Pi-PE: A Pipeline for Pulmonary Embolism Detection using Sparsely Annotated 3D CT Images

Deepta Rajan   drajan@us.ibm.com
David Beymer   beymer@us.ibm.com
Shafiqul Abedin   sabedin@us.ibm.com
Ehsan Dehghan   edehgha@us.ibm.com
IBM Research Almaden
650 Harry Road, San Jose, CA 95120

Abstract

Pulmonary embolisms (PE) are known to be one of the leading causes for cardiac-related mortality. Due to inherent variabilities in how PE manifests and the cumbersome nature of manual diagnosis, there is growing interest in leveraging AI tools for detecting PE. In this paper, we build a two-stage detection pipeline that is accurate, computationally efficient, robust to variations in PE types and kernels used for CT reconstruction, and most importantly, does not require dense annotations. Given the challenges in acquiring expert annotations in large-scale datasets, our approach produces state-of-the-art results with very sparse emboli contours (at 10mm slice spacing), while using models with significantly lower number of parameters. We achieve AUC scores of 0.94 on the validation set and 0.85 on the test set of highly severe PE. Using a large, real-world dataset characterized by complex PE types and patients from multiple hospitals, we present an elaborate empirical study and provide guidelines for designing highly generalizable pipelines.

1. Introduction

A pulmonary embolism (PE) manifests as blocks in pulmonary arteries triggered by blood clots, air bubbles, or accumulation of fat tissues that occur typically during surgery, pregnancy or cancer. PE is known to be one of the leading causes of cardiac-related mortality, where an early diagnosis and treatment is expected to have a significant impact in controlling the mortality rate. It is estimated that between $300,000$ to $600,000$ individuals are affected by PE every year in the US Beckman et al. (2010). Computed tomographic pulmonary angiography (CTPA) is the primary diagnostic exam to detect arterial diseases, given the high spatial resolution of CT scanners. Each CTPA study is a 3D image containing hundreds of slices, some of which show evidence of PE as irregularly shaped filling defects.

In practice, each occurrence of PE can belong to one of the following broad categories: peripheral, segmental, subsegmental, lobar, or saddle type, which can be typically determined based on its arterial location. In particular, subsegmental PE is considered to be the hardest to detect, since it often occurs subtly in subsegmental branches of the pulmonary artery. Consequently, radiologists are required to painstakingly examine every slice in a CT image for detecting PE, thus making this process highly cumbersome and time-consuming. Moreover, unlike other common diseases visualized in chest CTs such as lung nodules, which usually
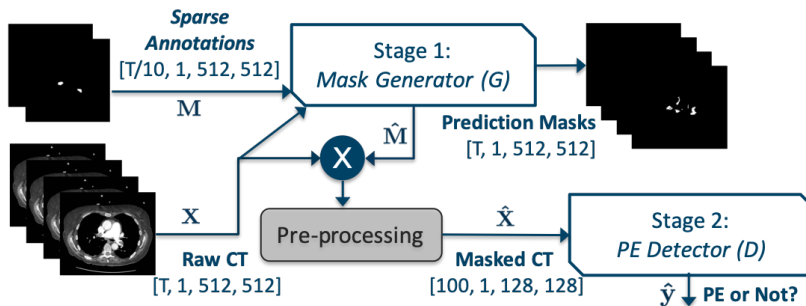
Figure 1: Illustration of the proposed two-stage pipeline for PE detection. `Stage 1` is comprised of a mask generator G, based on 2D context-augmented U-Net, and `Stage 2` is comprised of a PE detector D, involving a 2D Conv-LSTM model coupled with multiple instance learning. Both the raw input CTs (X) and sparsely annotated masks M are fed into G to produce prediction masks ($\hat{M}$). Subsequently, X and $\hat{M}$ are multiplied to obtain $\hat{X} = X \odot \hat{M}$ that is fed as input to `Stage 2`. Finally, D outputs a prediction $\hat{y} \in [1, 0]$ indicating the presence or absence of PE.

appear spherical, or emphysema, which can be observed across the entire lung, PE is known to appear much more asymmetrically only in isolated regions of pulmonary vasculature.

