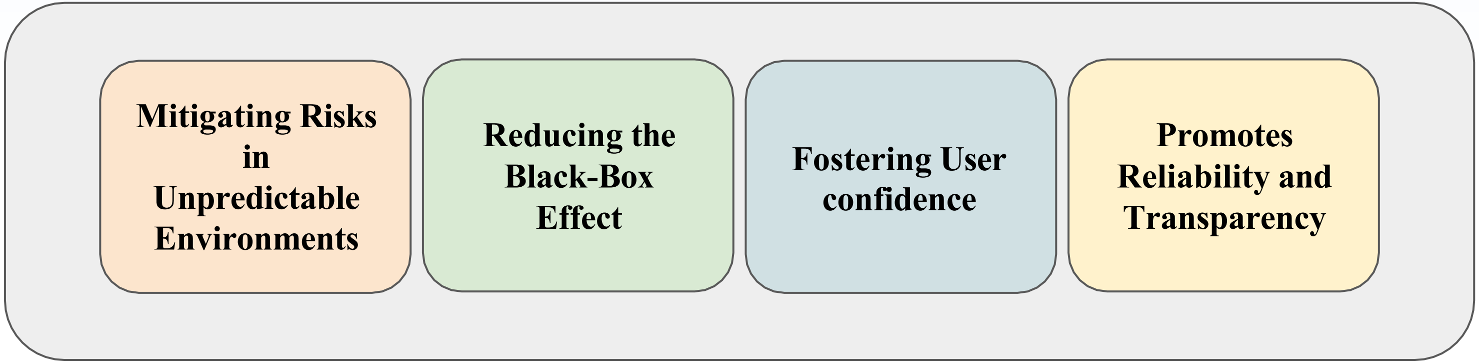
THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

An Investigation of Transparent Methods for Improved Human-AI Trust and Reliability in AI-Driven Autonomous Systems Applications

Authors: Shivangi Gupta, Vineetha Menon, Kristin Weger and Bryan Mesmer

Introduction

Explainability in AI systems refers to the model's ability to provide clear, interpretable insights into its internal workings, including how inputs contribute to the final outcome.



