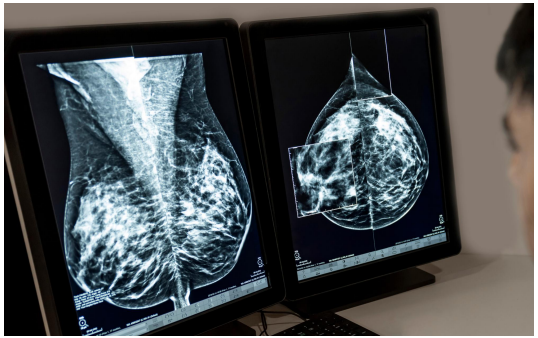
Introduction to explainable AI

Guido Gigante
Italian National Institute of Health – National Center for Radiation Protection and Computational Physics

Scuola dell'Equinozio, Pistoia (Italy), 23-25 September 2024

MAIBAI Project

- Develop a system to access and standardize data from multiple **medical imaging** databases.
- Categorize clinical data and **generate synthetic data** to improve AI training.
- **Develop and test AI tools** for breast cancer screening, **focusing on explainability and performance** in various scenarios.
- **Create an AI validation toolbox** and recommendations for the assessment of AI tools in disease screening.
- **Promote the adoption of the project's technology** and framework by relevant organizations and end-users.



What xAI is (“opening the black box”)

- Deep Neural Networks functions as “black boxes”
  - Intrinsically non-interpretable, “raw”, features (pixels, voxels, wavelets, *etc*.)
  - Highly non-linear (“hierarchical representations”)
- *xAI is a broad field of research in AI concerning development of methods to increase trust and understanding of ML model’s predictions*
  - xAI tries to make the “box” more transparent and understandable
  - “Why did you think that’s a cat?”
- The “shades-of-grey-box” problem
  - Trade-off between interpretability and representation power of ML models
  - Simple models (*e.g.*, linear models, decision trees) interpretable, yet admittedly not very powerful
  - Complex models (like deep neural networks) powerful, yet difficult “to understand”

What is an explanation?

- It depends on whom you ask
  - Developers (“is it working?”), expert users (*e.g.*, doctors; “which elements influenced the decision?”), end users (*e.g.*, patients; “what does it mean?”)
  - xAI mainly skewed towards developers
- What’s a good explanation?
  - (Most likely) not a simple rule
  - Visual, conceptual (text), exemplars?
  - Local vs global
  - Model specific vs model agnostic
  - Different methods can (and do) disagree
  - There are no clear measures to quantify the goodness of an explanation
  - (Is segmentation an explanation?)
- Is everything explainable?
  - Decision boundaries in 4+ dimensions
  - Blaise Pascal and the *esprit de finesse*
  - Split-brain patients and flashing words

Why xAI is important

- Helps ML developers understand system outputs simply and quickly
- Can suggest solutions and spot anomalies to investigate
- Makes possible to understand why a mistake was made
  - Dog in the snow = wolf
  - Or was not made (correct answer for the “wrong” reason)
- And train the system to stop it from happening again
- Helps to meet the “right to explanation” required by the GDPR
- Helps in avoiding discrimination and biases in decisions

Evaluating xAI

- The “human side”
  - Understandability: Can humans easily grasp the explanations provided by the xAI method?
  - Actionability: Do the explanations help users make informed decisions or take appropriate actions?
  - Trustworthiness: Do users trust the explanations and the underlying AI model?
  - Interaction: Can users interact with the systems, mimicking human communication patterns?
  - Learning: Is it easy for users to integrate the knowledge provided by the system?
- The “machine side”
  - Fidelity: How accurately do the explanations reflect the true workings of the AI model?
    - “Decision gradient” aligns with explanation? Counterfactuals do change the decision?
  - Sensitivity: Are explanations stable and consistent across similar inputs?
    - Perturbation analysis
  - Completeness: Do explanations cover all relevant factors?
    - Comparing explanations with domain knowledge or alternative xAI methods
- The “interface side”
  - Do humans change their attitude toward/their way of using the AI system?
  - Is it this change appropriate/useful/sound?
- It’s all quite difficult to quantify!

