**Aim 2 - Ontologies and Explainability of AI**

The goal of the second aim is to develop new theories and methods to annotate medical images with explanations of findings that are useful to the clinician.   A reading of journals such as *Radiology* shows two types of annotations are currently in use.   Articles about deep learning annotate images with heat maps that describe regions that lead to diagnoses.  Figure XXX shows an example of such an annotation from  [<https://doi.org/10.1148/radiol.2020203511>]  produced by Grad-cam [**DOI:** [10.1109/ICCV.2017.74](https://doi.org/10.1109/ICCV.2017.74)] by performing analysis of the weights and activations of the networks.  Similarly, Figure YYY shows a region of an X-ray from  [DOI: [10.1148/ryai.2020190043](https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1148%2Fryai.2020190043)] identified by LIME  [<https://doi.org/10.1145/2939672.2939778>] , a system capable of identifying important attributes or regions of any classifier.



Figure XXX1: An annotation of a chest X-ray produced by Grad-cam.



Figure XXX2: An annotation of a chest X-ray produced LIME.

In contrast, articles by radiologists with the goal of presenting novel findings, particularly on new diseases, such as COVID-19 or damage from vaping typically annotate images with arrows, circles, and most importantly labels that use a common vocabulary to describe the region of interest.   Figure XXX shows an example of such an annotation from [<https://doi.org/10.1148/radiol.2020200463>]. Annotations use terms such as ground-glass opacities, linear opacities, etc,.

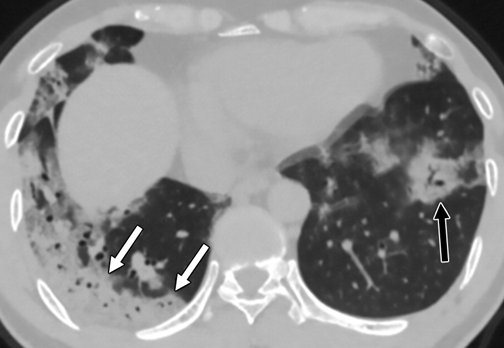


Figure xxX3: *An axial CT image obtained without intravenous contrast in a 42‐year‐old male in the “late” time group (10 days from symptom onset to this CT) shows bilateral consolidative opacities, with a striking peripheral distribution in the right lower lobe (solid arrows), and with a rounded morphology in the left lower lobe (dashed arrow).*

Rather than asking clinicians to learn about deep learning and heat maps, we propose to investigate a new generation of explainable AI methods whose goal is to annotate images exactly as radiologists do to inform other radiologists, clinicians, and interns about findings.  Most importantly, the goal is not only to identify important regions but also to describe what it is about that region that is significant using the vocabulary used by radiologists and clinicians. Indeed, algorithms for explaining deep learning such as LIME and Grad-cam might best be described as doing localization, finding an area of interest but not characterizing the region and why it is of interest in terms familiar to radiologists and clinicians.  Without such a characterization, the area highlighted could just as easily be an explanation for COVID-19 pneumonia or damage from vaping or chemotherapy or bacterial pneumonia. Indeed, most of the recent articles on deep learning analysis have a goal of distinguishing healthy patients from those with a single condition such as COVID-19 without considering the full set of possibilities encountered in a clinical setting.

We propose to use terms and concepts from Radlex [DOI: [10.1007/s10278-007-9051-6](https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1007%2Fs10278-007-9051-6), <https://doi.org/10.1148/ryct.2020200152> ] to annotate training data and learn to identify these features in CT and chest X-rays.  In this manner, we not only identify regions of interest, but also characterize these regions.  It is our ultimate goal to annotate images in a way that is indistinguishable from the annotations of expert radiologists.   In some ways, our goals are similar to those of Testing with Concept Activation Vectors (TVAC) which identify features to explain the output of learners.  However, TCAV produces a global explanation (i.e., in general late state COVID-19 is associated with linear opacities and lung opacities in more than 3 lobes) but does not identify regions in individual images where these features are present.   Our goal is to explain the classification of medical images in terms of these high-level features and to identify where they are located in the image.

**Background: Explainable AI**

There are two main approaches to identifying regions of interest in image analysis using deep learning.

* Model agnostic methods such as LIME manipulate the inputs (e.g., pixels) and measure how changes in the input affect the output.  If a change to an input has no effect on the output, it is not relevant to findings.  If a change has a major impact (e.g., changing the classification from interstitial pneumonia to normal), then the region is important to the classification.
* Other methods examine the activations and weights of the deep network to find regions of importance. Grad-cam and Layerwise Relevance propagation are examples of such methods.

