated features are then passed to the classifier front-end to classify the lesion type and adjust the bounding boxes.

3) CLASSIFIER FRONT-END

The classifier front-end consists of two sets of convolutional and batch normalization layers with rectified linear unit (ReLU) activation. For classification, we used a dense layer with softmax normalization to obtain the probabilities for each lesion type. The dense layer has $C+1$ neurons, where C is the number of lesion types: the additional class represents nonlesions. The bounding box prediction is designed to refine the proposal regions, which is treated as a regression task using the $4 * (C + 1)$ outputs from the dense layer. This layer predicts the position x and y coordinates, log(height), and log(width) for each class. Bounding box regression can be difficult for tasks involving objects with large size variations. Therefore, the height and width of the bounding box are converted to log scale, which is usually easier to regress. This preprocessing method is considered standard practice in object detection tasks [16].

We use categorical cross-entropy loss as the classification cost:

$$C_c = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J}\log P(\tilde{\mathbf{y}} = \mathbf{y}_{n,j} \mid \mathbf{x}_{n,j}), \tag{13}$$

where J is the maximum number of region proposals in the image.

The cost function for the bounding box prediction is

$$C_b = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} s(d_{n,j}), \tag{14}$$

where $d_{n,j}$ is the L1 norm (which is the same as Equation 2); $s(x)$ is the smoothed L1 loss (as shown in Equation 3); and C_b is the sum of the bounding box losses for all lesions.

4) MASK FRONT-END

Lesion segmentation is performed by the mask front-end, which—given a proposed lesion region from the RPN—outputs C+1 foreground-background segmentation

FIGURE 8. Illustration of the mask front-end for mask prediction in MaligNet.

masks for each class. The mask that corresponds to the lesion classification is used as the segmentation output. Our architecture, shown in Figure 8, consists of sets of convolutional and batch normalization layers with ReLU activation. After passing sigmoid activation, the output has a mask size of 14x14 pixels, which is the same as the original Mask R-CNN [22]. We adopt binary cross-entropy loss as the cost function:

$$C_m = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J}\mathbf{y}_{n,j}\log(s_{n,j})+(1-\mathbf{y}_{n,j})\log(1-s_{n,j}), \quad (15)$$

where $s_{n,j}$ is the foreground class score after the sigmoid function and $\mathbf{y}_{n,j}$ is the ground-truth mask.

5) UNIFIED LOSS

Our model includes many components, such as the LFPN (Figure 6), the RPN (Figure 3), the classifier, the bounding box front-end (Figure 7), and the mask front-end (Figure 8). Moreover, each component has a different objective function for supervised and unsupervised losses. Consequently, the loss calculation must be weighted to avoid loss values of each term that are too different. The weight multiplier λ_k is the hyperparameter of each loss.

The combination of supervised loss C_s is the summation of the weight multipliers with their losses, calculated by the following equation:

$$C_s = \lambda_{rc}R_c + \lambda_{rb}R_b + \lambda_{cc}C_c + \lambda_{cb}C_b + \lambda_{cm}C_m, \quad (16)$$

where R_c and R_b are the costs of the class and bounding box in the RPN, respectively, and C_c, C_b, and C_m are the costs of the lesion class, lesion bounding box, and lesion mask, respectively, in the classifier front-end and mask front-end. Each λ_k is a weight multiplier for each loss k.

For unsupervised data, the cost function is as shown below:

$$C_{us} = \lambda_{us}C_{LFPN,d}, \quad (17)$$

where λ_{us} is the weight of the unsupervised loss term and $C_{LFPN,d}$ is the cost function in the LFPN, which is shown in Equation 12.

We add an L2 regularization term to avoid overfitting and to improve generalizability. Therefore, the total cost is the sum of all cost functions with the L2 regularization term:

$$C_{total} = C_{us} + C_s + \lambda_{L2}\sum_{k=1}^{K}\boldsymbol{\omega}_k^2, \quad (18)$$

where λ_{L2} is the weight of regularization, and K is the number of trainable layers.

Each minibatch can include a mixture of supervised and unsupervised data depending on the random minibatch sampling. The unsupervised data (unlabeled) will be calculated as the only unsupervised cost (C_{LFPN}); the supervised cost of unlabeled data in the minibatch is zero.

Thus, our model is able to learn both supervised and unsupervised learning jointly, which is a form of semisupervised learning.

C. IMPLEMENTATION DETAILS

We chose the original ResNet-50 as our core network due to the limited amount of labeled data. Moreover, ResNet-50 is easy to upsample in an FPN. Because the output images from the chest detection stage have different sizes, we scale and resize both the image and mask to match the GPU's memory. During training, each minibatch contains both supervised and unsupervised data. We also tested a version that alternated between minibatches of supervised and unsupervised data but obtained the same results. We used two sets of NVIDIA GeForce 1080 Ti GPUs for each batch size, equal to eight per GPU. The hyperparameter details are listed in Appendices B and C.

V. EXPERIMENTAL SETUP

In this section, we provide information about the dataset and model evaluation. The dataset and its properties, including descriptions of the lesion types, are presented in Section V-A. There are two main tasks in our pipeline: lesion instance segmentation and bone cancer metastasis classification. To evaluate the performances of both tasks, we explain the details of the metrics in Section V-B.

A. DATASET AND PREPROCESSING

Our data included a total of 9,824 patients. The details of the patients' genders and ages are shown in Appendix A. The injection dose of 20 mCi/70 kg varied according to the patient's weight, and the uptake time was approximately 5 hours. The images are in DICOM format with a 16-bit depth. For chest detection, we used 680 images of the whole body for training, 200 for validation, and 240 for testing. For lesion instance segmentation, the dataset contained 19,648 chest images of which 1,088 were labeled images and the remaining 18,560 were unlabeled images.

The dataset was separated into training, validation, and testing data as listed in Table 1. The physician focused on four main lesion types: malignant (or cancerous), inflection/inflammation, degenerative change (bone deterioration), and posttrauma (broken regions caused by accidents). The numbers of each type of lesion are shown in Table 2.

The data collection was approved by the Institutional Review Board (IRB) of the Faculty of Medicine, Chulalongkorn University. The labeling was performed by five nuclear medicine physicians with 31, 28, 21, 11, and 8 years of diagnostic experience. Note that all the labeled data were labeled by nuclear medicine physicians without the use of medical records. Thus, the labeled lesion type might not

TABLE 1. The total labeled and unlabeled training data for lesion instance segmentation are separated into training data, validation data, and testing data.

Data type	Amount of data	Training data	Validation data	Testing data
Labeled data	1,088	741	231	116
Unlabeled data	18,560	14,786	3,774	0
Total data	19,648	15,527	4,005	116

TABLE 2. The total number of lesions per type.