Given the afore-mentioned challenges in detecting PE, computer-aided diagnostic tools Liang and Bi (2007) have become prominent. More specifically, data-driven approaches based on deep learning have produced promising results for automatic PE detection Tajbakhsh et al. (2015). The most successful solutions in medical image analysis often comprise multi-stage pipelines tailored for a specific organ/disease. Without loss of generality, such a pipeline in turn includes a segmentation stage for candidate generation, i.e. semantically meaningful regions that are likely to correspond to the disease occurrence, and a classification stage for the actual detection. A crucial bottleneck of this approach is the need for large annotated datasets. Acquiring expert annotations for 3D volumes, where every instance of disease occurrence is annotated (often referred to as *dense annotations*), is time-consuming and error-prone. Furthermore, there is a dearth of standard benchmark datasets for PE detection using CTPA, and most of the research in this space is conducted on custom datasets or small-scale challenge dataset de Radiodiagnóstico and the M+Visión consortium (2013). In practice, the small-data challenge is often combated by adopting a transfer learning strategy that refined classifiers pre-trained on natural image data Tajbakhsh et al. (2016). Recently, Huang et al. (2019) showed that such a strategy, where 500K video clips were used to pre-train a 77-layer 3D convolutional network, could be effectively fine-tuned for learning a PE classifier using dense annotations.

In this paper, we adopt an alternate approach of building an accurate PE detection system using only sparsely labeled CT volumes. More specifically, we develop a two stage detection pipeline (shown in Figure 1) designed exclusively using 2D CNNs, wherein the candidate generation state utilizes a novel context-augmented U-Net and the classifier stage employs a simple 2D Conv-LSTM model coupled with multiple instance learning Zhu et al. (2017); Ilse et al. (2018); Braman et al. (2018), compared to the 77−layer 3D CNN in Huang et al. (2019). We find that, even with significantly smaller number of parameters and with

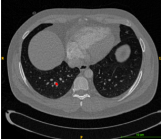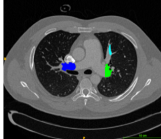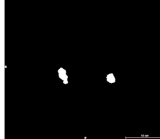| Subsegmental PE	Segmental PE	Lobar PE	Ground Truth	CA U-Net

Table 1: Example CT studies showing various types of PE. We also show the ground truth annotation and the prediction from `Stage 1` for the Lobar PE example.

no pre-training, our approach produces state-of-the-art detection results on a challenging, large-scale real-world dataset. Further, we study its generalization across hospitals/datasets, given the large disparity across image acquisition systems and protocols, and demonstrate the proposed approach to be highly robust.

Our contributions can thus be summarized as follows:

- We develop a novel two-stage approach for PE detection – `Stage 1` is comprised of a 2D U-Net based Ronneberger et al. (2015) mask generator and `Stage 2` utilizes a *ConvLSTM* Xingjian et al. (2015) based PE detector.

- Our approach does not require expensive *dense annotations* and operates exclusively on *sparse annotations* generated for every 10 mm of positive CT scans.

- We use a context-augmentation strategy that enables the 2D U-Net in `Stage 1` to produce high-quality masks.

- By modeling each 3D CT volume as a bag of instances, i.e., features for each 2D slice obtained using a Conv-LSTM, we propose to employ multiple instance learning, based on feature aggregation, to detect PE.

- For the first time, we evaluate our approach using a large-scale, multi-hospital chest CT dataset that well represents real-world scenarios through the inclusion of complex PE types and diverse imaging protocols.

- We present insights from an elaborate empirical study, while discussing the impact of different architectural design choices on the generalization performance.

- We show that our approach achieves state-of-the-art detection performance, with AUC scores of 0.94 on a validation set of all PE types and 0.85 on a test set of high-severity PE.

2. Dataset Description

We collected 1,874 PE positive and 718 negative anonymized, contrast-enhanced chest CT studies and their corresponding radiology reports. Note that, due to the specific anonymization protocol used in our data curation process, we are unable to determine if two studies belong to the same patient. Our dataset is curated to represent variations across multiple imaging centers (> 100) and different contrast-enhanced imaging protocols,
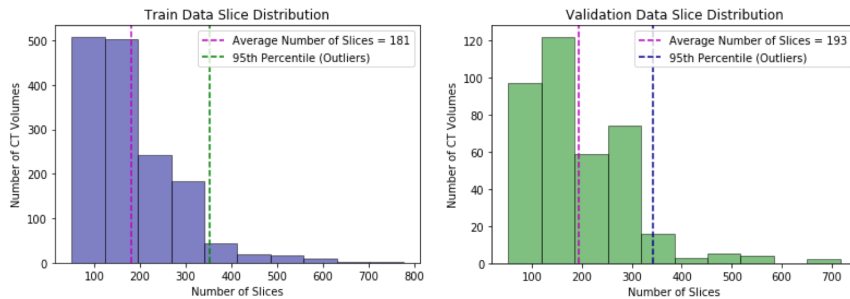
Figure 2: Histogram of number of image slices in each CT study across the train and validation datasets. Our PE detector (D) as part of `Stage 2` is designed to use CTs of 100 slices during training.

