SHAP

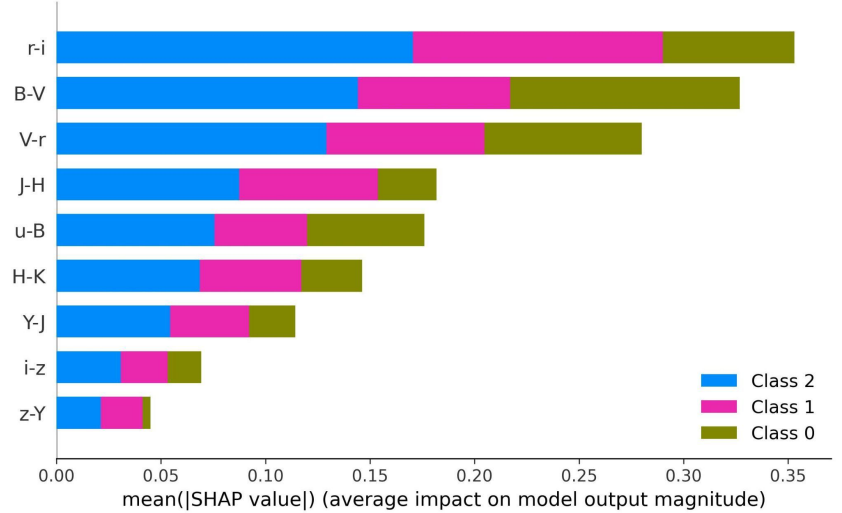
Rationale, Strengths, and Weaknesses

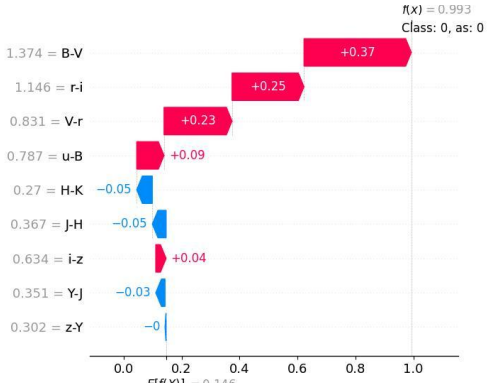
What is SHAP?

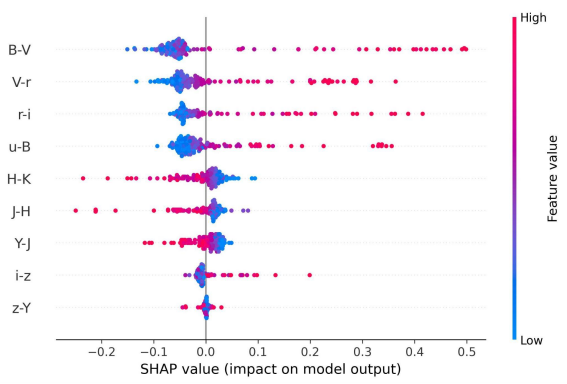
SHAP (SHapley Additive exPlanations) is an interpretability framework for machine learning models based on concepts from **cooperative game theory**. It seeks to **explain the output of any machine learning model** by attributing to each input feature (e.g., each pixel in an image) a **Shapley value**—a measure of its contribution to the prediction (introduced by Lloyd Shapley in 1953).

Shapley values were developed to **equitably distribute the total gains of a coalition** to individual players based on their contributions.

SHAP adapts this idea to machine learning by treating each **feature (e.g., pixel)** as a player and the **model output** as the "value" of their cooperation.