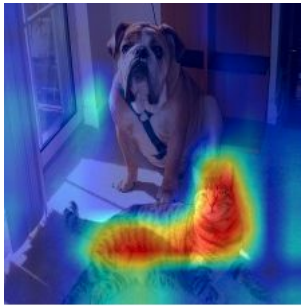
Taxonomy of xAI methods

- Importance/Attribution methods
  - What part of the input data influenced the decision (and how much)?
  - Region in an image, features, “representations”
- Exemplar methods
  - What training points impacted most the decision? - “Proponents” and “opponents”
  - Counterfactual explanations: “minimal” input change to make the machine change its mind
- Distillation methods
  - Train an interpretable (“white box”) model to reproduce the output of a complex model
  - “Structural” distillation (*e.g.*, NN into a random forest), “complexity” distillation (large NN into smaller NN)
- Intrinsic methods (“Interpretable AI”)
  - Models that provide, alongside the main output, an explanation
  - Optimizing both model performance and a certain quality of the explanations produced
  - Prototypes, templates, concepts

Importance examples

- CAM (Class Activation Maps), Grad-CAM, Integrated Gradients, Smooth Gradients
  - Aim to identify which regions in the image were relevant to a specific class
  - By looking at the magnitude of the gradients (class C_k, representation R_i) flowing through the network layers
  - Useful to measure how much each pixel/region activate the class predicted by the network
- Occlusion sensitivity
  - "Perturbing" (masking) the input
  - And looking at the effects on classification
  - Saliency = output variation

$$\frac{\partial C_k}{\partial R_i}$$



predicted class: cat

class dog

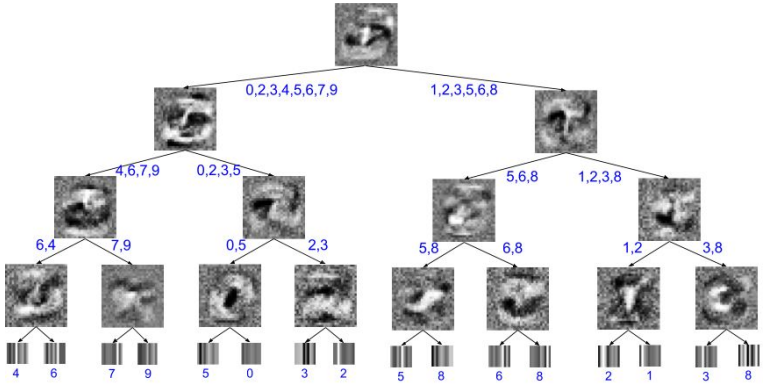
Exemplar example: Gradient tracing/TrackIN

- Data influence: How much has this training point influenced the prediction for this test point?
- Retrain the model without the training point and test
- This is prohibitively expensive!
- Resorting to approximations

$$\sum_{e} \nabla_{\mathbf{w}_e} \text{loss}(x_{\text{train}}) \cdot \nabla_{\mathbf{w}_e} \text{loss}(x_{\text{test}})$$

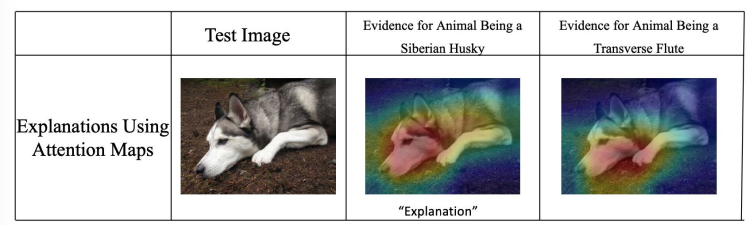
Distillation example

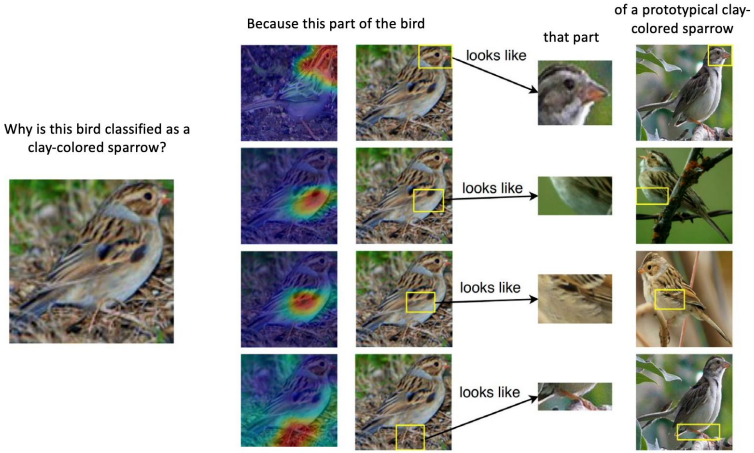
- DNN into a “soft decision tree”
- Each node is a “learned representation”
- Explanation = most probable path
- Why distillation?
  - Simply, the DNN is more powerful
  - And you can build better decision trees by using “**soft label**” generated by the DNN



