



National University of Science and Technology POLITEHNICA Bucharest

CHALLENGES IN EXPLAINABLE AI SYSTEMS

PROF. ADINA MAGDA FLOREA

UK – Romania Conference “Exploring our AI Potential”, 28 February 2024

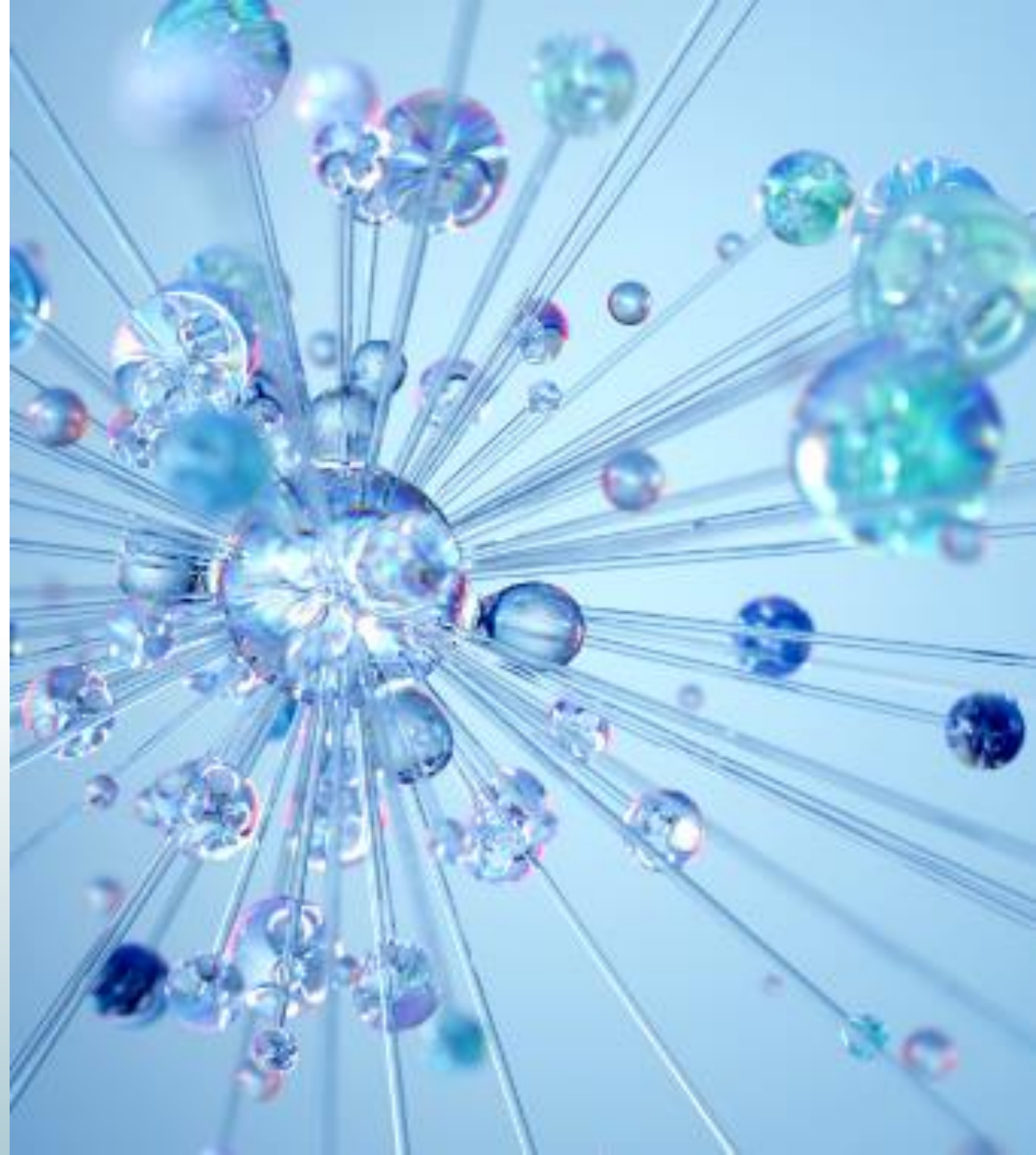


AI

Artificial Intelligence is transforming every aspect of the modern society and will have an ***enormous impact*** in the coming decades