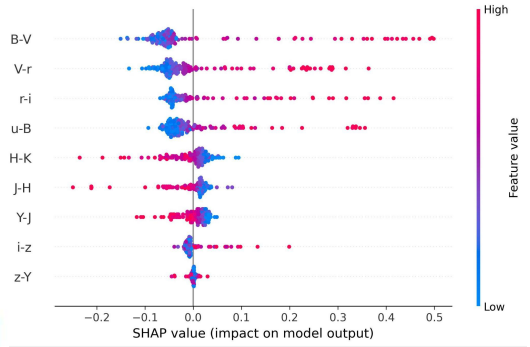


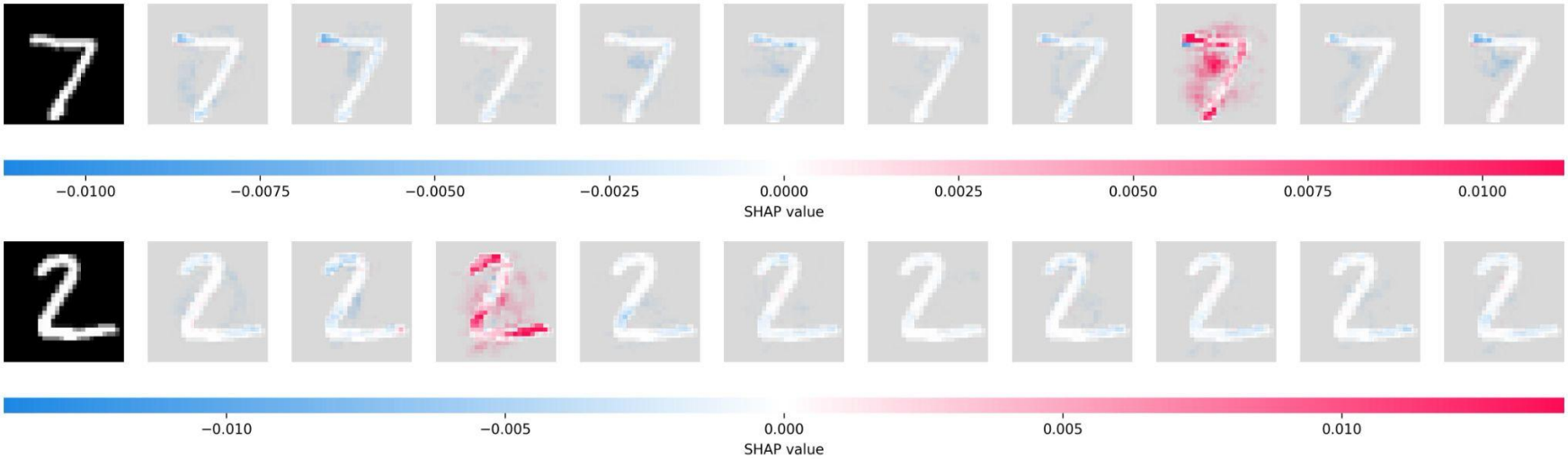


What is SHAP?

Shapley values were developed to **equitably distribute the total gains of a coalition** to individual players based on their contributions.
SHAP adapts this idea to machine learning by treating each **feature (e.g., pixel)** as a player and the **model output** as the "value" of their cooperation.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$