More recently, Pazzani and colleagues have developed methods for explaining contrasting categories (Pazzani, Feghahati, Shelton, & Seitz,  2018)  in deep learning (Krizhevsky, Sutskever,  & Hinton, 2012;  LeCun Bengio, & Hinton, 2015).  These methods find regions important for differential diagnosis, e.g., to distinguish bacterial from viral pneumonia.  Feghahati,  Shelton, Pazzani, & Tang, (2020) show how any of the approaches to  identifying features of importance can be converted to contrasting methods.

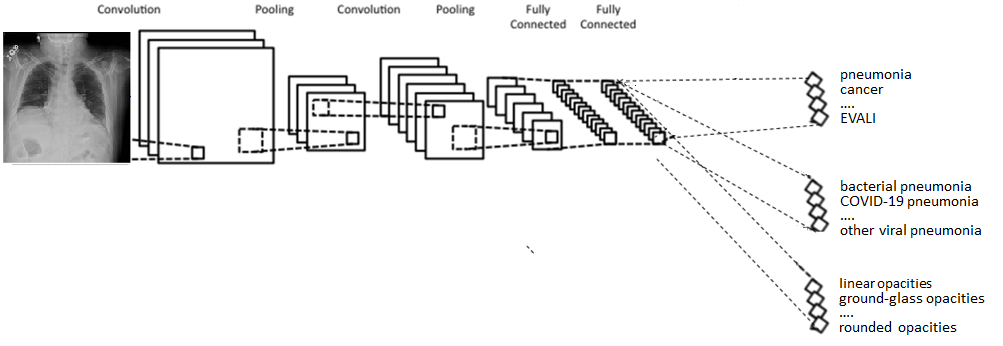
Although not originally developed with explanation in mind, U-nets (doi:10.1007/978-3-319-24574-4\_28) perform the same task as the above explainable AI algorithms, by annotating images with regions of importance.   One key difference is that the training data of U-net includes both an image and a segmentation map, where the segmentation map indicates regions of interest.  The U-net contains the same type of architecture as conventional convolutional neural networks that downsamples an image for categorization, but also contains additional structure that upsamples the image to reconstruct the image and the segmentation.  A system deployed at UCSD Health [a3hsiao@health.ucsd.edu](mailto:a3hsiao@health.ucsd.edu) uses the U-net architecture for

**Proposed Work**

We propose to explore five approaches to describe regions of medical images in terms meaningful to clinicians and radiologists.  The methods differ in how much labeling and segmentation of regions is required in the training data, allowing us to explore trade-offs between expert diagnoses and labels of segments and methods that require less effort but perhaps more data.   For example, an expert radiologist might identify the specific regions that contain features such as “linear opacities” and “ground-glass opacities” as well as a diagnosis, prognosis or categorization.   Alternatively, for other systems, the presence of features may be noted without requiring the radiologist to identify where the regions are.  This option allows for simple text analysis of radiological reports to identify features.  By exploring multiple approaches, we may find that some approaches work better in some situations while others are more suited for different tasks. Furthermore, a combination of approaches may work better than any individual approach. The approaches are described below.

**Mask R-CNN.**  We propose to use Mask R-CNN (He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969). and related algorithms (Cheng, T., Wang, X., Huang, L., & Liu, W. (2020). Boundary-preserving mask R-CNN. In European Conference on Computer Vision (pp. 660-676).) to identify  features of images.   These approaches can produce bounding boxes around the objects and features.  They may be most appropriate for identifying medical devices in images such as endotracheal tubes or pacemakers, but we will evaluate the effectiveness on other features as well.

