# Cross-modal weak supervision enables abnormality localization in full-body PET-CT

Sabri Eyuboglu*[1], Geoffrey Angus*[1], Bhavik Patel[2], Anuj Pareek[2], Guido Davidzon[2], Jared Dunnmon,[1] Matthew P. Lungren [2]

[1] Department of Computer Science, Stanford University, Stanford, CA 94305, USA

[2] Department of Radiology, Stanford University, Stanford, CA 94305, USA

* Equal contribution;

**Computational decision support systems could provide clinical value in full-body PET-CT workflows. However, the lack of labeled data and the sheer size of PET-CT studies make it challenging to apply existing supervised machine learning systems to these full-body scans. Leveraging recent advancements in natural language processing, we introduce a weak-supervision framework that extracts imperfect, yet highly granular regional abnormality training labels from the radiologist reports associated with each scan. Using these labels, we train a multi-task 3D convolutional neural network to detect abnormalities in 26 anatomical regions commonly of interest in PET-CT protocols. Our model can detect lymph node, lung and liver abnormalities with median areas under the curve of 87%, 85% and 92%, respectively. We show that performing multi-task pre-training on these core tasks can enable strong performance on rare abnormalities. Multi-task pre-training enables improvements of 28 and 17 AUC percentage-points in detecting abnormalities in the kidney and pancreas, respectively. Moreover, we show that a pre-trained model can learn to predict mortality within 90 days, a critical task for palliative care teams (AUC=79%). Our work introduces a new approach to weak supervision in medical imaging that enables the first deep learning model capable of making anatomically resolved abnormality predictions in full-body PET-CT scans.**

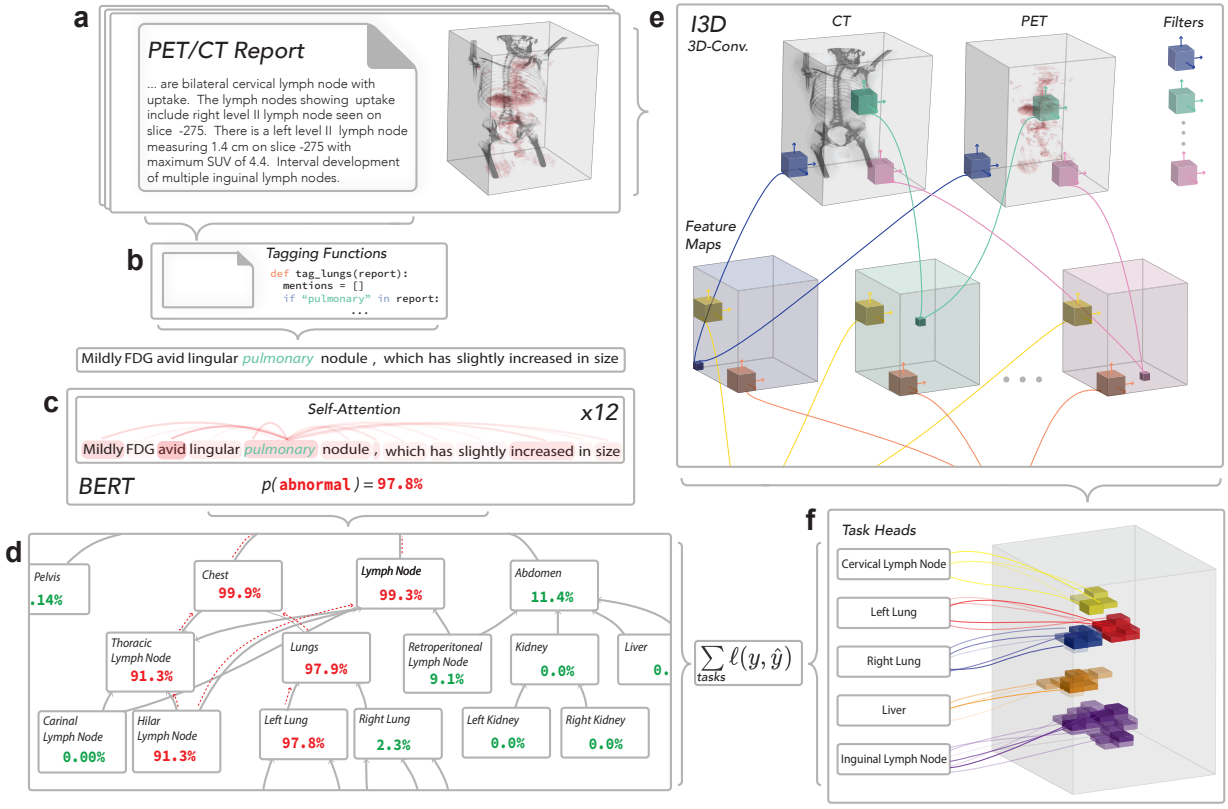


**Figure 1:** **(a)** Our dataset of $8,200$ PET-CT examinations. Each exam consists of (1) a 3D scan of 250 axial slices (each $224 \times 224$) and (2) a natural language, unstructured report written by the interpreting radiologist at the time of the study. Critically, there are no structured, ground truth labels for metabolic abnormalities in anatomical regions. **(b)** Tagging functions powered by regular expressions extract sentences that mention anatomical regions. **(c)** A language model predicts whether their is a metabolic abnormality in the tagged anatomical region. **(d)** Directed-acyclic graph of anatomical regions. **(e)** Schematic illustration of a 3D-convolutional neural network. **(f)** Each Task head uses an attention module to extract from the encoded scan the voxels most relevant to the anatomical region it is charged with.

