Quarter 2 Project Proposal

Brian Huynh
bth001@ucsd.edu

Joshua Lee
jdlee@ucsd.edu

Zoya Hasan
zohasan@ucsd.edu

Albert Hsiao
a3hsiao@health.ucsd.edu

1 Context

Modern healthcare generates enormous amounts of biological and clinical data, from imaging to lab values to physician notes. Bioinformatics models help transform this information into insights that can support diagnosis and treatment. Yet as machine learning becomes more capable, the central challenge is no longer just building accurate models, but ensuring that their predictions are transparent, trustworthy, and clinically interpretable.

This is especially critical in radiology, a field where deep learning models are increasingly used to detect diseases such as pulmonary edema, pneumonia, and heart failure from chest X-rays. Although these models often achieve impressive accuracy, they function as black boxes: they output predictions without expressing the clinical reasoning behind them. For physicians, this lack of explanation is not a minor inconvenience, rather it is a barrier to adoption. Physicians cannot act on a model's output unless they understand why it made its decision and whether its reasoning aligns with already established medical knowledge.

Existing explainability tools, such as SHAP values or gradient-based saliency maps, attempt to show which parts of the input influenced the model. However, these tools are still too technical and abstract for clinical decision-making. A heatmap or feature score does not tell a radiologist whether the model recognized diffused haziness, cardiomegaly patterns, pleural effusion margins, or other medically meaningful features. Nor do these methods relate model behavior to the language and reasoning found in radiology reports, which are the documents where clinicians articulate diagnoses, severity, and uncertainty.

Meanwhile, radiology reports themselves contain exactly the kind of reasoning that clinicians value: structured vocabulary, domain-specific interpretations, and descriptions of visual features linked to clinical outcomes. These reports represent a rich, deeper source of knowledge about how human experts explain imaging findings. Our project aims to bridge this gap by combining predictive modeling, SHAP explanations, and large language models capable of producing clear, clinically grounded interpretations. By training an LLM to translate technical model outputs into explanations that resemble real radiology-report reasoning, we make machine learning systems more understandable to the doctors who rely on them.

This work matters because interpretability is essential for safe and ethical deployment of AI in medicine. When doctors can see how an algorithm reached its conclusion, they can evaluate its correctness, catch failures, understand edge cases, and integrate model recommendations into clinical workflows with greater confidence. Improved interpretability also supports accountability, fairness, and patient trust, which are imperative concepts to responsible bioinformatics and healthcare technology.

In short, this project addresses one of the most pressing unmet needs in medical AI: turning black-box predictions into explanations that physicians can understand, evaluate, and ultimately trust. By leveraging radiology reports as a source of clinical reasoning and using modern LLMs to generate human-aligned explanations, this work contributes to the future of transparent, reliable, and clinician-centered artificial intelligence.

2 Methods

In previous papers, the interpretability problem has been addressed using attribution methods, such as Gradient-Base Saliency Maps & Perturbation-Based Saliency Maps, and non-attribution methods, such as attention networks, feature analytic methods, and generative methods. Attribution methods, in our case, would answer the question of which parts of the image contribute to the model's output; meanwhile, non-attribution methods focus on explaining the model rather than focusing on its input.

2.1 Attribution Methods

Popular model agnostic attribution methods are GradCAM and vanilla gradient maps, which generate heatmaps that show which regions of the input image contribute most to a model's output. These methods compute the gradient of the prediction with respect to each input pixel. For an image I, the vanilla gradient saliency map is defined as

$$G = \frac{\partial f(I)}{\partial I},$$

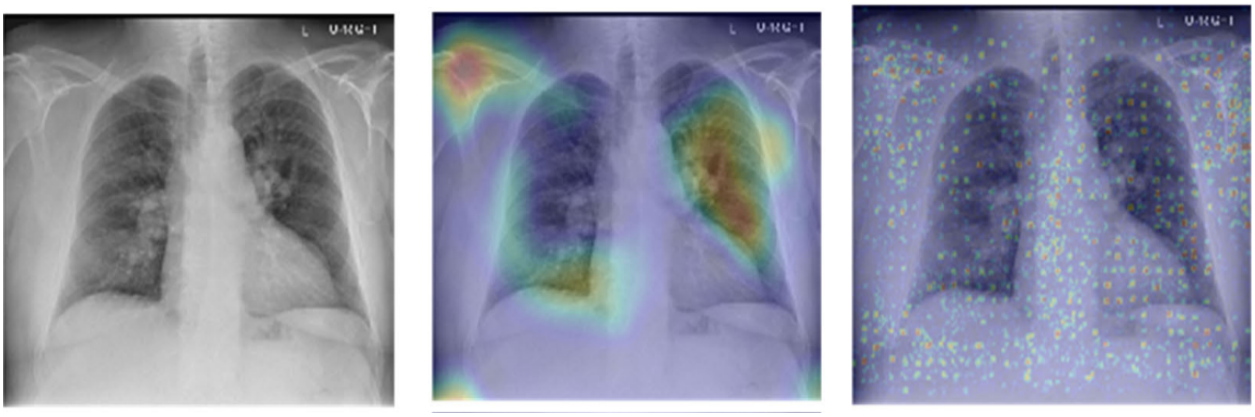
where $f(I)$ is the model's scalar output. In classification settings, this is often written for a specific class k as

$$G = \frac{\partial f^{(k)}(I)}{\partial I},$$

where $f^{(k)}(I)$ denotes the pre-softmax logit associated with class k and $k \in \{1, \ldots, K\}$. Because $\frac{\partial f}{\partial I}$ measures how sensitive the model output is to changes in each pixel, the magnitude $|G_{ij}|$ highlights pixels where the prediction is most locally sensitive. The resulting gradient values are typically normalized and overlaid onto the input image to form the vanilla gradient saliency map.

