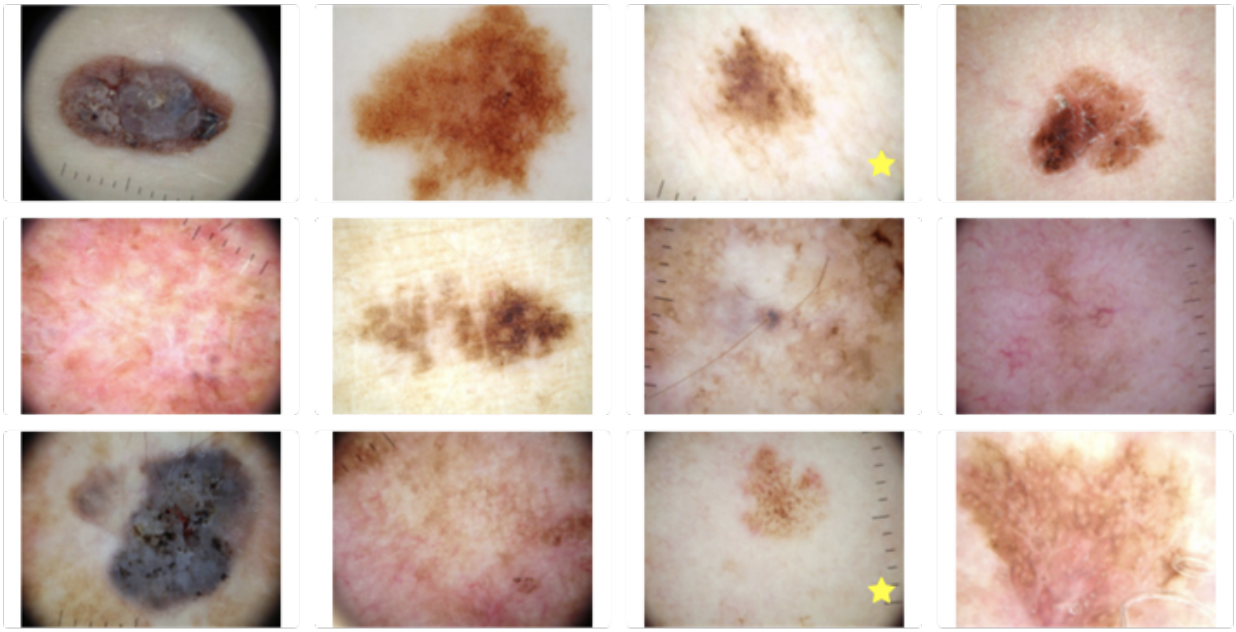
A Study of Different Machine Learning Interpretability Frameworks as Applied to the Biomedical Informatics Domain

First Authors: Sonish Sivarajkumar[1], Koushul Ramjauttan[2]
1 School of Computing and Information, University of Pittsburgh
2 University of Pittsburgh School of Medicine, Department of Biomedical Informatics

Abstract
Machine learning and deep learning have been widely adopted for applications in the medical domain. However, increased model complexity has often been used to achieve this boost in performance, converting such systems into "black box" approaches and causing uncertainty about how they operate and how the decisions are made. In clinical decision support systems, the interpretability of these models is critical to ensure that the chances of false positives are minimal. This ambiguity has made it difficult for machine learning systems to be implemented insensitively in the Medical domain, where the benefit of AI systems may be enormous.

Background
In 2017 Esteva et al. published their landmark paper in *Nature* (with currently 4656 citations) describing a deep neural network for diagnosing skin cancer with accuracy matching that of twenty-one board-certified dermatologists. However, later on, the authors realized that the model was actually using rulers (often present in images of malignant cancers) to make its predictions.



This case makes a strong argument for the need for more transparency in how a model is making its decisions and predictions and understanding the features that the model is focusing on. If such an approach had been applied in this work, the authors would have quickly realized that their model was disproportionately weighting the ruler part of the image instead of the lesions.

Introduction

Deep neural networks (DNNs) have reached/exceeded human performance for many complex tasks. These models are frequently implemented in a black-box way because of their complicated, non-linear structure, which means no information is offered about how they arrive at their predictions. In practice, the lack of transparency for AI explainability has led to criticism especially in the Medical domain, where transparency is vital for clinical decision support.

