Mentor(s) name: Daniel Flores Araiza, Gilberto Ochoa Ruiz

Project title: Robust Interpretability with Self-Explaining Deep Learning Models for Computer-Assisted Classification of Kidney Stones in Endoscopic Images

1. Motivation of the project: Developing "robust interpretability" in kidney stone classification using deep learning is crucial due to the high-stakes nature of healthcare decisions. Accurate classification of kidney stones is essential for determining appropriate treatment strategies. Still, traditional methods like Morpho-Constitutional Analysis (MCA) have limitations regarding invasiveness, time consumption, and the need for specialized expertise. While deep learning offers a promising alternative for more efficient and less invasive classification, the lack of interpretability in these models presents a significant challenge. In clinical settings, it's not enough for a model to be accurate; healthcare professionals must also understand the reasoning behind its predictions to trust and effectively use it. Therefore, enhancing robust interpretability in AI models for kidney stone classification advances the technology's efficacy and ensures its alignment with the critical needs for transparency, trust, and ethical decision-making in medical diagnostics.

2. Objective: Determine the robustness of different eXplainable-AI (XAI) models for classifying kidney stones, particularly the benefits and limitations of self-explainable models for robust explanations, and explore a training loss function based on metric learning (ICNN), to improve such robustness while maintaining clarity of explanations.

3. Approach: Train different XAI models, and a self-explainable architecture based on detecting prototypical parts per class will be evaluated on robustness and compared to the robustness of self-explainable models under a training loss function based on metric learning, based on the Intra-class and inter-Class Nearest Neighbors (ICNN), between the training data samples and prototypical parts for each class.

4. Experimental evaluation: Perform robustness and out-of-distribution tests on the different models trained on two ex-vivo kidney stones datasets. The robustness tests will consider the level of performance lost under specific levels of perturbations in the input image. In contrast, the OOD tests will indicate the level of performance retained by the different models under different types of photometric perturbations in the input. In addition, the explanation of the models input will be tested with the “insection” and “deletion” explanation metrics.

Data sources: Two ex-vivo kidney stones datasets (stones removed from the patient).

Preferred mentee's skills: Mentees from the computer science field with essential to intermediate knowledge of artificial intelligence, computer vision, and machine learning are preferred. Also, we expect mentees to be independent, proactive, and collaborative.

References

- Flores-Araiza, D., Lopez-Tiro, F., El-Beze, J., Hubert, J., Gonzalez-Mendoza, M., Ochoa-Ruiz, G., Daul, C., 2023. Deep prototypical-parts ease morphological kidney stone identification and are competitively robust to photometric perturbations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 295–304.
- David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S.Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018.
- Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song. Evaluation and improvement of interpretability for self-explainable part-prototype networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2011–2020, October 2023.
- Mendez-Ruiz, M., Gonzalez-Zapata, J., Reyes-Amezcua, I., Flores-Araiza, D., Lopez-Tiro, F., Mendez-Vazquez, A., Ochoa-Ruiz, G., 2023. Susana distancia is all you need: Enforcing class separability in metric learning via two novel distance-based loss functions for few-shot image classification. arXiv preprint arXiv: 2305.09062.