Lesion types	Number of lesions
Malignant	3,500
Inflection/Inflammation	290
Degenerative change	805
Post-traumatic	415
Total lesions	5,010

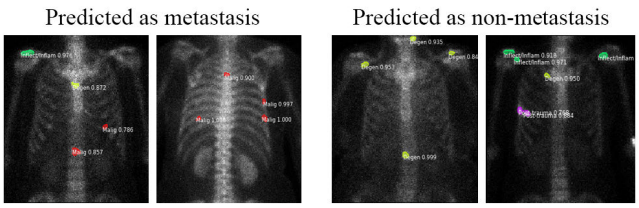

FIGURE 9. Example predictions for the bone cancer metastasis classification task. When at least one malignant lesion is predicted, the image is classified as metastasis.

reflect the true lesion type. Thus, supervised learning was applied using a test that was not a gold standard and may not reflect the true metastasis value of the hotspots.

We augmented the data with an affine transformation. Normally, bone scintigraphy requires adjusting the light and contrast such that the physician can observe the hotspots before labeling. Therefore, we also augmented the data by increasing and decreasing the light, contrast, and brightness for consistency with the physician's process.

B. EVALUATION METRICS

Our experiments were divided into two parts: lesion instance segmentation and bone cancer metastasis classification. Instead of using the mean average precision (mAP), which is a relative score metric used to evaluate object detection in natural images such as those in the Pascal VOC [12], COCO [36] and Open Images [34] datasets, we use mean precision, mean sensitivity, and mean F1-score to measure the lesion instance segmentation performance of our model. In the context of instance segmentation, we must not only correctly identify the object but also correctly locate its position. Thus, to calculate the mean precision and the mean sensitivity, we used the Jaccard index (Intersection over Union) to measure the overlapping region between the ground truth and the predicted area. The Jaccard index is defined as follows:

$$J(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}, \tag{19}$$

where A_1 is the area of the ground truth, and A_2 is the area of the prediction.

A prediction is considered correct if the Jaccard index exceeds a predetermined threshold. We chose a threshold of 0.5 because that threshold is sufficient for locating the lesion. For multiclass detection tasks such as ours, we can calculate the mean precision and mean sensitivity by taking the weighted averages of the precision and sensitivity values, respectively.

MaligNet is designed for lesion instance segmentation. Therefore, we cannot directly evaluate the performance of the model on the metastasis classification task. Instead, we convert the instance segmentation predictions to a binary prediction. If the model predicts malignancy for at least one lesion in the chest area, the image is classified as metastasis. However, we consider only cases in which at least one malignant prediction matches the ground truth (a true positive sample). In other words, a metastasis prediction caused by a false alarm in the instance segmentation task is not counted as a correct classification. Examples of interpretations in bone cancer metastasis classification are shown in Figure 9.

For cases where the model finds no malignant lesions in the image or finds another lesion type, such as degenerative change, inflection/inflammation, or post-traumatic, we assume that the model predicts a nonmetastasis status or a negative sample. We also evaluated our model of bone cancer metastasis classification in terms of various metrics, namely, accuracy, precision, sensitivity, specificity, and F1-score. The details of each type of measurement are described separately in the following sections.

VI. EXPERIMENTAL RESULTS

In this section, we report the experimental results by comparing the performances of each model. Our experiments are divided into three subtasks: chest detection, lesion instance segmentation, and bone cancer metastasis classification. We also conducted ablation studies to evaluate the impact of our semisupervised training results compared with those of other semisupervised methods.

A. RESULTS OF CHEST DETECTION

Anterior (front-side) and posterior (back-side) images are available for each patient. Because detecting the chest area in the whole image from bone scintigraphy is a simple task, the model provides accurate results: its min, mean, and max Jaccard indexes are 0.804, 0.933, and 0.987, respectively.

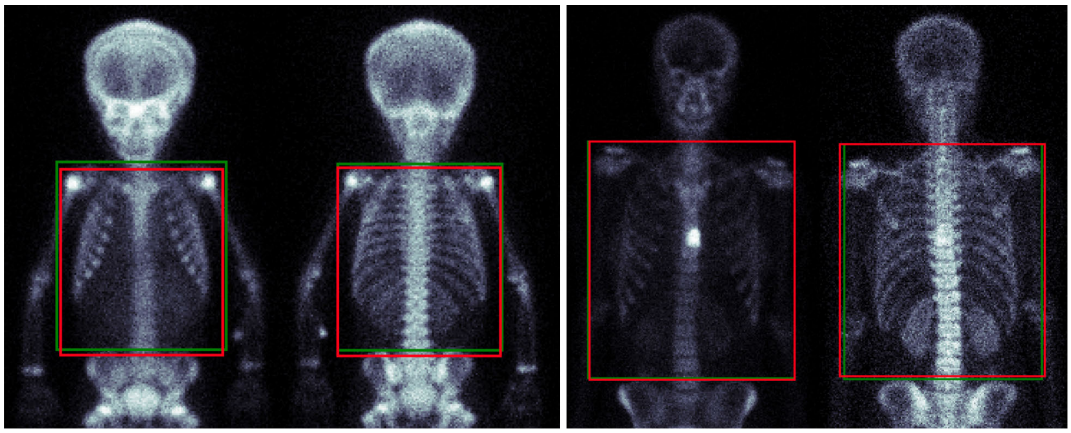

FIGURE 10. Some examples from the chest detection model. The first bone scintigram (the leftmost side) is an anterior view (front view), and the second bone scintigram is a posterior view (back view) of a pediatric patient. The third skeleton is an anterior view, and the last skeleton (the rightmost side) is a posterior view of an adult patient. The ground-truth boxes are indicated in green, while the outputs of the SSD model are indicated in red. The Jaccard indices from left to right are 0.895, 0.943, 0.987, and 0.914.

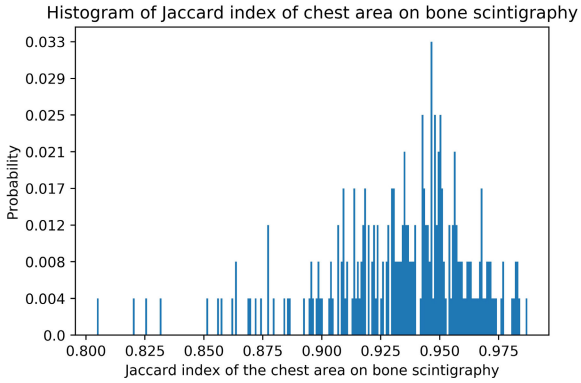

FIGURE 11. A histogram of the Jaccard index in chest detection for bone scintigraphy. The horizontal axis represents a comparison of the Jaccard index of the chest area between the ground truth and the model bounding box predictions. The vertical axis represents the frequency of each Jaccard index value in each bin of the histogram.