In an age where AI is becoming ubiquitous, the ***human dimension of AI*** applications is a challenge

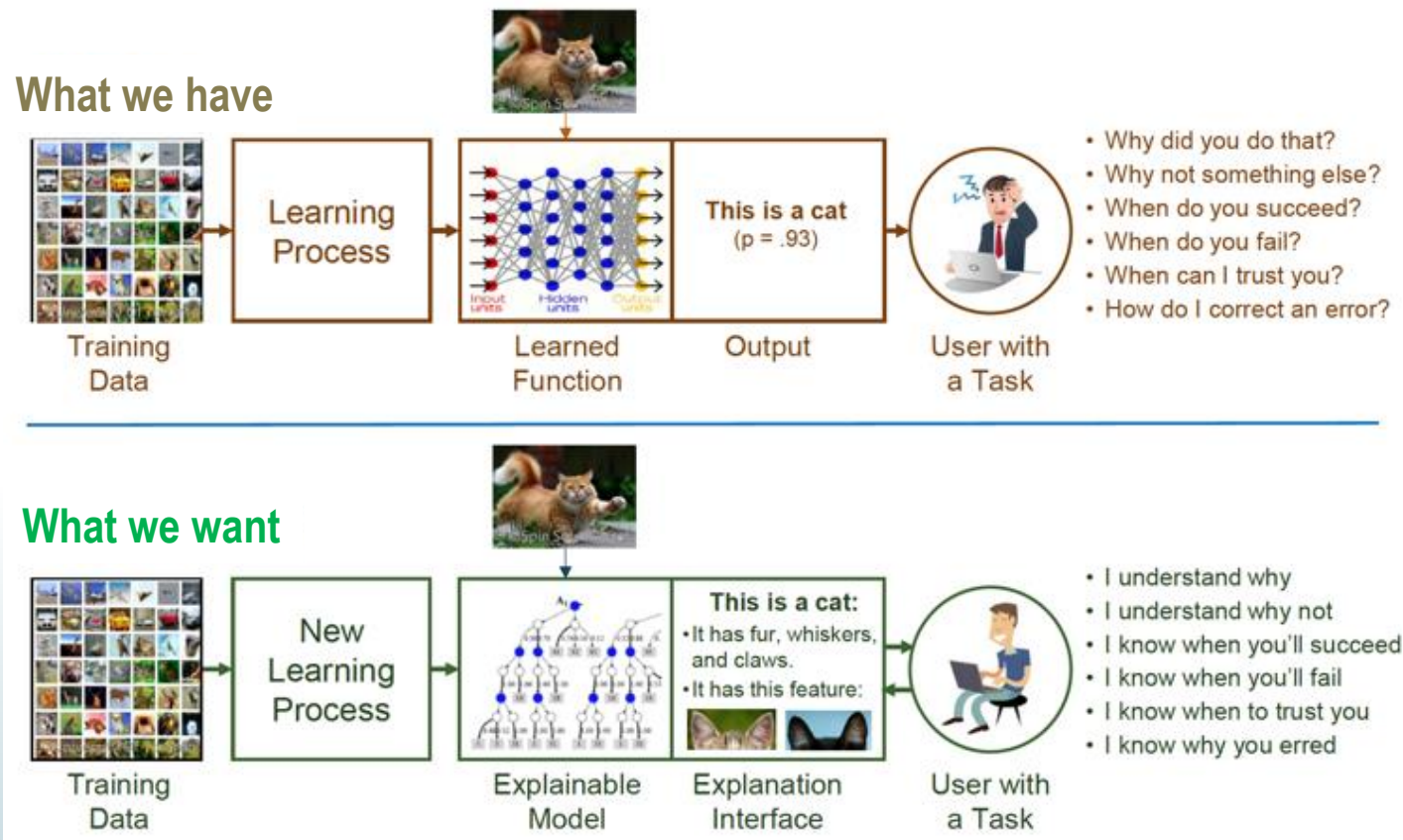
- While the “classic” AI models are easily interpretable, the last years have witnessed the rise of opaque decision systems such as Deep Neural Networks
- The success of Deep Learning stems from a combination of effective learning algorithms with a huge parametric space
- This space comprises hundreds of layers and millions of parameters, which makes DNNs be considered as complex “black-box” models
- There is a ***need to explain the internals of the model in a human interpretable manner***



EXPLAINABLE AI

XAI - Methods and techniques that make AI decisions understandable

- **Interpretability** = describing the model of an AI system in terms accessible to humans
- **Explainability** = explaining the functioning of models to understand reasoning and decision-making → evaluate correctness



(source: DARPA)

WHY XAI?