Moreover, model interpretability is critical in medical applications, for reasons including but not limited to:

- Legal compliance (e.g., European GDPR rules requires transparency and accountability for computational tools),
- Providing new insights into the underlying biological mechanisms responsible for a phenomenon,
- Giving additional feedback to clinicians such that they can make more informed decisions that don't rely solely on the final model prediction

To this end several frameworks have been proposed

- LIME (Ribeiro et al. 2016)
- SHAP (Lundberg & Lee 2017)
- Grad-CAM (Selvaraju et al. 2016)
- DeepLIFT (Shrikumar et al. 2017)
- LRP

In this project, we propose systematically studying and reviewing existing model interpretation techniques applied to the medical domain. We will focus on models used in biomedical image classification, Biomedical tabular data, and Clinical texts, thus making a comprehensive study on Machine Learning interpretability in the Biomedical domain.

Methods:

We are planning to analyze the most commonly used interpretability techniques, which encompasses five existing methods, including SHAP[1], LIME [2], DeepLIFT [3], Grad-CAM[4], and Layer-Wise Relevance Propagation [5], on all three datasets.

SHAP: Shapley Additive explanations (SHAP) is a game-theory-inspired technique that seeks to improve interpretability by estimating the significance values for unique predictions for each

attribute. all of which employ the same explanation model. SHAP values are widely accepted as a unified measure of feature importance that keeps three desirable properties: local accuracy, missingness, and consistency.

LIME: Local Interpretable Model-agnostic Explanations (LIME) method can generate interpretations for single prediction scores produced by any classifier using a simple yet powerful approach. Simulated randomly-sampled data around the input instance for which the prediction was formed are generated for any given instance. Following that, new predictions are formed for created instances using the probing model, weighted by their proximity to the input instance. Finally, this new dataset of modified cases is generated by a simple, interpretable model, such as a decision tree, and trained. The initial black-box model is interpreted as a result of understanding this local model.

DeepLIFT: DeepLIFT is a technique that defines the issue of importance in terms of differences from a "reference" condition, which is chosen based on the problem at hand. The input's reference state indicates a neutral input with no specific properties.

Contrastive explanations method (CEM) : Given any input and its corresponding prediction, this approach can determine which features must be present in order for that specific prediction to be made. CEMs may provide contrastive explanations for any black-box model. Grad-CAM [6] is a generalization of CAM that can produce graphical explanations for an image-based model.

Layer-wise Relevance Propagation LRP: Layer-wise Relevance Propagation (LRP) is a general method for interpreting DNN predictions by explaining them. LRP is a conservative approach, which means that the magnitude of each output y is conserved through the backpropagation process and equals the sum of the input layer's relevance map R.When applied to a variety of data kinds (images, text, audio, video, EEG/fMRI signals) and neural architectures, LRP is effective (ConvNets, LSTMs).

Data:
For a comprehensive analysis on interpretability in all modalities of medical data, we apply these interpretability techniques in 3 types of medical data: medical images, tabular data, and medical texts.

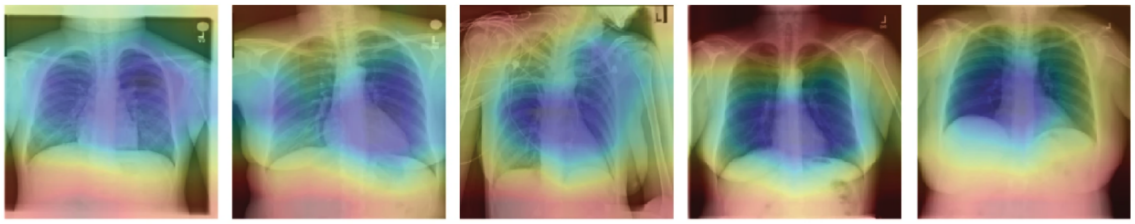
Medical Images:

1. pneumothorax images https://link.springer.com/article/10.1007/s10278-019-00299-9
2. Pathology images: https://kimialab.uwaterloo.ca/kimia/index.php/pathology-images-kimia-path960/
3. Skin cancer images: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