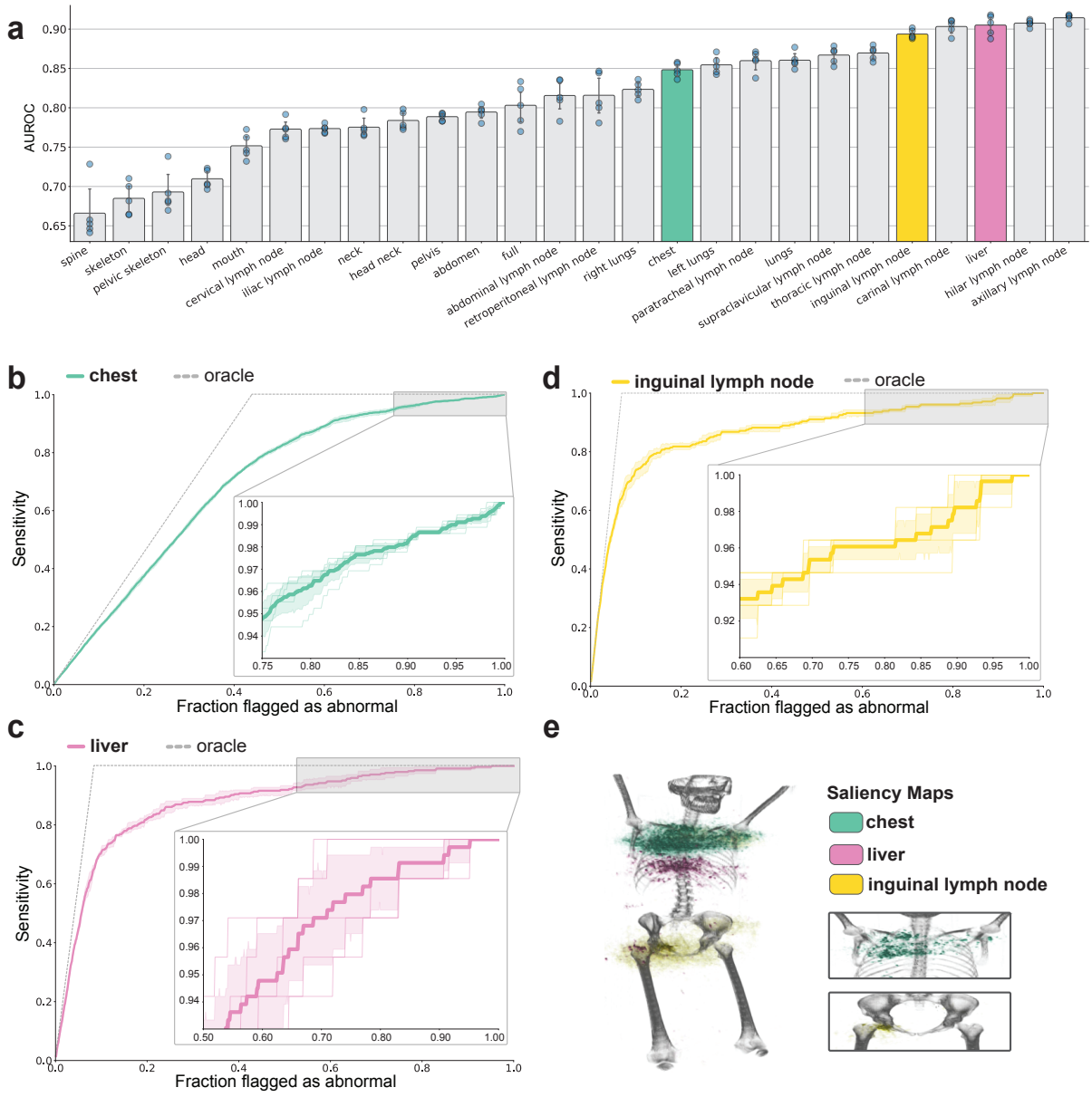


**Figure 2:** **(a)** Our model's predictive performance across anatomical regions. Each bar indicates the model's AUROC for detecting abnormal metabolic activity in a particular anatomical region. Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ samples from five random initializations. The individual AUROC result for each initialization is shown as a blue dot. Anatomical regions are sorted in order of increasing AUROC. **(b-d)** Sensitivity curve for abnormality detection in the **(b)** chest, **(c)** liver, and **(d)** inguinal lymph nodes. Each point on the curve indicates the sensitivity of our model at a prediction threshold where $x\%$ of the exams in our test set are flagged as abnormal. The sensitivity curve of a perfect detector is shown in grey. The shaded region represents confidence intervals (95%) computed using bootstrapping with $n = 1{,}000$ samples from five random initializations. The sensitivity curves for each of those initializations are shown in light blue. Sensitivity curves illustrate the utility of the model in a potential screening application. With our model, clinicians could ignore around 15% of exams while maintaining 99% sensitivity in liver abnormality screening. **(e)** 3D saliency map for abnormality detection in the chest (green), liver (pink), and inguinal lymph nodes (yellow). Colored volumes indicate regions where small perturbations to the input scan most effect the model's prediction for chest, liver, and inguinal lymph nodes.

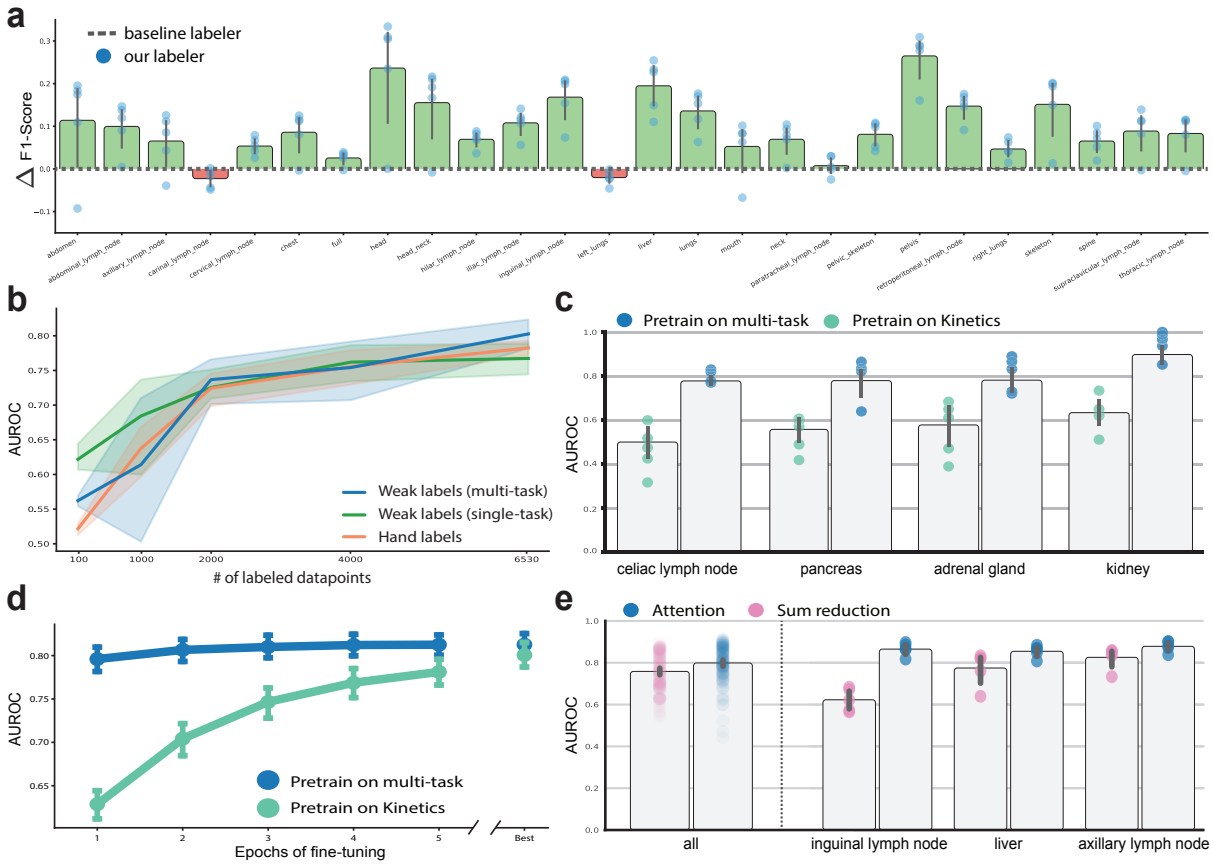


**Figure 3:** **(a)** Our labeling framework outperforms a regular expression baseline on 24 of the 26 regions specified in our test set. Shown is the average difference in F1 score of the between our model (blue) and the baseline (dotted line). **(b)** Model AUROC vs. the number of training data points. The performance of our weakly supervised model is statistically equivalent to that of a traditionally supervised baseline model trained on summary codes on the task of full body abnormality detection (DeLong $p > 0.33$). The shape of the curve suggests more data would lead to a continued increase in performance. **(c)** Multi-task learning improves performance on new, difficult tasks. We fine-tune our multi-task PET-CT to detect abnormalities in four unseen regions of the body: celiac lymph nodes, the pancreas, the adrenal gland, and the kidneys. For each of these regions, less than 5% of the exams have abnormal labels, leading to performance issues associated with large class imbalance. The plot shows that our multi-task learning approach mitigates some of these issues, outperforming single-task models trained solely on the new tasks. **(d)** Multi-task learning reduces training complexity. We fine-tune our multi-task PET-CT model in a single-task for each of the twenty-six core anatomical region. We do the same with a model trained on Kinetics [1], an out-of-domain dataset. The plot shows how mean AUROC across all anatomical regions improves with more epochs of fine-tuning. To the right, we show the best meaan AUROC after training to convergence. Confidence intervals (95%) were determined using bootstrapping with $n = 1{,}000$ samples from five random initializations. **(e)** The attention module incorporated into each task head leads to an improvement over a simple mean reduction.

# Introduction

Functional positron emission tomography (PET) computed tomography (CT) imaging has transformed cancer imaging. PET-CT imaging has enabled the detection of disease in the absence of clear-cut CT abnormalities, and, equally important, the presence or absence of metabolic activity in cases with ambiguous CT findings [2]. For example, PET-CT is more accurate than either PET or CT alone, demonstrating equal sensitivity and a better specificity even in the deep nodal regions of the abdomen and the mediastinum [3]. As a result of these improved capabilities and the clinical value of these imaging studies, over the last five years there has been a 16-fold increase in PET-CT examinations within the United States; unfortunately, however, this increase in imaging demand has coincided with a reduction in both imaging reimbursements and radiologist specialist availability [4]. It is estimated there will be a national shortage of radiologists in the U.S. of more than 5 thousand positions in the next five years while in the U.K. the situation is even more immediate, as 99 percent of NHS radiology departments currently report they were unable to meet their prescribed reading threshold while more than 230,000 NHS patients have had to wait a month or more for their imaging results.

Recent advancements in deep learning offer a path towards mitigating this projected imbalance between supply and demand in PET-CT interpretation. These techniques have demonstrated the potential to provide automated support to a wide variety of medical tasks, including detection and classification of arrhythmias, chest x-ray pathology, musculoskeletal disease on MRI, and lymph node metastases [5–9].