With the **adoption on a large scale** of DNN in many domains, e.g, healthcare, education, robotics, natural language understanding, economics, autonomous driving, we need XAI:

- **Build Trust** - Transparency in AI decisions
- **Compliance** - Meet regulatory requirements like GDPR - includes the right to explain the decision of an automated process
- **Fairness** - ensure that the ML models satisfy legal and moral standards
- **Debugging and Testing** - Enhances model performance
- **Research** - better understanding of the model and of inductive biases



XAI METHODS

Post hoc explanation methods - inspection of learned features, identification of feature importance and interaction of features, creating visual explanations in the form of saliency maps / heat maps

- **Model-agnostic explanations** - independent of the type of neural network, operating solely on the input and the output of the neural network

By perturbing the input, the user can inspect what are the changes in the output of the neural network.

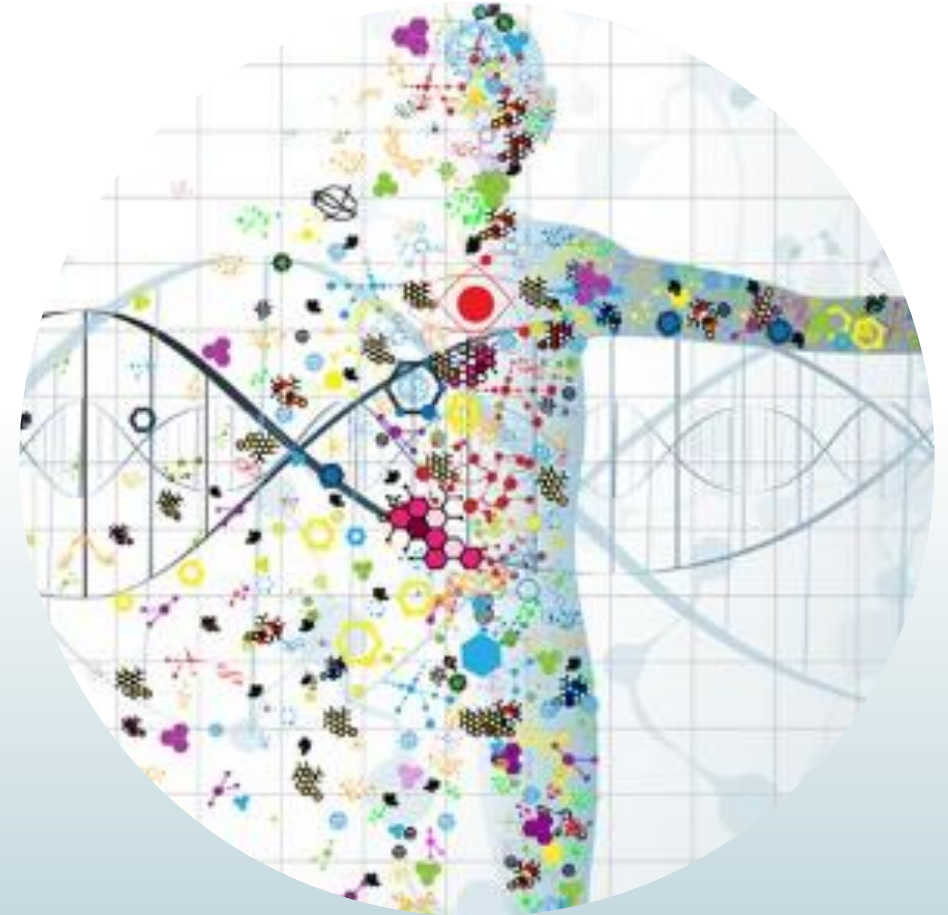
E.g, LIME – approximates locally with an interpretable model, perturbs the input data and observes the changes in predictions to understand how individual features affect the output

