Towards trustworthy AI-based algorithms in healthcare: A case of medical images.

L.M. Amugongo[1]
[1] *Faculty of Computing and Informatics, Namibia University of Science & Technology.*
* *lamugongo@nust.na*

@lamechthinkbig

OBJECTIVES

Over the last decade, there have been a lot of artificial intelligence (AI)-based solutions that have been proposed in healthcare. However, only a few of the solutions are clinically available. Lack of trust in healthcare AI-based solutions is tied to the technical characteristics of AI "black box" (as seen in Figure 1), and how these properties can be understood clinically or biologically. Explainable AI (XAI) has been suggested to improve the interpretability of AI-based solutions, providing qualitative and quantitative reasons for how AI models make their decisions. But, most of the XAI techniques produce a number and or use a threshold to determine whether the decisions made by the model are sound.

METHODS

- A public dataset of 5 856 routine clinical care X-Ray images were obtained from Kaggle.
  - Training: 5216 (Normal=1341 & Pneumonia=3875)
  - Test: 624 (Normal=234 & Pneumonia =390)
  - Validation: 16 (Normal=8 & Pneumonia=8)
- All images were screened to remove low quality images or images not readable.
- Two expert clinicians diagnosed the presence pneumonia.
  - A third expert used to account for diagnosis error.
- A clipbox was created to define the region of interest.
- CNN model was trained to distinguish between pneumonia and normal images.
- SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME).

Figure 1: Typically, designers machine Learning models cannot not explain how and why AI algorithm make specific decisions. Thus, they are referred as black boxes.