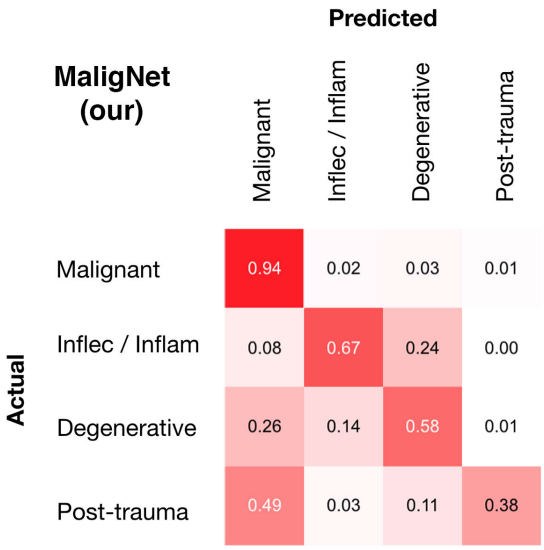

FIGURE 12. The normalized confusion matrix of the lesion classification task using MaligNet without self-training. The rows represent the true labels (ground truth), and the columns represent the predicted labels.

From Figure 11, the histogram shows that SSD provides excellent chest detection results based on the Jaccard index for at least every tested data point. Examples of chest detection results are shown in Figure 10.

B. RESULTS OF THE LESION INSTANCE SEGMENTATION TASK

In this section, we report the lesion instance segmentation results. The chest images from bone scintigraphy (the output results from chest detection), were used as the input data in this task. Data cleaning and augmentation were applied before performing the experiments. We evaluated our model on the lesion instance segmentation task for the four lesion types and compared the results with the baseline model (Mask R-CNN). Example results are shown in Figure 14 (hand-picked examples) and Figure 15 (random examples). We also studied the impact of each model technique on its overall performance. We conducted the experiments with the techniques applied separately and combined, as shown in Table 3. Moreover, we applied self-training with both the baseline and our model to compare the effect of each technique.

Using global features allows the model to use high-level features and semantically strong features to make prediction decisions, which increases the lesion classification accuracy. Moreover, applying the ladder network in the FPN makes the model capable of learning the representation of the images via unsupervised learning, which improves the model in every comparable configuration. Utilizing the LFPN for semisupervised training over the standard Mask R-CNN, MaligNet capitalizes on the large amounts of unlabeled data (14,786 images) and significantly increases model performance, reaching an F1-score of 0.835. Furthermore, combining global features further improves the F1-score of MaligNet

TABLE 3. Comparison between each model and technique for the lesion instance segmentation task. The global features in this table are the output features from layer C6 in Figure 5.

Method	Mean precision	Mean sensitivity	Mean F1-score
ResNet-50 + FPN (Mask R-CNN)	0.827	0.811	0.816
ResNet-50 + FPN w/ global features	0.829	0.826	0.824
ResNet-50 + LFPN w/o global features	0.838	0.839	0.835
ResNet-50 + LFPN w/ global features (MaligNet)	**0.852**	**0.856**	**0.848**
ResNet-50 + FPN + self-training	0.849	0.843	0.840
ResNet-50 + LFPN w/ global features + self-training	**0.867**	**0.844**	**0.851**

TABLE 4. Comparison between MaligNet and baseline for lesion classification on the lesion segmentation task.

Lesion types	Model	Accuracy	Precision	Sensitivity	Specificity	F1-score
Malignant	MaligNet	**0.886**	**0.912**	**0.941**	**0.710**	**0.926**
	Mask R-CNN	0.839	0.864	0.937	0.519	0.899
Inflection/Inflammation	MaligNet	0.950	0.432	0.667	**0.962**	0.525
	Mask R-CNN	**0.953**	**0.447**	**0.739**	**0.962**	**0.557**
Degenerative change	MaligNet	**0.905**	0.662	**0.584**	0.954	**0.621**
	Mask R-CNN	0.891	**0.667**	0.342	**0.974**	0.452
Post-trauma	MaligNet	**0.952**	**0.737**	**0.378**	**0.991**	**0.500**
	Mask R-CNN	0.936	0.476	0.278	0.980	0.351

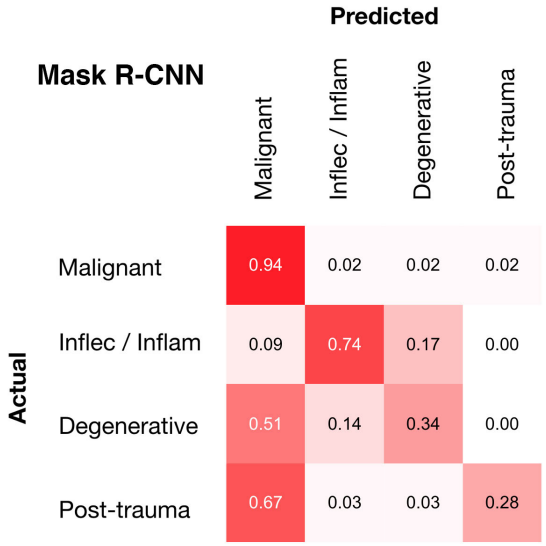

FIGURE 13. The normalized confusion matrix of the lesion classification task using Mask R-CNN. The rows represent the true labels (ground truth), and the columns represent the predicted labels.

to 0.848. We also show the results of lesion classification using MaligNet compared with the baseline model in the confusion matrices in Figure 12 and Figure 13. Furthermore, we compare the results of both models on each lesion class, as shown in Table 4.

The results show that MaligNet tends to predict malignancy extremely well. Other classes are rarer in the training data, which reduces the accuracy. Moreover, post-trauma is similar to malignancy; thus, it is difficult to distinguish this class from malignant lesions, which also reduces the accuracy.

C. THE IMPACT OF DATA

To study the effectiveness of unsupervised learning in our semisupervised approach in leveraging unlabeled data, we trained models with varying amounts of labeled and unlabeled data and measured their performances.

1) EFFECT OF THE AMOUNT OF LABELED DATA

In this experiment, we varied the amount of labeled training data while keeping the amount of unlabeled data fixed and measuring the F1-score. The results are shown in Figure 16. Using unsupervised data, MaligNet w/o global features improves every time the amount of training data is increased, thus it outperforms the Mask R-CNN baseline model by an average of 1.51%. Adding the global features improves the performance even further, reaching a relative F1-score average improvement of 2.40%. At the same F1-score level, our proposed method reduces the amount of labeled data required by an average of 20.11%. This result can be used as an anecdotal reference when determining the trade-off between spending more time labeling the data and making use of semisupervised methods.

2) EFFECT OF THE AMOUNT OF UNLABELED DATA

We also studied the effect of varying the amount of unlabeled data. As shown in Figure 17, the performance increases as we include more unlabeled training data. However, at higher amounts, the gain from adding more data decreases. This is expected because the unlabeled data are used to learn better representations. When the model has captured sufficient variation from the unsupervised data, adding more unsupervised data will have little to no effect.