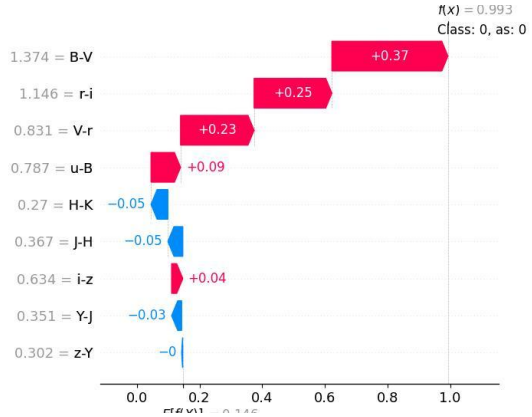


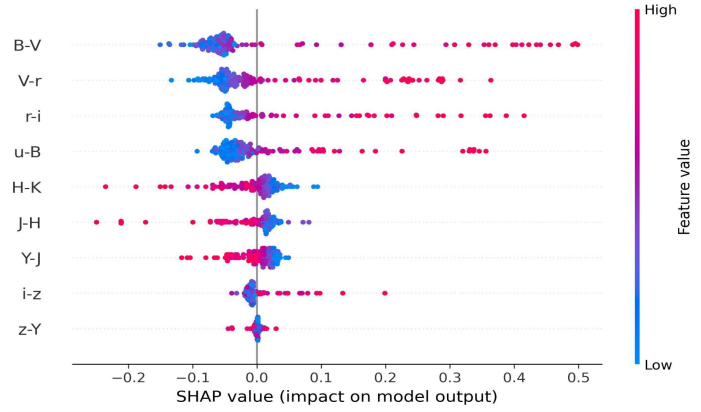
Strengths of SHAP

Collective and individual evaluations.

Individual: what happen in this particular example, correctly or incorrectly predicted.

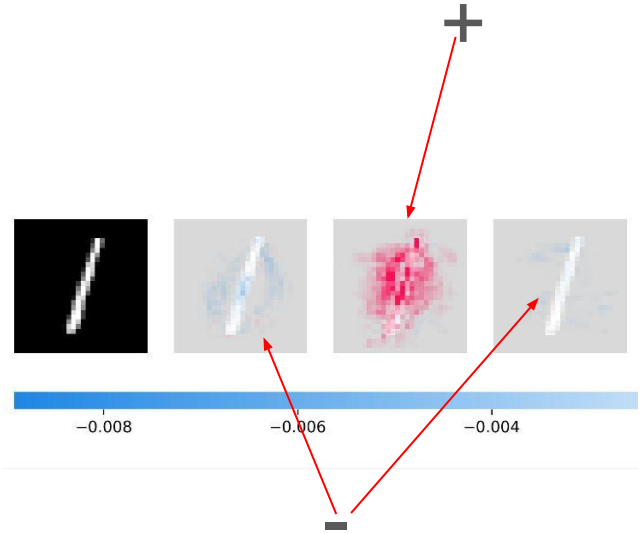
Collective: explain how the neural networks behaves, which variables are more relevant, and how much.

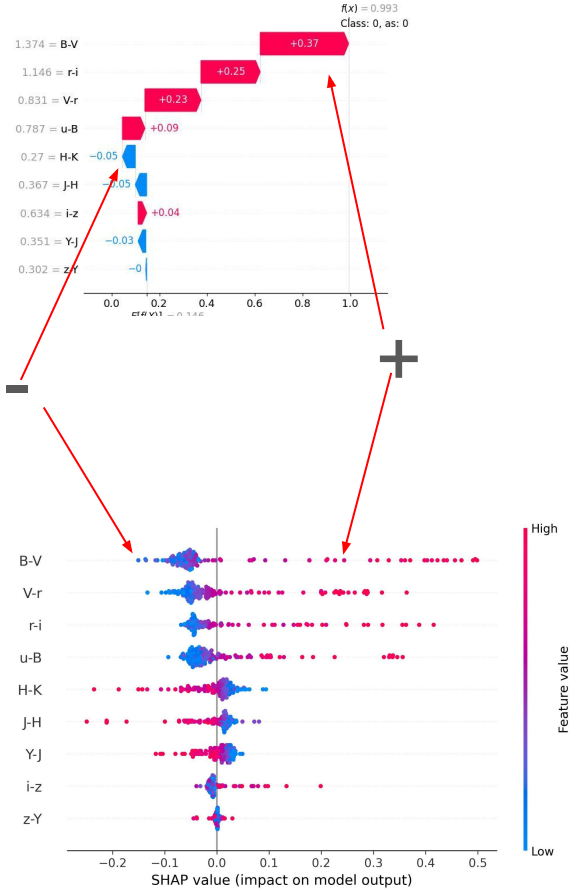




Strengths of SHAP

Mark positive and negative contributions.





