A Critical Overview of Interpretability and Explanation in Machine Learning

Lizzie Kumar
9/27/19

Agenda

1. Intro

    - **Why is this a hot topic**? What are the motivating problems?

    - **What do people mean** when they talk about “interpretability” and “explainability”?

2. Overview of research

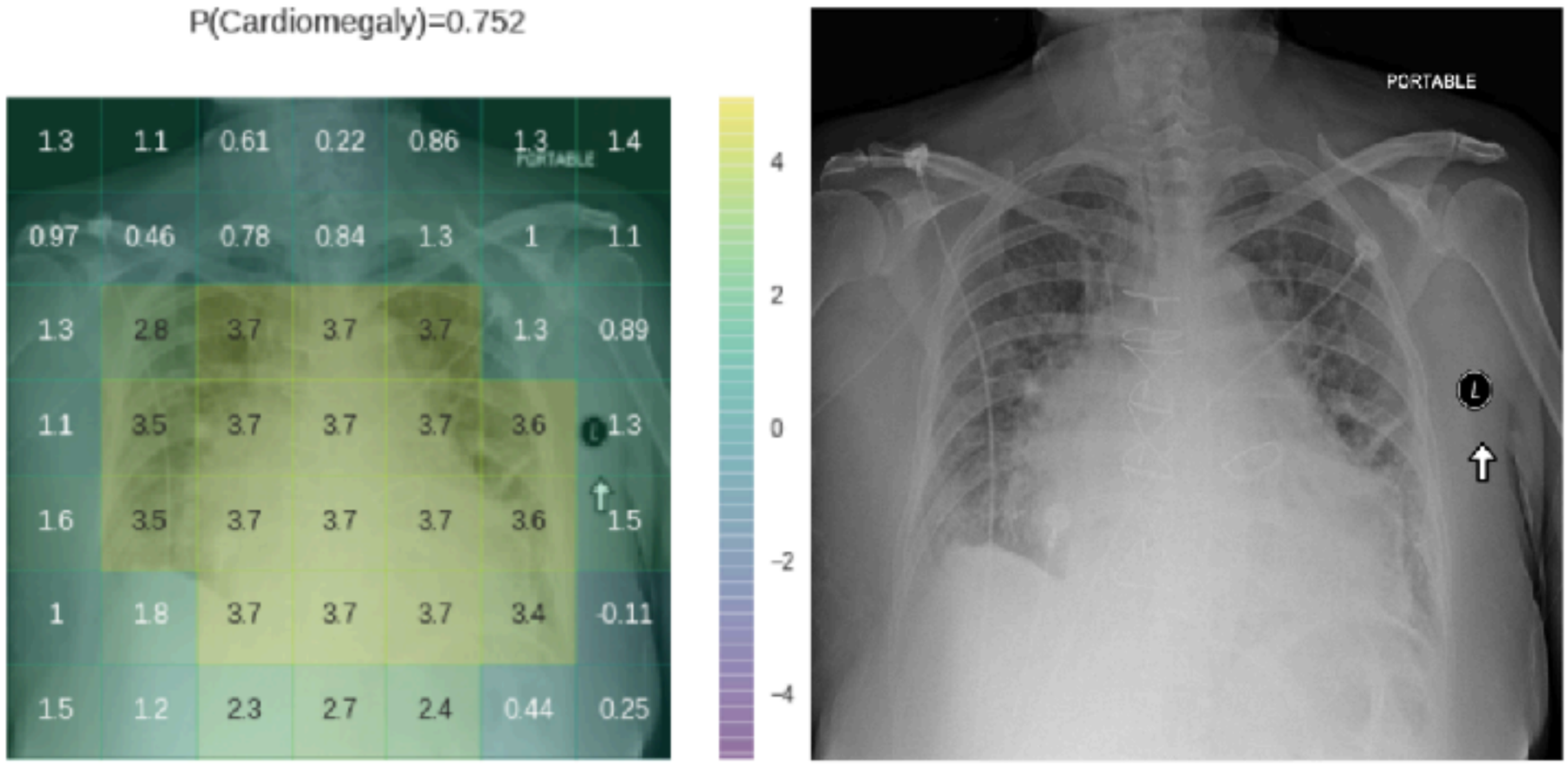
    - **What are some methods** to learn interpretable models or explain uninterpretable models?

3. Hot takes and critical analysis

    - Is the problem at hand fundamentally **misspecified**?