Mitigating Risks in Unpredictable Environments

Reducing the Black-Box Effect

Fostering User confidence

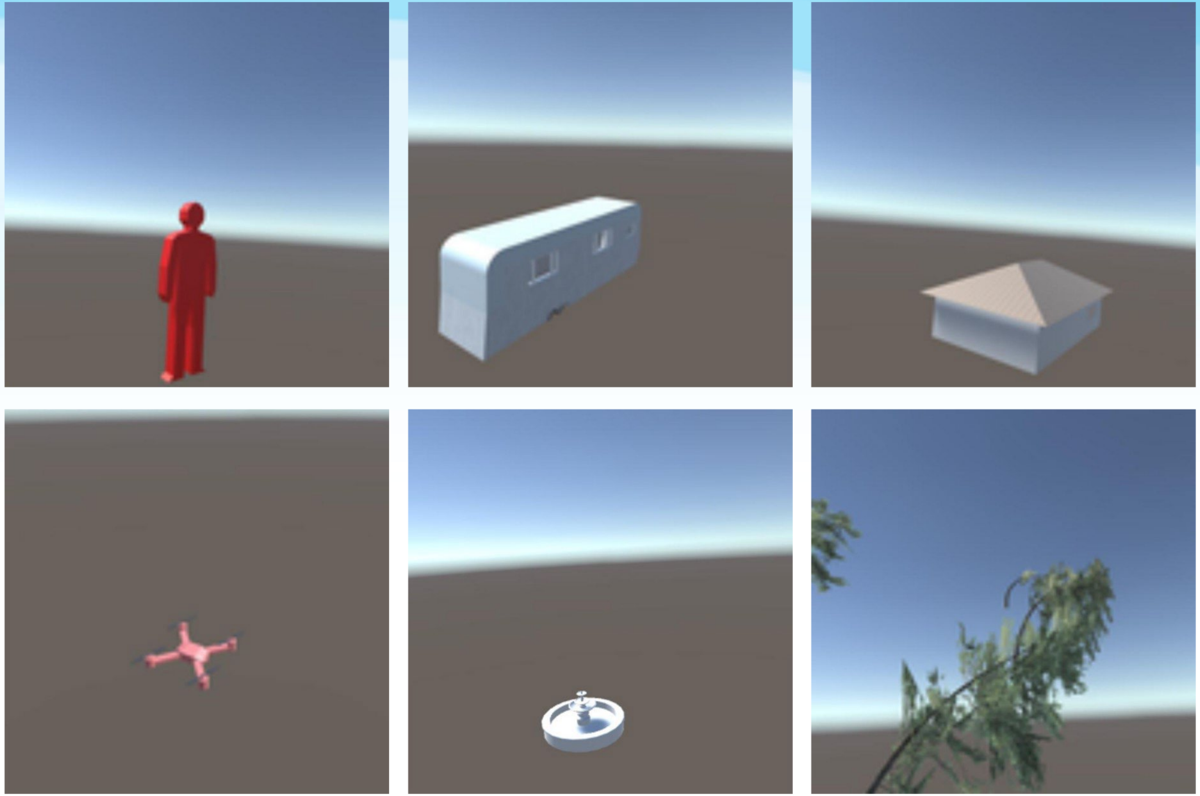
Promotes Reliability and Transparency

Research Objective

- **Why is explainability crucial for ensuring the reliability of AI models in critical applications?**
- **How can we utilize explainability approaches to enhance the reliability of AI models in target detection within dynamic environment simulations?**
- **In what ways does explainability influence the robustness and reliability of AI systems?**

Dataset description

- 6 classes in the dataset
  - Person (445 images)
  - Trailer (849 images)
  - House (803 images)
  - Drone (82 images)
  - Fountain (78 images)
  - Foliage (827 images)
- Total data: ~3100 images
- 94:6 train-test split
  - 23% for training
  - 71% for validation
  - 6% for testing



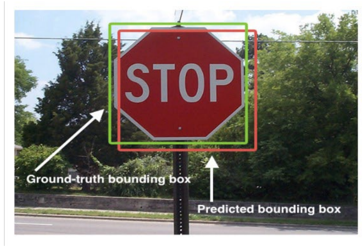
Theoretical overview

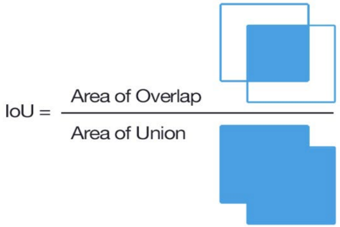
Faster R-CNN (Region-Convolutional Neural Network)

- It is a deep learning model designed for object detection. It combines region proposal generation with object classification and bounding box regression in a unified framework.
- It uses a Region Proposal Network (RPN) to quickly identify regions of interest, which are then passed through a CNN to classify objects and refine their location.

Performance metrics to evaluate AI

- **Precision:** It is calculated using Intersection over Union(IoU). It indicates the overlap of the predicted bounding box coordinates to the ground truth box. Higher IoU indicates the predicted bounding box coordinates closely resemble the ground truth box coordinates.