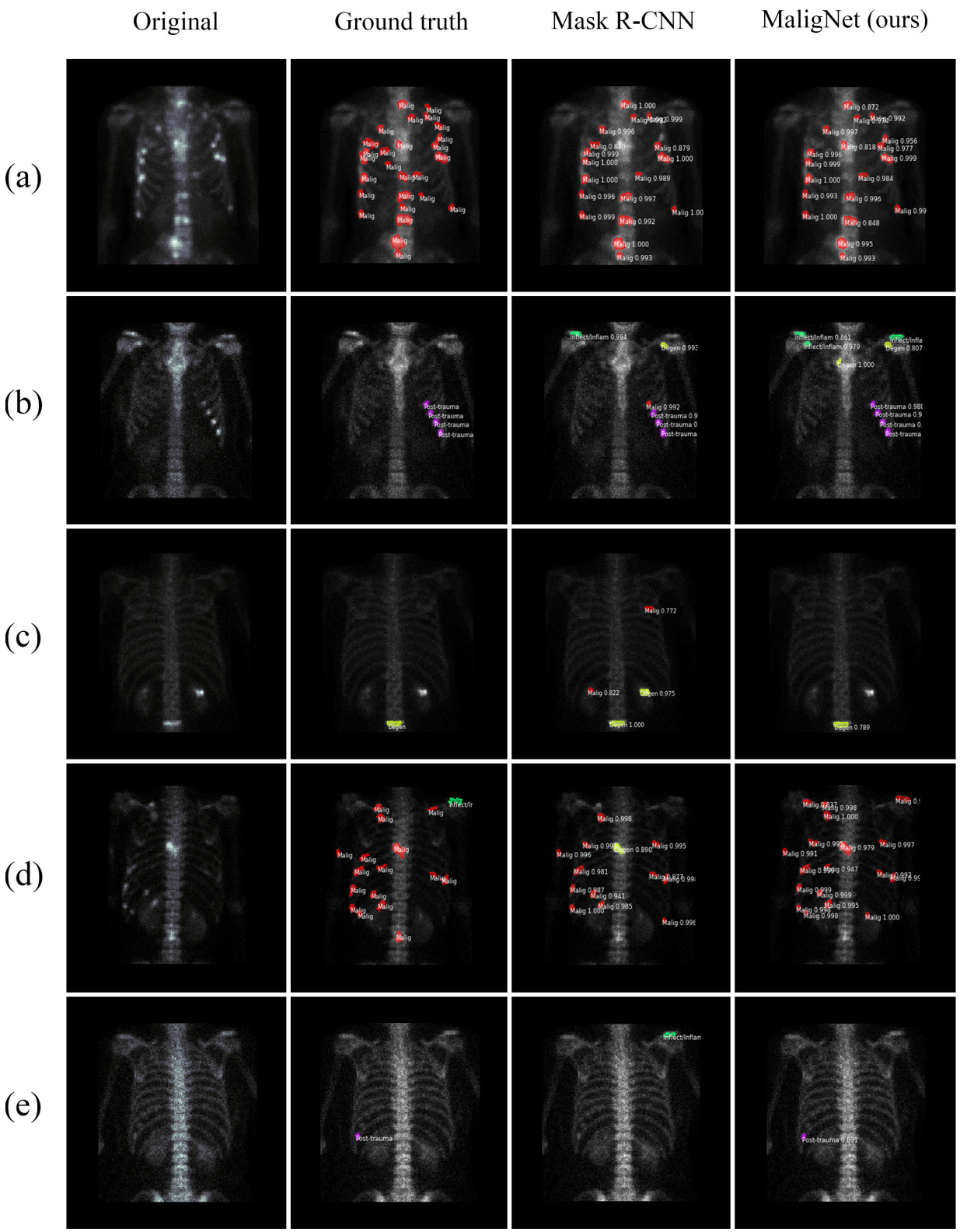

FIGURE 14. Hand-picked examples of the comparison results: the leftmost image is the original bone scintigram, the second image is the ground-truth image, the third image is the result of Mask R-CNN, and the rightmost image is the result of MaligNet (ours). Each row represents a different subject, and each column represents different image sources. A red region represents a malignant lesion, a green region represents an inflection/inflammation lesion, a yellow represents a degenerative change lesion, and a purple region represents a post-trauma lesion.