    - Can interpretability and explainability be **misleading**?

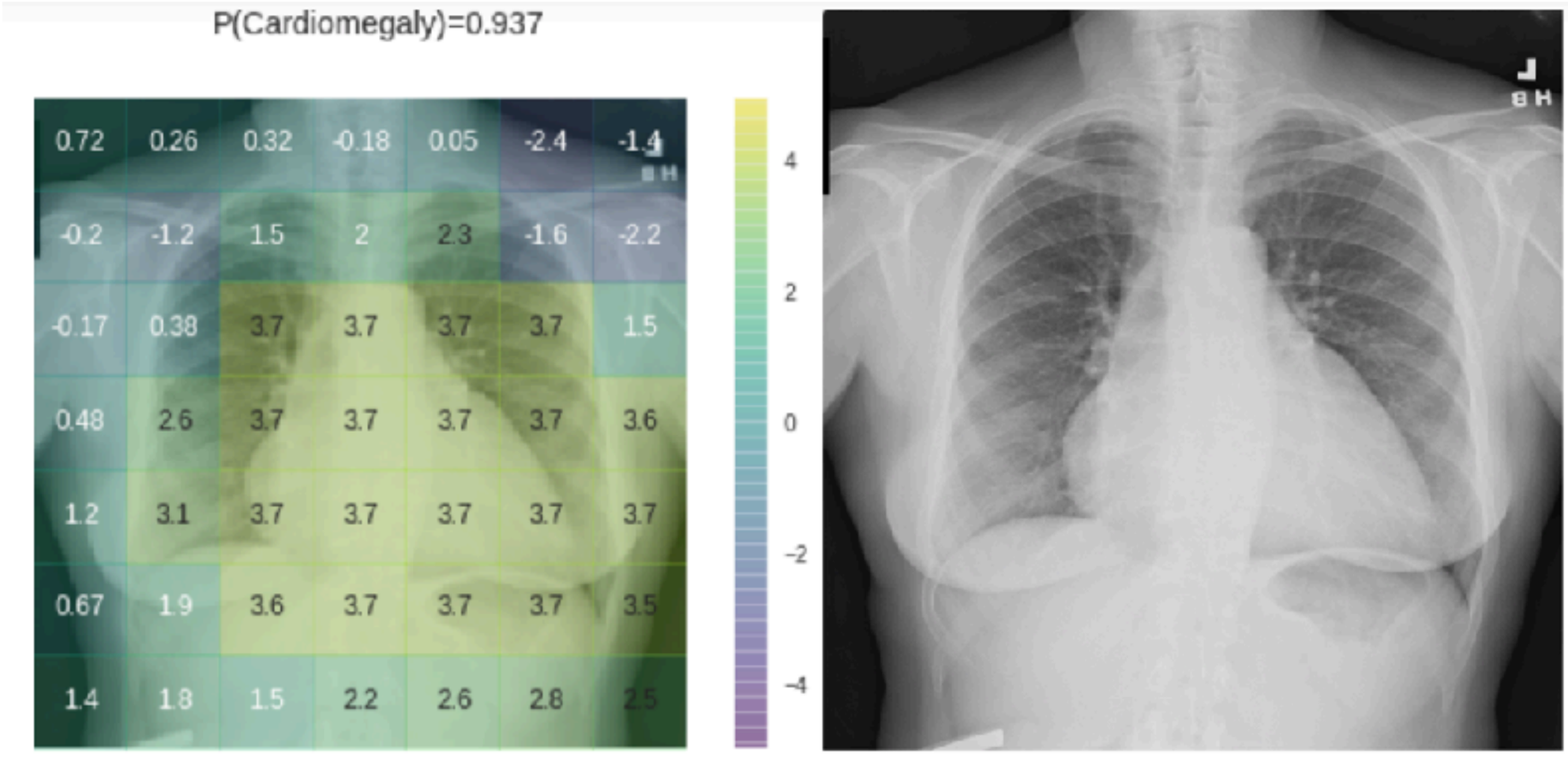
Motivating Example



Patient with cardiomegaly: heatmap of CNN on left, original image on right

In this heatmap, positive values with green/yellow shading indicate regions of the image that contributed to a positive prediction of cardiomegaly. Negative values with blue/purple shading contributed to a negative prediction of cardiomegaly.