Strengths of SHAP

- **Theoretical guarantees**: Based on Shapley values from game theory.
- **Feature attribution** is **fair** and **consistent** (under certain assumptions).
- **Works with many types of models** (deep learning, tree-based, etc.).
- **Additive explanations**: easy to interpret and visualize (e.g., SHAP image plots).

Weaknesses of SHAP

Computationally expensive, especially with high-dimensional data like images.

Feature independence assumption: Many SHAP variants (e.g., KernelSHAP) assume input features are independent, which is **not true for images**, where pixels are highly correlated.

Choice of baseline (background data) heavily influences the explanation quality.

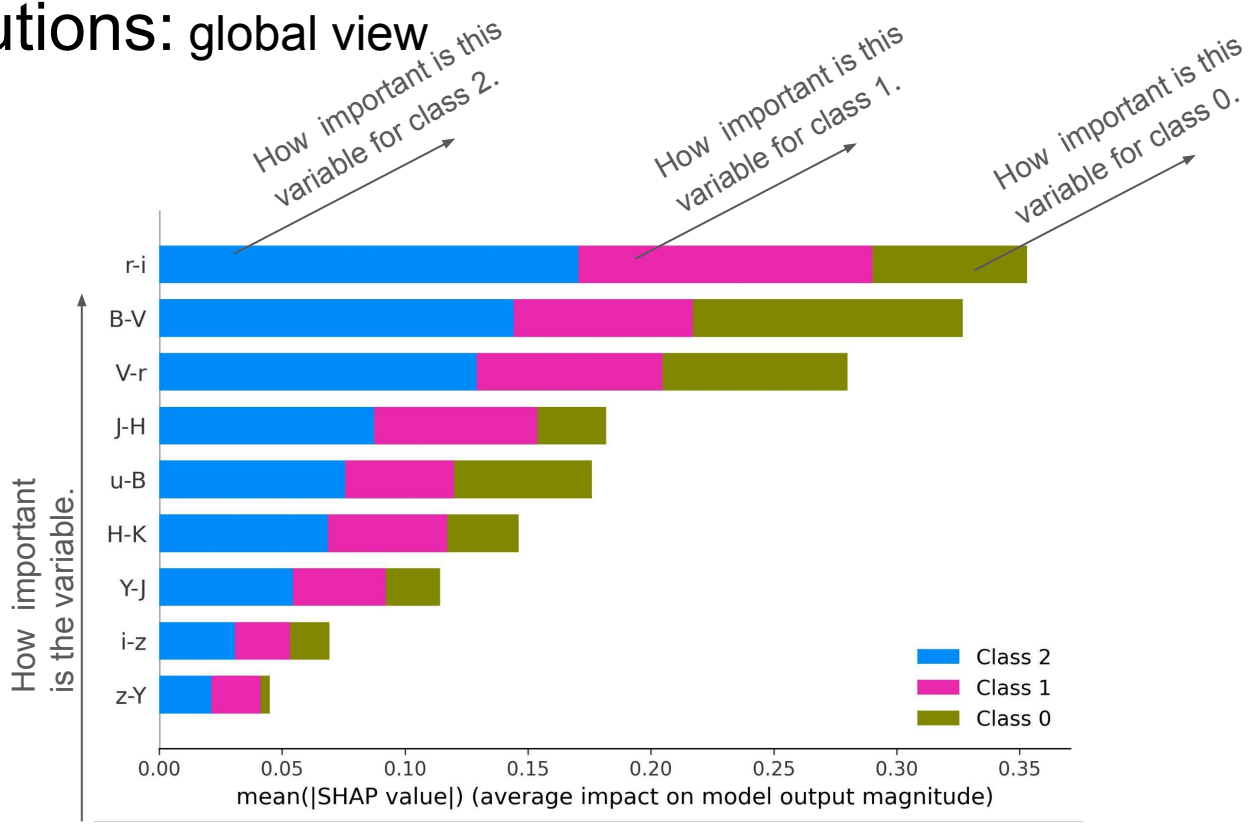
For images, **spatial structure is not directly preserved** unless specific adaptations are used (e.g., segmentation-based SHAP).