Intrinsic example

- Heat-maps look good, don't they?
  - Explanation seems more like "attention", not related to the specific class
  - They can be quite "unsensitive" (distillation can help with that)
- "This looks like that" — prototypes
  - The network dissects the image by finding prototypical parts
  - And combines evidence from the prototypes to make a final classification
  - Explanation mimicking (in part) the way an ornithologist would reason





C.Chen et al., "This Looks Like That: Deep Learning for Interpretable Image Recognition", NeurIPS 2019

SHAP

What is SHAP?

Here SHAP is just an algorithm that, given a model f (XGBoost, Deep Network, etc.) and given one individual i (one set of values for the predictors $\mathbf{X}_i$ - for example: age 37, BMI 28.4, systolic blood pressure 135), estimates the contribution of each predictor to the specific prediction y_i.

In simple words, SHAP gives a principled estimation of, for example, **how much the prediction y_i would change if the age of individual i were not observed**

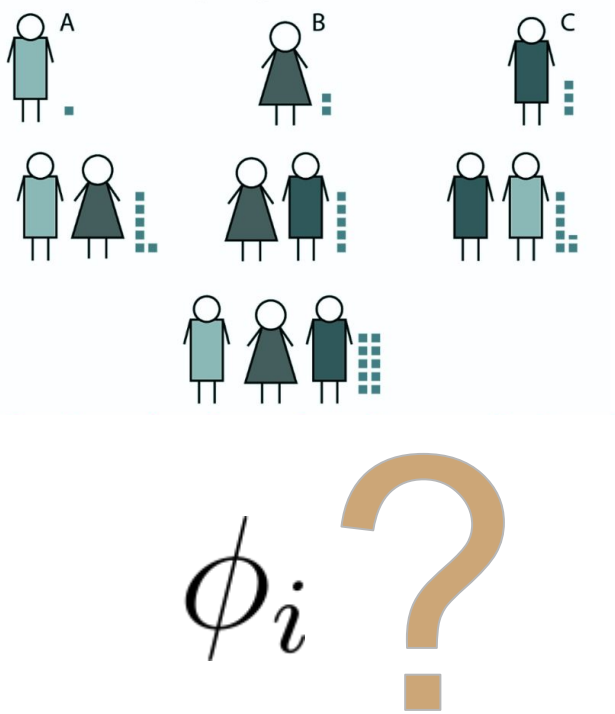
$\Delta y_i = \text{SHAP}(\text{age}_i)$

$$\text{SHAP}(\text{age}_i) = f(\mathbf{X}_i|\theta) - f(\mathbf{X}_i \setminus \text{age}_i|\theta)$$

SHAP is an Importance Method

Shapley value (1/2)

- A game theory concept introduced by Lloyd Stowell Shapley (Nobel Prize in economic sciences)
- **The Shapley value gives a “fair” estimate of the contribution of each player participating in a collaborative work**
- There is a game, there are N players; for each possible team T of k (≤ N) players there is a payoff V(T) (that measures the “value of the team”)
- Knowing V(T) for all the possible teams, how can we best measure the “value” of player i?



Shapley value (2/2)

Infinite choices... Add some sensible constraints, like:

$$\sum_{i \in \text{all players}} \phi_i = v(\text{all players})$$

$$v(T \cup i) = v(T \cup j) \text{ for every team } T \implies \phi_i = \phi_j$$

$$v(T \cup i) = v(T) \text{ for every team } T \implies \phi_i = 0$$

$$\phi_i(\alpha\, v + w) = \alpha\, \phi_i(v) + \phi_i(w)$$

Then, there is only one possible solution:

$$\phi_i = \sum_{T \subseteq \text{all players} \setminus i} \frac{|T|!\,(N - |T| - 1)!}{N!} \Big(v(T \cup i) - v(T)\Big)$$

SHAP (1/2)

- In SHAP the players are the predictors
- To define the Shapley value for a predictor, we need to define the function v(T) for every possible team, starting from our function $f(\mathbf{X}|\theta)$ and a data point $\mathbf{X}_k$ we want to analyse
- "How much the prediction would change if the i-th predictor in X_k were not observed"

$$v(T) = \mathrm{E}[f(x)|x_i = X_i \text{ for } i \in T]$$

SHAP (2/2)