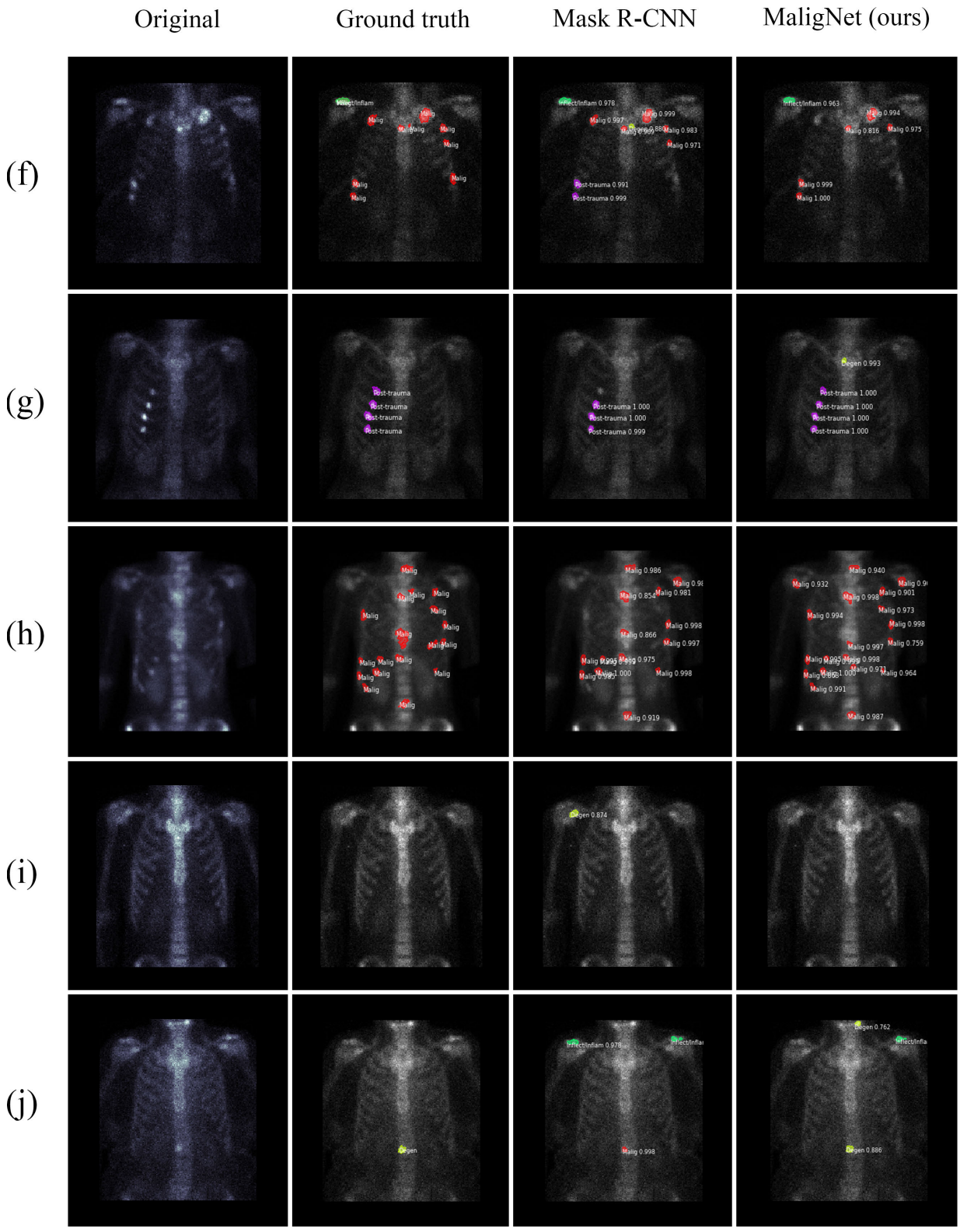

FIGURE 15. Random examples of the comparison results: in each row, the leftmost image is the original bone scintigram, the second image is the ground-truth image, the third image is the result of Mask R-CNN, and the rightmost image is the result of MaligNet (ours). Each row represents a different subject, and each column represents different image sources. A red region represents a malignant lesion, a green region represents an inflection/inflammation lesion, a yellow represents a degenerative change lesion, and a purple region represents a post-trauma lesion.

TABLE 5. The results of the self-training approach with different confidence thresholds.

Confidence threshold	Filter out	No. images	No. lesions	Mean precision	Mean sensitivity	Mean F1-score
All	-	18,560	52,756	0.815	0.826	0.818
>0.8	Lesions level	18,560	47,140	0.834	0.819	0.822
	Images level	15,423	33,572	0.837	0.836	0.831
>0.85	Lesions level	18,560	42,998	0.827	0.837	0.828
	Images level	12,761	21,964	0.849	0.843	**0.840**
>0.9	Lesions level	18,560	38,018	0.829	0.825	0.823
	Images level	10,443	13,871	0.836	0.833	0.830

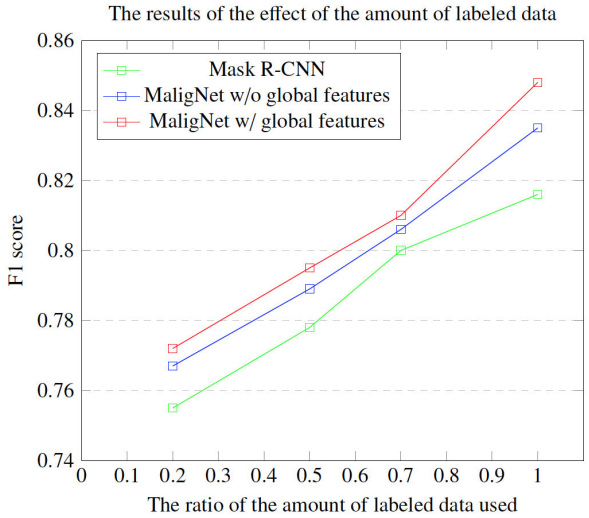

FIGURE 16. The effect of the amount of labeled data in lesion instance segmentation measured by F1-score. In the experiments the amount of labeled data was increased at various ratios. MaligNet can also use unlabeled data, whereas Mask R-CNN is fully supervised.

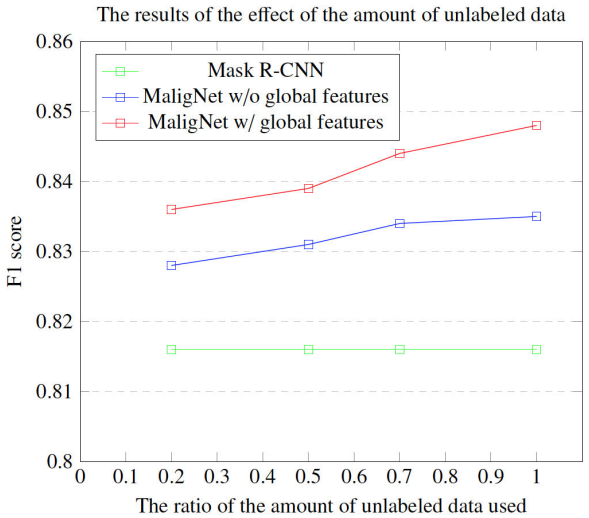

FIGURE 17. The effect of the amount of unlabeled data on lesion instance segmentation as measured by F1-score. We conduct the experiment by increasing the amount of unlabeled data while keeping the amount of labeled data fixed. Because Mask R-CNN cannot use unlabeled data, the performance remains constant.

TABLE 6. The results of MaligNet on bone cancer metastasis classification.

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score
Mask R-CNN	0.707	**0.941**	0.608	**0.919**	0.739
MaligNet	**0.741**	0.863	**0.657**	0.857	**0.746**

D. RESULTS OF THE BONE CANCER METASTASIS CLASSIFICATION TASK

Bone cancer metastasis classification from lesion instance segmentation is more difficult than direct classification. Rather than distinguishing between metastases and non-metastases, the model must also locate the positions of malignant lesions. The results in Table 6 show that MaligNet has higher accuracy, sensitivity, and F1-scores than does the baseline model. Although our model has lower precision and specificity, for our application, sensitivity is preferred over other metrics.

Our model requires a slightly longer inference time (0.76 ms for Mask R-CNN vs 0.87 ms for MaligNet). Even though the F1-score only increases slightly, the sensitivity, which is the main metric used for screening, improves by an absolute 5% without requiring more labeled data. Note that we could also apply model optimization and compression techniques, such as network pruning [20], weight quantization [26], binarized neural networks [8], and deep compression [19], to reduce the inference time; however, model accuracy is the main concern of this work.

E. COMPARISON WITH THE SELF-TRAINING METHOD

One popular semisupervised approach is *self-training*. Self-training produces virtual labels for unlabeled data by treating model predictions as the ground-truth label. The original labeled data are then combined with the unlabeled data (along with the labels produced by the model) to train a better model.

A confidence threshold value can be used to filter unlabeled data about which the model is not certain. We can treat the softmax output probability from the model as a confidence level and use only the data above a certain confidence level. We set the minimum confidence threshold for Mask R-CNN postprocessing (which removes clutter and merges overlapping regions) to 0.7 according to [48], which results

in confidence values ranging only from 0.7 to 1.0. We tried different confidence levels at 0.05 increments and report only the values that show a local maximum in Table 9.

We filtered the data in two ways: lesion level and image level. For the lesion level, we excluded lesions below the threshold value. For the image level, we excluded images containing at least one lesion with a score lower than the threshold. This approach filters out approximately 30% of the images. The results are shown in Table 5. The best results were obtained using a threshold value of 0.85 and image-level filtering. This approach improved the Mask R-CNN baseline results from 0.816 to 0.840.

VII. DISCUSSION

A. THE EFFECT OF APPLYING EACH TECHNIQUE IN MALIGNET

MaligNet reached an F1-score of up to 0.848 over the baseline of 0.816. Based on the results in Table 3, we studied the effectiveness of LFPN and global features. The results show that the effectiveness of LFPN makes our model more accurate than using global features. However, applying both techniques outperforms using only one technique.

B. THE LIMITATION OF USING UNLABELED DATA

The usefulness of unlabeled data is limited. At some point, the amount of unlabeled data reaches a saturation point, and the model efficiency does not greatly increase beyond that point. In contrast, increasing the labeled data can still improve the F1-score. However, labeling bone scintigraphy data is a time-consuming task. Thus, MaligNet is a good choice for utilizing unlabeled data to save significant amounts of time and resource savings.

C. ANALYSIS OF THE PREDICTION RESULTS

As shown in Figure 14.b and Figure 15.f, applying global features appears to help the model categorize lesions more accurately. MaligNet uses not only lesion features but also global information to categorize the lesion types. However, caution is needed when applying global features to avoid relying too much on global features rather than lesion features. As a result, MaligNet tends to predict malignant lesions more often than other types. For this reason, MaligNet has higher sensitivity than the baseline, as shown in Figures. 14.a and 14.d. This occasionally causes a false positive, as shown in the examples presented in Figure 15.h.

D. DIFFERENCE BETWEEN THE LFPN AND SELF-TRAINING

Self-training is a useful approach for performing semi-supervised learning. We trained MaligNet using the self-training method, which improves the model even further. As shown in Table 3, the F1-score improves from 0.848 to 0.851 after self-training. Self-training and the LFPN can be considered different ways to learn from unlabeled data. The LFPN, which is similar to an autoencoder, tries to learn a better data representation, while self-training provides discriminative information that helps the classification task. When using only the LFPN, our method achieves a slightly lower F1-score than when using self-training. However, with the former, MaligNet can be trained in one step on both types of data simultaneously, which takes less training time than self-training. The training and inference times without self-training for Mask R-CNN were 19.6 hours and 0.76 milliseconds, respectively, while MaligNet required 23 hours and 0.87 milliseconds, respectively. Models with self-training required twice the amount of training time. Moreover, our method was more accurate when we combined both techniques.

VIII. CONCLUSION AND FUTURE WORK

Almost all object detection or instance segmentation models are designed to be trained via supervised learning, which requires large amounts of labeled training data. However, our medical image dataset consists of only a small amount of labeled data; consequently, it can easily lead to model overfitting. We focused on using unlabeled data to leverage its utility and achieve the most effective model possible given the limited amount of labeled data. Therefore, we proposed MaligNet, a ladder network extension of Mask R-CNN for lesion instance segmentation in bone scintigraphy that uses semisupervised learning for training.

MaligNet is a single network that is simple, effective, flexible, and lightweight. Normally, semisupervised models must be trained using multiple steps. However, MaligNet is an end-to-end solution that can be trained in one step with both labeled and unlabeled data simultaneously, which reduces the training time. Our input data are bone scintigraphy images, which have similar patterns, characteristics, and compositions, unlike general images. For this reason, the LFPN can take advantage of the specificity of the data, enabling the model to learn good representations of bone scan images from unlabeled data. Furthermore, applying global features helps to classify the lesion types based on the overall composition of the image, more closely mimicking the diagnostic approach of physicians.

We evaluated the model using the mean precision, mean sensitivity, and mean F1-score in the lesion instance segmentation task and the accuracy, precision, sensitivity, specificity, and F1-score in bone cancer metastasis classification. MaligNet significantly outperforms the baseline model—by up to 2.33% without using global features and by 3.92% when global features are included.

We plan to compare our results with those of a nuclear medicine physician as a gold standard to determine the difference in decision making between a machine and physician for performance improvements. In further analyses, we plan to visualize the model to determine what the model sees and the reasons why it makes its categorizations. We also plan to apply our model to other domains, e.g., MRI and CT. Finally, we believe that our method provides an alternative approach for handling unlabeled data and will be useful in applications for other works.

TABLE 7. Details of the patients' gender and age statistics for each dataset type.

Dataset type	Male images	Female images	Min age	Mean age	Max age
Supervised training data	274	467	2	59.16	96
Supervised validation data	86	145	5	58.94	96
Supervised testing data	38	78	2	59.02	90
Unsupervised data	7624	10,936	2	57.40	97

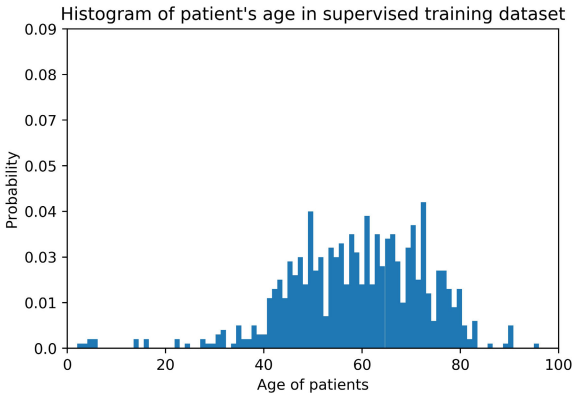

FIGURE 18. A histogram of patient age at bone scintigraphy in the supervised training dataset.

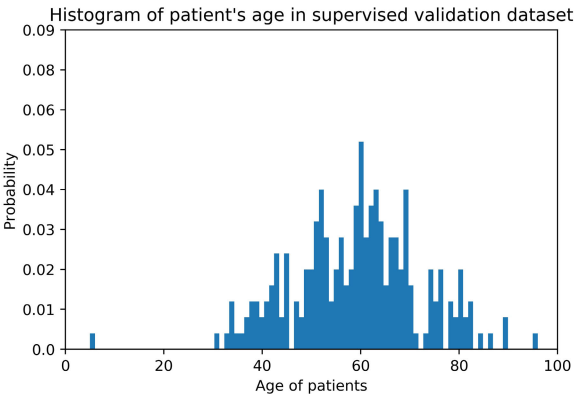

FIGURE 19. A histogram of patient age at bone scintigraphy in the supervised validation dataset.

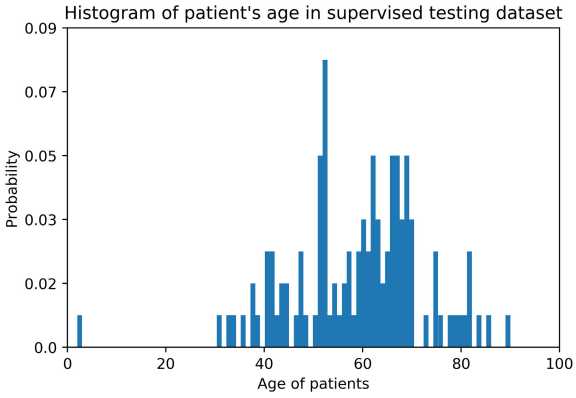

FIGURE 20. A histogram of patient age at bone scintigraphy in the supervised testing dataset.