The equation

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

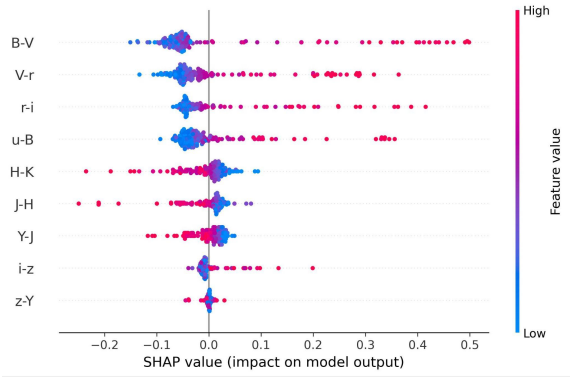
- *M* is the total of variables.
- *S* is the total of active variables.
- The subtraction is the contribution with the variable *i* active or not.

Global contributions: global view

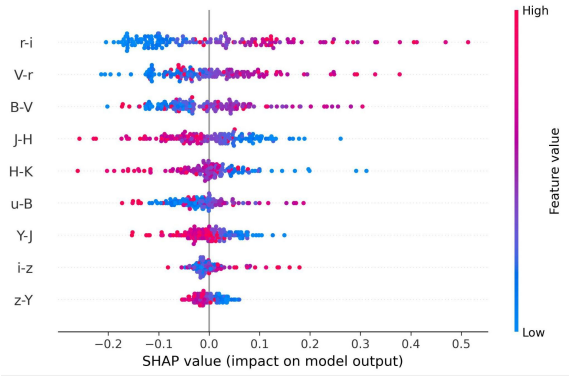


Global contributions: individual view.

for class 0



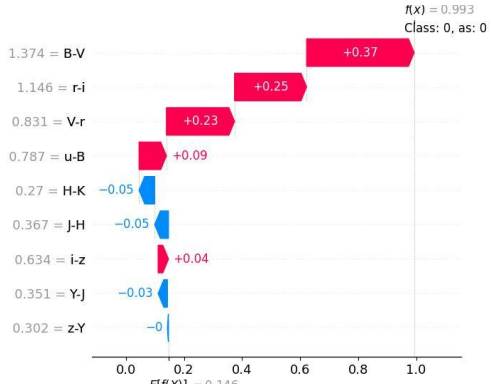
for class 1



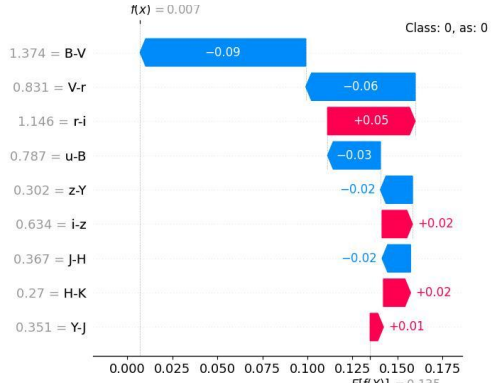
for class 2

Individual contributions. Example class 0 predicted as 0.