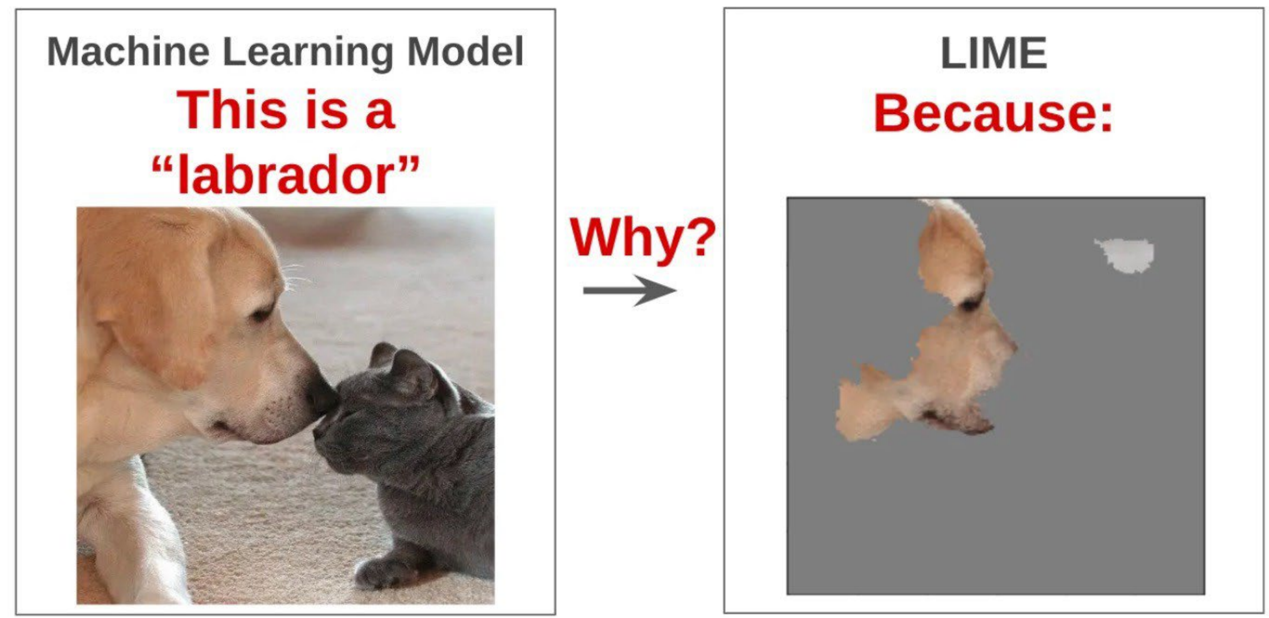


- **Recall:** It is a ratio of True Positive and the sum of True Positive and False Negative.

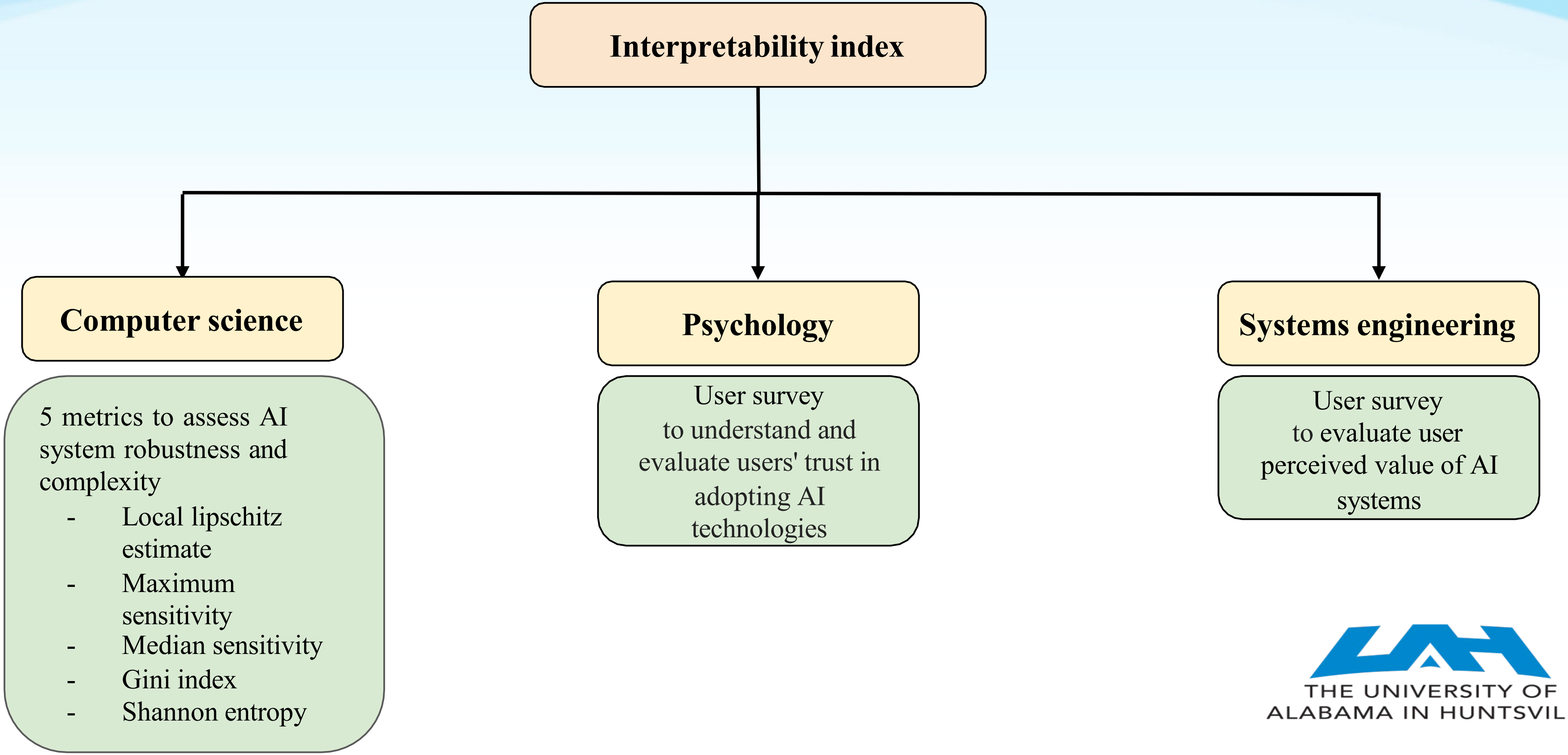
Theoretical overview (Explainability Method)

