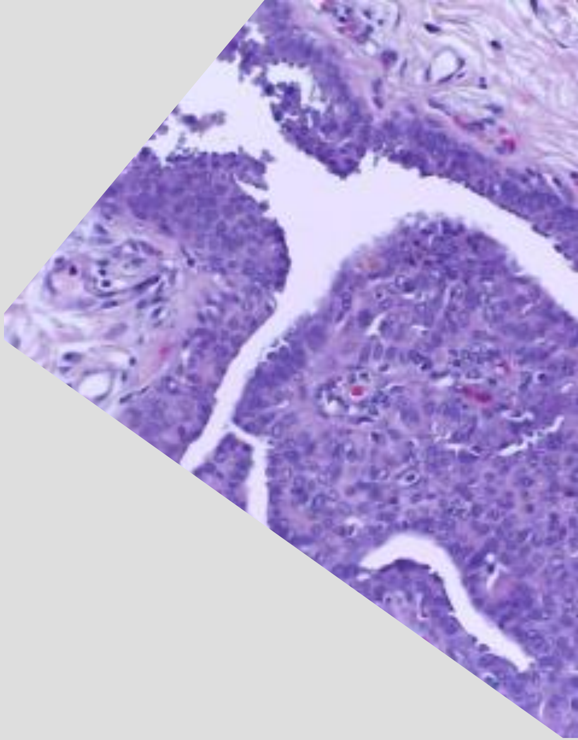
Model Agnostic Interpretability for Multiple Instance Learning

ICLR 2022

Joseph Early
AIC Research Group
University of Southampton, UK

J.A.Early@soton.ac.uk @JosephAEarly



UKRI Engineering and Physical Sciences Research Council

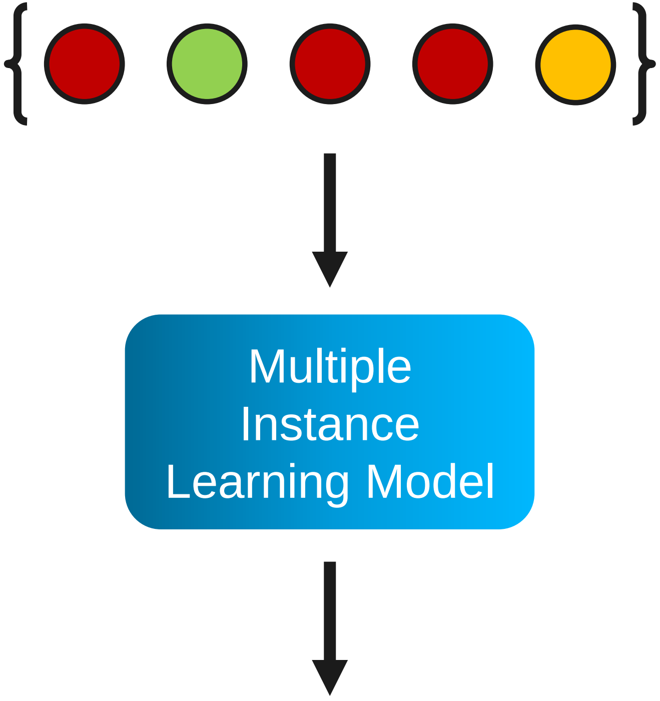
UKRI Trustworthy Autonomous Systems Hub

UNIVERSITY OF Southampton

The Alan Turing Institute

Multiple Instance Learning

- Making predictions from bags of instances
- Each bag has a single label, and instance labels are not given
- The outcome is often only determined by a few "*key instances*"



Multiple Instance Learning Model

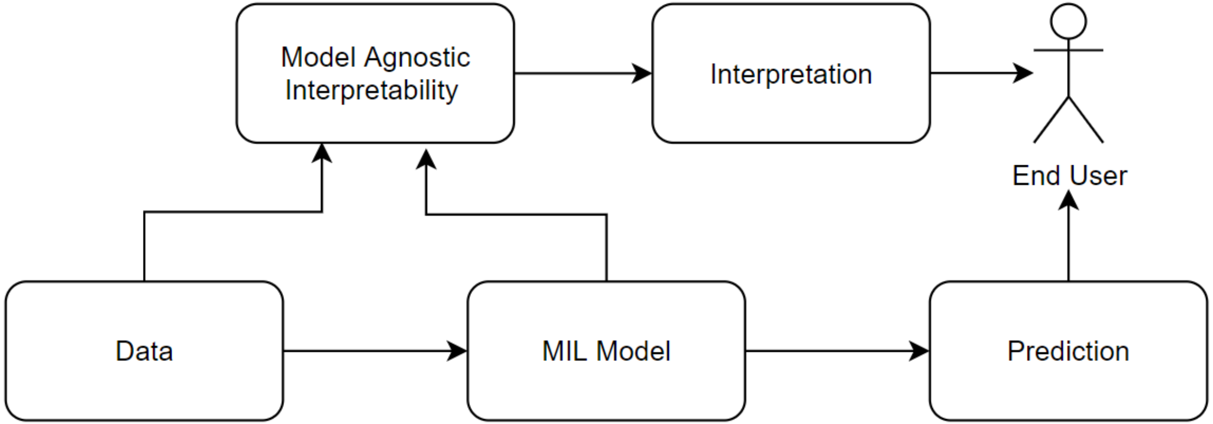
Prediction: [0.01, 0.67, 0.32]

Interpretability for MIL

- We want to answer two questions:

  Which are the key instances in a bag?