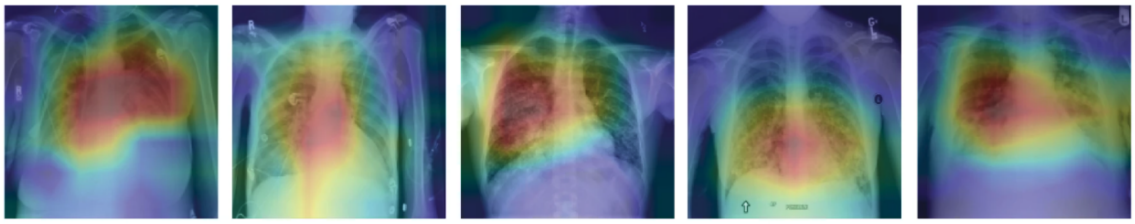
Tabular Data:
We plan to use the cardiovascular disease dataset[7]. The cardiovascular disease dataset is an open-source dataset that contains 70,000 patient records (34,979 with cardiovascular disease and 35,021 without cardiovascular disease) and 11 features. The four demographi features include Age, Height, Weight and Gender. There are 4 examination features -Systolic Blood Pressure, Diastolic Blood Pressure, Glucose, and Cholesterol levels. The three social, historic features include Smoking, Alcohol intake, Physical activity.
https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Medical Texts:
For evaluating the interpretability techniques on Medical texts, we use COVID-19 Open Research Dataset Challenge (CORD-19)[6]. This is a public dataset released as a part of an AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House. This data was collected from PubMed, bioRxiv, medRxiv preprint servers, and the World Health Organization (WHO) Covid-19 Database.CORD-19 consists of over 52K papers with over 41K full texts. We plan to perform text classification tasks on these clinical texts and perform model interpretation.

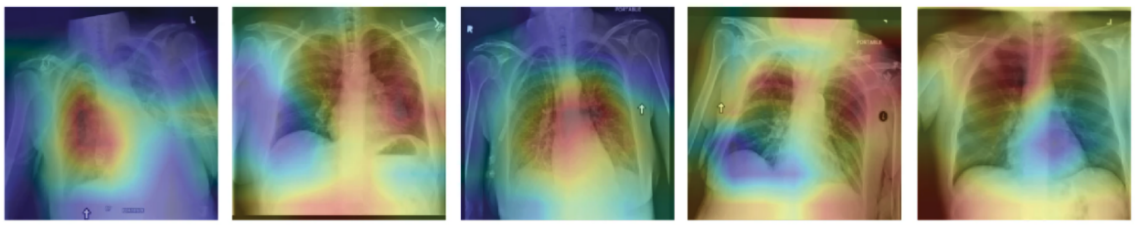
Examples of using explanation frameworks to understand ML models



(a) Heatmaps for low value images mislabeled as pneumonia



(b) Heatmaps for low value images mislabeled as no pneumonia



(c) Heatmaps for high value images mislabeled as pneumonia

Figure 1: Example visualization produced by Grad-CAM attention map - revealing the specific regions that a CNN focused on to make its final prediction.