Motivating Example



Patient with cardiomegaly: heatmap of CNN on left, original image on right

Again, we see the CNN classifying this one positively because it's looking at the heart — that's impressive! But this time, it downweighted cardiomegaly based on the areas that indicate the scanner on which it was acquired (indicated by the negative, blue squares surrounding the laterality marker in the upper right corner). This is a regular scan, not a portable one.

Premise

- If we are able to understand “how” a model “thinks,” we can weigh in on **whether the logic is reasonable or justified**

Interpretability

- Possible definitions:
    - The degree to which a human can understand the cause of a decision (Miller 2017)
    - The degree to which the impact of each feature on the model's prediction is easy to understand (Poursabzi-Sangdeh et al. 2017)
        - **Simulability**: A human can consistently predict the model's result (Kim et al. 2016)
        - **Scrutability**: The rules that govern decision-making are not so complex, numerous, and interdependent that they defy practical inspection and resist comprehension (Selbst and Barocas 2018)
- *Interpretability is a domain-specific notion, so there cannot be an all-purpose definition. Usually, however, an interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge* — (Rudin 2019)

Explanation

- Possible definitions:
    - A human-interpretable description of the process by which a decision-maker took a particular set of inputs and reached a particular conclusion (Wachter et al. 2017)
    - Any of numerous ways of exchanging information about a phenomenon (Mittelstadt et al. 2018)
        - Anything that answers one these questions: What were the main factors in a decision? Would changing a certain factor have changed the decision? Why did two similar-looking cases get different decisions, or vice versa? (Doshi-Velez et al. 2017)
- *Analytically, explanation is infinitely variable, and there can be many valid explanations for a given phenomenon or decision. For example, a partial list of reasons for a glass having shattered include: a) because it hit the ground; b) because it was dropped; c) because the holder was startled (and that's why it was dropped); d) because gravity pulled it toward the earth; … These are all valid explanations, some nested within others, and some having nothing to do with each other* (Selbst and Barocas 2018)

Methods

- Learning interpretable models
    - Additive models
    - Tree-based models
    - Prototype classifiers
    - …
- Explaining black-box models
    - Global feature importance
    - Example-based explanations
    - Sensitivity analysis
    - Interpretable approximations

Additive Models

$$g(\mathrm{E}(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

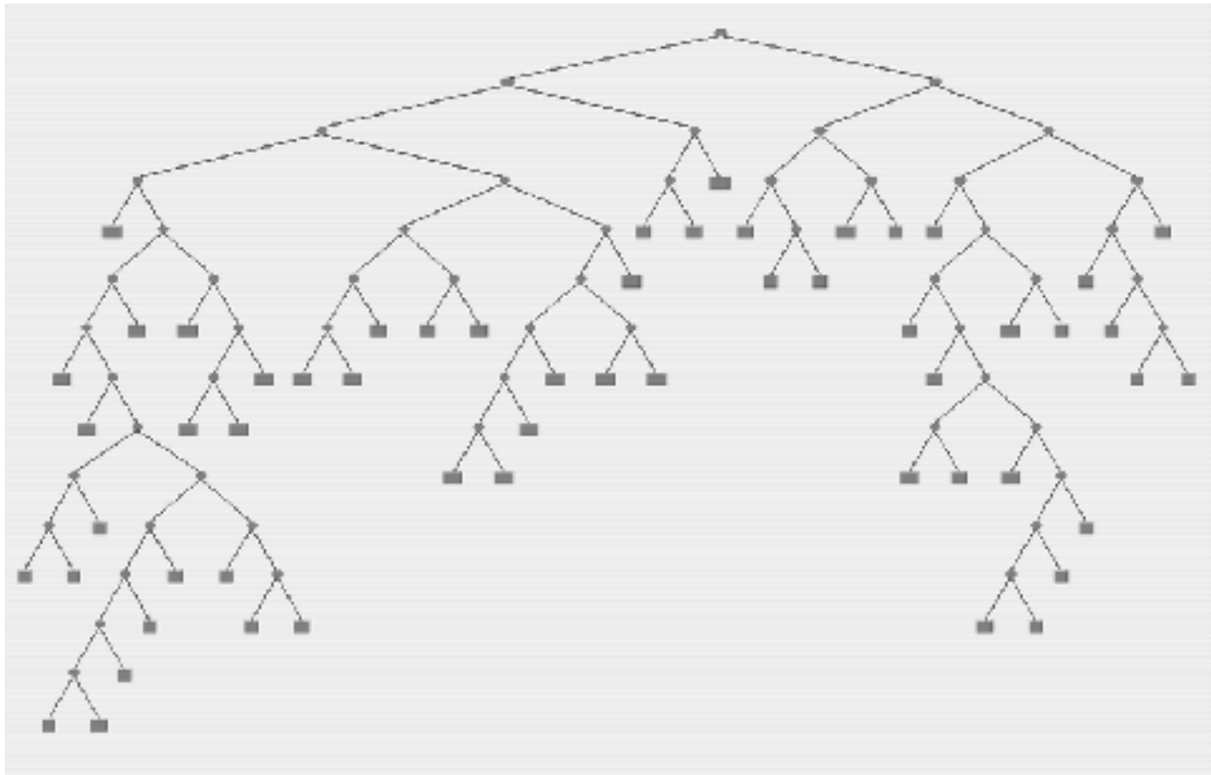
$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