XAI METHODS

- **Model-specific explanation** - specific to a particular classes of models, e.g., use attributes that are specific to a type of neural network
 - tailored to provide insights into the functioning and decision-making processes of specific types of machine learning models
 - take advantage of the particular structure and internal mechanics of the models to generate explanations
- **Visual explanations** are commonly the output- visualize and understand what each neuron in a deep network is "looking" at when making a decision
 - E.g., Activation Maximization, Feature Visualization, Class Activation Maps (CAM, Grad-CAM)

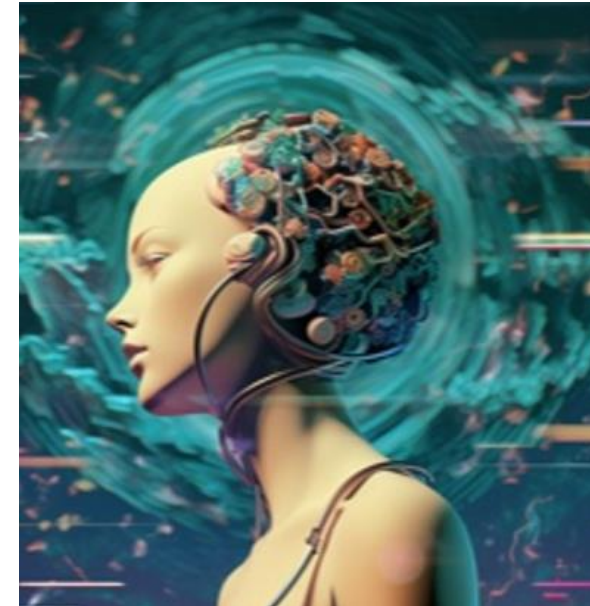
CHALLENGES IN XAI

- Trade-off between accuracy and explainability
- Models explainable by design
- Metrics to evaluate the explanations
- Lack of standardization – difficult to compare different XAI techniques or to establish benchmarks
- Scalability - efficiently providing explanations in real-time or for very large datasets



CHALLENGES IN XAI

- Complexity of explanations
- Different explanations for different stakeholders
- Correlate explainability with interpretability of the model
- Link between causality and XAI, e.g., image analysis typically consists of correlations rather than causation, while causality describes the relation between cause and effect



International Center of Excellence in Artificial Intelligence at POLITEHNICA Bucharest

Under the patronage of the National Commission of Romania for UNESCO



INTERNATIONAL CENTER OF
EXCELLENCE IN ARTIFICIAL
INTELLIGENCE

AIMAS LABORATORY

- Multi-agent systems
- Deep learning
- Ambient assisted living
- Social robots
- Computer Vision
- Context aware computing



