UROP RESEARCH PROJECTS

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Explainable AI (XAI) Frameworks and Applications

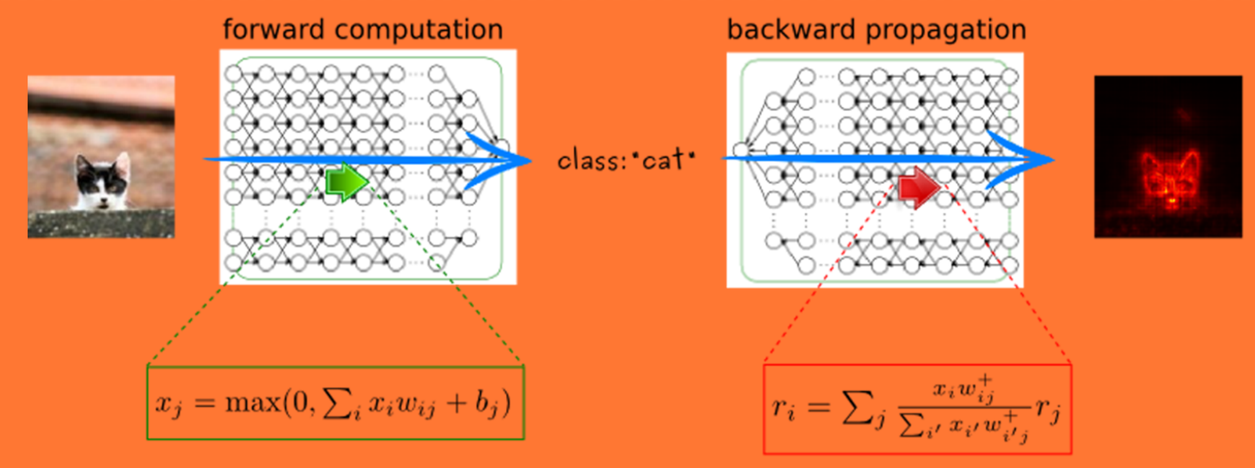
Authors
Bryan Tan (Chen Zhengyu), CSD
Sahana Katragadda, CSD
Lee Pei Xuan, CSD

Supervisor:
Assist Prof. JIANG Wenchao, ISTD.

Demo: Real-time LRP Processing

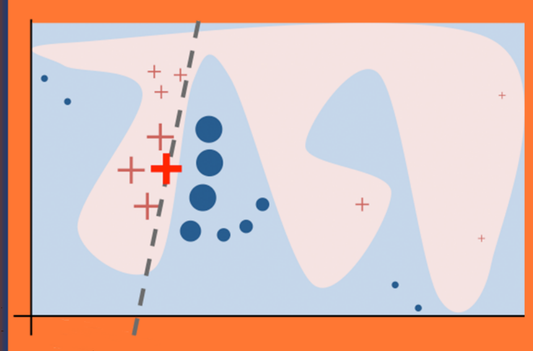
Approach 1: Explanation through Backward Propagation

Layer-wise Relevance Propagation (LRP) is a technique that brings such explainability by propagating the prediction backward in the neural network. Pixels that are important to a model's decision appear brighter in the resultant heatmaps.



(a) (b) (c)

In the example, LRP highlights weights from the VGG16 convolutional neural network model pre-trained on the ImageNet dataset. In (a), LRP correctly highlights the outline of a cat as being important in determining its object type. In (b), LRP shows that pixels in the foreground surrounding the benches and tents are considered more important. For (c), pixel importance is more evenly spread since the model is not trained on such images.

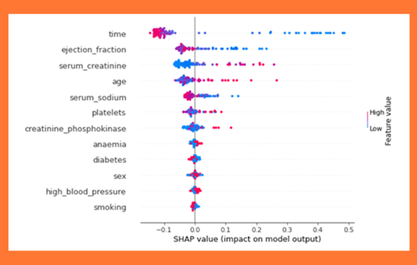


In the figure, the black-box model's decision function is complex (blue/ pink background) but there exists a explanation that is locally faithful.

Glassbox models such as Explainable Boosting Machine provide further insight into which clinical parameter has the greatest risk of heart disease

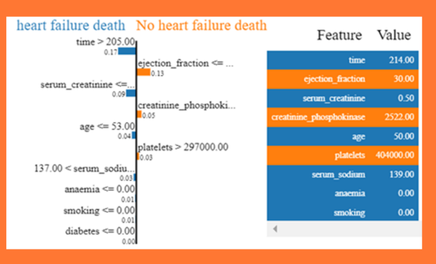
Approach 2: Explanation through Local Surrogate Models

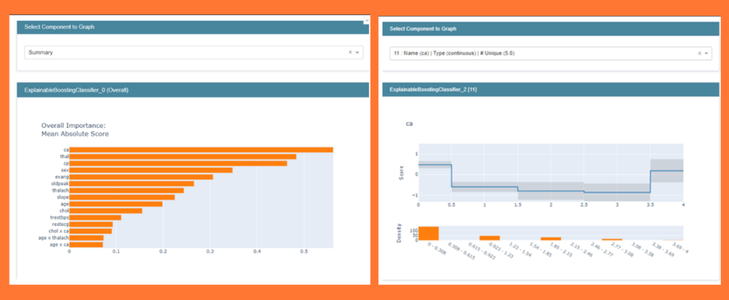
Local function explanations could be more accurate than global explanations. Among them are Local Interpretable Model Agnostic (LIME) and Shapley Additive Explanations (SHAP) that build local surrogate models to black-box ML models. LIME assumes a linear local model while SHAP adopts Shapley value with a fair contribution of variables.



Example is using LIME (below) and SHAP (left) to explain ML model predictions on heart failure.

Both models find the same top 3 features. But LIME tends to be biased to binary data while SHAP is more fair to all variables.

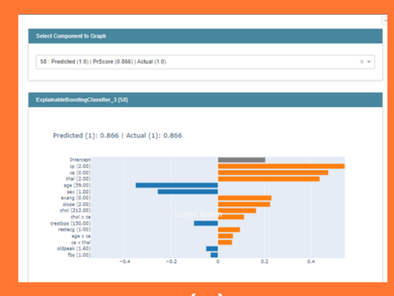




(c) (d)

Approach 3: Explanation through Glassbox Models

These are the global predictions using Explainable Boosting Machine (EBM) The clinical parameter, *ca* , has the highest risk in predicting heart disease as seen in (c) and (d). *ca* is the number of blood vessels.



(e)

Local predictions show how much the clinical parameter , *ca*, increases the risk for each specific patient as seen in (e).

The XAI approaches studied in the project will

1 Enhance trust in AI models

Through XAI, we can visualize whether the models are looking at the expected areas or not, and argue the accountability of an AI model. It is critical for the adoption of AI models in fields where precision and accountability are essential, such as healthcare.

2 Empower people to take action

XAI goes beyond prediction accuracy and tells the importance of each feature to the final prediction. It will hence empower people to take corresponding actions to improve their situations.

LEAVE YOUR MARK