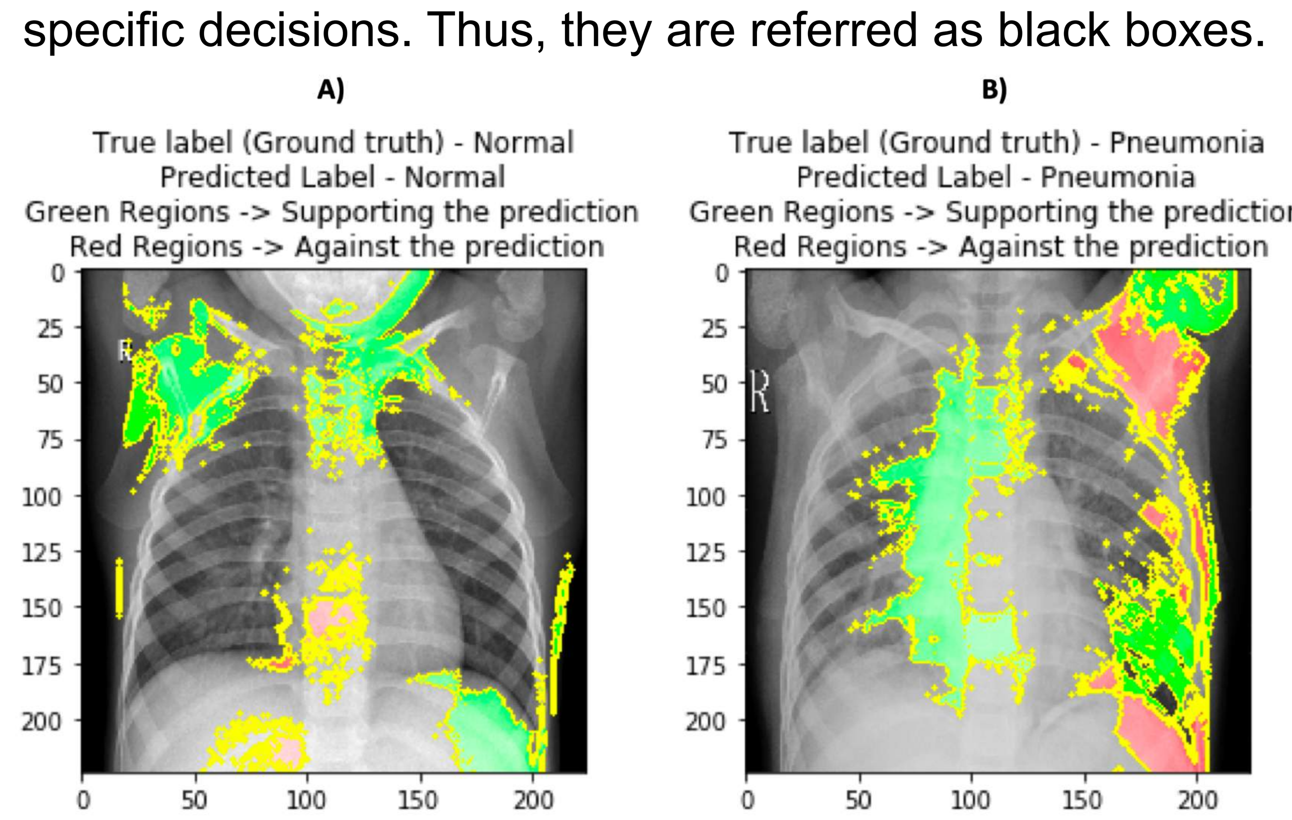

Figure 3: Lime explanation, the regions highlighted in green shows the regions that contribute the most to the prediction and the region in least contributing to the prediction in red. A) Image correctly classified as normal and B) image correctly classified as Pneumonia.

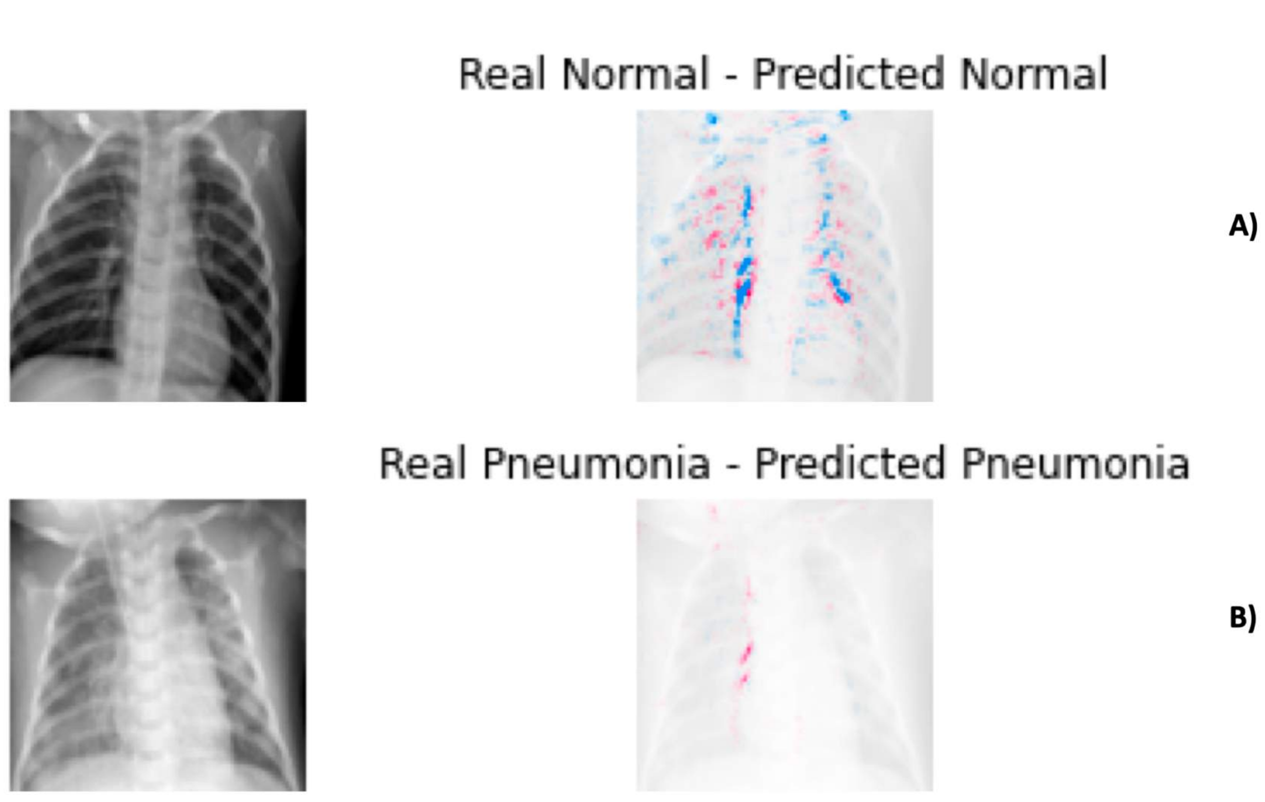

Figure 2: SHAP explanations, highlighting regions that contributed the most to the prediction (in red) and regions contributed against the prediction in blue. A) Image correctly classified as normal and B) image correctly classified as Pneumonia.

RESULTS

As seen in Figure 2 and 3, both LIME and SHAP highlight the regions that contribute to the explanations. However, they do not provide qualitative nor quantitative explanations that help clinicians trust the decisions of algorithms.

