LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS FOR MEDICAL IMAGE SEGMENTATION

Amaan Jogia-Sattar[1], Audrey Kim[2], Rui Nie[3]
(Authors are listed in alphabetical order)
[1]University of California, Berkeley, [2]Smith College, [3]University of Michigan - Ann Arbor

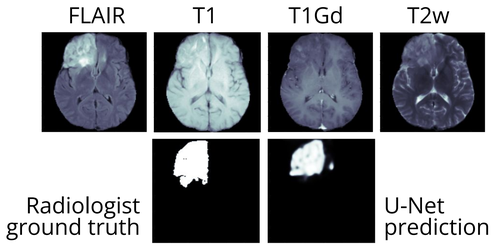
INTRODUCTION

- The field of **eXplainable Artificial Intelligence (XAI)** attempts to address the AI black-box problem by uncovering unknown mechanisms and limitations of complex models.
- Existing XAI frameworks often have narrow compatibility and cannot be applied to medical image segmentation tasks.
- **Local Interpretable Model-Agnostic Explanations (LIME)** is an explainability framework intended to accommodate a vast array of black-box models.

We modified LIME to work with multiple MRI sequences for slice-wise tumor segmentation.

MATERIALS

Dataset
- MRI scans via The Cancer Genome Atlas
- 144 slices × 144 pixels × 144 pixels × 4 sequences
- Patient of Interest: 'TCGA-HT-7874,' slice 75



U-Net
- Convolutional neural network architecture
- Performs 2D slice-wise tumor segmentation with multisequence MRI input

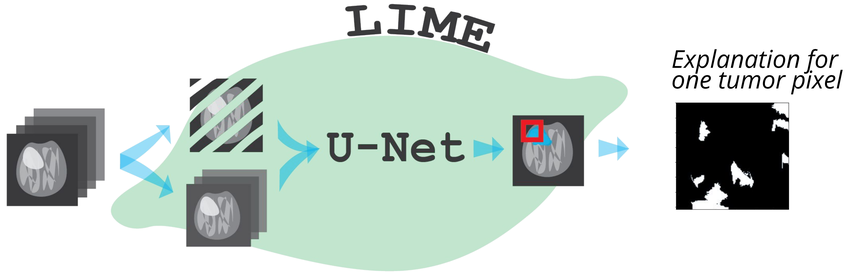
LIME

Identifies important features for individual predictions

Local: investigates specific instance/prediction
Interpretable: linear, sparse surrogate model
Model-agnostic: perturbations around a local input allow compatibility across classifiers
Explanations: surrogate model feature weights approximate model behavior

$$\xi(x) = \text{argmin}_{g \in G} \ \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$
$$\mathcal{L}(f, g, \Pi_x) = \sum_{z, z' \in \mathcal{Z}} \Pi_x(z) \left(f(z) - g(z')\right)^2$$

GENERATING EXPLANATIONS



Perturb one sequence → pass all four to U-Net → explain one pixel's prediction
Repeat binary classification (pixel = tumor/non-tumor) to simulate segmentation

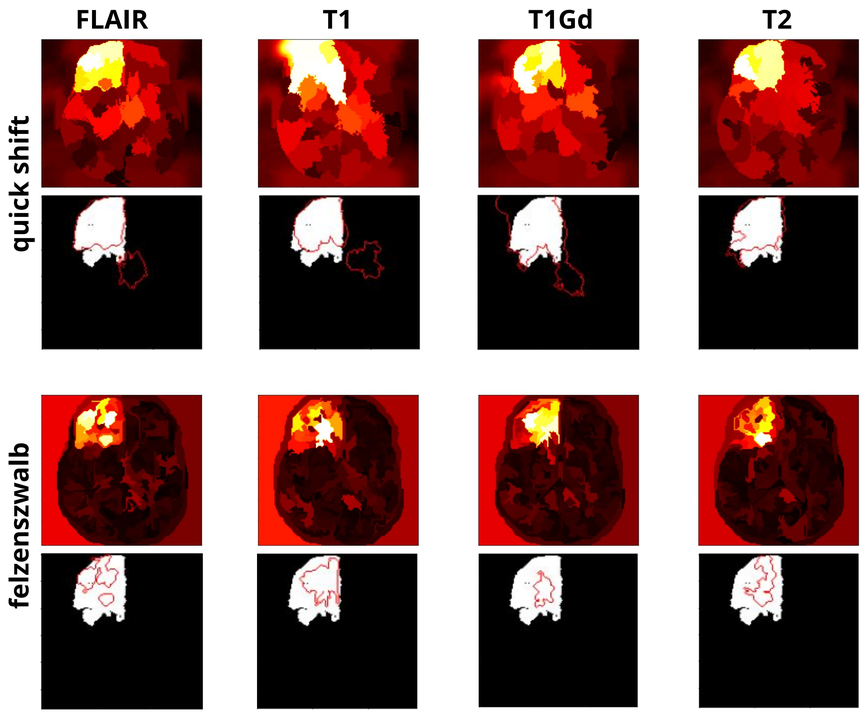
Superpixel Generation Algorithm Comparison

quick shift
- Forms a tree of density-dependent links to nearest neighboring pixel
- Superpixel clusters occupy joint spatial and color dimensions

felzenszwalb
- Graph-based segmentation approach
- Pixels-as-vertices representation, connecting edges
- Segmentation based on vertex-wise affinity

RESULTS

Explaining U-Net: LIME-determined important features



	FLAIR	T1	T1Gd	T2
quick shift	74.9%	72.0%	89.7%	74.9%
felzenszwalb	26.1%	34.1%	12.8%	22.2%

Table: Percentage of tumor pixels included in explanations when threshold = 0.5

CONCLUSION

Takeaways
- Modified LIME to accommodate segmentation tasks by simulating binary classification problems. Allowed for inputs of multisequence imaging data types with limitations.
- *Quick shift*-segmented explanatory regions for tumor-identified pixels incorporate a larger percentage of total tumor pixels relative to *felzenszwalb.*

Future Endeavors
- Explore the merits of *slic* segmentation algorithm, which utilizes the LAB color space as opposed to RGB.
- Train U-Net on contextual information (e.g. clinical observations) as opposed to lone ground truth segmentations and tumor vs. non-tumor labeling.
- Develop metrics for assessing U-Net accuracy and determine if particular MRI sequences result in more optimal diagnoses.
- Attempt global explanation using a set of local instances.

REFERENCES

alexandrusocolov. 2020. "Modifying LIME for Neural Networks on Medical Imaging." Jupyter Notebook. https://github.com/alexandrusocolov/LIME-for-medical-imaging.

"Frontiers | Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation." n.d. Accessed July 27, 2022. https://www.frontiersin.org/articles/10.3389/fnins.2021.740353/full.

"Quick Shift and Kernel Methods for Mode Seeking | SpringerLink." n.d. Accessed July 27, 2022. https://link.springer.com/chapter/10.1007/978-3-540-88693-8_52.

Ribeiro, Marco Tulio Correia. (2016) 2022. "Lime." JavaScript. https://github.com/marcotcr/lime.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Visani, Giorgio. 2021. "LIME: Explain Machine Learning Predictions." Medium. January 28, 2021. https://towardsdatascience.com/lime-explain-machine-learning-predictions-af8f18189bfe.

ACKNOWLEDGEMENTS:
Dr. Nikola Banovic, Snehal Prabhudesai, Dan Barker, Dr. Bhramar Mukherjee