Label	Train	Validation	Test	Total
Positive	1,053	264	385	1,702
Negative	473	118	127	718

Table 2: Sample sizes of the custom CT dataset sourced from multiple hospitals.

namely PE and CTA. In comparison, currently reported studies in the literature including the state-of-the-art PENet Huang et al. (2019) focus exclusively on the PE protocol, and study generalization to only two hospitals. Consequently, the problem setup we consider is significantly more challenging and further, we do not assume access to dense annotations. Table 2 shows sample sizes used for train, validation and test phases of our algorithm. Further, the number of slices in each of the volumes can vary significantly, as illustrated in Figure 2.

2.1. Sparse Annotations

As part of the data preparation, we adapt the NLP pipeline described in Guo et al. (2017) to identify PE positive studies from a patient's radiology report, while detecting the radiologist recommended phrase 'no evidence of pulmonary embolism' to identify PE negative studies. The positive studies were further vetted by board-certified radiologists who annotated the scans by drawing a contour around every embolism occurrence on slices approximately 10mm apart. This process naturally results in multiple unannotated slices between every pair of annotated slices, depending on the slice spacing. We refer such CT studies to be *sparsely* annotated. While each study was annotated by only one clinical expert, a total of 17 radiologists served as annotators in the process. Out of the 1,874 positive studies that were processed, 172 of those were discarded due to reasons including the lack of definitive evidence for presence of PE (discrepancy between annotator and the reporting radiologist), insufficient contrast, metal or motion artifacts, etc.

3. Proposed Methodology

3.1. Approach Overview

We develop a two stage approach for PE detection from CT images. While the first stage processes the raw CT volumes to produce a mask that identifies candidate regions that are likely to correspond to emboli regions, the latter stage operates on the masked volume from `Stage 1` to perform the actual detection. In contrast to existing solutions, our approach relies exclusively on 2D convolutions and does not require dense annotations. As illustrated in Figure 1, `Stage 1` is implemented using a novel context-augmented 2D U-Net, and for `Stage 2`, we adopt a multiple instance learning (MIL) formulation, wherein each 3D volume X is viewed as a bag of instances defined using the individual 2D slices $x_1, \ldots, x_T$. Here, T denotes the total number of slices in X. Broadly, existing MIL methods focus on inferring appropriate aggregation functions either on (i) the instance-level predictions $(y_1, \ldots, y_T)$ to produce bag-level prediction y Zhu et al. (2017); Braman et al. (2018), or (ii) the instance-level latent features $\{z_1, \ldots, z_T\}$ to construct the bag-level feature z, which can be subsequently used to obtain the prediction y Ilse et al. (2018). We adopt the latter approach, where the instance features are obtained using a 2D Conv-LSTM model and the feature aggregation is carried out using different functions including mean, max and learnable attention modules.

3.2. Stage 1: Candidate Mask Generation

The role of `Stage 1` is to segment an image and identify PE candidates which are localized regions with semantics indicative of the disease. As an initial preprocessing step, each input CT scan is resampled to a volume with 2mm slice spacing. The architecture for the mask generator G is a standard 2D U-Net Ronneberger et al. (2015), an encoder-decoder style network comprised of a contracting path to downsample the input image while doubling the number of channels, followed by an expansive path to upsample the image. Though using a 2D U-Net significantly simplifies the computation, processing each 2D slice independently fails to leverage crucial context information in the neighboring slices. In order to circumvent this, we propose to extract *slabs* of 4 neighboring slices from either side of each 2D slice, to form a stack of 9 slices. We treat the raw intensities from each the 9 slices as the channel dimensions, thus producing slabs of size $(9, 512, 512)$ representing number of channels, height and width. We refer to this architecture as the context-augmented U-Net (CA U-Net). We observed from our experiments that this simple augmentation strategy consistently produced high-quality masks (see example in Table 1).