LIME (Local Interpretable Model-agnostic Explanations)

A model-agnostic approach that explains individual predictions by approximating the model's behavior in the neighborhood of a specific input, identifying key features that most influence the decision in a locally interpretable way.



Theoretical overview



Interpretability index

Computer science

5 metrics to assess AI system robustness and complexity

- Local lipschitz estimate
- Maximum sensitivity
- Median sensitivity
- Gini index
- Shannon entropy

Psychology

User survey to understand and evaluate users' trust in adopting AI technologies

Systems engineering

User survey to evaluate user perceived value of AI systems

Theoretical overview

Interpretability index

- Interpretability index (I-index) is used to give a holistic value that encompasses all the measures from different domains.

- Interpretability index is calculated using weighted softmax function.

$$I - index = \sum_{i=1}^{N} \frac{w_i * e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

Here, ‘w’ is the weight associated with each metric used and ‘x’ is the calculated value.

Theoretical overview

Metrics from computer science perspective utilized to evaluate the quality of explanations

- **"Robustness criteria"**

- It checks how stable the explanation is when the input picture is significantly modified, guaranteeing that the model's classification is not affected.
- Robustness metrics - Local Lipschitz estimate, Median sensitivity and Maximum sensitivity
- Lower score -> more robust (stable) explanation