- Generally considered interpretable because the relationship between individual inputs and output is simple

- Do not automatically capture interaction terms, so model capacity is limited

- Is a high-dimensional linear model really interpretable?

Decision Trees



- Easy to navigate
- Bad at dealing with linear/smooth variation
- Are deep trees really interpretable?

Others

- Rules
  - IF-THEN prediction rules
  - Sequential covering
    - Bayesian Rule Lists
- Prototype classifiers
  - KNN
  - Neural networks

Global feature importance

- Permutation feature importance (Breiman 2001)
    - “If a model’s accuracy doesn’t change when a feature is scrambled, that feature was not important”
- Partial dependence (Friedman 2001)
    - “If we calculate a model’s prediction, on average, for each value of this feature, we can see how it broadly relates to the output”
- Model Class Reliance (Fisher et al. 2019)
    - “Existing FI measures do not generally account for the fact that many prediction models may fit the data almost equally well… we define model class reliance (MCR) as the highest and lowest degree to which any well-performing model within a given class may rely on a variable of interest for prediction accuracy”

Interpreting learned features

- How does a neural network “see”? What inputs will maximize the activation of a certain neuron?

Example-based explanations

- Counterfactual explanations
  - Find an example X’ close to X where Y’ would have occurred instead (Wachter et al. 2017) - related to recourse
    - “The decision was made because X is not X’ ”
- Influential training set examples
  - Influence functions (Koh and Liang 2017)
    - “The decision was made because X is close to X’ ”

Sensitivity

- Individual conditional expectation plots (Goldstein et al. 2017)
    - “Holding all other features constant, how does the prediction for this instance change while varying the feature of interest?”
- Saliency
    - Generally gradient-based
    - “Which parts of the input is the model sensitive to?”

Interpretable approximations

- Local (instance-based)
    - LIME (Ribeiro et al. 2016)
    - SHAP (Lundberg and Lee 2017)
    - “How much of the output is each feature ‘responsible’ for?”
- Global

Criticism of post-hoc explanations

- **Saliency**: existing saliency methods are independent both of the model and of the data generating process. Consequently, methods that fail the proposed tests are inadequate for tasks that are sensitive to either data or model, such as, finding outliers in the data, explaining the relationship between inputs and outputs that the model learned, and debugging the model

- **Counterfactuals, approximations, feature importance**: Permuting input data can lead the model to extrapolate beyond the actual domain of the problem (Hooker and Mentch)

- **Approximations**: Explanations must be wrong; if the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation (Rudin)