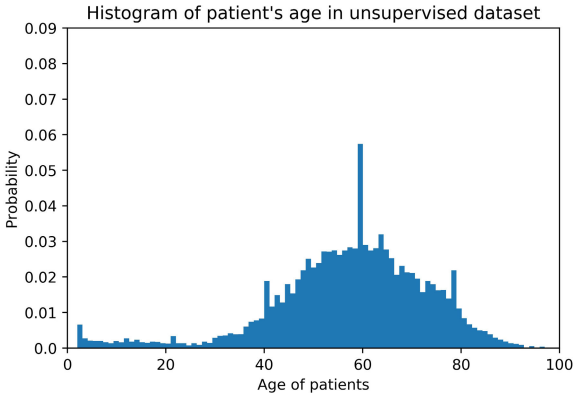

FIGURE 21. A histogram of patient age at bone scintigraphy in the unsupervised dataset.

TABLE 8. Final values of the hyperparameters used in the chest detection experiment.

Parameters	Parameter used
Image size (width,height)	(512, 512)
Core network	VGG-16
Batch size	16
Optimizer	Adam
Learning rate	0.001
Weight decay	0.0005
L2 regularization	0.0005
IoU threshold	0.45
Anchor box scaling factors	[0.07, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 1.05]
Anchor box steps	[8, 16, 32, 64, 128, 256, 512]
Anchor box offsets	[0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]

ACKNOWLEDGMENT

The authors would like to thank Mr. Thanayut Wiriyatharakij for his help with data collection. We would also like to thank Vivattanachai Sangsa-nga, Thanyaporn Phinthuphan, Penpicha Sangsa-nga, and Panyawut Sri-iesaranusorn for their help with the early versions of our labeling tool.

APPENDIXES

APPENDIX A: DETAILS OF PATIENT GENDER AND AGE IN THE DATASET

The details of the patients' gender data and age statistics, which are divided into supervised training, validation, testing, and unsupervised datasets, are shown in Table 7. We also display the age range of the patients in a histogram in Figure 18 - 21.

APPENDIX B: HYPERPARAMETERS OF THE SINGLE SHOT MULTIBOX DETECTOR (SSD) IN THE CHEST DETECTION EXPERIMENTS

In Section IV-A, our SSD has pretrained weights from ImageNet and is retrained with our data using the hyperparameters, as shown in Table 8.

TABLE 9. Final values of the hyperparameters used in the lesion instance segmentation experiment from the parameter search.

Parameters	Parameter search	Final parameters
Image size (width,height)	(320, 320),(512, 512)	(320, 320)
Core network	ResNet-34, ResNet-50, ResNet-101, Xception	ResNet-50
$\lambda_{rc}, \lambda_{rb}, \lambda_{cc}, \lambda_{cb}, \lambda_{cm}, \lambda_{us}$	1.0	1.0
Gaussian noise ratio	0.015, 0.03, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5	0.3
Batch size	2, 8, 16 (maximum batch size)	16
Optimizer	SGD, Adam	Adam
Learning rate	0.0001, 0.001, 0.002, 0.005	0.001
Weight decay	0.01, 0.001, 0.0001, 0.01, 0.02	0.0001
RPN NMS threshold	0.6, 0.7, 0.9, 0.99	0.7
Train ROI per image	80, 100, 200, 300	200
RPN anchor scales	(32,64,128,256,512)	(32,64,128,256,512)
RPN anchor ratio	[0.5, 1, 2]	[0.5, 1, 2]
RPN anchor stride	1	1
RPN anchor per image	256	256
Max ground-truth instances	50, 100, 200, 300	100
Detection minimum confidence	0.7	0.7

APPENDIX C: HYPERPARAMETERS OF MALIGNET IN THE EXPERIMENTS OF INSTANCE SEGMENTATION

All experiments in Section VI-B use the same hyperparameters. We trained the network from scratch without pretrained weights. We attempted to optimize model performance by searching for the optimal hyperparameters to the greatest extent possible. The final hyperparameter values are shown in Table 9. Because all the cost functions are self-normalized and the costs do not vary largely, we set λ equal to one for all the experiments. For the batch size hyperparameter, we found that larger batches led to more accurate results; however, due to GPU resource limitations, 16 is the maximum batch size that can be used.

REFERENCES

[1] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow, "Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative," *Med. Image Anal.*, vol. 52, pp. 109–118, Feb. 2019.

[2] R. Azmi, N. Norozi, R. Anbiaee, L. Salehi, and A. Amirzadi, "IMPST: A new interactive self-training approach to segmentation suspicious lesions in breast MRI," *J. Med. Signals Sens.*, vol. 1, no. 2, p. 138, 2011.

[3] L. Belcher, "Convolutional neural networks for classification of prostate cancer metastases using bone scan images," Dept. Astron. Theor. Phys., Lund, Sweden, Tech. Rep., 2017.

[4] T. Bradshaw, T. Perk, S. Chen, H.-J. Im, S. Cho, S. Perlman, and R. Jeraj, "Deep learning for classification of benign and malignant bone lesions in [F-18] NAF PET/CT images," *J. Nucl. Med.*, vol. 59, no. 1, p. 327, 2018.

[5] M. Bustamante, V. Gupta, D. Forsberg, C.-J. Carlhäll, J. Engvall, and T. Ebbers, "Automated multi-atlas segmentation of cardiac 4D flow MRI," *Med. Image Anal.*, vol. 49, pp. 128–140, Oct. 2018.

[6] V. Cheplygina, M. De Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[7] C. B. Confavreux, J.-B. Pialat, A. Bellière, M. Brevet, C. Decroisette, A. Tescaru, J. Wegrzyn, C. Barrey, F. Mornex, P.-J. Souquet, and N. Girard, "Bone metastases from lung cancer: A paradigm for multidisciplinary onco-rheumatology management," *Joint Bone Spine*, vol. 86, no. 2, pp. 185–194, Mar. 2019.

[8] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1," 2016, *arXiv:1602.02830*. [Online]. Available: https://arxiv.org/abs/1602.02830

[9] J. Dang, "Classification in bone scintigraphy images using convolutional neural networks," M.S. thesis, Lund Univ., Math. (Fac. Eng.), Lund, Sweden, 2016.

[10] B. D. D. Vos, J. M. Wolterink, P. A. D. Jong, M. A. Viergever, and I. Išgum, "2D image classification for 3D anatomy localization: Employing deep convolutional neural networks," *Proc. SPIE*, vol. 9784, Mar. 2016, Art. no. 97841Y.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[13] S. Geng, S. Jia, Y. Qiao, J. Yang, and Z. Jia, "Combining CNN and mil to assist hotspot segmentation in bone scintigraphy," in *Proc. Int. Conf. Neural Inf. Process.* New York, NY, USA: Springer, 2015, pp. 445–452.

[14] S. Geng, J. Ma, X. Niu, S. Jia, Y. Qiao, and J. Yang, "A MIL-based interactive approach for hotspot segmentation from bone scintigraphy," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 942–946.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1440–1448.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[18] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Med. Image Anal.*, vol. 52, pp. 199–211, Feb. 2019.

[19] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*. [Online]. Available: https://arxiv.org/abs/1510.00149

[20] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

[21] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine-GAN: Semantic segmentation of multiple spinal structures," *Med. Image Anal.*, vol. 50, pp. 23–35, Dec. 2018.