While deep learning models could provide clinical value within PET-CT interpretation workflows, this potential remains untapped to date due to a lack of labeled datasets of sufficient size and schema granularity. The availability of large labeled datasets has been of critical importance to the progress of machine learning in medical imaging, and datasets that support existing studies often contain hundreds of thousands of examples. However, given the relative infrequency of the diagnostic and the requirement for highly-trained clinicians to provide useful labels, it is challenging to curate large labeled PET-CT datasets at such scale. Furthermore, even if such datasets were to become available, common changes in the underlying data distribution (scanner type, imaging protocol, postprocessing techniques, patient population, clinical classification schema, etc.) could rapidly cause the models these data support to become obsolete. Thus, we need to not only develop

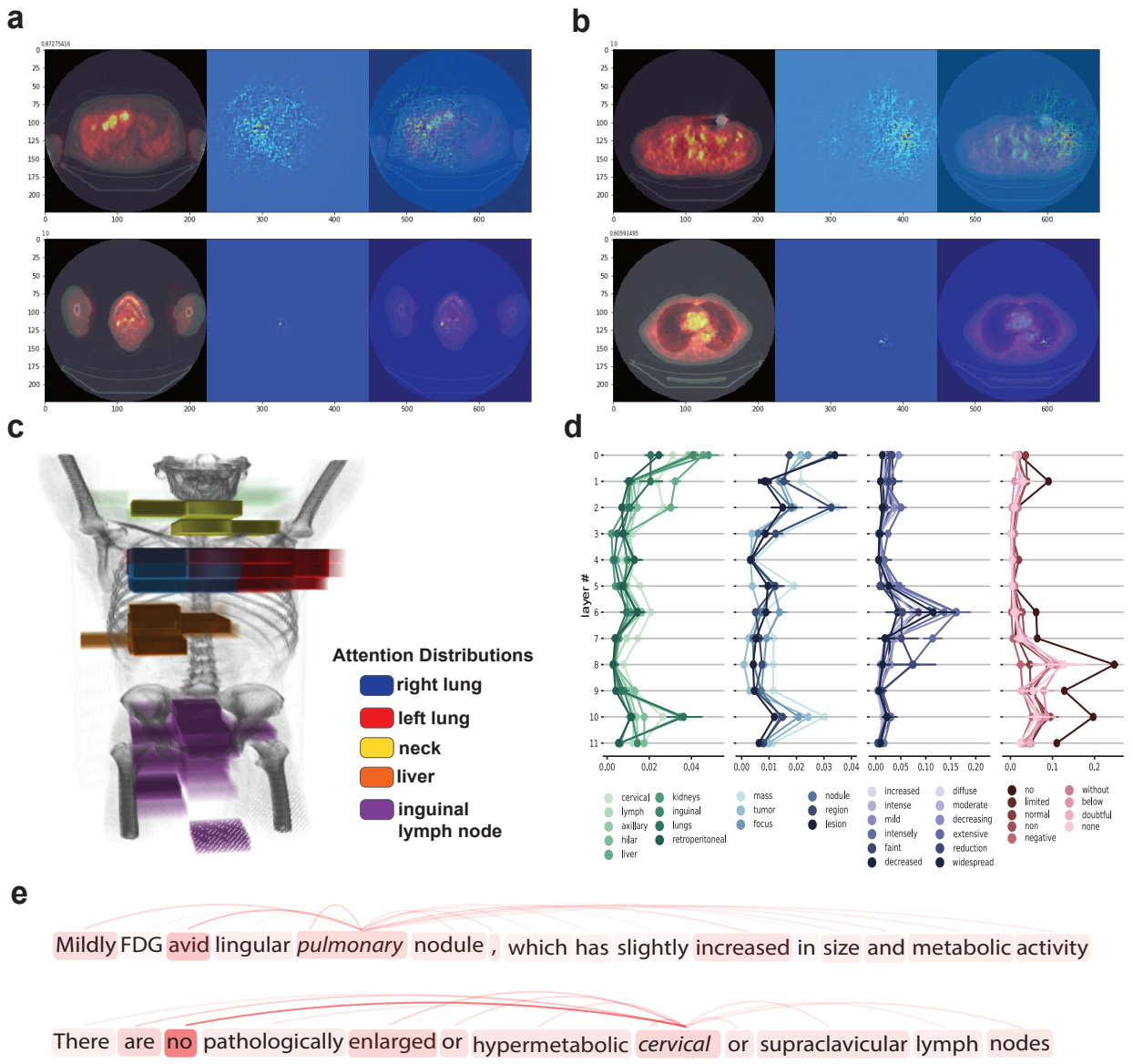


**Figure 4:** **(a)** Two-dimensional saliency maps demonstrating visual examples of predictions made by the model. Saliency maps are produced by computing the gradient of the scan model prediction with respect to the input scan. Pixels with high absolute gradient magnitude are indicated in red. Example of accurate identification and localization of multiple hypermetabolic colorectal liver metastases (top left panel) with saliency map (top middle panel) and overlap of saliency map and image (top right panel). Similarly the left level I cervical hypermetabolic lymph node (bottom left panel) is correctly localized by the saliency map (bottom middle panel) confirmed by overlap of the saliency map with the clinical image (bottom right panel) **(b)** Example of false negative prediction of a left hypermetabolic axillary lymph node in a patient with metastatic breast cancer and a left chest wall metallic port catheter (top left panel) that, while the saliency maps do localize to the area of the abnormality, the threshold for abnormal detection is not reached (top middle and right panel) which is felt due, in part, to the overlying streak artifact from the port. A false positive localization by the model is shown in the left posterior lower lobe of the lung where no clinical abnormality is seen (bottom left panel) despite the fact that the saliency map localized strongly to that area (bottom middle and right panels). **(c)** The soft-attention mechanism in region-specific decoder modules learn to attend to the appropriate part of the body. The opacity of each voxel is proportional to the attention score output by the soft-attention mechanism. **(d)** Layer-wise attention distribution in the report model for four semantic groups. Each point indicates the mean attention that the tagged token applied to the word on layer $k$. Error bars represent $95\%$ confidence intervals determined using bootstrapping with $n = 1{,}000$ samples from all mentions of the word in our training set. From left to right the semantic groups are (1) anatomical regions, (2) physical features, (3) qualifiers and (4) negative constructions. **(e)** A visualization of the attention distributions in the report model for the mentions "pulmonary" (top) and "cervical" (bottom). Opacity of lines are proportional to mean attention across all twelve transformer layers.

effective deep learning modeling techniques for PET-CT, but also establish reliable, repeatable, and robust mechanisms for rapidly creating datasets to support training new models as they are needed in practice. Compared to other imaging modalities, however, whole body PET-CT imaging exams are both up to three times larger in terms of pixel volume and contain two separate series with different types of information: functional (PET) and anatomical (CT), each with different resolutions and fundamental physics. Additional challenges include significant signal-to-noise imbalance amongst different anatomic areas, a wide variety of disease processes, and highly variable feature patterns.

The purpose of this study is to combine techniques from natural language processing (NLP) and computer vision with recent advances in multi-task learning to enable scalable end-to-end prioritization of PET-CT examinations via anatomically resolved discrimination between normal activity and pathology. In this work, we demonstrate that the leveraging of large collections of unlabeled image-report pairs, modern techniques from Natural Language Processing (NLP) and computer vision, and recent advances in multi-task learning can enable scalable end-to-end prioritization of PET-CT examinations via anatomically resolved discrimination between normal activity and pathology. We first curate a large dataset of over 8,200 PET-CT scans and associated radiology reports to develop an NLP approach for automatically labeling abnormalities in different anatomical regions using only the radiology report. After training this NLP model on a small amount of hand-labeled data, we combine it with a hierarchical ontology of disease locations we have developed to create weaker, noisier labels for the presence of pathology in each of 26 different anatomical regions. Our NLP labeling model achieves an accuracy of 86% compared to human labelers. We use these generated multi-task labels to train a 3-D multi-modal Convolutional Neural Network (CNN) for PET-CT scans with a spatial attention layer for each task that ensures that the prediction for each anatomical area is able to learn the relevant spatial locations to which to attend.

In a similar vein, deep learning also has the potential to accelerate the workflow for interpretation of imaging multi-modal volumetric examinations such as PET-CT; for example to manage the priority of unread cases in the imaging backlog by automatically identifying clinically urgent cases, based on a quick analysis of the incoming images, and ensuring they are elevated to the top of the reading queue for immediate interpretation. In this way increased efficiencies could be achieved through triage and automated localization of abnormalities and improve radiologist

workflows.