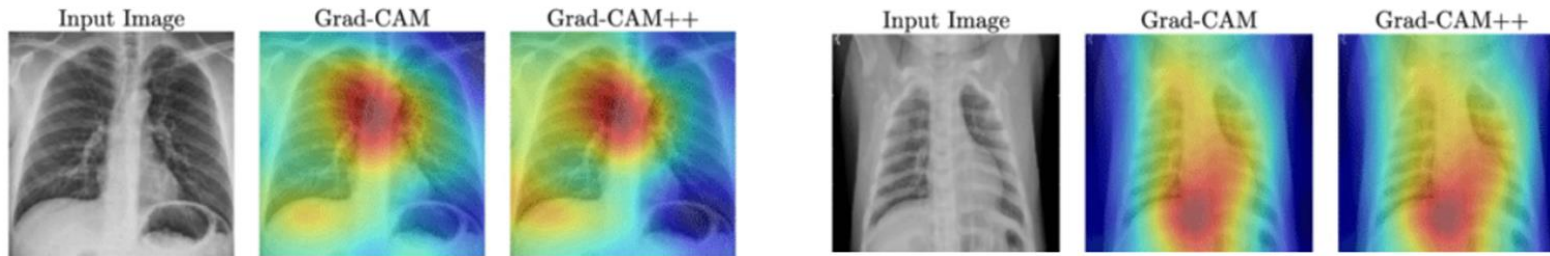
AIMAS LABORATORY



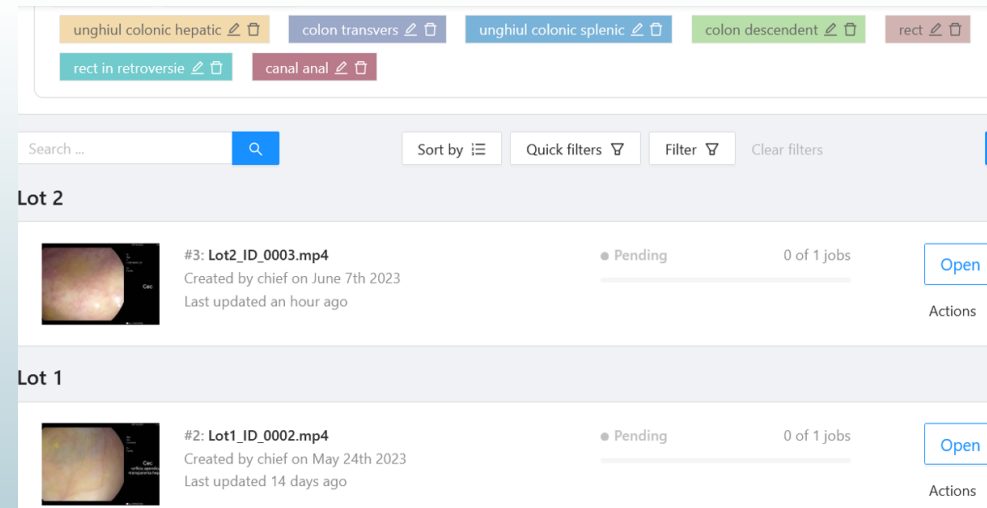
CNCC - Centrul Național de Competență în domeniul Cancerului

- Explanation by visualization in medical imaging

PNRR Project, 2023-2025



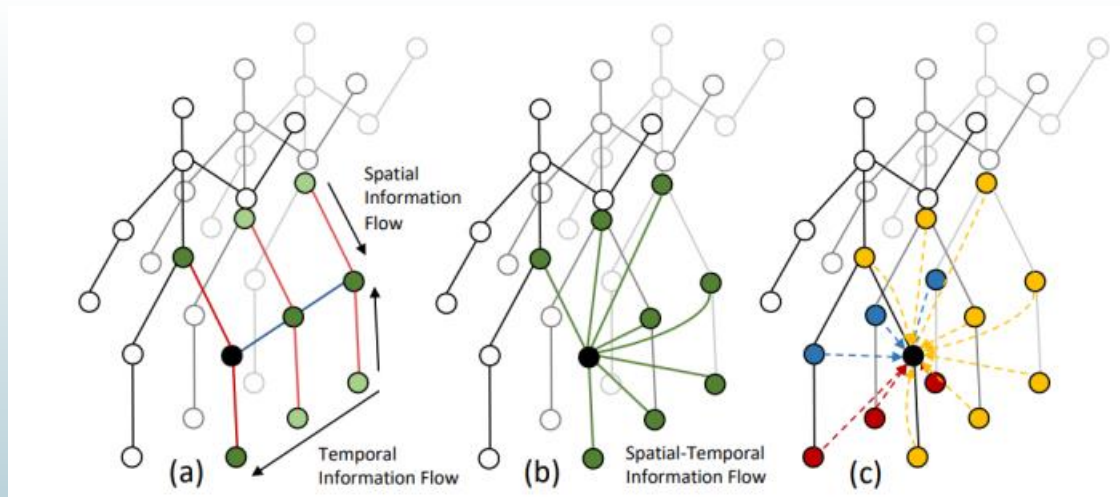
- Quality of a colonoscopy – Challenge – justify the decision (Mihai Trăscău)



- Based on our previous research results in HAR - a neural model with multiple input branches based on TCN and GCN layers (Mihai Nan)

Spatio-Temporal NN with Handcrafted Features

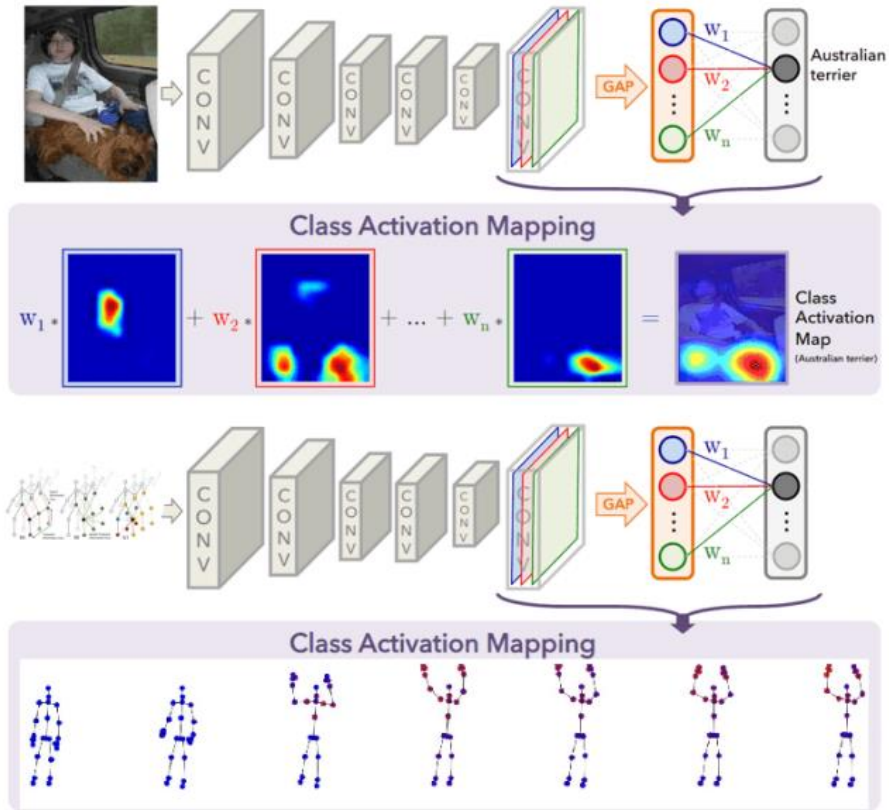
- Explainability of HAR - Allows finding an activation map for explaining the recognized action (Mihai Nan)



- Use the features resulting from the application of the spatio-temporal module and the weights of the last two linear layers
- Determine for each frame which are the most important joints considered by the network in terms of activations
- Pictured them in an Activation Map

Adapting the CAM Method for HAR

(Mihai Nan)



• Explanation of the prediction – *brushing teeth*



Our challenges in ENFIELD

- **Explainability and interpretability of spatio-temporal models for decision making**

Collaboration with Technical University of Eindhoven, Netherlands, Institute for Systems and Computer Engineering, Technology and Science, Portugal

- **Extracting Explanation Symbols from Visual Data**

Collaboration with Technical University of Eindhoven, Netherlands

- **Interpretable Decision Making in the Medical Domain**

Collaboration with Technical University of Chemnitz, Germany

AIMAS LABORATORY

Visualization for XAI image recognition - Evaluating different methods for generating saliency maps for target class golden retriever (Andrei Dugăescu)



Grad-CAM



Grad-CAM++



Eigen-CAM



Eigen Grad-CAM



Eigen Layer-CAM



XGrad-CAM



Score-CAM



Ablation-CAM



Layer-CAM



Full Grad-CAM

AIMAS LABORATORY

XAI - Influence of adversarial examples in image recognition (Andrei Dugăescu)

Real inputs

(a) Original Image.
Prediction: **Jay**

(b) Explanation Jay
(Original input)

(c) Explanation Black Grouse
(Original input)

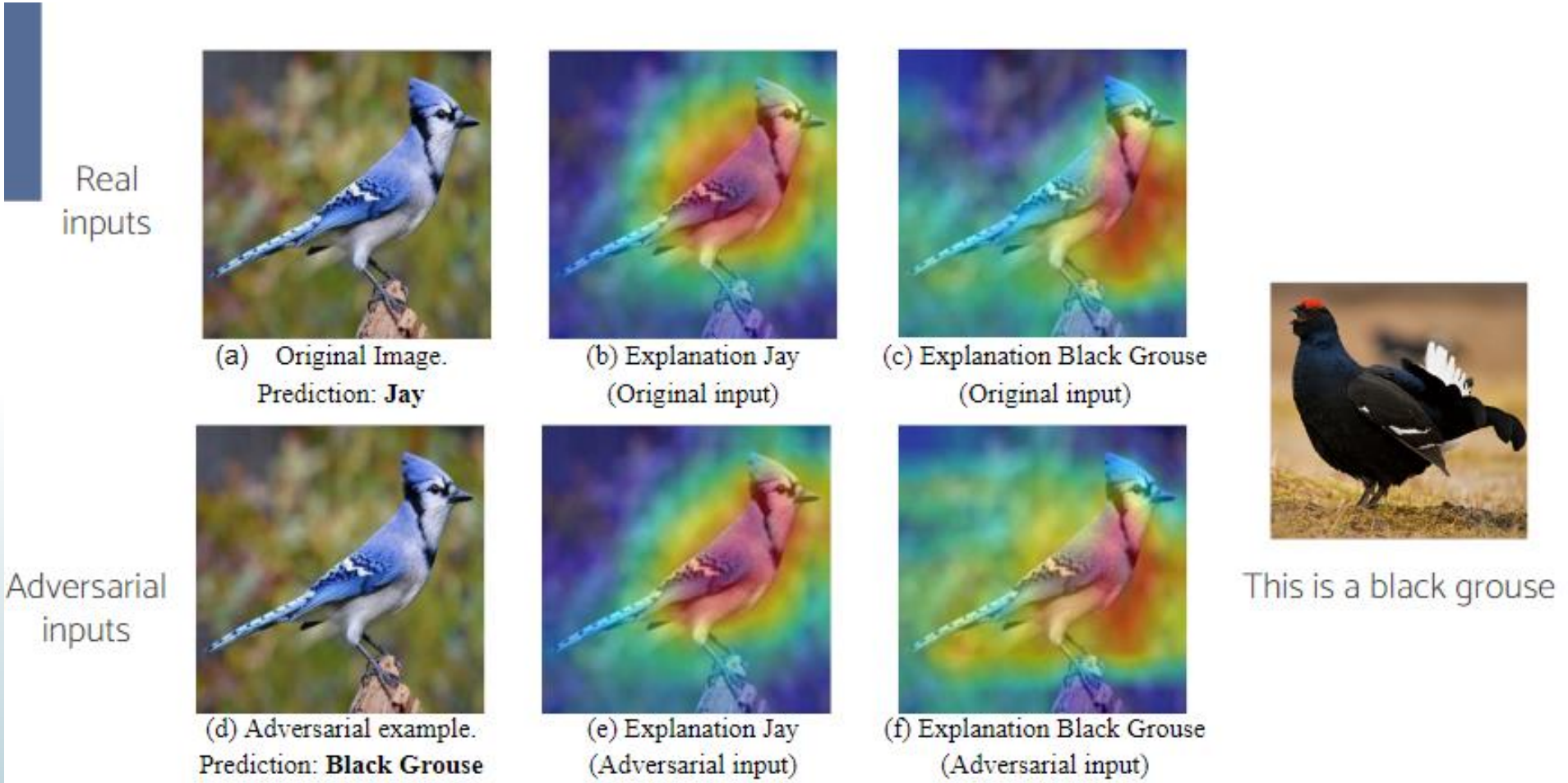
Adversarial inputs

(d) Adversarial example.
Prediction: **Black Grouse**

(e) Explanation Jay
(Adversarial input)

(f) Explanation Black Grouse
(Adversarial input)

This is a black grouse



AIMAS TEAM



Adina Magda Florea



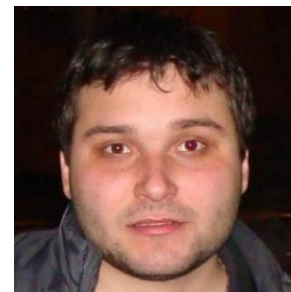
Irina Mocanu



Andrei Olaru



Alexandru Sorici



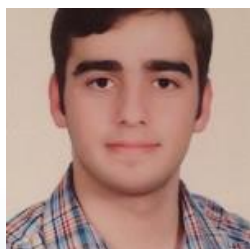
Mihai Trăscău



Ștefania Ghiță



Mihai Nan



Alex Awada



Dan Novischi



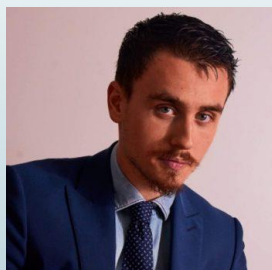
David Iancu



Andrei Dugăsescu



Ana Maria Simion



Teodor Poncu



Vlad Florea



Florin Dumitrescu



Bogdan Ceachi



Andrei Mihalea



THANK YOU

CONTACT

aimas.cs.pub.ro
info@aimas.pub.ro