[22] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] M. P. Heinrich, O. Oktay, and N. Bouteldja, "OBELISK-net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions," *Med. Image Anal.*, vol. 54, pp. 1–9, May 2019.

[25] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[26] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, 2017.
[27] S. Ibrahim and M. Altaei, "Satellite image classification using multi features based descriptors," Al-Nahrain Univ., Baghdad, Iraq, Tech. Rep., Apr. 2018, p. 9.
[28] T. Ibrahim, L. Mercatali, and D. Amadori, "Bone and cancer: The osteoncology," *Clinical Cases Mineral Bone Metabolism*, vol. 10, no. 2, p. 121, 2013.
[29] *Primary Bone Cancer*, Nat. Cancer Inst., Bethesda, MD, USA, 2018.
[30] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.
[31] S. K. Kang, H. Choi, S. Park, S.-K. Kim, T.-S. Kim, and J. S. Lee, "Unsupervised lesion detection in bone scintigraphy using deep learning-based image inpainting technology," *J. Nucl. Med.*, vol. 60, no. 1, p. 403, 2019.
[32] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Med. Image Anal.*, vol. 54, pp. 88–99, May 2019.
[33] T. Küstner, S. Müller, M. Fischer, J. Weib, K. Nikolaou, F. Bamberg, B. Yang, F. Schick, and S. Gatidis, "Semantic organ segmentation in 3D whole-body MR images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3498–3502.
[34] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, and T. Duerig, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," 2018, *arXiv:1811.00982*. [Online]. Available: https://arxiv.org/abs/1811.00982
[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, vol. 1, Jul. 2017, p. 4.
[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2014, pp. 740–755.
[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 21–37.
[38] D. J. Magee, J. E. Zachazewski, W. S. Quillen, and R. C. Manske, *Pathology and Intervention in Musculoskeletal Rehabilitation*, vol. 3. Amsterdam, The Netherlands: Elsevier, 2015.
[39] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using CNNs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2016, pp. 230–238.
[40] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, and K. Shpanskaya, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: https://arxiv.org/abs/1711.05225
[41] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3546–3554.
[42] S. E. A. Raza, L. Cheung, M. Shaban, S. Graham, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Micro-net: A unified model for segmentation of various objects in microscopy images," *Med. Image Anal.*, vol. 52, pp. 160–173, Feb. 2019.
[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[44] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[46] Q. Tong, C. Li, W. Si, X. Liao, Y. Tong, Z. Yuan, and P. A. Heng, "RIANet: Recurrent interleaved attention network for cardiac MRI segmentation," *Comput. Biol. Med.*, vol. 109, pp. 290–302, Jun. 2019.
[47] B. M. W. Tsui, R. N. Beck, K. Doi, and C. E. Metz, "Analysis of recorded image noise in nuclear medicine," *Phys. Med. Biol.*, vol. 26, no. 5, pp. 883–902, Sep. 1981.
[48] Waleedka. (2017). *Mask RCNN*. [Online]. Available: https://github.com/matterport/Mask_RCNN
[49] S. Wang, K. He, D. Nie, S. Zhou, Y. Gao, and D. Shen, "CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation," *Med. Image Anal.*, vol. 54, pp. 168–178, May 2019.
[50] M. Winkels and T. S. Cohen, "Pulmonary nodule detection in CT scans with equivariant CNNs," *Med. Image Anal.*, vol. 55, pp. 15–26, Jul. 2019.
[51] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai, I. Eric, and C. Chang, "Gland instance segmentation using deep multichannel neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2901–2912, Mar. 2017.
[52] D. Yang, S. Zhang, Z. Yan, C. Tan, K. Li, and D. Metaxas, "Automated anatomical landmark detection ondistal femur surface using convolutional neural network," in *Proc. IEEE 12th Int. Symp. Biomed. Imaging (ISBI)*, Apr. 2015, pp. 17–21.
[53] M. Zeng, T. Yu, X. Wang, L. T. Nguyen, O. J. Mengshoel, and I. Lane, "Semi-supervised convolutional neural networks for human activity recognition," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 522–529.

TERAPAP APIPARAKOON received the B.Eng. degree from the Faculty of Computer Engineering, Chulalongkorn University, Bangkok, Thailand, in 2018, where he is currently pursuing the M.Eng. degree with the Faculty of Computer Engineering. His research interests include computer vision, medical image analysis, and face recognition.

NUTTHAPHOL RAKRATCHATAKUL received the B.Eng. degree in computer engineering from Chulalongkorn University, Bangkok, Thailand, in 2018. His research interests include computer vision, medical image analysis, and remote sensing.

MAYTHINEE CHANTADISAI received the degree of Thai Board of Nuclear Medicine from the Medical Council of Thailand and the Faculty of Medicine, Chulalongkorn University, in 2013. Her research interests include precision medicine, theranostics, and cardiovascular nuclear medicine imaging.

USANEE VUTRAPONGWATANA received the degree of Thai Board of Nuclear Medicine from the Medical Council of Thailand and the Faculty of Medicine, Siriraj Hospital, and Mahidol University, in 2010. Her research interest includes clinical applications of nuclear medicine.

KANAUNGNIT KINGPETCH received the degree of Thai Board of Nuclear Medicine from the Medical Council of Thailand and the Faculty of Medicine, Chulalongkorn University, in 2000. Her research interest includes clinical applications of nuclear medicine in thyroid cancer.

SASITORN SIRISALIPOCH received degree of the Thai Board of Nuclear Medicine from the Medical Council of Thailand and the Faculty of Medicine, Chulalongkorn University, in 1993. Her research interest generally includes clinical applications of nuclear medicine.

YOTHIN RAKVONGTHA received the B.Eng. degree (Hons.) from Chulalongkorn University, Thailand, the M.S. degree from the University of California at Los Angeles, Los Angeles, USA, and the Ph.D. degree from The University of Texas at Arlington, USA, all in electrical engineering. He is currently the Director of Chulalongkorn University Biomedical Imaging Group and an Assistant Professor with the Division of Nuclear Medicine, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Thailand. Prior to joining Chulalongkorn University, he was an Instructor of radiology with the Gordon Center for Medical Imaging, Massachusetts General Hospital, and Harvard Medical School, USA. His research interest focuses on biomedical imaging.

TAWATCHAI CHAIWATANARAT received the degree of Thai Board of Nuclear Medicine from the Medical Council of Thailand and the Faculty of Medicine, Chulalongkorn University, in 1990. His research interests include clinical applications of nuclear medicine and medical image analysis.

EKAPOL CHUANGSUWANICH received the B.S. and M.S. degrees in electrical and computer engineering from Carnegie Mellon University, in 2008 and 2009, respectively, and the Ph.D. degree from MIT, in 2016. He then joined the Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory. He is currently a Faculty Member of the Department of Computer Engineering with Chulalongkorn University. His research interests include speech processing, assistive technology, and health applications.