# Results

## Modeling Approach

Our model is a weakly supervised, multi-task CNN for PET-CT trained to detect the presence of an abnormality in each of 26 anatomical regions commonly of interest in PET-CT protocols. Our model consists of a common encoder that supports 26 binary classification task heads, one for each region. Each of these task heads is composed of a task-specific soft attention mechanism and a linear classifier. The model is initially trained on all tasks jointly. Then, single-task instances of the model are fine-tuned for each task. We train this model using labels determined by our proposed NLP-based label generation framework, which ingests unstructured radiologist report data and outputs probabilistic labels describing the abnormality of predefined regions in every PET-CT scan. Our labeling framework is underpinned by a text classification model trained to predict whether or not the tagged anatomical region in any given sentence is abnormal. Combined with a highly granular entity tagging ontology, this approach allows us to train a BERT-based NLP model to extract anatomically resolved abnormality labels with only a small amount of training data. We evaluate our modeling and supervision approaches through a series of experiments meant to examine three critical aspects of the study: (1) the absolute model performance on the task of abnormality localization and its efficacy as a tool for clinical imaging triage, (2) the utility of our multi-task learning approach for this task, and (3) the viability of our weak supervision technique as an approach to train multi-task abnormality detection models with little to no hand labeled data.

## Multi-task Weak Supervision Yields Clinically Impactful PET-CT Abnormality Detection Performance

In Fig. 2a, we report the overall performance of our weakly supervised PET-CT models on the task of PET abnormality detection for 26 anatomical regions. These results come from models supervised with labels predicted by our NLP-based label generation framework alone — they use no hand labeled image data. The models were trained in two stages using an approach manner commonly seen in multi-task learning settings. First, a single, multi-task PET-CT model was learned — its task heads, each assigned to detect abnormalities in a single region, produced predictions and updated their weights in tandem. Then, each of the task heads was fine-tuned in a single task setting. 22 of the 26 tasks achieve a mean AUROC $> 0.75$ over five different random seeds and 10

tasks achieve a mean AUROC $> 0.85$, many of which are defined over critical anatomical regions, such as the lungs, the liver, and the thoracic, inguinal, and carinal lymph nodes. On the aggregate task of abnormality detection in the full body, we achieve a median AUROC of 0.803, which is not statistically different than the median baseline model trained using a large hand-labeled dataset (DeLong $p > 0.3$), a result further analyzed in Fig. 3b. Confidence intervals (95%) in these and all subsequent experiments were determined using bootstrapping with $n = 1000$ samples from five random seed initializations. .

In Figs 2b-d, we further report the results that follow from analysis of sensitivity (true positive rate) in three clinically important regions: the chest, the liver, and the inguinal lymph node. One of the primary purposes of this study is to build an effective triage tool for workflow prioritization in PET-CT screening— in order to evaluate model efficacy, we frame the test dataset as a worklist, which we sort using the predicted probability of abnormality output by our model. We compare this sorting metric to a random sort baseline. When we sort the predictions by decreasing predicted abnormality, the model achieves a sensitivity of 0.82, 0.93, and 0.92 in the top 50% of the exams for the chest, the liver, and the inguinal lymph node, respectively. A random sort baseline – which is similar to what is done in current clinical practice, where images are often read in the order they are received – achieves a sensitivity of 0.50.

We evaluate the ability of our model to generalize into other clinically relevant domains, specifically mortality prediction given PET-CT imaging data. We frame mortality prediction as a binary classification task, where the model classifies whether or not the patient's date of death is within $x$ days of the study given only the PET-CT scan as input. Our experiments set $x$ to be $\{45, 90, 180, 365\}$ days. Our mortality prediction model is a model pretrained on abnormality detection in the 26 anatomical regions and fine-tuned on a small patient mortality dataset that pairs each study with a date of death. Our model achieves a mean AUROC of 0.799 when predicting patients whose date of death is within 90 days after the time of study. Fig. **??** showcases model performance for each scenario.

In order to quantify the performance gained through use of our model, we compare the aforementioned model to a baseline model pretrained solely on the Kinetics dataset. Fig. **??** compares the AUROC of our model against that of the baseline model for each of the 4 mortality thresholds. Our model outperforms the baseline model in each scenario, in particular achieving

a statistically significant result when predicting mortality with a threshold of 90 days (t-test $p < 0.05$).

## Multi-task Formulation Improves Performance, Enables Localization, and Reduces Cost

We next analyze the performance gains associated with our multi-task learning approach by analyzing the utility of the weights learned by the multi-task model. Fig. 3d compares the amount of additional fine-tuning required to achieve convergence on tasks covering each of the 26 anatomical regions when the model is initialized using our pretrained multi-task model weights versus the same model architecture pretrained on the Kinetics activity detection dataset [1]. Our approach leads to reduced training requirements across all 26 anatomical regions. Across all 5 seeds and 26 tasks, 68% of single-task models pre-trained on our multi-task model achieved their best performance within the first 4 epochs. In comparison, only 7% of single-task models pre-trained on Kinetics converged as quickly— 60% required 8 or more epochs to train to convergence.

We additionally fine-tune our multi-task PET-CT model to make predictions on each of four unseen regions of the body: celiac lymph nodes, the pancreas, the adrenal gland, and the kidneys. Each of these regions exhibit large class imbalance (1:20). In Fig. 3c, we see that our weakly supervised multi-task pretraining mitigates some of the performance issues commonly associated with large class imbalance, substantially outperforming the models pretrained on Kinetics. In terms of mean AUROC, we see a 56% improvement (0.502 to 0.780) on the celiac lymph node task, a 40% improvement (0.559 to 0.782) on the pancreas task, a 34% improvement (0.580 to 0.779) on the adrenal gland task, and a 41% improvement (0.635 to 0.899) on the kidney task against baseline models.

Finally, we evaluate the utility of the soft attention module in each of the task heads. We compare a multi-task PET-CT model using soft attention with a near-identical multi-task PET-CT model that uses a naive sum reduction module in place of soft attention. Neither of these models are fine-tuned. In 22 of the 26 tasks, the soft attention model outperformed the sum reduction model. In Fig. 3e, we show the overall performance (far left), as well as three regions in which soft attention improved performance.

## Weak Supervision Reduces Labeling Cost While Maintaining Performance

We evaluate the effectiveness of weak supervision by (1) comparing the performance of weakly supervised models against a traditionally supervised baseline and (2) comparing the performance of our report model-based labeling procedure against a regular expression baseline.

We evaluate the impact of training set size on the performance of both weakly supervised and traditionally supervised models using an identical 3D CNN architecture on the task of full body abnormality detection. We train two weakly supervised models: a single-task model trained exclusively on the full body task, and a multi-task model trained on all 26 regional tasks before being fine-tuned on the full body task. These models are trained using only weakly labeled training data, i.e. no hand-labeled PET-CT scans, built with our labeling procedure using a set of exams disjoint from the test set. We train these two models on PET-CT exams uniformly random sampled from the training data. We compare the two against a traditionally supervised baseline model. Our baseline model was trained using the full body summary codes, a proxy for hand labels that is commonly associated with PET-CT exams. Note that our hand-labeled baseline model is itself a new result, as we mined summary codes generated during years of clinical practice at our institution to create the dataset required to train this model. In order to mitigate performance issues related to class imbalance (1:9) among the summary code labels, we train the baseline model by oversampling normal exams with replacement at a rate of 3 normal exams for every 7 abnormal exams, a rate found via empirical tuning on the validation set. We train these models on training datasets of sizes $\{100, 1000, 2000, 4000, 6530\}$. When trained on all 6530 exams, we find that our weakly supervised, multi-task model achieves a mean AUROC of 0.803 is statistically indistinguishable from the baseline model (0.783 to 0.803), and that it provides a 5 point (0.751 to 0.803) gain over the weakly supervised single-task model. Importantly, the medians of both weakly supervised models are statistically equivalent to the median baseline model (multi-task: DeLong $p > 0.33$, single-task: DeLong $p > 0.12$), and we see that the median multi-task weakly supervised model outperforms the median single-task weakly supervised model (DeLong $p < 0.05$). The performance curve over the size of the training dataset for all three models can be seen in Fig. 3b.