Each *downblock* in our U-Net architecture contains 2D convolution layers with a 3x3 kernel, a batch normalization layer, a ReLU activation layer and a maxpool layer with a stride of 2 to downsample the image. While, each *upblock* upsamples and then concatenates features at the same level or depth of the network, followed by a convolutional layer coupled with batch normalization and ReLU activation. The depth of the network G was fixed at 4. Upon training, G produces output probabilities for each pixel in the middle slice of the slab, indicating the likelihood of being PE candidate. The training objective was to achieve a high *dice coefficient*, a metric which describes the pixel-wise similarity between prediction masks $(\hat{M})$ and ground truth annotation masks (M), and has a range of $[0-1]$. It is defined as $DC = (2 * \sum_{i=1}^{N} \hat{m}_i m_i)/(\sum_{i=1}^{N} \hat{m}_i^2 + \sum_{i=1}^{N} m_i^2)$, where N is number of voxels,
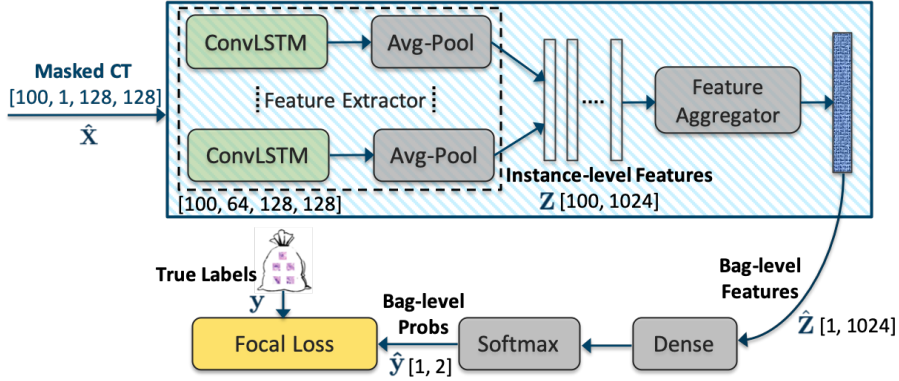
Figure 3: Architecture of the proposed `Stage 2` PE detector. Each instance $(\hat{x}_1, \ldots, \hat{x}_{100})$ is transformed using a single *ConvLSTM* layer followed by an *AvgPool* layer to obtain instance-level features $(Z = z_1, \ldots, z_T)$. A feature aggregation function (e.g. *max*) then produces a bag-level feature that can be subsequently used for the actual classification.

$\hat{m}_i \in \hat{M}$ and $m_i \in M$ Milletari et al. (2016). In practice, we adopt the continuous dice loss as $\mathcal{L}_{dice} = 1 - DC$.

3.3. Stage 2: Pulmonary Embolism Detection

As described earlier, to perform the actual PE detection, we treat each CT volume as a bag of multiple 2D slices (instances). Hence, the goal of `Stage 2` is to assign a prediction label to a bag indicating the presence or absence of PE. Multiple instance learning is a well-studied problem, where each instance is processed independently, and their features (or predictions) can be aggregated for obtaining bag-level predictions. However, we argue that processing each slice independently in a 3D volume can produce noisy predictions since the local context is not included. More specifically, we utilize a Conv-LSTM Xingjian et al. (2015) model to produce instance features that automatically incorporates context from its neighboring slices, and perform feature aggregation similar to any MIL system.

As illustrated in Figure 3, the PE detector D contains an instance-level feature extractor followed by an MIL module. The feature extractor is a 2D Conv-LSTM architecture that effectively captures spatio-temporal correlations in a CT volume and produces meaningful instance-level features. All input-to-state and state-to-state transitions use a convolution operation containing 64 filters, a 3x3 kernel and a padding size of 1. The input to D are the masked CTs, denoted as $\hat{X}$, that is obtained as follows: First, the prediction masks, $\hat{M}$, from `Stage 1` are multiplied with raw CT volumes X to create masked CT volumes. We then reduce the z-dimension of the masked volumes for computational efficiency. To this end, we use a lung segmentation algorithm to detect the boundary axial slices (z_{start}, z_{end}) that span the lung region. We then extract $T = 100$ middle slices from within this range, crop to reduce image height and width to $(384, 384)$ and finally resize to $(128, 128)$, thus transforming X to produce $\hat{X} \in \mathbb{R}^{100,1,128,128}$. Each instance $(\hat{x}_1, \ldots, \hat{x}_{100})$ is transformed