  What outcomes do they support?



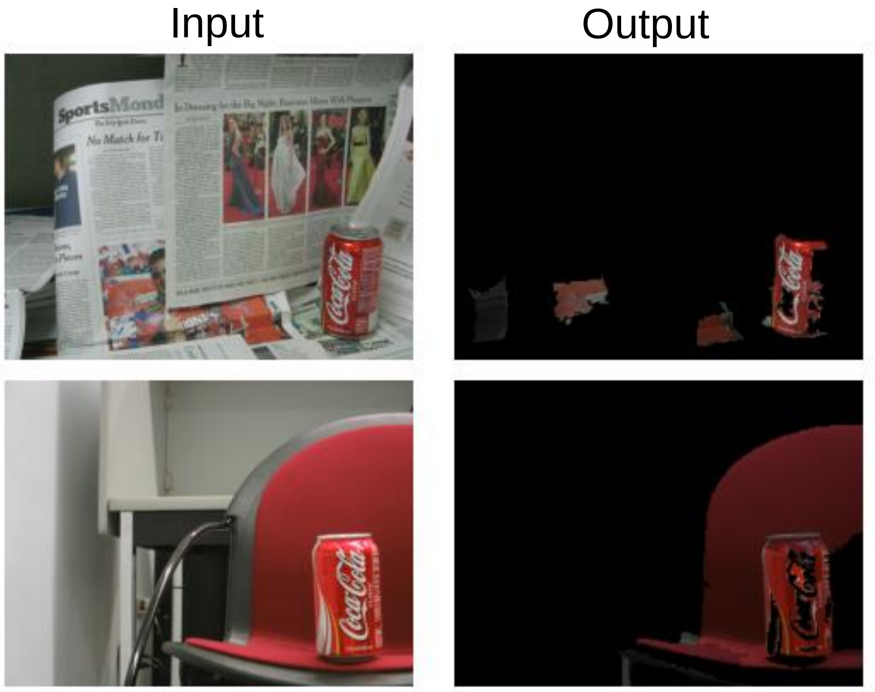
Multiple Instance Learning Local Interpretations (MILLI)

- Approximate the true MIL model with an interpretable surrogate model that is locally faithful
- Sample coalitions (sub-bags) of instances to take interactions between instances into account
- Use an adaptive sampling approach that can be tailored to certain instances and coalition sizes

Results

- – Evaluated nine methods across seven datasets and four models
- – Datasets included complex objects, handwritten digits, colorectal cancer tissue classification
- – Models included MIL Attention and MIL GNN
- – MILLI has the best overall performance as well as being more sample efficient

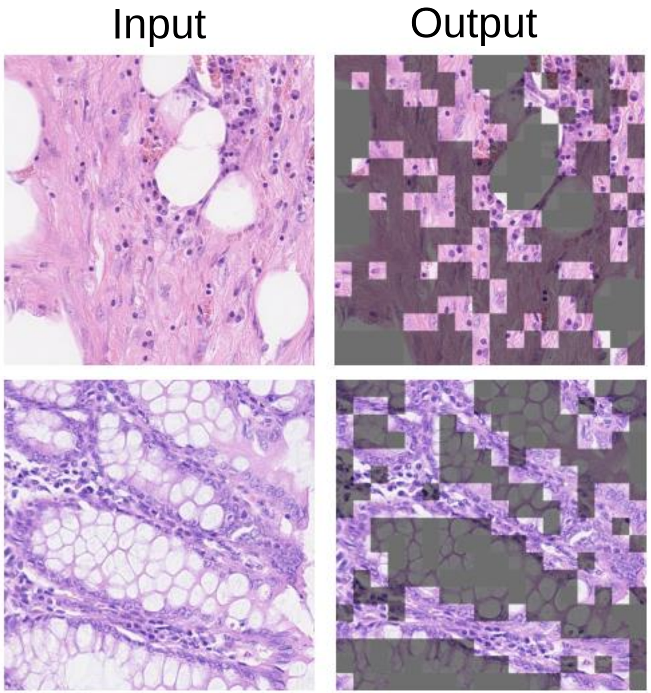
Example Outputs



Input

Output

SIVAL Dataset



Input

Output

CRC Dataset

Summary

J.A.Early@soton.ac.uk
@JosephAEarly

- We developed a model-agnostic interpretability method for multiple instance learning
- Our method outperforms the state-of-the-art methods on a range of benchmark datasets by up to 30%

Christine Evers
University of Southampton
AIC Research Group

Gopal Ramchurn
University of Southampton
AIC Research Group