Fig. 3a compares the performance of our labeling procedure with a regular expression baseline. The regular expression baseline is a labeling procedure that, for each exam, scans for mentions of predefined regional terms in the associated radiologist report. If a regional term is detected, the

sentence is scanned for “negation” keywords (e.g. “without”, “no”, “none”). If there are no negation keywords found in the sentence containing the detected regional term, the procedure gives an abnormal label to the region and all of its parents (as defined by the term ontology). If a negation is found, or the entire report is scanned without a mention of the region or any of its children, the procedure gives a normal label to the region. The report model is trained using only 3272 phrases across 332 exams, hand labeled by non-experts. Our report model saw a 28% improvement in mean F1 score (0.522 to 0.666) over the baseline and achieved a mean AUROC of 0.900. At our chosen operating point, our model achieves a mean sensitivity of 0.870 and a mean specificity of 0.822, outperforming the regular expression baseline, which achieved a sensitivity of 0.833 and a specificity of 0.669.

# Discussion

The purpose of this work was to build a model that could accurately identify and localize abnormalities in full-body PET-CT scans. We also aimed to show that abnormality localization in PET-CT could be achieved with only a small number of hand labels, and to outline a set of repeatable techniques to build weakly supervised, anatomically resolved classifiers for high-dimensional volumetric imaging. We found that our models are able to achieve high ROC-AUC values for a large number of pathology prediction tasks, achieving a mean ROC-AUC of $> 0.85$ in critical regions such as the lungs, the liver, and the thoracic lymph nodes. Further, the spatial attention layers improve performance while enabling the model to correctly isolate the set of voxels relevant for each region-specific pathology prediction task. We find that multi-task pretraining allows for substantial increases in performance (up to 56%) on additional, low-resource anatomical prediction tasks compared to pretraining on natural videos, and that models fine-tuned from this multi-task representation converge far more quickly than those pretrained on common activity detection datasets, which represent the closest analog to volumetric imaging in traditional computer vision. We further find that our weakly supervised model, which required only a small amount of hand-labeled NLP data to create, is not statistically different than a model trained using hand-provided clinical summary codes recorded in the course of clinical practice. While automated segmentation and deep learning tasks involving CT of the chest or abdomen have been described, none have been shown for whole body PET-CT imaging. In particular, no previous work has attempted automated detection and localization of abnormalities on PET-CT imaging, which likely relates both to the difficulty of the task and the difficulty of acquiring sufficient amounts of data. For example, there are on average 60 million pixels in a whole body PET-CT examination and often less than 0.01% of pixels represent an important abnormality. When combined with relatively high signal-to-noise ratios (in PET) and the wide variety of possible disease presentations (on CT, PET, or both), these factors make it difficult to build effective automated detection models. Our empirical results suggest that the combination of modeling and supervision strategies we have used to overcome these challenges have yielded algorithms that may be able to provide clinical value. Study level diagnosis may not improve clinical efficiency alone, as false positives may lead to increased time searching for abnormality - instead pathology localization can potentially decrease time for interpretation and highly specific models for normal examinations could operationalize preliminary interpretations,

draft reports, and improve workflow for radiologists and the clinical referring services.

From a technical standpoint, the results presented in this work tie together a variety of recent developments in machine learning practice. First, weak supervision has demonstrated a remarkable capacity to leverage large amounts of unlabeled data along with relatively noisy labels to build models that perform as well as those trained with hand-labeled datasets, but with a dramatic reduction in labeling cost. The advantage of using such a scheme is not only that we can leverage existing resources like text reports to provide labels for our models, but that schema and labeling strategies can be rapidly updated and tested to achieve models that best fit clinical workflows and requirements. The utility of the particular NLP approach we propose here is that it leverages recent advances in self-supervised language modeling [10] to provide a representation that can be adapted to create a reliable sentence-level abnormality detector with a relatively small amount of hand-labeled data. We observe in Fig. 3a that our approach improves on a simple rules-based abnormality detector by an average of 28%, and provides meaningful gains on 24 of 26 anatomical subclasses. Combined with the comprehensive ontology we created to capture different terms used in the PET-CT report, this approach presents a powerful technique for determining which regions of a full-body scan should be labeled as abnormal when performing abnormality detection on volumetric imaging. The high level of accuracy our labeling framework achieves supports our conclusion that models trained using these labels can perform well in practice, and we observe this to be the case in our experiments. Furthermore, we find that comparing our weakly supervised models – where generating training data using our model took several weeks in total – with those using hand-provided summary codes collected over nearly seven person-years yielded statistically equivalent binary triage classifiers. These results suggest that even for complex volumetric imaging analyses, weak supervision can provide a powerful tool for supervising medical imaging classification models.

Our use of weak supervision is the key to leveraging recent advances in multi-task machine learning to build clinically impactful models in this work. If labeling for one task is daunting, labeling for multiple tasks – in our case, 26 – can be practically impossible. However, because our label model is task independent, using our approach we can add new tasks without any additional labeling. Importantly, this also provides additional value in that most of our PET-CT exams are positive from a binary standpoint – it is very rare to order a PET-CT for a normal patient. However,

most patients do not have an abnormality in a given anatomical region. Thus, this framework allows us to recover a large number of negative examples for each task by defining each one more narrowly, rather than being forced to oversample rare examples that are negative for any abnormality.

As a result of our multi-task learning formulation, we are not only able to learn representations for each task in a parameter-efficient way, but perhaps more importantly we are able to make progress towards notion of model interpretability and auditability. Because each task head has a spatial attention layer and each task is defined as detecting pathologies in a given anatomic region, we are able to build a useful common representation of the PET-CT scan, but also assess the degree to which our task heads are using the correct spatial regions to make their classification. As shown in Fig. 2e and Fig. 4c, our models not only achieve high empirical performance on a variety of tasks, but the areas in which their saliency and spatial attention are concentrated are well-aligned with the anatomical region each is meant to analyze. We show this to be the case for chest, liver, and inguinal lymph node tasks, and it remains the case for the remaining 23 tasks. As shown in Fig. 3, we also find that this attention mechanism can provide modest empirical performance improvement over a naive sum reduction, with a 3 point increase in mean AUROC taken over all tasks and all seeds (t-test $p < 0.005$).

Finally, we show that using weakly supervised learning for model pretraining on the PET-CT domain not only decreases the cost of model fine tuning, but also enables models trained on resource-poor tasks to achieve high levels of empirical performance. We see in Fig. 3c that weakly supervised pretraining dramatically increases performance on celiac lymph node, pancreas, adrenal gland, and kidney tasks by up to 56%. We also find that weakly supervised pre-training causes models to converge to these higher levels of performance more quickly than pretraining on a common dataset describing action prediction from video. This indicates that using weak supervision for in-domain pretraining may be a far more useful approach to model pretraining than is pretraining on datasets from other domains such as ImageNet or Kinetics. Examples of specific error and success cases can be found in Fig. 4a-b.

This work has limitations in addition to those associated with a retrospective study. First, we had a small training dataset, and did not investigate the effect of video pre-training as a function of the training data size. As investigated in other medical imaging tasks, we expect that the effect of

pre-training will diminish as the target data size increases. Second, our clinical training data was from a single center, performed on devices from a single vendor using our standard institutional technique for image acquisition, limiting generalizability.

In conclusion, we report that we have developed an approach to training large, multi-task abnormality localization PET-CT models with little to no hand labeled image data. We validate our approach through various analyses of our PET-CT model, demonstrating the efficacy of a labeling framework built on text classification of unstructured reports and the subsequent viability of weak supervision and multi-task learning. Such a model allows us to train clinically useful, anatomically resolved pathology detection models for automated triage and workflow prioritization. Perhaps more importantly, we outline a technique that can be adapted to train models in any medical imaging modality commonly associated with qualitative analysis in the form of unstructured report text. Future work involves deeper analysis of model behavior and the acquisition of a prospective dataset for further model evaluation. We also found that combining modern machine learning techniques from weak supervision, NLP, computer vision, and multi-task learning allows us to train clinically useful, anatomically resolved pathology detection models for PET-CT while using only modest labeling resources — approximately $1,279$ sentences ($< 0.3\%$ of the sentences in our full dataset) labeled for their description of abnormality over the course of only a dozen non-expert person-hours. Such models would not only enable automated, anatomically-resolved triage of PET-CT studies that would provide value to overburdened radiologists, but the supervision techniques we propose would support consistent model refinement and retraining to support the variability and dynamism inherent in clinical practice.