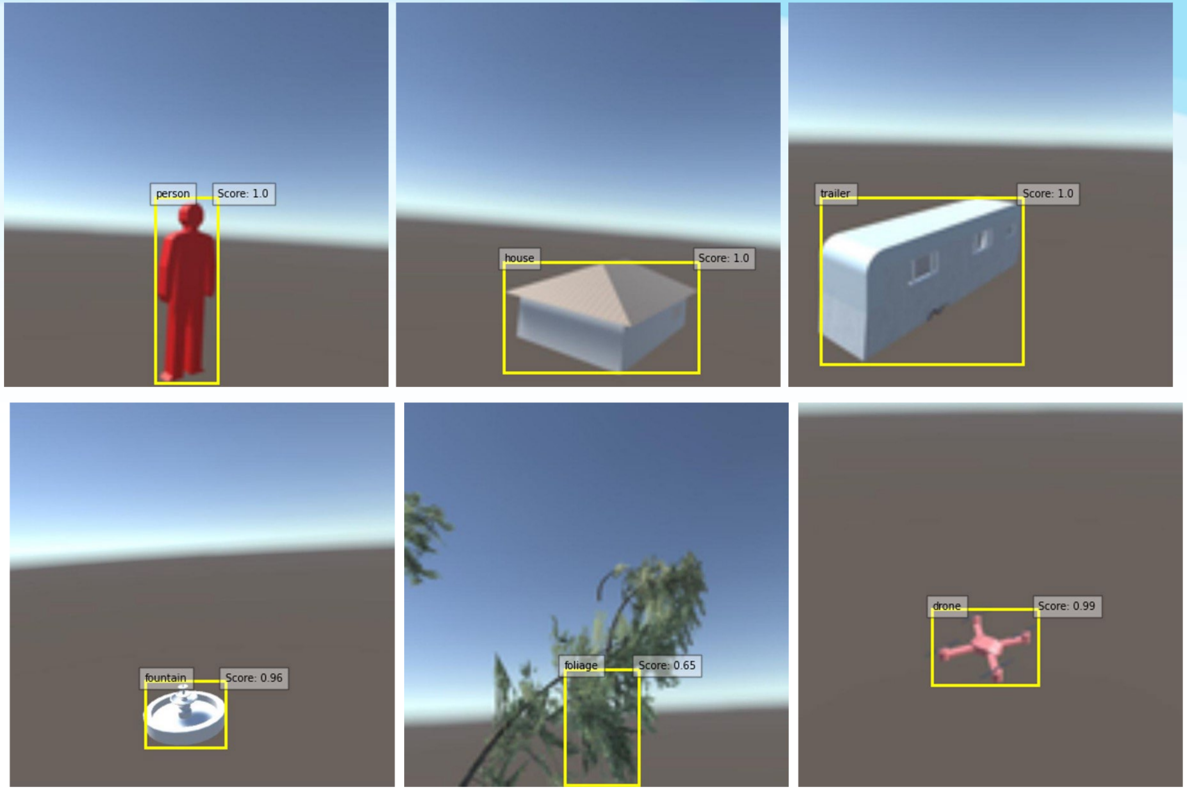
- **"Complexity criteria"**

- Complexity is defined as the measure of interpretability in an explanation, i.e., the measure of how easily a user can simulate and/or understand it.
- Complexity metrics - Shannon-entropy and gini index
- Lower score -> less complex explanation

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

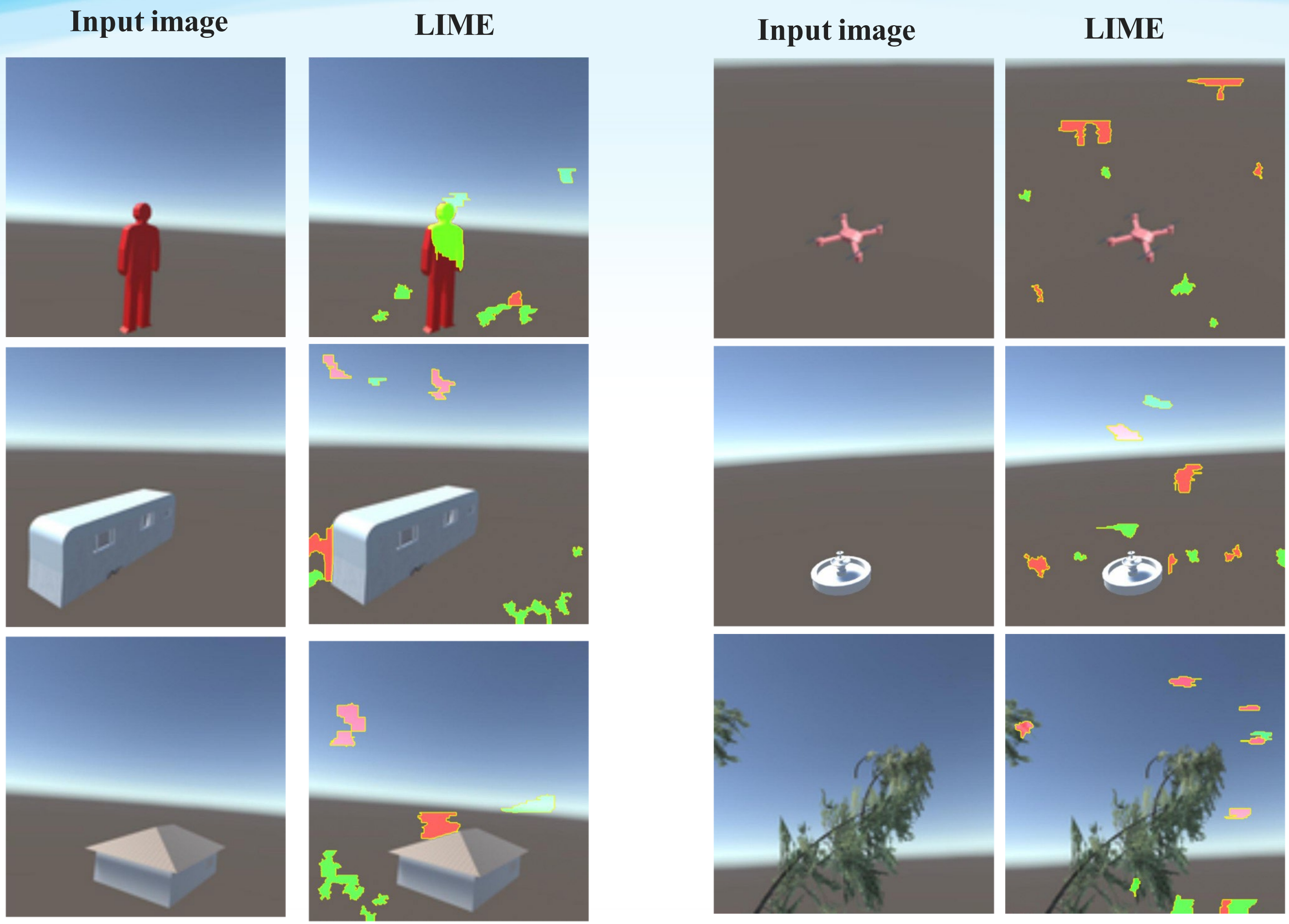
Experimental results

Class	Average precision	Recall
person	0.815	0.858
trailer	0.838	0.885
house	0.808	0.852
fountain	0.694	0.752
drone	0.769	0.825
foliage	0.053	0.261
Overall	0.664	0.742