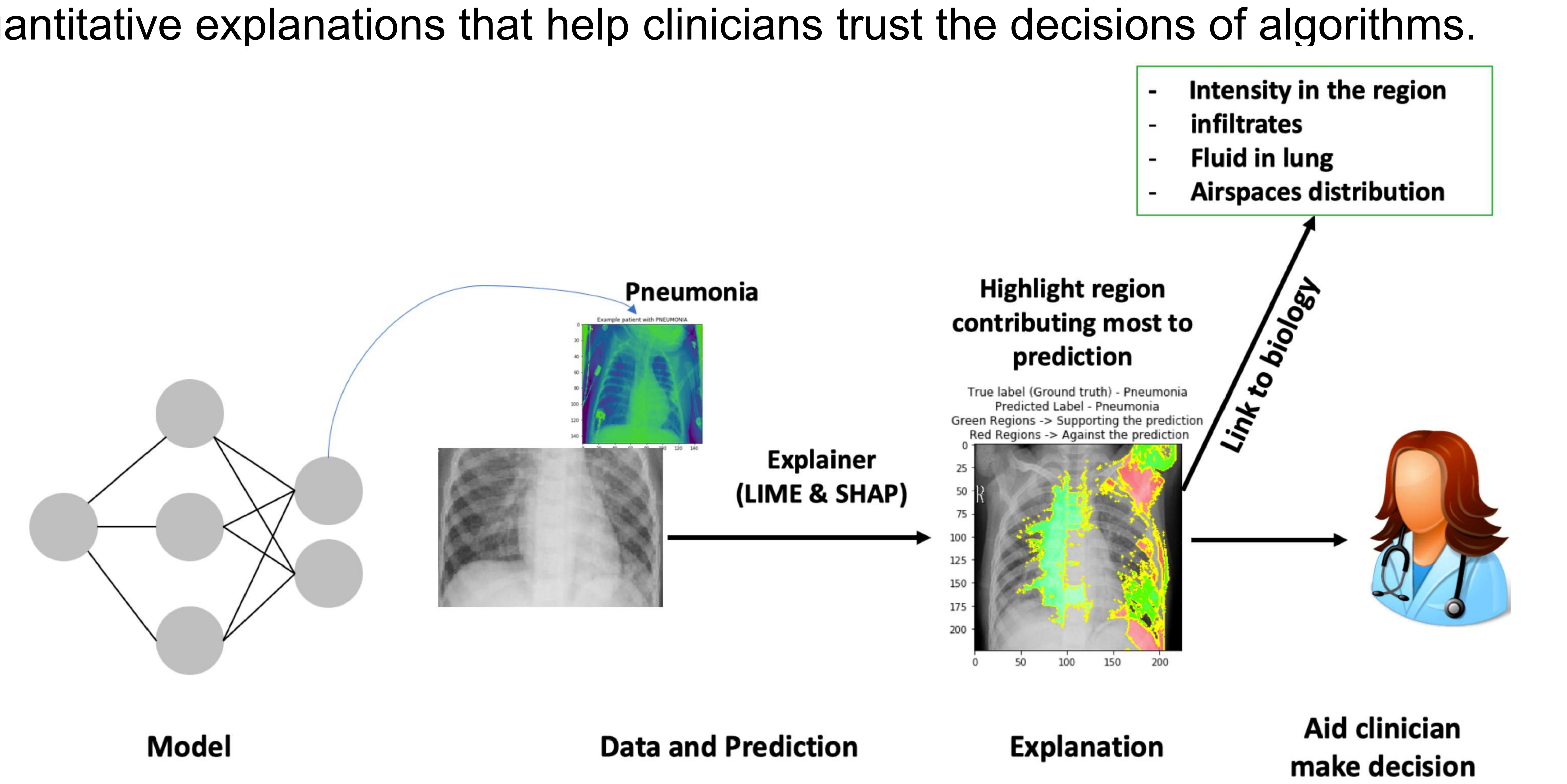

Figure 4: Prediction explanation for an individual patient. A model predicts that a patient has cancer, and existing explainer tools (LIME and SHAP) highlight the region(s) that contribute the most to the prediction. Though a clinician can further investigate the highlighted region. Highlighting a region is insufficient explanation, due to the complex underlying biology. Therefore, when dealing with medical imaging XAI tools should be able to link the qualitative and quantitative explanations to the underlying biological phenotypes in medical images.

CONCLUSION AND FUTURE WORK

When dealing with medical imaging, a number and or highlighting the region that contribute to prediction is insufficient explainability, due to the complex underlying biology.

- Link explanation to biology, as explained in Figure 4.