# Methods

**Dataset.** We use a dataset of 8,251 FDG PET-CT exams from 4,749 patients. The exams were administered at Stanford hospital between 2003 and 2010. Each exam includes PET and CT axially-oriented image sequences that span from the upper thigh to base of skull. Each PET-CT exam also includes a summary code and a free-text report written by the interpreting radiologist at the time of the examination. The summary code describes overall patient status and takes on a value of 1, 2, 4, or 9, where 1 indicates no evidence of abnormality and 2, 4, and 9 indicate the presence of at least one abnormality.

PET-CT is a combination of two different imaging modalities. The PET scan, which captures metabolic activity within the body, is composed of a sequence of 128 $\times$ 128 pixel images. The CT scan, which captures anatomical structure, consists of 512 $\times$ 512 pixel images. Figure 1a shows an example PET-CT scan, visualized such that the image frames are stacked vertically into a three-dimensional volume. The length of the image sequences used in this study range from 69 to 307 ($\mu$ =212.39, $\sigma$ = 23.77) images.

A PET-CT report is an unstructured text document containing the clinical context of the patient's exam and the findings of the interpreting radiologist. The report typically consists of 4 sections: (1) the "clinical history" section, which specifies the indication of the study as well as any prior disease and/or treatment, (2) the "procedure" section, which describes the techniques used in administering the exam, (3) the "findings" section, which details clinically significant observations made in each region of the body, and (3) the "impression" section, which serves as a summary of the most significant observations made in the report [11].

We split the 4,749 patients in our dataset into training, validation and test sets via uniform random sampling. We split our dataset by patients, not individual exams, to ensure that two exams from the same patient will not straddle the train-test divide. The validation and test sets were sampled at random with uniform probability to capture the class distribution expected in a clinical setting. **Mortality Dataset.**

**Weak supervision.** Recent work has established the effectiveness of weak supervision: a machine learning paradigm wherein supervised machine learning models are trained with imperfect, yet cheaply generated training labels. With weak supervision we can reduce or even eliminate the need for costly hand labeled data [12–14].

In this study, we work from within the weak supervision paradigm and implement a labeling framework that ingests radiology reports — rich, yet unstructured bodies of text describing the findings of the interpreting radiologist — and outputs probabilistic labels for each anatomical region in the PET-CT scan (e.g. lungs, liver). Our labeling framework leverages (1) a custom ontology of anatomical regions relevant to PET-CT (See Figure 1d), (2) programmatic functions that tag anatomical regions in the reports (See Figure 1b), and (3) a text classification model, which we call the *report model*, that determines whether the tagged regions are described as metabolically abnormal in the report (See Figure 1c). We then use the labels generated from the reports to train a large convolutional neural network, which we call the *scan model*, to predict abnormalities in the full PET-CT scans. Although we depend on reports to generate training labels, at test time the trained scan model can detect abnormalities in PET-CT scans without an accompanying report.

**Regional ontology.** We construct an ontology of $94$ anatomical regions relevant to PET-CT. Anatomical regions include coarse, high-level regions like "chest", "abdomen", and "pelvis" as well as fine-grained regions like "left inguinal lymph node" and "upper lobe of the right lung". Our ontology is a directed-acyclic graph where nodes represent regions and edges connect regions to sub-regions. For example, edges lead from "lungs" to "left lung" and from "thoracic lymph node" to "hilar lymph node". To determine which anatomical regions to include in the ontology, we perform a systematic analysis of region mentions in PET-CT reports. Specifically, we compute $k$-gram counts ($k = 1, 2, 3$) across all reports in our training dataset. Then, if a $k$-gram refers to an anatomical region and appears at least $35$ times in our dataset, we add the anatomical region to the ontology.

**Tagging functions.** Each anatomical region in the ontology is accompanied by a set of tagging functions that search a report for mentions of the region. A tagging function could be a simple regular expression query, or a complex set of rules that capture more elaborate descriptions of the region. Given a report and a region from the ontology, tagging functions allow us to extract a list of sentences that mention that region. In our experiments, we run the tagging functions on the findings and impression sections of each report.

Note that the mention of an anatomical region does not necessarily imply that there is an abnormality in that region. In fact, the majority of mentions in our dataset occur in the context of a neutral finding. Existing approaches depend on words of negation and uncertainty (e.g. "not",

"no", "unlikely") to classify mentions as neutral or negative. However, compared to other modalities, the language in full-body PET-CT reports is quite nuanced, so these approaches produce a large number of false positives (a baseline labeler that uses this approach achieves a sensitivity of $83.3\%$, but a specificity of only $66.9\%$). For example, it requires a nuanced understanding of PET-CT language to know that the sentence "intense physiologic uptake in the cerebral cortex" is describing a neutral finding.

**Report model.** Rather than rely on hard-coded rules to classify mentions, we leverage an expressive language model that can capture the complexity of PET-CT reports. Our language model, which we call the *report model*, is trained to predict whether an anatomical region is mentioned in the context of an abnormal finding. With a trained report model, we can assign a probability of abnormality to each mention returned by the tagging functions.

Recent work in natural language processing has shown that pre-training large-scale language models on lengthy corpora of unlabeled text can enable strong performance on down-stream tasks with relatively little labeled training data [15–18]. One such language model known as BERT (bidirectional encoder representations from transformers) learns word representations by conditioning on context to both the left and the right of the word [15]. Our report model is based on the BERT language model.

The model accepts as input one or more sentences of natural language. As a preprocessing step, we split up the sentences into a list of *wordpieces* [19]. A wordpiece is a sequence of a few characters that make up part or all of a word. The model treats each wordpiece as an indivisible unit. Wordpieces allow us to operate on rare, out-of-vocabulary words. For example, the sentence:

"FDG uptake in subcarinal and contralateral mediastinal lymph nodes."

might be represented with the following wordpieces:

[FDG uptake in sub -carinal and contra -lateral media -stinal lymph nodes .]

Notice that rare, complex words like "subcarinal" and "contralateral" are split into wordpieces while common words like "lymph" and "uptake" are kept intact. Wordpieces are particularly useful for PET-CT reports where prefixes like "sub-" and "hyper-" often play important semantic roles.

BERT is typically used with a vocabulary of $30,000$ wordpiece tokens optimized for generic text (BookCorpus and English Wikipedia) [20]. Because PET-CT reports are filled with highly specialized vocabulary, using BERT's out-of-the-box wordpiece tokens will mean splitting up many important, domain-specific words. To account for specialized PET-CT text, we add an additional $3,000$ wordpiece tokens generated using a data-driven approach. Specifically, we use a greedy algorithm to find the set of $3,000$ wordpieces that minimize the number of tokens required to reconstruct the reports in our dataset [19, 21]. We use Google's SentencePiece library to generate these wordpiece tokens.

We can formalize our report model as a function $\mathcal{G}$ (parameterized by $\theta_{\mathcal{G}}$) that maps a sequence of wordpiece tokens $(x_1, x_2...x_n)$ to an equal-length sequence of hidden representations $(\mathbf{z}_1, \mathbf{z}_2...\mathbf{z}_n)$,

$$\mathcal{G}(x_1, x_2...x_n; \theta_{\mathcal{G}}) = (\mathbf{z}_1, \mathbf{z}_2...\mathbf{z}_n). \tag{1}$$

The report model has a Transformer architecture, which uses self-attention to draw relations between tokens in the input. Because we use an implementation identical to the original, we refer the reader to the original Transformer manuscript for details on the architecture [22].

To predict whether or not a token $x_i$ occurs in the context of an abnormal finding, we pass its hidden representation $\mathbf{z}_i$ through a single fully-connected layer with a sigmoid activation. This gives us a probability

$$P(x_i \text{ is abnormal}|(x_1, x_2...x_n)) = \sigma(\mathbf{w}^T\mathbf{z}_i + b)) \tag{2}$$

where $\mathbf{w}$ is a weight vector and $b$ is a bias term.

**Report model training.** To train our report model, we need a dataset of sentences that mention anatomical regions and are labeled as negative, neutral or positive. Formally, we can train the model with labeled examples $((x_1, x_2...x_n), (y_1, y_2...y_n))$ where $y_i \in \{0, 1, -\}$ takes on a value of "0" if token $x_i$ forms part of a mention that is neutral or negative, "1" if token $x_i$ forms part of a mention that is positive, and "$-$" if $x_i$ is not part of any mention. For example, the sentence

$$\mathbf{x} = [\text{Abnormal FDG uptake in the left lung}]$$

would be labeled

$$\mathbf{y} = [- \; - \; - \; - \; - \; 1 \; 1]$$

because only the words "left" and "lung" are part of the anatomical mention.