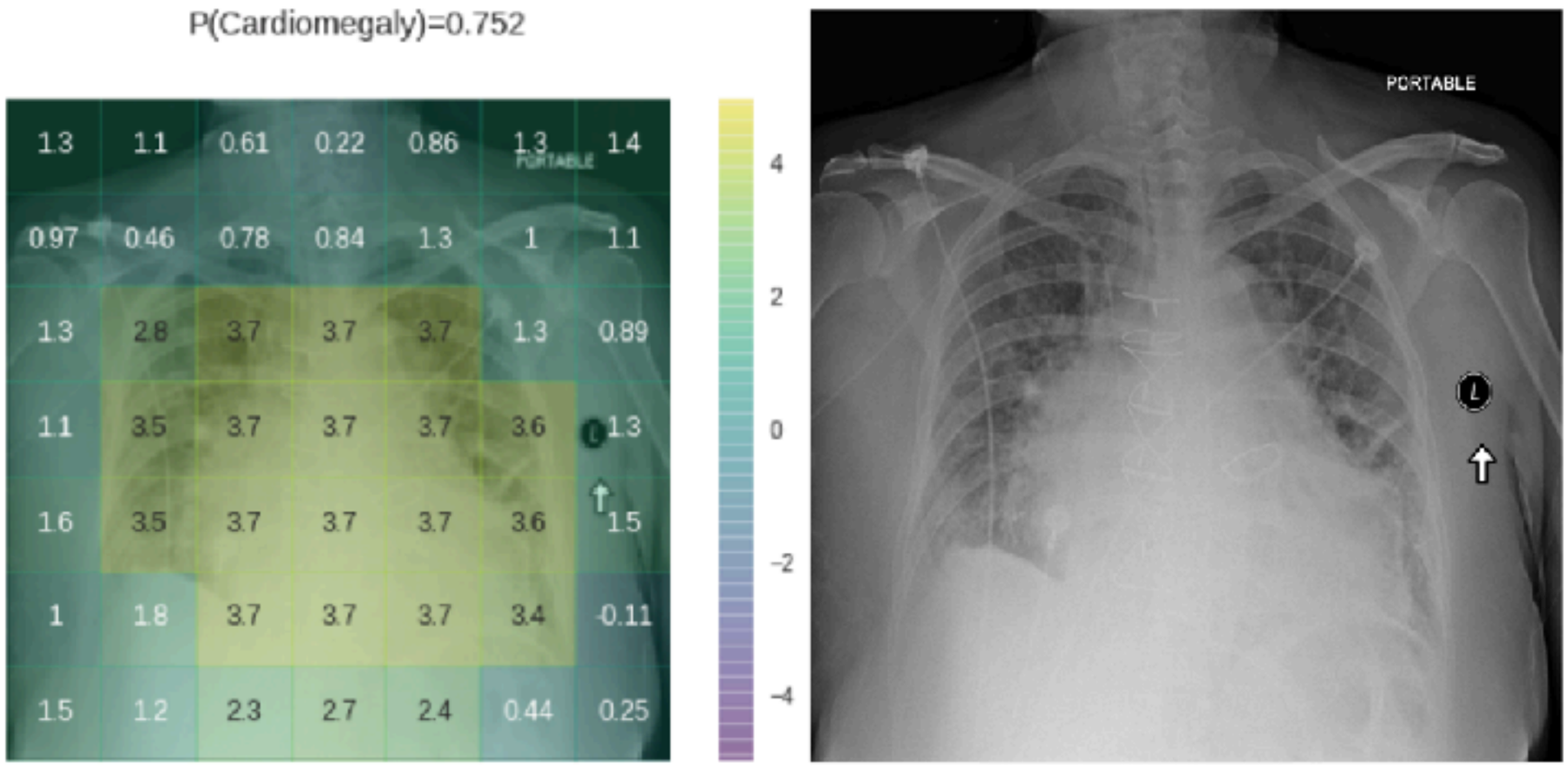
Criticism of interpretability

- “Interpretable” models do not necessarily have the properties we want them to have (MSR)

- Just because we can comprehend decision logic, it does not mean we can effectively evaluate it (Selbst and Barocas)

- The need for interpretability stems from an incompleteness in the problem formalization (Doshi-Velez and Kim)

Motivating Example



Patient with cardiomegaly: heatmap of CNN on left, original image on right

In this heatmap, positive values with green/yellow shading indicate regions of the image that contributed to a positive prediction of cardiomegaly. Negative values with blue/purple shading contributed to a negative prediction of cardiomegaly.

Motivating Example

Confounding variables can degrade generalization performance of radiological deep learning models

John R. Zech[1*], Marcus A. Badgeley[2*], Manway Liu[2], Anthony B. Costa[3], Joseph J. Titano[4], Eric K. Oermann[3]

A cross-sectional design was used to train and evaluate pneumonia screening CNNs on 158,323 chest x-rays from NIH (n=112,120 from 30,805 patients), Mount Sinai (42,396 from 12,904 patients), and Indiana (n=3,807 from 3,683 patients). In 3 / 5 natural comparisons, performance on chest x-rays from outside hospitals was significantly lower than on held-out x-rays from the original hospital systems. CNNs were able to detect where an x-ray was acquired (hospital system, hospital department) with extremely high accuracy and calibrate predictions accordingly.

Takeaways

- Interpretability and explainability are an amalgamation of a huge number of different qualities

- Research in interpretable or explainable machine learning should have well-specified objectives and evaluation metrics that can adequately address these