- The expectation is very hard to compute
- TreeSHAP, for tree-based models (random forest, XGBoost, *etc*.), is an exception
- Different instances of SHAP for different problems make different simplifying assumptions:

$$\mathrm{E}_{z|t}[f(z,\, t)] \simeq \mathrm{E}_{z}[f(z,\, t)] \simeq f(\mathrm{E}[z],\, t)$$

$$z = \{X_i \text{ for } i \notin T\}$$
$$t = \{X_i \text{ for } i \in T\}$$

Individualists, synergies, and conflicts

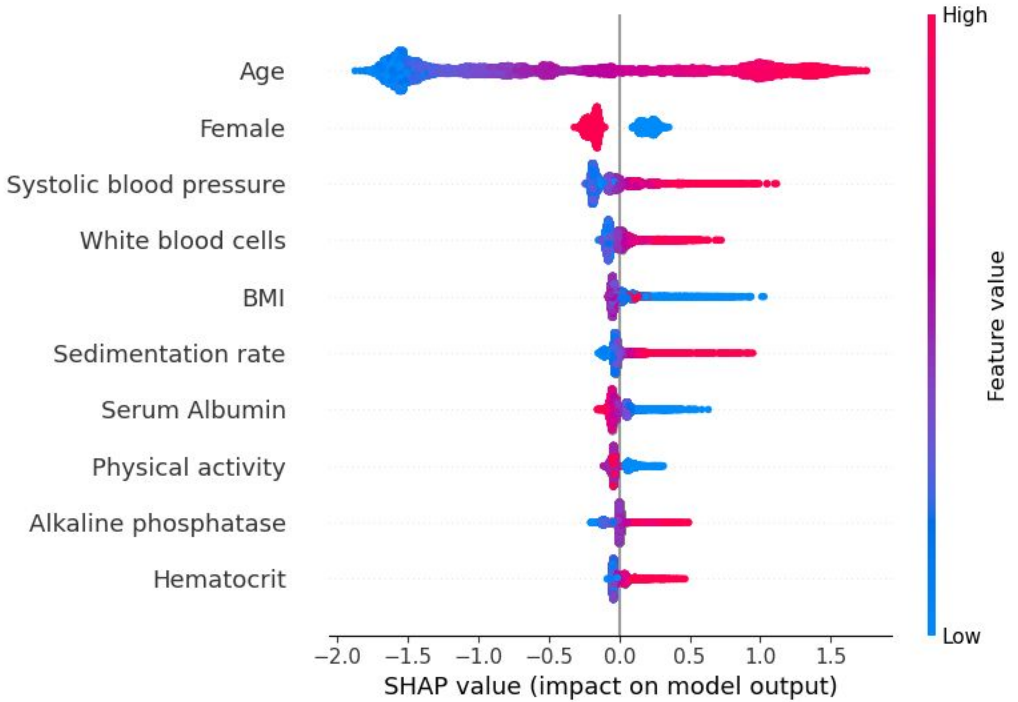
- An individualistic player C is one that gives the same "boost" v_C to every team ($\mathbf{v_{AC} = 0}$ for all players A, with A != C)
- Synergy: player A adds v_A; player B adds v_B; $V(A + B) > v_A + v_B$
  - A and B boost each other game ($\mathbf{v_{AB} > 0}$)
- Conflict: player A adds v_A; player B adds v_B; $V(A + B) < v_A + v_B$
  - A and B compete for the same "role" ($\mathbf{v_{AB} < 0}$)
- **Main effect** and **interaction values**

$$V(T) = \sum_{A \in T} v_A + \frac{1}{2} \sum_{(A,B) \in T} v_{AB}$$

$$\text{SHAP}_A = v_A + \frac{1}{2} \sum_B v_{AB}$$

A quick example on mortality

- Age is by far the most significant predictor
- Sex (second place) acts largely as a constant factor
- There is, quite consistently, a monotonic relationship between feature value (*e.g.*, Age) and the effect on prediction (SHAP)



DeepSHAP (building on DeepLift)