During training, we use a loss function that ignores wordpiece tokens that are not part of any anatomical mention (i.e. labeled with "−"). Formally, our optimization objective is

$$\theta_{\mathcal{G}}^{*} = \underset{\theta_{\mathcal{G}}}{\operatorname{argmin}} \sum_{k=1}^{M} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[y_i^{(k)} \neq -] \mathcal{L}(y_i^{(k)}, \hat{y_i}^{(k)}) \tag{3}$$

where $\mathcal{L}$ is cross-entropy loss and $\hat{y_i} = \sigma(\mathbf{w}^T \mathbf{z}_i^{(k)} + b)$, $m$ is the number of examples in the training set, and a superscript $(k)$ is a reference to the $k^{\text{th}}$ training example.

To efficiently generate a mini-dataset of labeled mentions, we implement a lightweight, labeling GUI. Using it, two non-expert annotators (G.A. and S.E.) were able to label the mentions in a sample of 1,229 sentences ($< 0.3\%$ of the sentences in our full dataset) over the course of twelve person-hours. The GUI integrates with Jupyter Notebooks and could be used to easily label data for a different task.

Prior to training, we pre-train our report model with masked language modeling (MLM) and next-sentence prediction (NSP), the two pre-training tasks proposed in the original BERT manuscript [15]. In MLM, we randomly mask wordpiece tokens in the input and task the model with recovering the masked token using just the context around the mask. To do so, we pass each hidden representations $\mathbf{z}_i$ through a MLM task head. In NSP, we feed the model two sentences and task it with predicting whether or not the first sentence preceded the second in the corpus. We take the weights from $\text{BERT}_{\text{base}}$ pre-trained on BookCorpus and English Wikipedia [15]. Then, we perform domain-specific pre-training on our training dataset of PET-CT reports.

We train all of our models with an Adam optimizer and an initial learning rate of $\alpha = 0.0001$ [23]. This is annealed with a StepLR learning rate scheduler with $\gamma = 0.5$ and a step size of $20$ epochs. We do not employ any loss-based regularization. We train the report model with a batch size of $16$. In each epoch, we sample $100$ mentions with replacement from the training set. We train the multi-task model for $85$ epochs. We perform early stopping with AUROC on a validation set of $61$ sentences.

**Generating Labels.** Using a trained report model, we can generate labels for our training and validation datasets. For each report, we run the tagging functions from the region ontology. This gives us a list of sentences that mention anatomical regions of interest. We then tokenize each of those sentences into a sequence of wordpieces $(x_1, x_2 ... x_n)$ and feed them through our report

model. This yields a sequence of predictions ($\hat{y_1}$ , $\hat{y_2}$...$\hat{y_n}$) where $\hat{y_i}$ represents the probability that token $x_i$ occurs in the context of an abnormal finding (i.e. $P(x_i \text{ is abnormal}|(x_1, x_2...x_n))$, see Eq. 2). Because mentions can span multiple tokens, we reduce multiple predictions to a single probability for the whole mention by taking the mean probability across the tokens in the mention.

Note that some anatomical regions may be mentioned more than once in a report and others not at all. To reconcile the predictions made by the report model into a single probability for each anatomical region we leverage the regional ontology. Specifically, by propagating probabilities up the regional ontology we collect for each region $t$ a list of probabilities $(p_1, p_2...p_{n_t})$ for all mentions of $t$ and its children. We then compute the probability that at least one of those mentions occurred in the context of an abnormal finding,

$$P(t \text{ is abnormal}) = 1 - \prod(1 - p_i) \tag{4}$$

The label propagation process is illustrated in Fig. 1d.

After label propagation we are left with a single probability $\bar{y}_t = P(t \text{ is abnormal})$ for each anatomical region $t$ in the ontology. Below, we show how we can use these predictions as probabilistic labels and train a model to detect abnormalities in PET-CT scans.

**Multi-task learning.** A considerable body of research has focused on using multi-task learning to reduce generalization error in computer vision and natural language machine learning models. Multi-task learning has proven particularly useful in settings where labeled training examples are limited [24, 25]. In this work, we leverage the noisy, probabilistic labels generated by the report model to train a *scan model* that maps a full-body PET-CT scan to a probability of abnormality in one or more regions of our ontology. We use a simple multi-task architecture comprised of a shared encoder module $\mathcal{F}$ (parameterized by $\theta_{\mathcal{F}}$), and $T$ region-specific decoders $\{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$ (parameterized by $\{\theta_{\mathcal{D}_1}, \ldots, \theta_{\mathcal{D}_T}\}$) [26]. For each region $t$, the model outputs the probability that there is a metabolic abnormality in that region. We can formalize the prediction for some input scan $\mathbf{X} \in \mathbb{R}^{2\times224\times224\times l}$ and region $t$ as

$$P(t \text{ is abnormal}|\mathbf{X}) = \mathcal{D}_t(\mathcal{F}(\mathbf{X}; \theta_{\mathcal{F}}); \theta_{\mathcal{D}_t}). \tag{5}$$

To train the model, we perform the following optimization

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{k=1}^{M} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\bar{y}_t^{(k)}, \hat{y_t}^{(k)}) \tag{6}$$

where $M$ is the number of samples in our dataset, $\mathcal{L}$ is cross-entropy loss, $\bar{y}_t$ is the probability of abnormality output by the report model (See Eq. 4), and $\hat{y}_t^{(k)}$ is the probability of abnormality output by the scan model (See Eq. 9).

**Scan model.** For our shared encoder module $\mathcal{F}$, we use an Inflated Inception V1 3D CNN (I3D) pre-trained on the Kinetics dataset (with optical flow) [1]. We remove the final classification layer so that the encoder outputs a 3-dimensional encoding of the input scan. The encoding consists of $d = 1024$ channels, each of shape $7 \times 7 \times (\frac{l}{6})$, where $l$ is the number of slices in the original exam. Formally, the encoder module $\mathcal{F}(\mathbf{X}; \theta_{\mathcal{F}})$ outputs a tensor $\mathbf{A} \in \mathbb{R}^{d \times 7 \times 7 \times \frac{l}{6}}$. The encoding $\mathbf{A}$ can be viewed as a volume where each voxel is a vector $\mathbf{a}_{i,j,k} \in \mathbb{R}^d$. We visualize this encoding on the right hand side of Fig. 1f.

Each region-specific decoder $\mathcal{D}_t$ is composed of a soft-attention mechanism and a single linear classification layer. Intuitively, the attention mechanism allows each task head to "focus" on specific regions of the scan. To perform soft-attention by computing the dot product between each voxel $\mathbf{a}_{i,j,k} \in \mathbb{R}^d$ in the encoding and a learned weight vector $\mathbf{w} \in \mathbb{R}^d$

$$s_{i,j,k} = \mathbf{w}^T \mathbf{a}_{i,j,k} \tag{7}$$

yielding a score $s_{i,j,k}$ for each voxel. We apply softmax across the scores $\alpha = \text{Softmax}(s_{i,j,k})$, and use them to compute a linear combination of all the voxels in the scan encoding

$$\mathbf{a} = \sum_{i,j,k} \alpha_{i,j,k} \mathbf{a}_{i,j,k}. \tag{8}$$

Intuitively, the larger the $\alpha_{i,j,k}$, the more attention is paid to the voxel at coordinates $(i, j, k)$. The linear combination $\mathbf{a} \in \mathbb{R}^d$ is then fed to a final linear classification layer with a sigmoid activation. Altogether, each region-specific decoder outputs a single probability of abnormality

$$P(t \text{ is abnormal} | \mathbf{A}) = \mathcal{D}_t(\mathbf{A}; \theta_{\mathcal{D}_t}). \tag{9}$$

**Scan model training.** We train the scan model with an Adam optimizer and an initial learning rate of $\alpha = 0.0001$ [23]. This is annealed with a StepLR learning rate scheduler with $\gamma = 0.5$ and a step size of 16. We do not employ any loss-based regularization. Due to machine constraints, we train the scan model with batch size of only 2. In each epoch, we sample $2,000$ exams with replacement from the training set.

