# Model-Agnostic XAI Models: Benefits, Limitations and Research Directions

Author: Mikolaj Knap Supervisor: Chhagan Lal Responsible Professor: Mauro Conti

TU Delft

## Contributions

- A detailed examination into the inner workings of 5 model-agnostic XAI techniques, as well as their inherent advantages and disadvantages
- A comparison study between the investigated XAI techniques using a list of pertinent metrics gathered from literature
- Future proposals for potential areas of improvement for the evaluated XAI techniques, and what directions future research should take when extending these models
- General future research directions proposed for the XAI research field

## Background Information

- The ever increasing presence of AI/ML algorithms in sensitive and safety-critical fields has spurred a massive amount of research into the field of explainable artificial intelligence (XAI) models
- These XAI model's aim to introduce explainability into these black-box AI/ML systems, therefore providing an element of accountability into the actions of an AI/ML system
- The model-agnostic category of XAI techniques, allows for the generation of explanations behind the predictions of any ML/AI system regardless of the internal structure of the system

## Research Questions

1. What are the current limitations and benefits of state-of-the-art model-agnostic XAI techniques?
2. What metrics can be used to compare the current state-of-the-art models?
3. How do the investigated XAI models perform on a broad evaluation against this series of metrics?
4. What future research directions should be considered to improve and alleviate the limitations present in current XAI models?
5. Beyond specific XAI technique research directions, what are other general research directions to explore in the XAI field?

## XAI Model Comparison

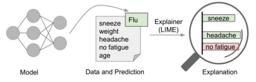
- Metrics used: Scope, Approach, Consistency, Resistance to Adversarial Attacks, Time, Interpretability and Privacy
- Metrics gathered from either XAI implementation papers or XAI survey evaluations (experienced difficulty in directly comparing XAI techniques due to lack of available research)

| XAI Technique | Scope | Approach | Consistency | RAA | Time | Interpretability | Privacy |
|---|---|---|---|---|---|---|---|
| LIME (2016) [7] | Local | Perturbation | Inconsistent [23] | None [12] | Medium [7] | Medium [7] | None |
| Anchors (2018) [9] | Local | Perturbation | Inconsistent [14] | ? | Medium [9] | High [9] | None |
| SHAP (2017) [8] | Local | Perturbation | Inconsistent [11] | None [12] | High [13] | Medium [8] | None |
| Counterfactual Explanations (2017) [16] | Local | Contrastive | Inconsistent [18] | None [18] | Low [16] | High [17] | None [19] |
| Contrastive Explanations (2019) [21] | Local | Contrastive | ? | ? | High [22] | High [20] | None |

## XAI Models Investigated

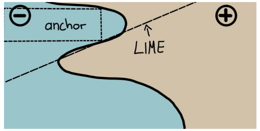
### LIME

- Creates a local explainable model g(x) for an individual ML model's prediction (visualized with a flu prediction [1])
- Explainable model generated by LIME is only locally faithful and cannot be applied globally to the ML model
- Uses a perturbation strategy to optimize it's generated explainable model g(x)



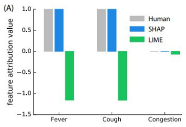

### Anchors

- Generates a set of if-then rules to explain a ML model's prediction
- This set of rules is referred to as an anchor, and they are generated through a perturbation strategy
- These anchors include the notion of coverage and show the area within which they remain faithful (can be seen in figure [2])



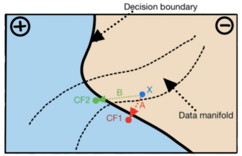

### SHAP

- Calculates the Shapley values of the individual features that affect a ML model's prediction
- Each value is a calculation of the weight a feature has on the final prediction
- Example of Shapley weights is shown below [3]



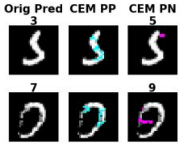

### Counterfactual Explanations

- Creates counterfactuals for the prediction of a ML model
- These counterfactuals aim to change the prediction of the ML model for an input by applying minimal changes to the initial input (visualized below [4])




### Contrastive Explanations

- Identifies the Pertinent Negatives and Positives (PNs and PPs) for a prediction
- PNs are the features whose absence determines a prediction, while the PPs are the features whose presence determines a prediction (example of PPs and PNs in image classification[5])




## Future Improvements

- Across all XAI techniques, more analysis and research should be conducted into improving and evaluating the individual XAI technique's consistency
- XAI techniques currently don't have much research done into the implementation of resistance against adversarial attacks
- In general the XAI field has a lack of large scale evaluations into the interpretability, performance and time complexity of models so therefore this is a potential research direction for the future
- Future improvements proposed for specific models are further expanded on within the paper

References:
[1] M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should i trust you?": Explaining the predictions of any classifier,” 2016.
[2] C. Molnar, Interpretable Machine Learning. 2 ed., 2022.
[3] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017.
[4] S. Verma, J. Dickerson, and K. Hines, “Counterfactual explanations for machine learning: A review,” 2020.
[5] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, “Explanations based on the missing: Towards contrastive explanations with pertinent negatives,” 2018.