by the *Conv-LSTM* model as follows:

$$
\begin{aligned}
\text{(input gate):} \quad & i_t = \sigma(W_{xi}\hat{x}_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
\text{(forget gate):} \quad & f_t = \sigma(W_{xf}\hat{x}_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
\text{(cell state):} \quad & c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}\hat{x}_t + W_{hc}h_{t-1} + b_c) \\
\text{(output gate):} \quad & o_t = \sigma(W_{xo}\hat{x}_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o) \\
\text{(hidden state):} \quad & h_t = o_t \circ \tanh(c_t)
\end{aligned}
\tag{1}
$$

The features are then average-pooled using a kernel of size 32 to produce dense 1024-dimensional features $z_1, \ldots, z_{100}$ for all slices in $\hat{X}$. In order to perform feature aggregation for MIL, we explored the use of *max*, *mean* and learnable *self-attention* functions. The self-attention function used is similar to the one described in Song et al. (2018), and was implemented with multiple attention heads.

$$
\hat{z} = \sum_{k=1}^{100} a_k z_k; \quad a_k = \frac{\exp(U^T \tanh(V z_k^T))}{\sum_j \exp(U^T \tanh(V z_j^T))}.
\tag{2}
$$

Here, a_k denotes the attention coefficients and U, V denote the learnable parameters for the attention module. The aggregated feature from the multi-head self-attention was further projected using a *linear* layer to obtain the final bag-level features. For training the detector model D, we also explored using the standard binary cross-entropy (BCE) loss and the focal loss Lin et al. (2017) defined as:

$$
\ell(\hat{y}, y) = -y(1 - \hat{y})^\gamma \log(\hat{y})
\tag{3}
$$

At inference time, we apply the preprocessing steps of cropping and resizing to $(128, 128)$ spatial resolution, but make predictions for moving windows of $T = 100$ slice (with 25 slice overlap) and use the maximum detection probability as the final prediction for the test CT scan.

4. Empirical Results

In this section, we present a detailed empirical analysis conducted to evaluate the performance of the proposed pipeline and study its behavior with respect to different architectural choices. In particular, we share insights from ablation studies focused on effect of the number of instances T used in Stage 2 PE detection, the strategies used for feature extraction and aggregation, and finally the choice of loss function used for training Stage 2.

4.1. Experiment Setup

All our experiments are based on modifying the PE detector in Stage 2, while retaining the Stage 1 model to be the same. Details on sample sizes used in our empirical study are provided in Table 2. Typically, for successful adoption of detection algorithms in clinical practice, they are expected to have a high recall rate on the abnormal cases (also referred to

Metrics	PE Detector: Feature Extractor + Aggregation Strategy + Loss Function					
	C+SA+B	CL+SA+B	CL+MSA+B	CL+AP+B	CL+MP+B	CL+MP+F
Acc.	0.80	0.83	0.83	0.84	0.86	**0.88**
AUC	0.86	0.88	0.89	0.90	0.91	**0.94**
F1	0.86	0.88	0.87	0.88	0.90	**0.91**

Table 3: Validation performance comparison using different combinations of feature extractor, aggregation strategy and loss function. Here, CL = Conv-LSTM, C = Conv. with no LSTM, SA = Self-attention, MSA = Multi-head Self Attenion, B = BCE Loss, F = Focal Loss, MP = Max pooling aggregation and AP = Average pooling aggregation.
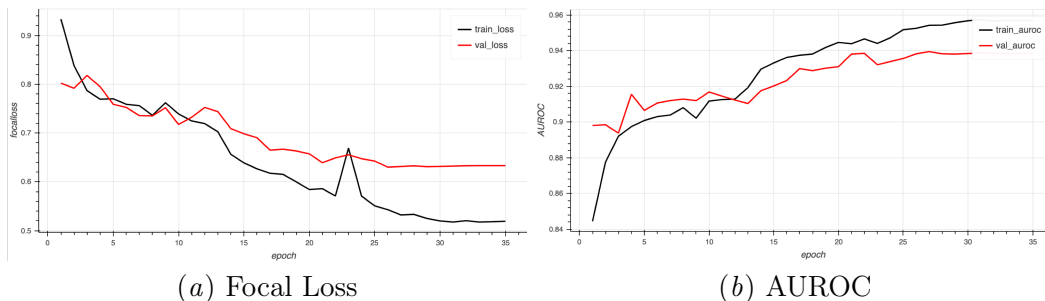