We train a multi-task scan model on the $26$ regions for which there is at least one positive example for every nine negative examples (i.e. fraction of positive examples $\geq 10\%$). The multi-task model is trained for $15$ epochs. We then perform single-task fine-tuning for the $26$ multi-task regions as well as four "rare" regions with class balance less than $10\%$. Single-task fine-tuning is performed for $5$ epochs. Note, that fine-tuning typically converges after $1-2$ epochs, so training for $5$ epochs is not strictly necessary in practice (see Fig. 3d).

We preprocess each PET-CT scan by: (1) upscaling PET images and downscaling CT images to a common resolution of 224 $\times$ 224 pixels, (2) normalizing each sequence so that its pixels have a mean value of 0 and a standard deviation of 1, (3) stacking the now equally-sized PET and CT image sequences to create a two channel image sequence.

During training, we apply two basic data-augmentation transforms to each scan. We randomly crop the image sequence to a 200 $\times$ 200 pixel region, then resize to 224 $\times$ 224 pixels. We additionally jitter the brightness of the image sequence by adjusting brightness throughout the sequence by a factor $\gamma \sim$ Uniform(0.0, 0.25).

**Model evaluation.** Our test set of 800 exams was labeled by four board-certified radiologists (G.D., B.P., A.P., M.L.). Each exam received 30 regional abnormality labels based on the contents of its associated radiologist report. Models were evaluated a single time using this test set. No validation was performed with the test set.

We evaluate the performance of our label generation framework using positive predictive value (precision), sensitivity (recall), F1-score, and area under the ROC curve (AUROC). We use F1 score to make direct performance comparisons against a regular expression baseline.

We evaluate the performance of our PET-CT model using positive predictive value (precision), sensitivity (recall), F1-score, and area under the ROC curve (AUROC).

**Model interpretation.** We present three distinct ways to interpret the predictions of the PET-CT model— the three-dimensional saliency map shown in 2e), the two-dimensional saliency map shown in Fig. 4a-b, and the three-dimensional attention map shown in Fig. 4d. The three-dimensional saliency maps produce a high-level visualization of model behavior. They are generated using Guided Backpropagation [27]. The gradients from each of the task heads are often very different numerical scales. In order to visualize the task gradients jointly, we preprocess the task gradients. Let $\mathbf{X}'_t$ be the gradient of scan $\mathbf{X}$ with respect to the prediction of task head $t$. We first

normalize $\mathbf{X}'_t$ by subtracting the minimum value and dividing by the maximum value. This yields $\tilde{\mathbf{X}}_t^{(i)\prime}$, a tensor with values on the range $[0, 1]$. We manually set thresholds $\beta_1, \ldots, \beta_T$ such that for the scan gradient $\tilde{\mathbf{X}}_t^{(i)\prime}$ of each task head $t$, we create a saliency map $\delta_t^{(i)}$ with the same shape as $\mathbf{X}'_t$:

$$\delta_t^{(i)} = \begin{cases} \tilde{\mathbf{X}}_t^{(i)\prime} & \text{if } \tilde{\mathbf{X}}_t^{(i)\prime} \geq \beta_t \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

The two-dimensional saliency maps (produce a slice-wise interpretation of model behavior for a single task head $t$. We use the aforementioned normalization scheme to compute $\tilde{\mathbf{X}}_t^{(i)\prime}$ and automatically set $\beta_t = \min \tilde{\mathbf{X}}_t^{(i)\prime}$.

We can additionally visualize the attention scores computed per attention head for a single task. The visualization simply maps the scalar values computed in $\alpha$ to their respective voxels in $\mathbf{a}$. An example of an attention visualization can be seen in Fig. 4d.

**Related work.** Weak supervision is a broad term for techniques used to train models without hand labeled data. Distant supervision is one such technique that leverages noisy labels in order to train models for a closely related task. This is a technique commonly seen in NLP tasks due to frequent correlation between easily identifiable tokens and high-level semantic meaning [25, 28]. Ratner et al. build upon the distant supervision paradigm and propose an unsupervised framework that uses generative modeling techniques to denoise labels derived from programmatic, coarse-grain labeling functions [29]. Their framework has been used to achieve state of the art performance on numerous NLP benchmarks, and has additionally shown itself to be useful in the medical imaging domain. Fries et al. effectively classify aortic valve malformations in unlabeled cardiac MRI sequences [12]. Dunnmon et al. identify abnormalities in 2-D chest radiographs (CXR), knee extremity radiographs, 3-D head CT scans (HCT), and electroencephalography (EEG) signals [30].

Our proposed training methodology is related to model distillation, which leverages one model's predictions as soft targets to train a smaller, less memory-intensive classifier [31]. The primary difference between the two methods is in its higher-level objective: our modal is to allow for cross-modal learning (knowledge transfer to achieve a task with different input modalities versus knowledge compression to learn the same task with fewer parameters). The use of trained models to label unlabeled data to train "student" models has been immensely successful in computer vision tasks such as object detection and human keypoint detection [32].

# References


1. Carreira, J., Zisserman, A., Com, Z. & Deepmind, . Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Tech. Rep.

2. El-Galaly, T. C., Gormsen, L. C. & Hutchings, M. Pet/ct for staging; past, present, and future. In *Seminars in nuclear medicine*, vol. 48, 4–16 (2018).

3. Hutchings, M. *et al.* Position emission tomography with or without computed tomography in the primary staging of hodgkin's lymphoma. *Haematologica* **91**, 482–489 (2006).

4. National, C. P. F., National Academies of Sciences, E., Medicine *et al.* Appropriate use of advanced technologies for radiation therapy and surgery in oncology: Workshop summary (2016).

5. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* **25**, 65 (2019).

6. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine* **15**, e1002686 (2018).

7. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031* (2019).

8. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine* **15**, e1002699 (2018).

9. Bejnordi, B. E. *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318**, 2199–2210 (2017).

10. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).

11. Niederkohr, R. D. *et al.* Reporting Guidance for Oncologic 18f-FDG PET/CT Imaging. *Journal of Nuclear Medicine* **54**, 756–761 (2013).

12. Fries, J. A. *et al.* Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature Communications* **10** (2019).

13. Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D. & R, C. Data Programming: Creating Large Training Sets, Quickly. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R. (eds.) *Advances in Neural Information Processing Systems 29*, 3567–3575 (Curran Associates, Inc., 2016).

14. Reed, S. *et al.* Training Deep Neural Networks on Noisy Labels with Bootstrapping. *arXiv:1412.6596 [cs]* (2014). ArXiv: 1412.6596.

15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (2018). ArXiv: 1810.04805.

16. Phang, J., Fvry, T. & Bowman, S. R. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *arXiv:1811.01088 [cs]* (2018). ArXiv: 1811.01088.

17. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training 12.

18. Peters, M. E. *et al.* Deep contextualized word representations. *arXiv:1802.05365 [cs]* (2018). ArXiv: 1802.05365.

19. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]* (2016). ArXiv: 1609.08144.

20. Zhu, Y. *et al.* Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 19–27 (IEEE, Santiago, Chile, 2015).

21. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016).

22. Vaswani, A. *et al.* Attention Is All You Need. Tech. Rep.

23. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2014). ArXiv: 1412.6980.

24. Luo, Y., Tao, D., Geng, B., Xu, C. & Maybank, S. J. Manifold Regularized Multitask Learning for Semi-Supervised Multilabel Image Classification. *IEEE Transactions on Image Processing* **22**, 523–536 (2013).

25. Rei, M. Semi-supervised Multitask Learning for Sequence Labeling. *arXiv:1704.07156 [cs]* (2017). ArXiv: 1704.07156.

26. Caruana, R. Multitask Learning 35.

27. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional Net. *arXiv:1412.6806 [cs]* (2014). ArXiv: 1412.6806.

28. Go, A., Bhayani, R. & Huang, L. Twitter Sentiment Classication using Distant Supervision 6.

29. Ratner, A. *et al.* Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment* **11**, 269–282 (2017). ArXiv: 1711.10160.

30. Dunnmon, J. *et al.* Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *arXiv:1903.11101 [cs, eess, stat]* (2019). ArXiv: 1903.11101.

31. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]* (2015). ArXiv: 1503.02531.

32. Radosavovic, I., Dollar, P., Girshick, R., Gkioxari, G. & He, K. Data Distillation: Towards Omni-Supervised Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4119–4128 (IEEE, Salt Lake City, UT, USA, 2018).