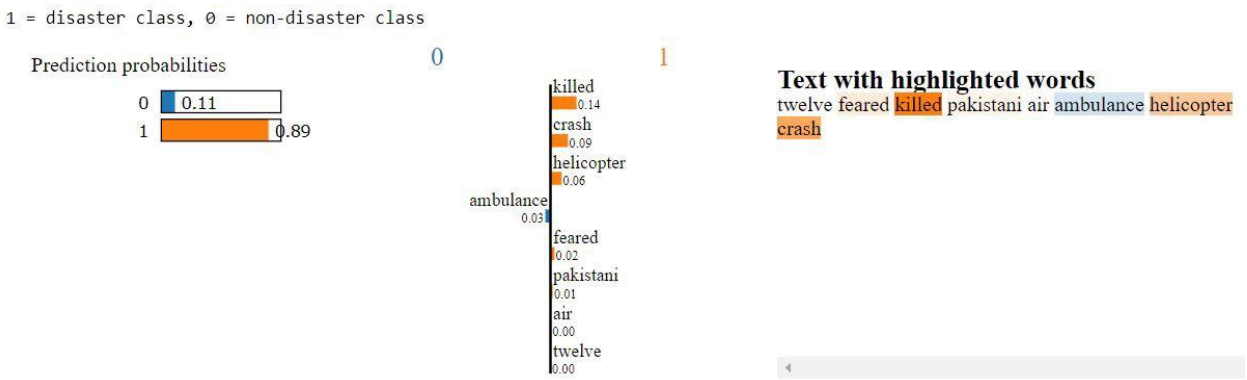


Figure 2: Understanding language model text weights

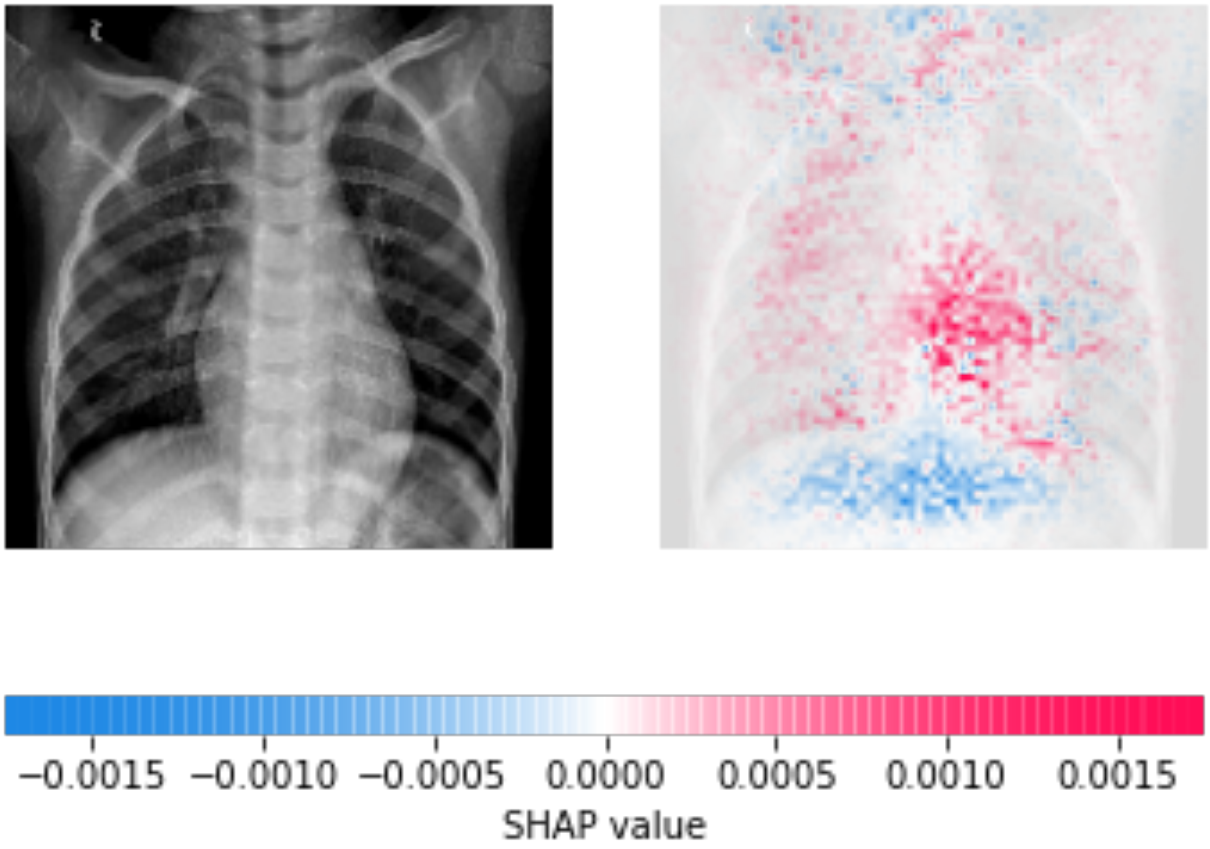


Figure 3: Using SHAPley values to get pixel level feature maps used by a CNN for diagnosis

References

1. Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances In Neural Information Processing Systems*, *30*. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
2. Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Retrieved 17 February 2022, from https://arxiv.org/abs/1602.04938
3. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Retrieved 17 February 2022, from https://arxiv.org/abs/1704.02685
4. Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., & Goldberg, Y. (2021). Contrastive Explanations for Model Interpretability. Retrieved 17 February 2022, from https://arxiv.org/abs/2103.01378
5. Binder, A., Montavon, G., Bach, S., Müller, K., & Samek, W. (2016). Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. Retrieved 17 February 2022, from https://arxiv.org/abs/1604.00825
6. Lu Wang, Lucy et al. “CORD-19: The Covid-19 Open Research Dataset.” ArXiv arXiv:2004.10706v2. 22 Apr. 2020 Preprint.
7. Rajib Kumar Halder, November 10, 2020, "Cardiovascular Disease Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/7qm5-dz13.

Sources:

- **https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools**
- **https://www.nature.com/articles/s41598-021-87762-2**
- **https://www.oak-tree.tech/articles/healthcare-ai-primer**
- **https://medium.com/@kalia_65609/interpreting-an-nlp-model-with-lime-and-shap-834ccfa124e4**