While these attribution methods effectively highlight "where" the model is attending within the image, they do not specify "what" features within those highlighted regions (e.g., shape,

opacity, or texture) drive the prediction. As a result, they provide spatial localization but lack the semantic detail necessary for clinical interpretability.



(a) Original Scan (b) GradCAM (c) Vanilla Gradient

Figure 1: Comparison of original scan, GradCAM, and Vanilla Gradient. Adapted from Hasenstab et al. (2023)

2.2 Non-attribution Methods

Non-attribution approaches aim to explain the model rather than assign importance to the individual. For example, attention networks integrate attention weights directly into a neural architecture. This highlights internal features that contribute to a prediction. However, attention as an interpretability tool is limited. Attention weights do not necessarily correspond to causal importance. Also, deep attention layers exhibit sparsity. This makes them difficult to analyze in large CNN architectures.

Feature analytic methods such as Principal Component Analysis (PCA) and t-SNE reduce the dimensionality of the hidden CNN feature space, allowing researchers to create visualizations. Using those visualizations, researchers can identify groups of cases that are incorrectly clustered or too trivial to detect. Despite PCA's and t-SNE's power to create "human-friendly" features, they still fall short in identifying specific aspects of an X-ray scan that influence the model's predictions.

Lastly, a generative approach such as Generative Adversarial Networks (GANs) can produce synthetic images that closely resemble those seen during training. GANs comprise two models: the generator and the discriminator. The generator model is tasked with generating images similar to those in the training dataset from noisy inputs, while the discriminator model distinguishes between generated images and real inputs. Using GANs, methods such as activating specific neurons associated with a given pixel can help reveal which parts of the neural network correspond to parts of an image(Simonyan, Vedaldi and Zisserman (2014)) However, the issue with this is that the synthetic images aren't realistic, making their interpretation difficult. (Hasenstab et al. (2023))

2.3 Our Approach

In our work, we address these limitations by including an additional modality that prior interpretability methods lacked, which are paired radiology reports. Unlike attribution maps or GAN-based approaches, radiology reports provide grounded descriptions of clinical findings (e.g., pulmonary edema, pneumonia, or an enlarged heart), offering a natural-language explanation of the visual patterns associated with predicted BNPP values. The addition of X-ray reports allows us to move beyond the simple method of identifying "where" the model is attending; instead, we can link the highlighted regions to the medical features they exhibit.

Our approach can be broken down into three stages. First, for each input image, we compute attribution maps using GradCAM or vanilla gradient to locate which regions influence the BNPP value prediction. Second, we process the paired radiology report using a pre-trained clinical language model (e.g., BioClinicalBERT, BlueBERT). Lastly, we integrate these two sources of information using a vision-language model that associates the model's highlighted image regions with the clinical findings that are extracted from the report. By following these steps, we will generate a natural-language explanation that incorporates both the model's internal focus and the medically relevant features of the X-ray scan.

Finally, when incorporating attribution maps with clinical language, our approach directly addresses the deficiencies of existing interpretability methods. It provides explanations that are both spatially localized and clinically meaningful, offering a clearer understanding of why the model predicts specific BNPP values for a given individual.

3 Primary Output

Our primary output of our Quarter 2 project will be a research report accompanied by a supporting website that showcases our results, interpretability findings, and LLM-generated explanations. Since our project goes beyond modeling and goes into explainable AI, our output will focus on comparing traditional interpretability techniques with LLM-based methods.

Due to our project not only analyzing data, but also generating data, we will showcase the analyses through statistical summaries of error patterns and model performance, traditional explainability outputs, and LLM-generated explanations. For the traditional explainability outputs, we will include activation maps, SHAP (SHapley Additive exPlanations) value plots, and feature ablations. For the LLM-generated explanations, we will potentially include synthesized clinician-facing explanation templates, LLM interpretations of SHAP or Grad-CAM outputs, natural-language summaries of model predictions, and structured narrative descriptions of biomarker- and imaging-derived signals.

Because the LLM component produces new data in the form of text explanations, we will analyze these outputs using not only qualitative and quantitative evaluation methods, but potentially human-in-the-loop (HITL) assessments, if feasible, by comparing the generated LLM summaries with expert criteria. For the qualitative evaluation, we're going to evaluate

the LLM's clinical relevance, coherence, and consistency. On the quantitative evaluation side, we are going to evaluate through agreement measures between the explanations from the LLM and model-derived features as well as similarity metrics through metrics such as cosine embedding similarity, Bilingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

The final deliverables will integrate these analyses into the written report. The website will have visualization panels, example predictions, LLM-generated explanations, and comparison tools that will allow users to interactively interpret model behavior.

References

Hasenstab, Kyle A., Justin Huynh, Samira Masoudi, Guilherme M. Cunha, Michael Pazzani, and Albert Hsiao. 2023. "Feature Interpretation Using Generative Adversarial Networks (FIGAN): A Framework for Visualizing a CNN's Learned Features." *IEEE Access* 11: 5144–5160. [Link]

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." [Link]