(a) Focal Loss
(b) AUROC

Figure 4: Training behavior of the `Stage 2` PE detection model. Using the focal loss produces significantly better generalization when compared to the conventional BCE loss.

as sensitivity). However, in order to obtain a well-rounded evaluation of the performance we report the following metrics: accuracy (*Acc*), sensitivity or recall (*Rec*) and precision (*Prec*).

$$\text{Acc} = \frac{tp + tn}{tp + fp + fn + tn}; \quad \text{Rec} = \frac{tp}{tp + fn}; \quad \text{Prec} = \frac{tp}{tp + fp}; \quad \text{f1} = 2.\frac{\text{Prec.Rec}}{\text{Prec} + \text{Rec}},$$

where tp, fp, fn, tn correspond to the number of true positives, false positives, false negatives and true negatives respectively. Further, to obtain an overview on the performance we use the $f1$-score and the area under receiver operator curve (AUROC).

Training: We trained the model D for `Stage 2` using an adaptive learning rate of $1e - 3$, which is subsequently reduced based on plateauing behavior of the validation loss. Other hyperparameters include a batch size of 8, the number of instances (T) set to 100 (unless specified otherwise), and the *Adam* optimizer with a weight decay of 0.01. All implementations were carried out in Pytorch, and we performed multi-gpu training using 4 NVIDIA GTX GPUs.

4.2. Ablation Studies

In this section, we provide details on the various ablation studies carried out to understand the effect of each architectural component towards the validation performance of the PE detector.

- **Study 1 - Effect of number of instances**: Given the limited GPU memory sizes, and the large sizes of CT volumes, we varied the number of instances (T) that were selected

8

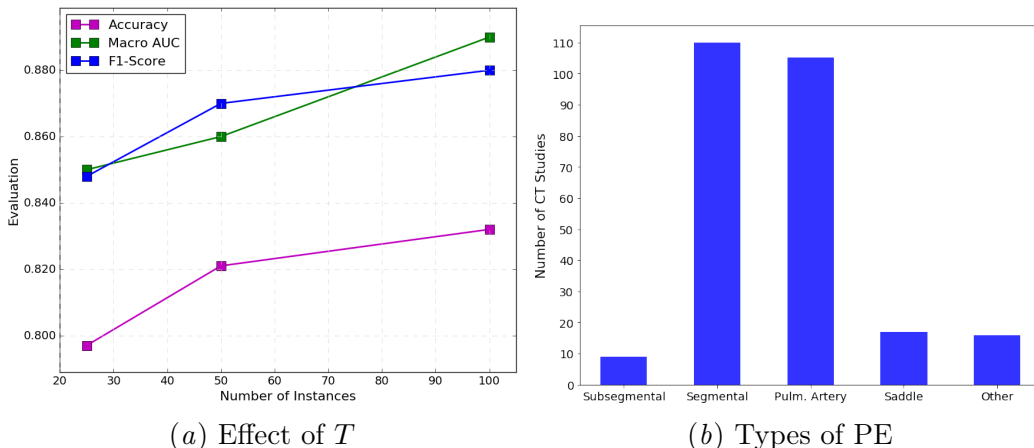(a) Effect of T (b) Types of PE

Figure 5: Performance characteristics - (a) Increasing number of instances (T) from 25 to 100 steadily improves model performance, (b) Histogram of various types of PE.

from the masked volume to invoke `Stage 2` and studied its effect on the performance. We found that increasing T expectedly improved the classifier performance as shown in Figure 5(a).

- **Study 2 - Feature Extraction and Aggregation**: We studied the effect of using LSTM for feature extraction by training a model with Conv-LSTM (CL) layer + Self-Attention (SA) and compared it to using only Conv. (C) + Self-Attention (SA). As expected, the Conv-LSTM model appears to extract more representative features from the slices, compared to treating each of the slices to be independent, as seen in Table 3. A similar empirical analysis on the choice of feature aggregation strategy was carried out. Surprisingly, using *max* pooling achieved the best performance when compared to even the self-attention module with learnable parameters. This is likely due to the fact that the LSTM already captures dependencies between instances in the bag, thus not requiring a dedicated attention module.