- There exists a “chain rule” to compute the SHAP value layer by layer

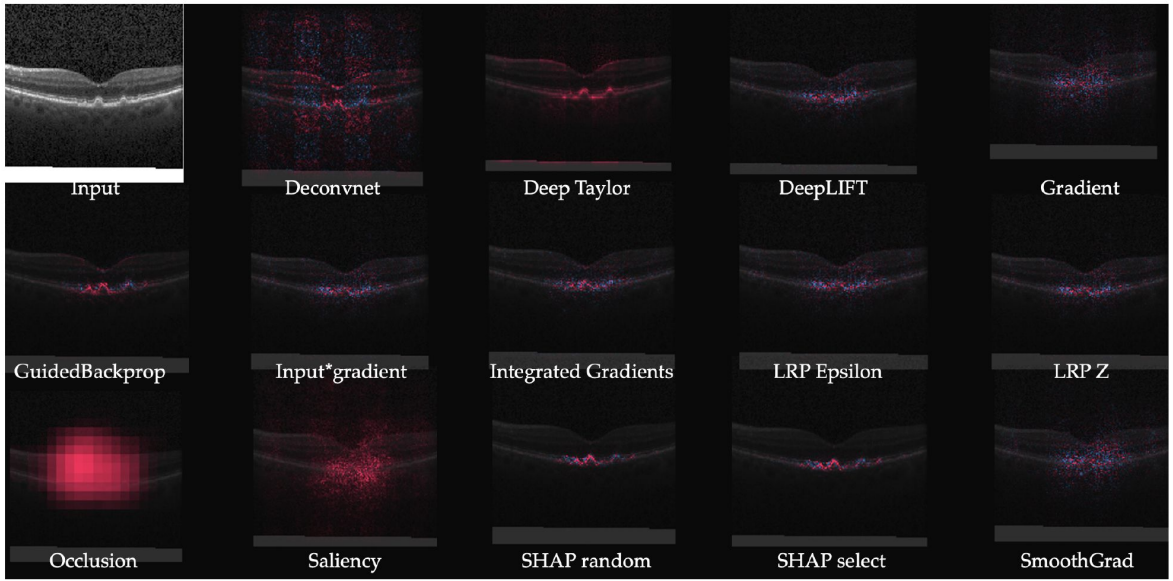
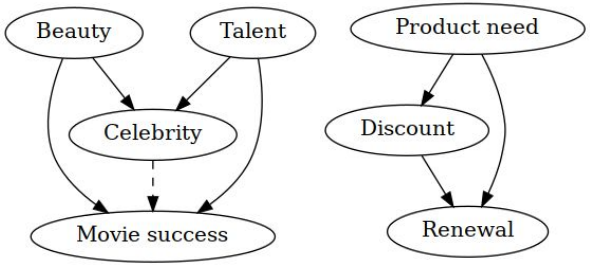


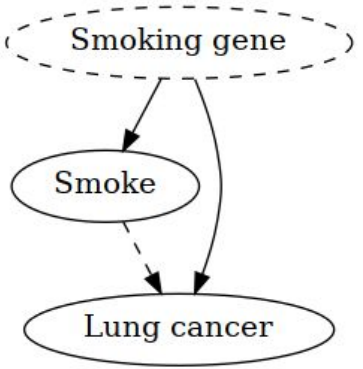
Figure 4. Example of heat maps from a retinal OCT image [40].

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020)

Explanation is not causation

- AI doesn't try to build a "physical model" of the data
- AI just "wants to predict"
- AI can discover, and take advantage of, highly complex and non-linear "correlations" among predictors, and between predictors and response
  - A good collider can be better than the real causes (and more parsimonious)
  - Direct and mediated effects entangled in explanations
- xAI just unearths these "correlations on steroids"
  - Don't forget that <u>you are explaining your model, not reality</u>
- And, of course, unobserved confounders are the hardest problem with or without AI





Yet...

- It is difficult to dismiss the predictive power of AI
  - Prediction is important *per se* (Where and when can I intervene most effectively?)
  - And there are anyhow many factors that we cannot control, but are determinant
- And the patterns of “influence” uncovered by SHAP are just too intriguing
- At the very least, AI + xAI represents today one of the most promising instruments for scientific *exploration*
- As for *explanation*:
  - Strongly nonlinear causal effects (*e.g.*, U-shaped, one predictor “gating” the effects of another) will be mostly under-detected and under-measured by traditional methods
  - xAI can make the most out your prior knowledge
  - There are attempts to make AI (and xAI) more “causality-aware”
  - **Double Machine Learning** tries to extract casual knowledge from observational data
- But we leave that for another occasion

Thank you!

Contact me: guido.gigante@iss.it