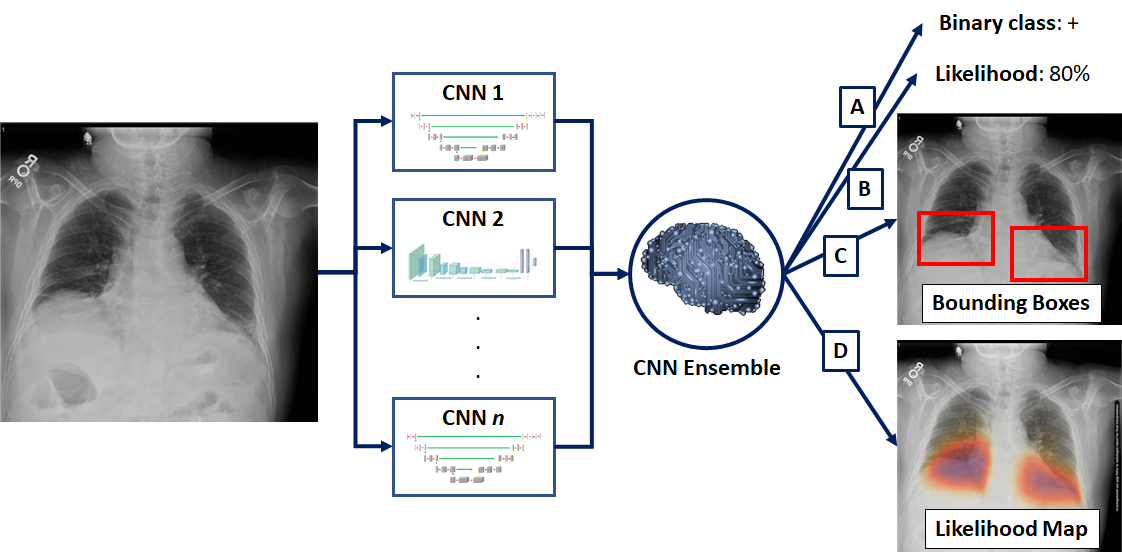
**Multi-task learning with CNN.** Multi-task learning has been shown to improve the accuracy of solving multiple related tasks. We propose to explore multi-task learning with convolutional neural nets in the context of medical images in solving the problem of both a diagnosis or categorization and identifying underlying features as an output.  In this case, the features and diagnosis are both outputs of the network, but it is not required to provide as input an identification of the regions that contain each feature. Rather, one of the XAI algorithms surveyed above may identify output features. Figure XXX2 illustrates a proposed network with groups of output nodes.  For a specific example, positive feedback would be presented on pneumonia, COVID-19 pneumonia, and ground-glass opacities, allowing the trained system to identify images with these features.

****

**Multi-task learning with U-nets.**  U-nets may easily be extended to do multitask learning. In this case, rather creating a segmentation map with a binary classification for each pixel (i.e, relevant to the diagnosis or not relevant), a separate class may be used  for each feature, such as linear opacity or ground-glass opacity.   In this case, the training must not only identify whether a feature is present or absent in the image but also the location of that feature.  Obviously, this requires more effort on labeling the training data than multi-task learning with CNNs. We will explore the trade-off between the accuracy of these two methods and the amount of effort required to prepare data training.  In any event, to evaluate the approaches, we will need some testing data with the location of features identified to evaluate our models.

**Multi-Task Ensemble Learning**

We propose to explore and evaluate techniques for ensemble learning. One of the major challenges and benefits of developing individual component CNNs is that different methods perform best on different problem. In medical imaging and medicine in general, diagnostic tasks are often multifaceted. Diagnosis is often composed of multiple component tasks, each of which the human brain can quickly balance and shift between. We propose to use CNN ensembling as a mechanism of integrating multiple CNNs. Instead of combining multiple CNNs to optimize performance on a single task, as it is more often used, we propose to leverage ensembling to concurrently accomplish multiple related tasks. This concept, when applied to detection of pneumonia, is diagrammed in figure x.

****

**Figure x.** *Strategies for CNN algorithms to assist in imaging diagnosis.* In an example case of a patient with bilateral lower lobe pneumonia, radiologist knowledge can be used direct CNN training at multiple levels of granularity, from provision of (a) whole-image binary class labels, (b) whole-image likelihood of diagnosis, (c) bounding box localization of areas affected with feature labels, or (d) a heat map assignment of areas affected. Most current algorithms rely on binary class labels, though a few public machine learning challenges have begun to incorporate bounding box ground truth labels. An ensemble strategy may be used to combine multiple independent CNNs to weigh the results of component networks and generate each of the output inferences.

**Discovery of New Features.   The ensemble approach described above introduces the possibility of discovering new features.** Whenever possible, we intend to label regions with known feature labels such as those in RadLex. However, we leave open the possibility that novel features without existing names contribute to a classification.   A novel feature may be detected, for example, when a method such as LIME finds a region that is important to a classification but a feature identifier in a method such as multi-task learning does not identify and label the region.   If this is noted in multiple examples, we propose to use the similarity methods described in Section N to group similar regions together to ask an expert radiologist to label the features. This can be evaluated by leaving out a known label from the training to see if the system can identify it.