- **Study 3 - Loss Functions**: We also observed that using Focal (F) loss, with $\gamma = 2$ in Equation 3, significantly boosts the detection performance by countering the inherent imbalance in the dataset as opposed to using the Binary Cross-Entropy (B) loss.

4.3. Test Performance - Variations across PE Types

Our dataset contains several kinds of PE with varying levels of severity, a distribution of which is shown in Figure 5(b). We report performance of our pipeline on low severity types such as subsegmental and segmental PE, as well as high severity types, namely saddle and main pulmonary artery shown in Figure 7(a). As expected, our pipeline picks up evidence for high severity PE more easily by achieving an AUC score of 0.85, while obtaining an AUC of 0.70 in detecting low severity PE that are harder to find. When compared to the PENet model Huang et al. (2019), our approach achieves improved test accuracies on a dataset characterized by larger amounts of variability, while using a significantly reduced
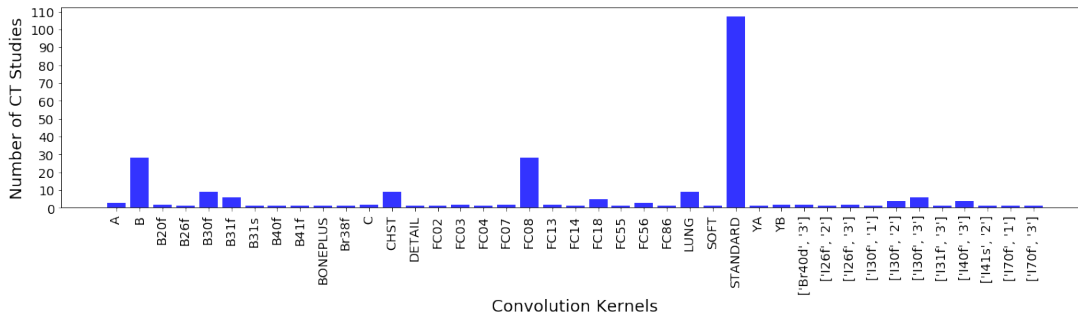
Figure 6: Test set distribution – convolution kernels used for reconstructing the CT volumes.

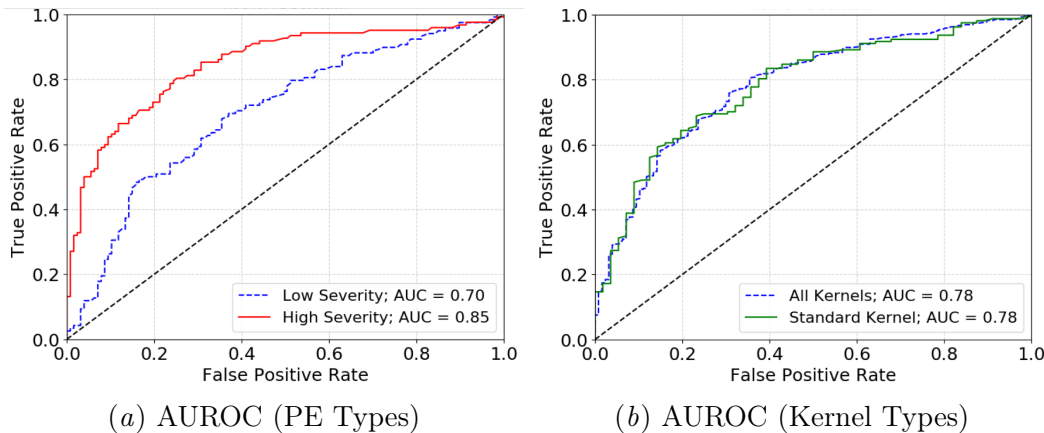(a) AUROC (PE Types)   (b) AUROC (Kernel Types)

Figure 7: Test set performance – (a) AUROC when dataset includes all types of PE, versus only mild PE, (b) AUROC when dataset includes CT studies with diverse convolution kernels versus cases reconstructed using only the GE standard kernel.

number of parameters. Note that PENet has $28,398,705$ parameters, while our model only has $3,168,116$ parameters where `Stage 1` has $1,966,450$ and `Stage 2` has $1,201,666$.