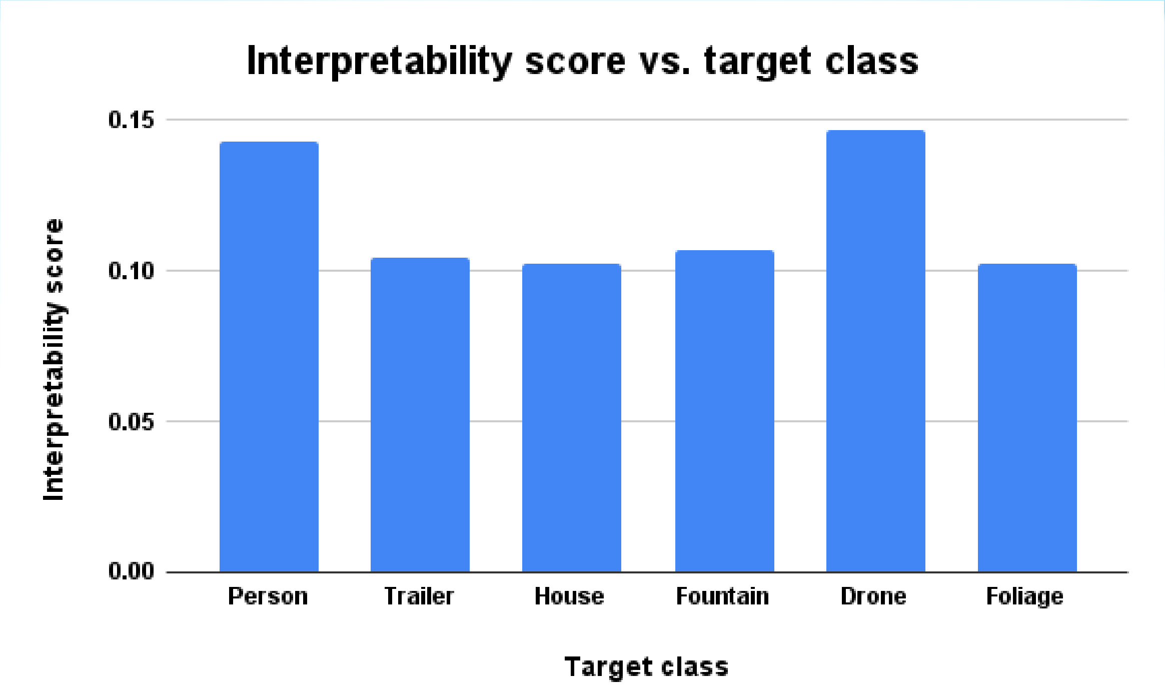


THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

Experimental results (XAI Visual representations)



Experimental results (XAI statistical measure)



The interpretability score plot illustrate the robustness and complexity of explainability method “LIME”. The plot displays the calculated interpretability scores for different target classes used during model training.

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

Summary

- Higher interpretability score -> more robust and less complex explainability method

- The order of interpretability scores is as follows: Drone > Person > Fountain > Trailer > House/Foliage

- "Person" class is the most easily understood or interpretable within the context of a model or analysis. This interpretability is supported by both statistical measures and visual explanations.

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

Conclusion

- Explainability in AI systems aid in detecting errors, improving model reliability and accountability.
- Interpretability scores offer a consistent approach to evaluate how understandable AI models are, bridging the gap between technical performance and human usability.
- Interpretability scores provide a unified way to assess explainability methods used to evaluate AI systems.
- Interpretability scores enhance trust and transparency in AI models.
- Interpretability scores promote ethical AI usage, especially in high-stakes applications where decision-making transparency is critical.

THANK YOU!!

For any questions about our work, please contact the authors at sg0097@uah.edu or vineetha.menon@uah.edu

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE