XAI : Explainable Artificial Intelligence

Turning Black Box into Glass Box

Benyamin Ghahremani Nejad, Matin Hossein Pour

Spring 2021

University of Birjand

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

Let’s classify Huskies and Wolves

Introduction

XAI
Benyamin Ghahremani Nejad & Matin Hossein Pour

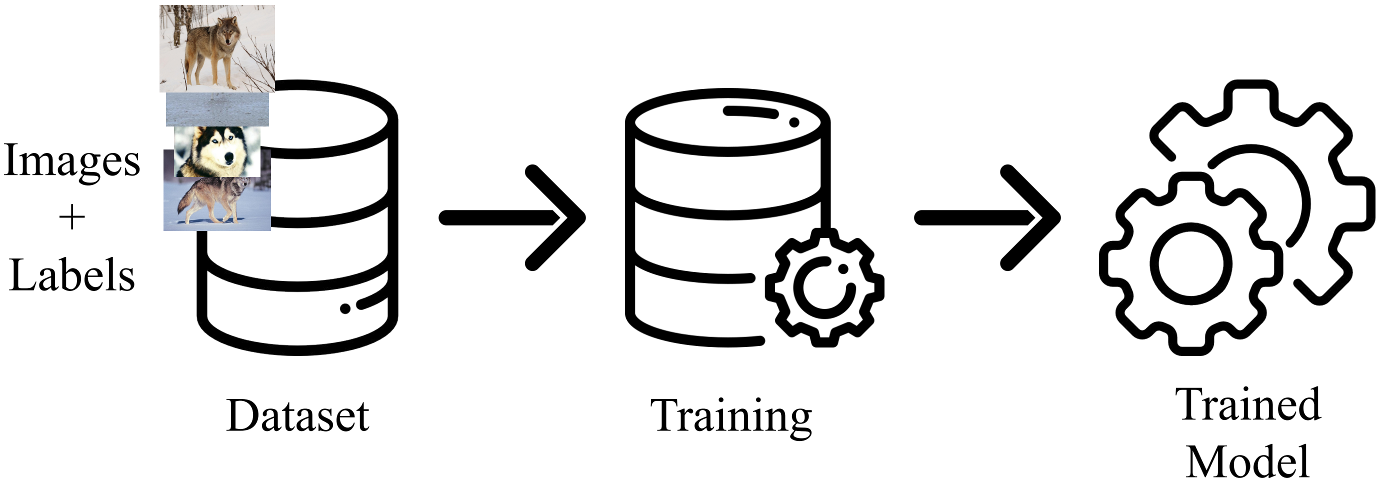
Let's classify Huskies and Wolves

Classify Huskies and Wolves

Differences?

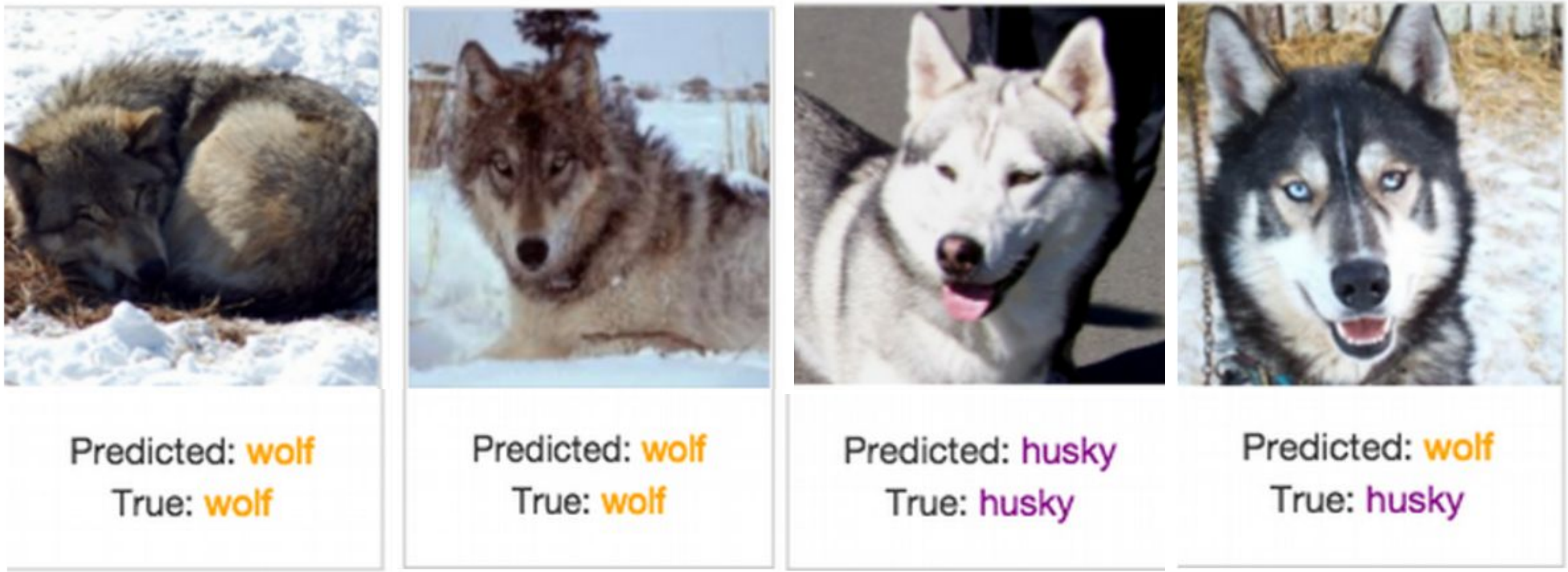
Classify Huskies and Wolves

Building a Classification Model



Test The Model

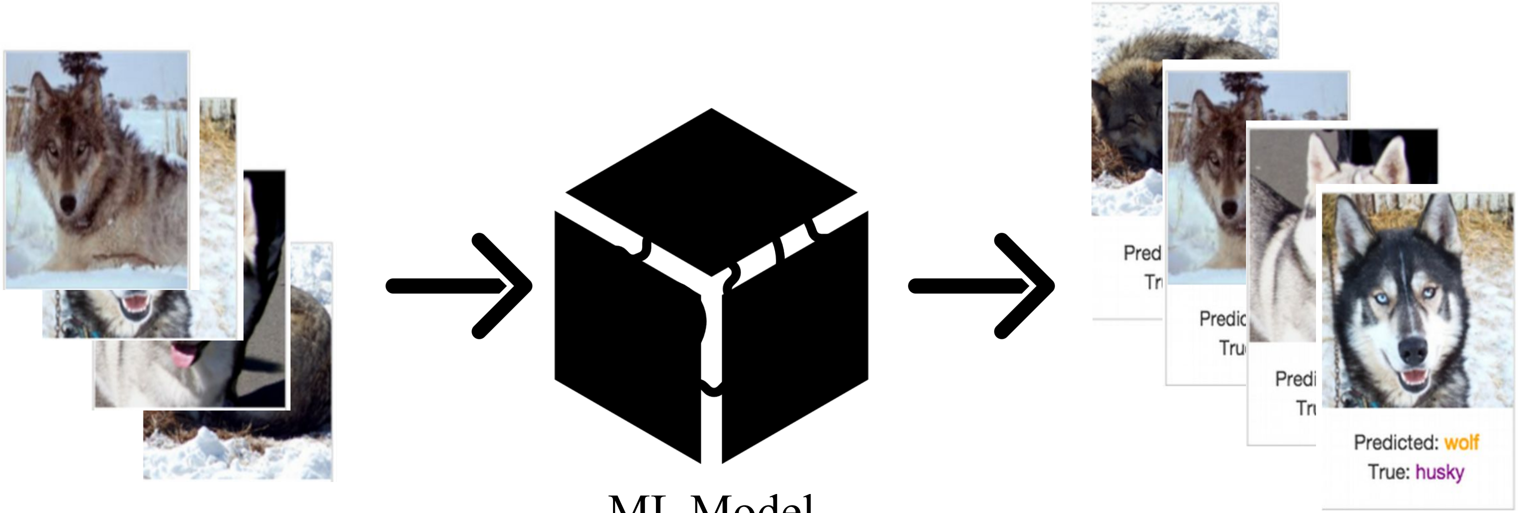
Which is a Husky and which is a Wolf ?



Test The Model

Benyamin Ghahremani Nejad & Matin Hossein Pour

How does the model reach its decisions?



Input

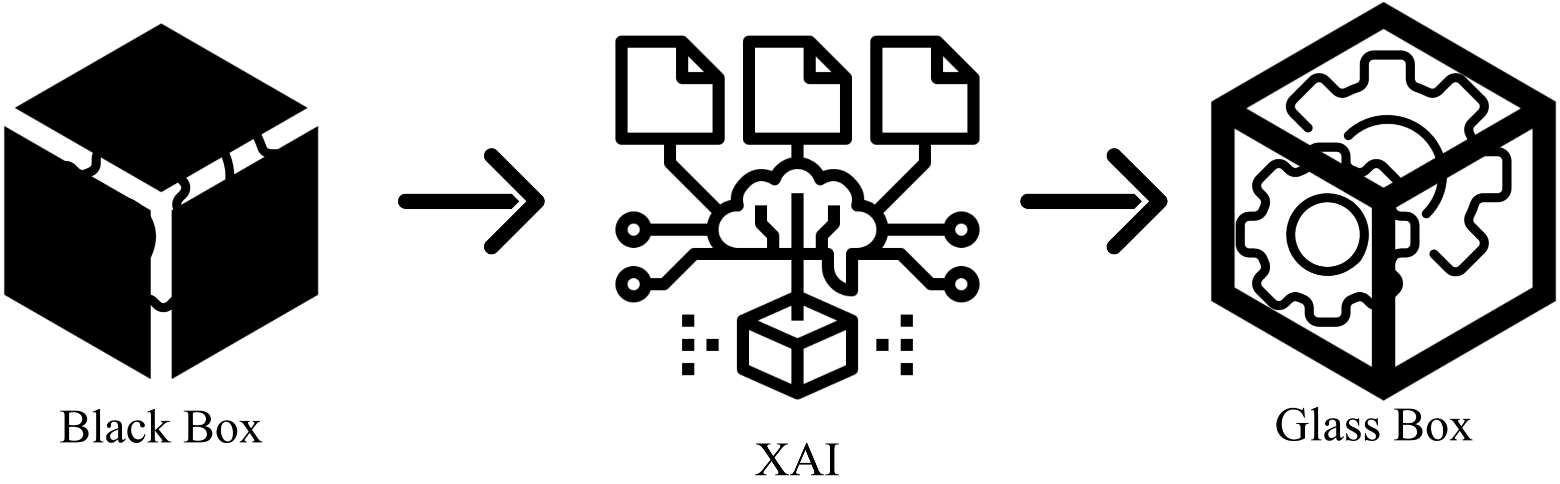
ML Model

Output

How Does The Model Reach Its Decisions?

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour



Definition

Explainable AI (XAI) has developed as a subfield of AI, focused on exposing complex AI models to humans in a systematic and interpretable manner.

An explanation is the answer to a why-question (Miller 2017).

- Why did you reach the decision?
- Why is there a 95% accuracy?
- What is the reason behind it?

Categories & Methods

Two Main Categories

Intrinsic Interpretable Models

These methods are intrinsically interpretable ML models by their own structures.

Post-Hoc Explanation

In contrast to the interpretable methods, we need some standalone algorithms to explain the BB ML/DL methods

Intrinsic Interpretable Methods

- Logistic/ Linear Regression
- Decision Trees
- K-Nearest Neighbors
- Rule-base Learners
- Bayesian Models

Post-Hoc Explanation

	Model-Specific	Model-Agnostic
Local	Use attention mechanisms to show how the model selectively focuses on features in high-dimensional input for an instance	Develop interpretable surrogate models with local fidelity in the vicinity of an instance
Global	Enforce interpretability constraints into the structure and learning mechanisms of deep learning models	Develop interpretable global surrogate models based on input-output associations predicted by a black-box model

Post-Hoc Explanation Methods

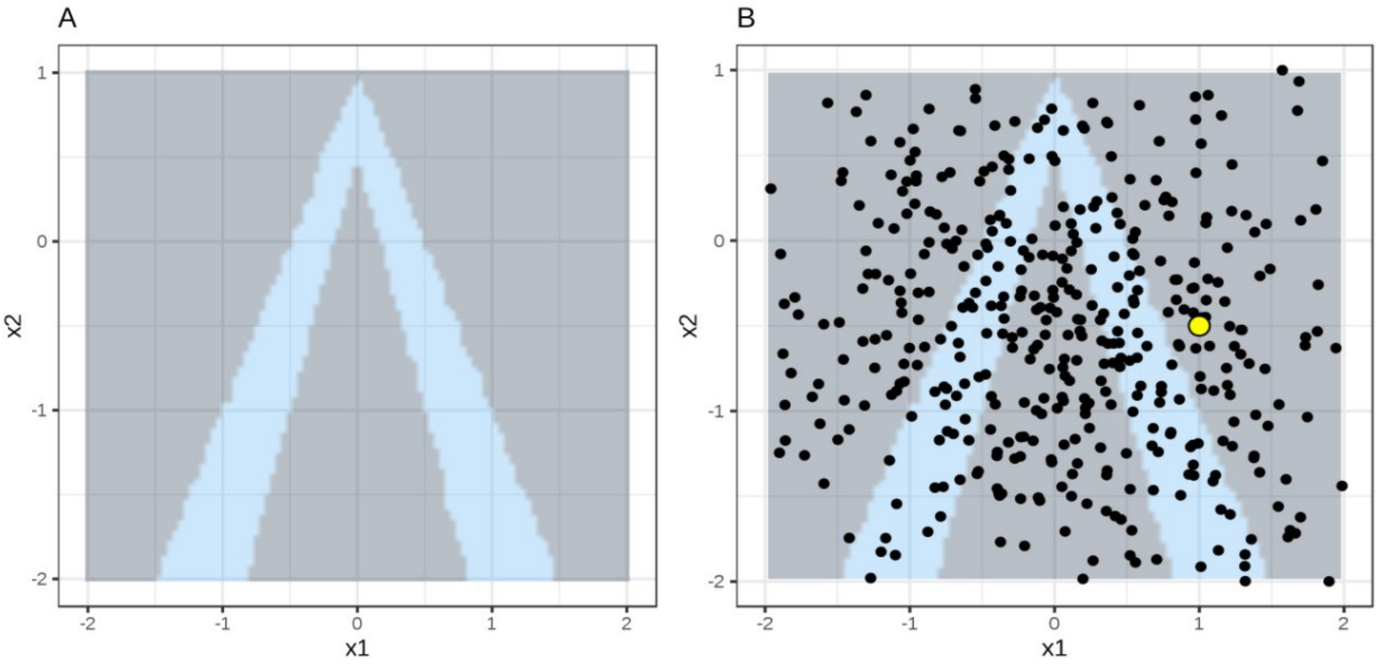
	Model-Specific	Model-Agnostic
Local	• Attention-based Models	• LIME • SHAP • Anchor
Global	• Intrinsically or Inherently Interpretable Models	• Global Surrogate Models • PDP • ICE

LIME

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

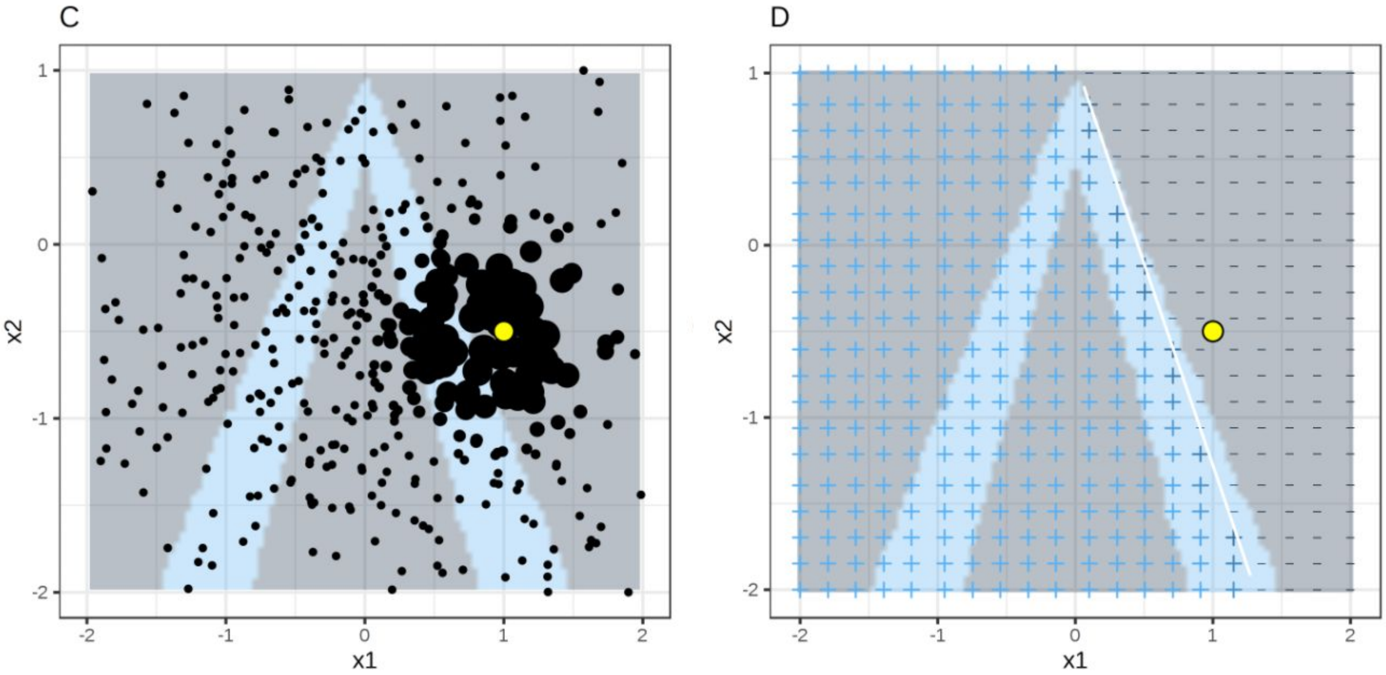
Local Interpretable Model-Agnostic Explanations



LIME

Benyamin Ghahremani Nejad & Matin Hossein Pour

Local Interpretable Model-Agnostic Explanations



LIME

Local Interpretable Model-Agnostic Explanations

LIME algorithm for tabular data. A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark) or 0 (light). B) Instance of interest (big dot) and data sampled from a normal distribution (small dots). C) Assign higher weight to points near the instance of interest. D) Signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).

LIME

Local Interpretable Model-Agnostic Explanations

Our test Instance

Complex Model

Model Complexity

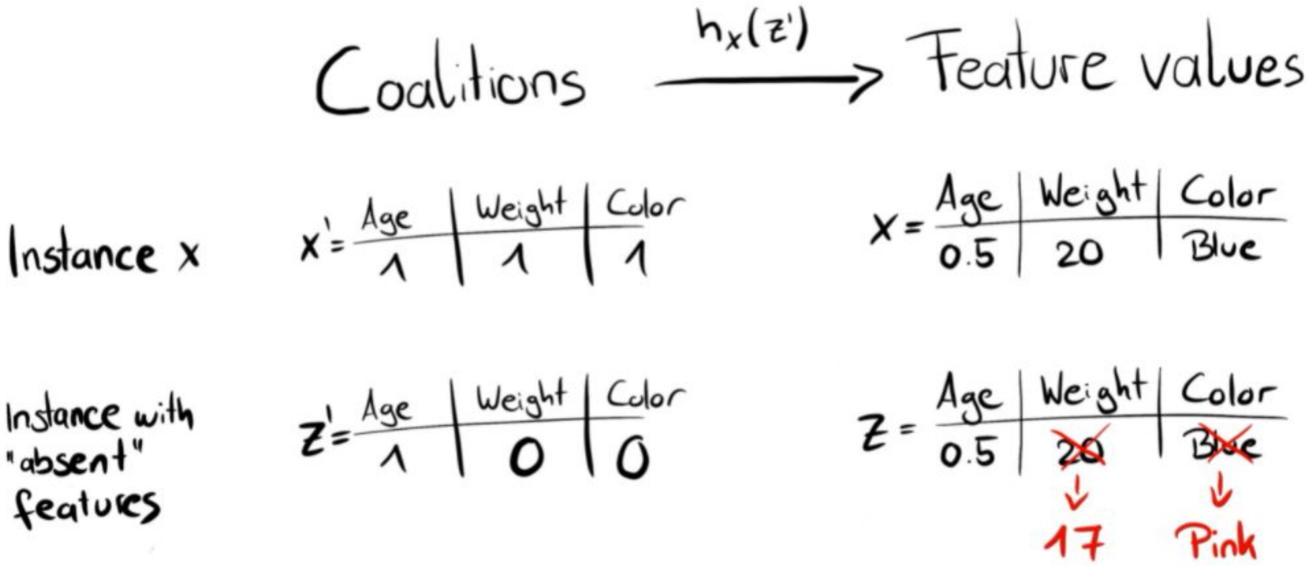
$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

g is a model from intrinsic interpretable family models G

Proximity Measure

SHAP

SHapley Additive exPlanation



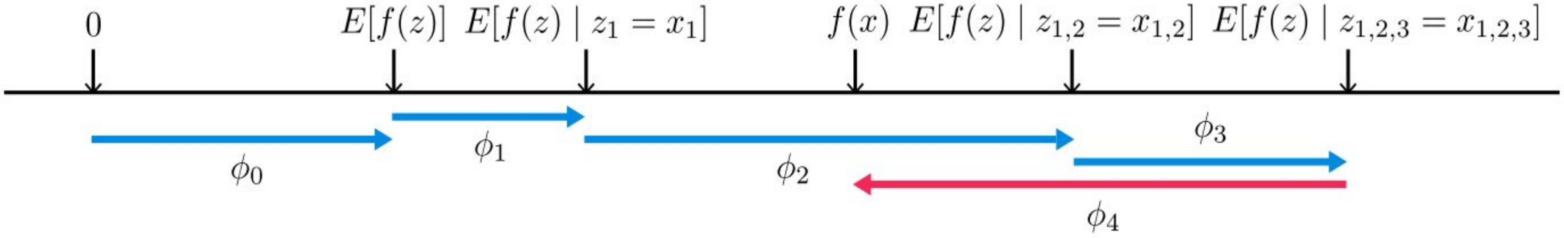
SHAP

SHapley Additive exPlanation

Function hx maps a coalition to a valid instance. For present features (1), hx maps to the feature values of x. For absent features (0),hx maps to the values of a randomly sampled data instance.

In the coalition vector, an entry of 1 means that the corresponding feature value is "present" and 0 that it is "absent".

SHapley Additive exPlanation



SHAP

XAI

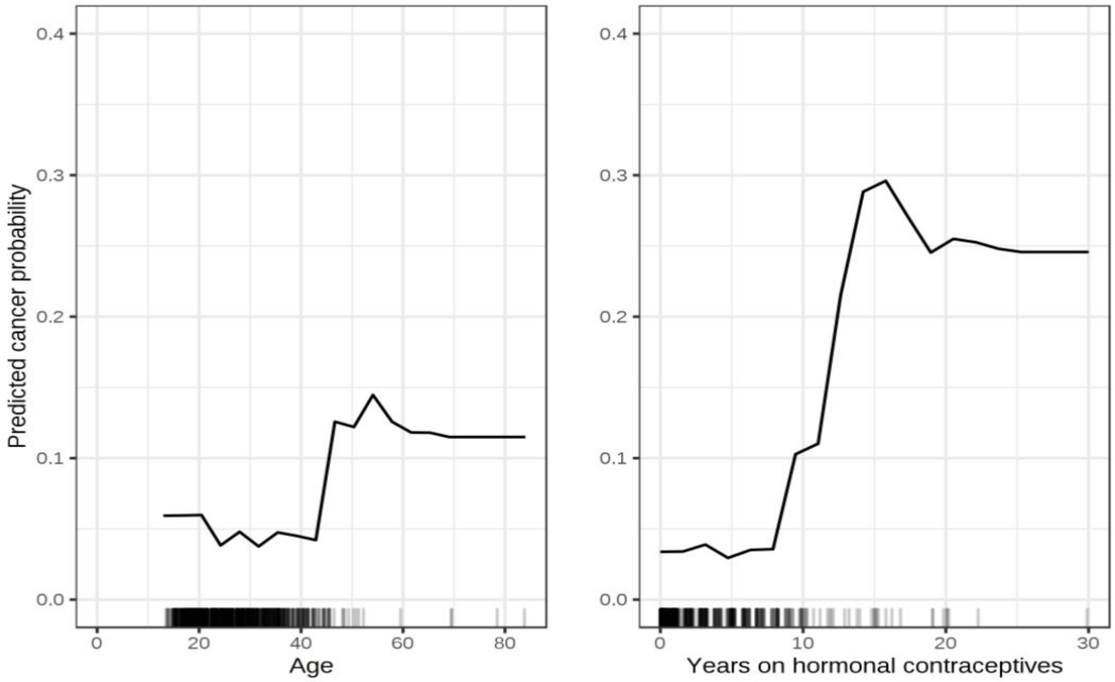
Benyamin Ghahremani Nejad & Matin Hossein Pour

SHapley Additive exPlanation

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

PDP

Partial Dependence Plot



PDP

XAI

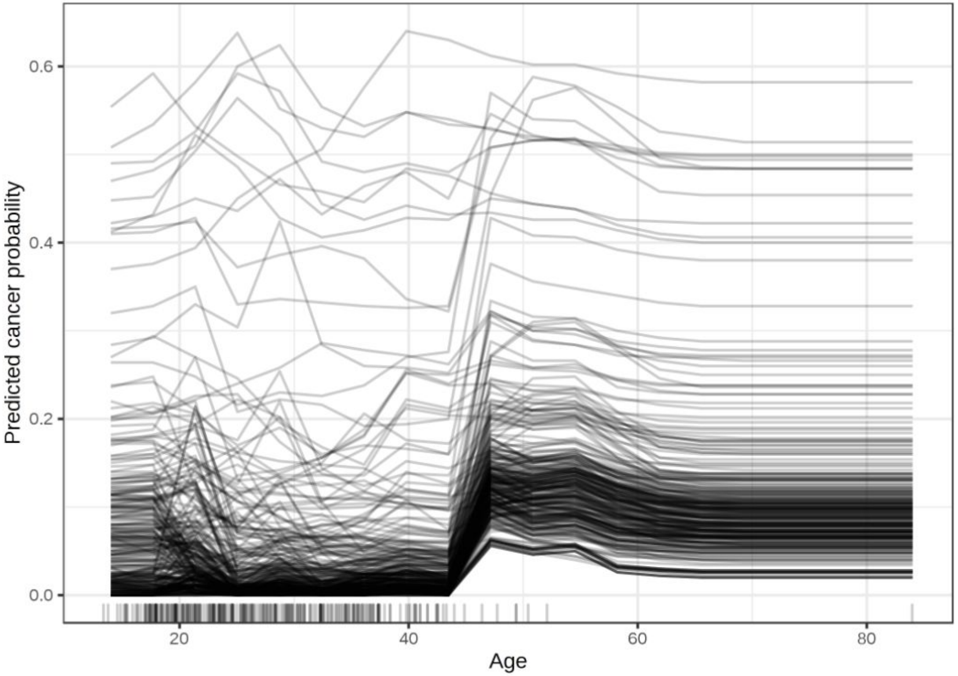
Benyamin Ghahremani Nejad & Matin Hossein Pour

Partial Dependence Plot

PDPs of cancer probability based on age and years with hormonal contraceptives. For age, the PDP shows that the probability is low until 40 and increases after. The more years on hormonal contraceptives the higher the predicted cancer risk, especially after 10 years. For both features not many data points with large values were available, so the PD estimates are less reliable in those regions.

ICE

Individual Conditional Expectation



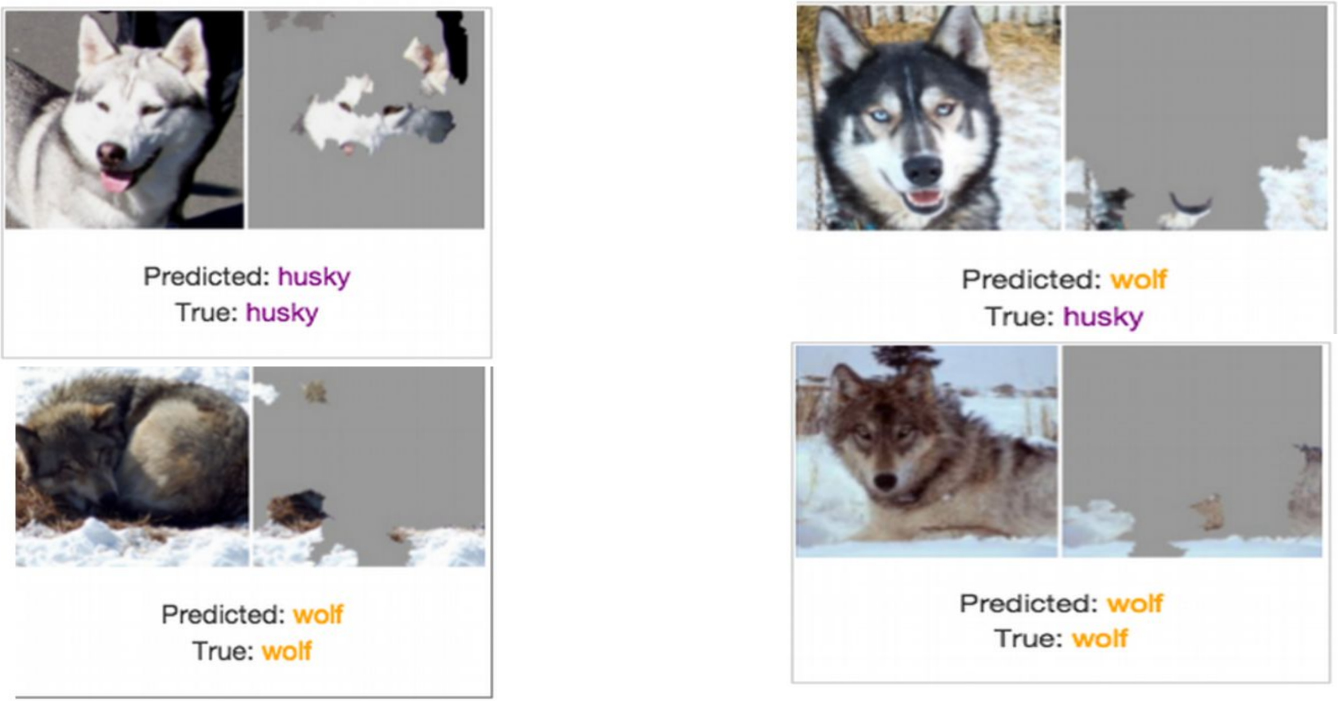
ICE

Individual Conditional Expectation

ICE plot of cervical cancer probability by age. Each line represents one woman. For most women there is an increase in predicted cancer probability with increasing age. For some women with a predicted cancer probability above 0.4, the prediction does not change much at higher age.

Importance

Let’s get back into the husky and wolf classification example



Importance

Do you trust the model?

When should I trust the model?

Who is responsible for the decisions?

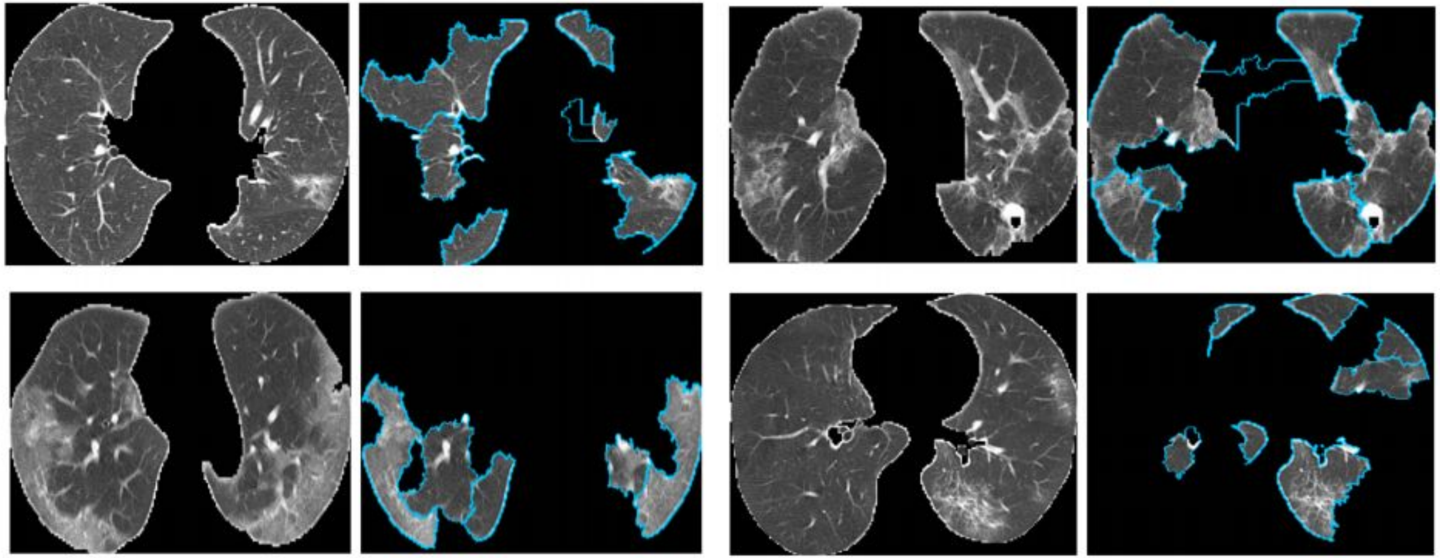
Critical Situations

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

Medical Diagnosis

Covid-19 Detection Through Chest X-Ray



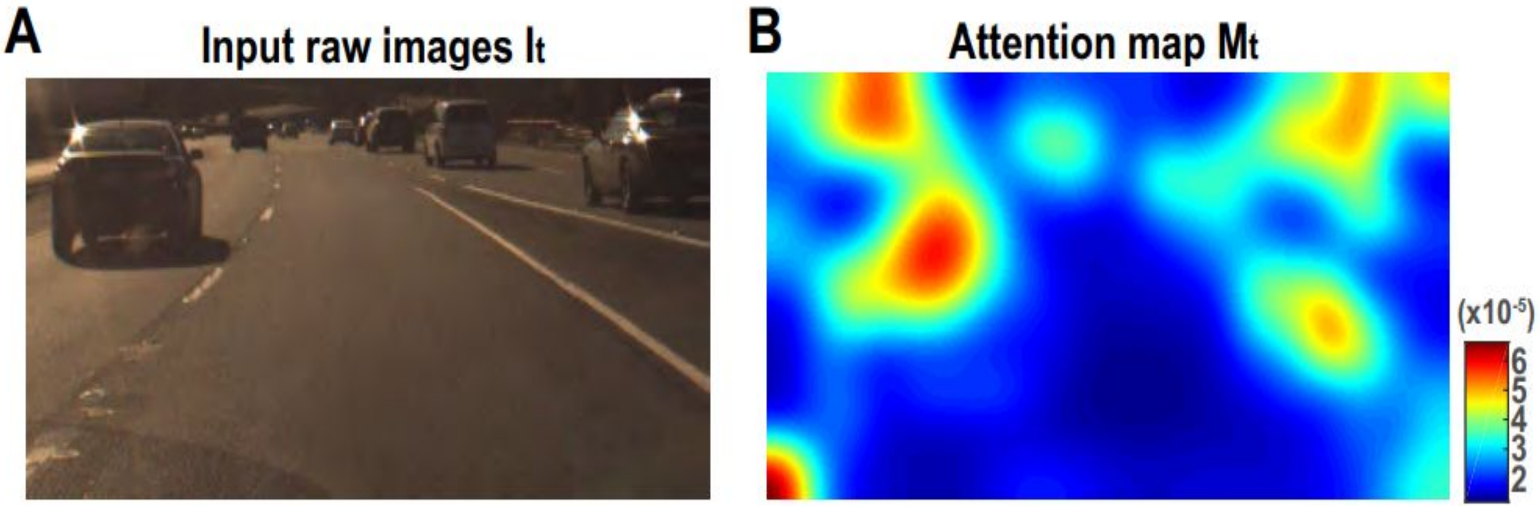
Critical Situations

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

Autonomous Driving

Object Detection



Importance

Improvement of The System

The first step towards improving an AI system is to understand it’s weaknesses.