4.4. Test Performance - Variations across CT Convolution Kernels

In addition, our dataset is comprised of CT images reconstructed using different convolutional kernels, whose choice typically controls the image resolution and noise-levels. Figure 6 shows the distribution of kernels for our dataset, where despite most cases using the 'GE Standard' kernel, the dataset includes volumes reconstructed using a wide variety of other kernels. From Figure 7(b), we find that our pipeline is robust to variations in kernels by consistently achieving an AUC of 0.78 on all cases.

5. Relation to Existing Work

In medical imaging applications, commonly deployed disease detection algorithms often involve multi-stage pipelines comprising both segmentation and classification models Ardila et al. (2019). An early work on PE detection used custom feature extraction based on

hierarchical anatomical segmentation of organs such as vessels, pulmonary artery, and aorta Bouma et al. (2009). Though it appears natural to directly build a classifier model on the 3D volumes, in practice, algorithms that first identify semantically meaningful candidate regions, and subsequently extract discriminative features from those regions to perform detection are found to be more effective. These methods are inspired by the success of such two-stage methods in object detection, examples include region-based RCNN Girshick (2015). However, it is important to note that, adapting those techniques to problems in medical imaging have proven to be less trivial mainly for two reasons. One, these solutions require large datasets with ground truth in the form of bounding boxes that characterize regions of interest (ROIs) or dense annotations, which are usually harder to obtain in the clinical domain. Second, the models need to be capable of handling the heavy imbalance between number of positive cases against the more prevalent negative ROIs. Consequently, weakly supervised approaches have gained research interest. Methods that leverage information ranging from single-pixel labels Anirudh et al. (2016) to approximate segmentation labels from class activation maps have been proposed Zhou et al. (2018). However, in the context of PE detection, most existing methods have relied exclusively on supervised learning with dense annotations, and the state-of-the-art solutions such as PENet Huang et al. (2019) utilize transfer learning from pre-trained models for effective detection.

6. Conclusion

In this work, we present a generalizable two-stage pipeline for detecting pulmonary embolisms (PE) observed in 3D CT images. The pipeline comprises of a context-augmented U-Net model to generate segmentation masks, and a convolutional LSTM based classifier used in a MIL setting to detect PE. The proposed approach achieves state-of-the-art results on a challenging real-world dataset while alleviating need for dense annotations of CTs and using models with substantially lower number of parameters compared to prior art. We achieve AUC scores of 0.94 on the validation set and 0.85 on a test set of high-severity PE. Further, our insights from the rigorous ablation studies conducted provide guidelines for designing effective disease detection pipelines.

Acknowledgments

References

Rushil Anirudh, Jayaraman J Thiagarajan, Timo Bremer, and Hyojin Kim. Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 978532. International Society for Optics and Photonics, 2016.

Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end

lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954, 2019.

Michele G Beckman, W Craig Hooper, Sara E Critchley, and Thomas L Ortel. Venous thromboembolism: a public health concern. *American journal of preventive medicine*, 38 (4):S495–S501, 2010.

Henri Bouma, Jeroen J Sonnemans, Anna Vilanova, and Frans A Gerritsen. Automatic detection of pulmonary embolism in CTA images. *IEEE transactions on medical imaging*, 28(8):1223–1230, 2009.

Nathaniel Braman, David Beymer, and Ehsan Dehghan. Disease detection in weakly annotated volumetric medical images using a convolutional LSTM network. *arXiv preprint arXiv:1812.01087*, 2018.

Unidad Central de Radiodiagnóstico and the M+Visión consortium. ISBI 2013 CAD-PE challenge dataset, 2013. URL http://www.cad-pe.org.

Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Yufan Guo, Deepika Kakrania, Tyler Baldwin, and Tanveer Syeda-Mahmood. Efficient clinical concept extraction in electronic medical records. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn L. Ball, Norah Borus, Andrew Huang, Bhavik N. Patel, Pranav Rajpurkar, Jeremy Irvin, Jared Dunnmon, Andrew Y. Ng, and Matthew P. Lungren. PENet - a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging, 2019. URL https://ssrn.com/abstract=3384889.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

Jianming Liang and Jinbo Bi. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 630–641. Springer, 2007.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 62–69. Springer, 2015.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5): 1299–1312, 2016.

SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017.