Learning from The System

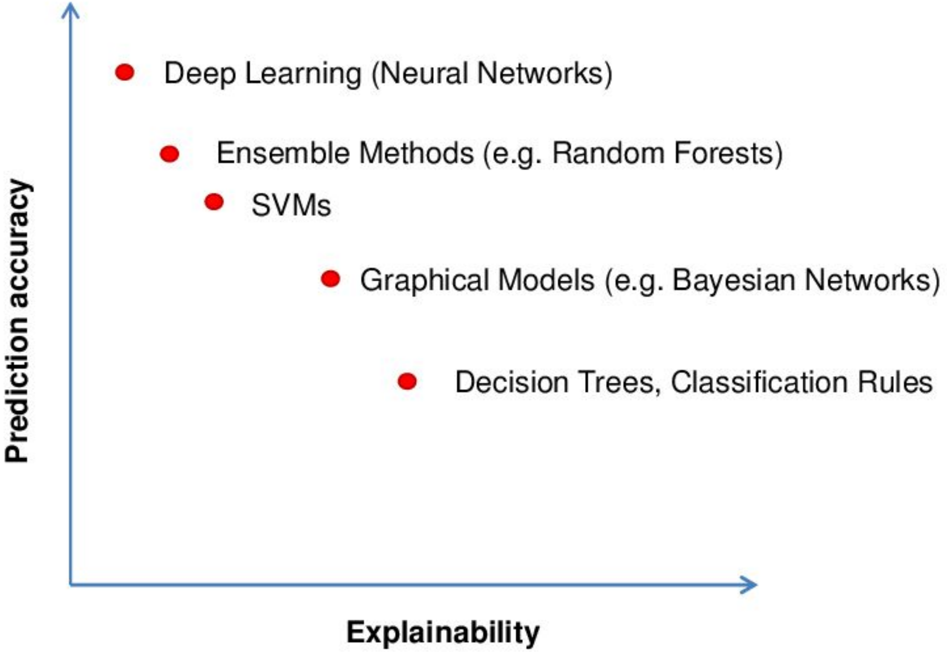
Extract knowledge from the vast amount of data and crealation between them

Compliance to Legislation

legal aspects have recently received increased attention.

Challenges

Performance vs. Explainability



Challenges

Evaluation

Are all models in all defined-to-be-interpretable model classes equally interpretable?

HUMAN-LIKE EXPLANATIONS

Future Works

Ethical AI

Responsible AI

AGI

References

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

https://christophm.github.io/interpretable-ml-book/

Adadi, A. and M. Berrada (2018). "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6: 52138-52160.

Bach, S., et al. (2015). "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." PLOS ONE 10(7): e0130140.

Barredo Arrieta, A., et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58: 82-115.

Došilović, F. K., et al. (2018). Explainable artificial intelligence: A survey. 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE.

References

Galhotra, S., et al. (2021). "Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals." arXiv pre-print server.

Gunning, D. and D. Aha (2019). "DARPA's explainable artificial intelligence (XAI) program." AI Magazine 40(2): 44-58.

Holzinger, A. (2021). "Explainable AI and Multi-Modal Causability in Medicine." i-com 19(3): 171-179.

Holzinger, A., et al. (2017). "What do we need to build explainable AI systems for the medical domain?" arXiv preprint arXiv:1712.09923.

Lundberg, S. and S.-I. Lee (2017). "A unified approach to interpreting model predictions." arXiv preprint arXiv:1705.07874.

Marco, et al. (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." arXiv pre-print server.

References

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

Mohseni, S., et al. (2018). "A human-grounded evaluation benchmark for local explanations of machine learning." arXiv preprint arXiv:1801.05075.

Rai, A. (2020). "Explainable AI: from black box to glass box." Journal of the Academy of Marketing Science 48(1): 137-141.

Ribeiro, M. T., et al. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence.

Samek, W., et al. (2021). "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications." Proceedings of the IEEE 109(3): 247-278.

Samek, W., et al. (2017). "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." arXiv pre-print server.

Sheikh, et al. (2021). "Explainable Artificial Intelligence Approaches: A Survey." arXiv pre-print server.

References

XAI

Benyamin Ghahremani Nejad & Matin Hossein Pour

Introduction

Categories & Methods

Importance

Challenges & Future Works

References

Stiglic, G., et al. (2020). "Interpretability of machine learning‐based prediction models in healthcare." WIREs Data Mining and Knowledge Discovery 10(5).

Sturm, I., et al. (2016). "Interpretable deep neural networks for single-trial EEG classification." Journal of Neuroscience Methods 274: 141-145.

Tjoa, E. and C. Guan (2020). "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI." IEEE Transactions on Neural Networks and Learning Systems: 1-21.

Van Der Waa, J., et al. (2021). "Evaluating XAI: A comparison of rule-based and example-based explanations." Artificial Intelligence 291: 103404.

Wan, A., et al. (2021). "NBDT: Neural-Backed Decision Trees." arXiv pre-print server.

Ye, Q., et al. (2021). "Explainable AI For COVID-19 CT Classifiers: An Initial Comparison Study." arXiv preprint arXiv:2104.14506.

THANKS!

Do you have any questions?

Benyamin Ghahremani Nejad
University of Birjand

Github: github.com/BenyaminGhN
Linkedin: benyamin-ghahremani-nejad
Email: benjaminghahremani@gmail.com
Telegram: @BenyaminGhahremani

Matin Hossein Pour
University of Birjand

Email: matin192hp@gmail.com
Twitter: @MHoseeinpour
Telegram: @matin_hp_192
Linkedin: matin-hosseinpour