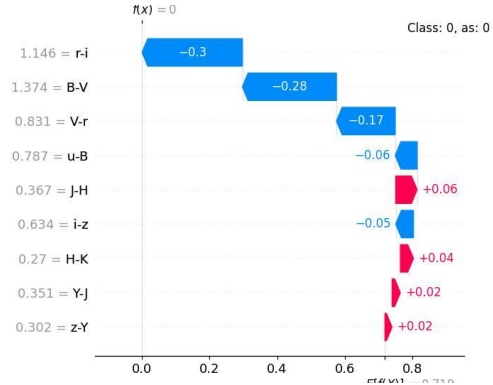
for class 0



for class 1

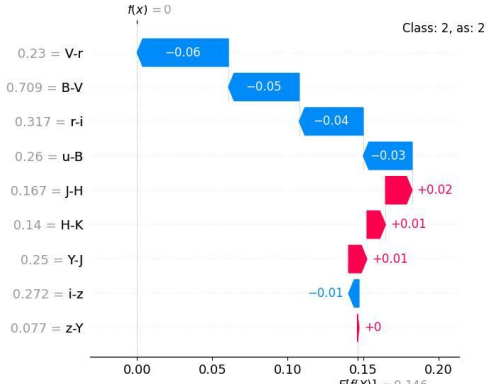


for class 2

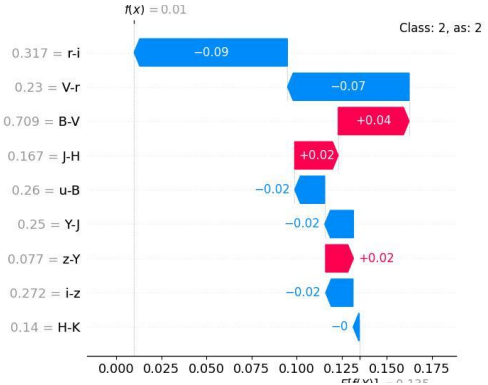


Individual contributions. Example class 2 predicted as 2.

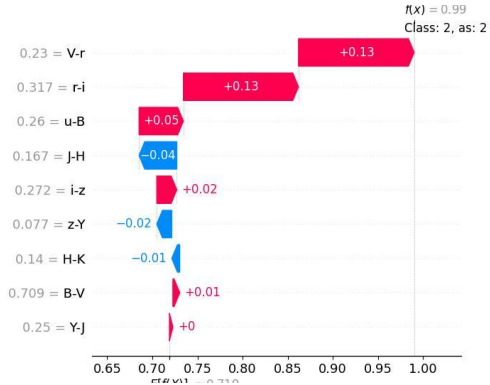
for class 0



for class 1



for class 2



Summary

SHAP: SHapley Additive exPlanations

Core

Interpretability framework for machine learning models based on cooperative game theory: Shapley value.

Advantages

Theoretical foundation, feature attribution is fair and consistent, works with many types of models, additive explanations.

Limitations

Computationally expensive, feature independence assumption, spatial structure not directly preserved.

Feature attribution is **fair** and **consistent** (under certain **assumptions**). Which ones?

SHAP attributions are considered **fair** and **consistent** under **four key assumptions** or axioms.

1. Efficiency (Local Accuracy)

The total contribution from all features must add up to the difference between the model's prediction for a specific input and the average model prediction (baseline).

$$\sum_{i=1}^{n} \phi_i = f(x) - E[f(x)]$$

Where:

- ϕ_i is the SHAP value for feature *i*,
- *f(x)* is the model's output for input *x*,
- *E[f(x)]* is the expected output over the background distribution.

2. Symmetry (Feature Symmetry)

If two features contribute equally to all possible subsets, then they receive equal importance.

$$\text{If } f(S \cup \{i\}) = f(S \cup \{j\}) \text{ for all } S, \text{ then } \phi_i = \phi_j$$

Implication: Equivalent features are treated equally — no arbitrary bias in attribution.

Feature attribution is **fair** and **consistent** (under certain **assumptions**). Which ones?

SHAP attributions are considered **fair** and **consistent** under **four key assumptions** or axioms.

3. Dummy (Null Feature)

If a feature does not change the model's prediction in any coalition, its attribution is zero.

$$\text{If } f(S \cup \{i\}) = f(S) \text{ for all } S, \text{ then } \phi_i = 0$$

Implication: Only influential features get credit — irrelevant features are ignored.

4. Additivity (Linearity)

For two models *f* and *g*, the SHAP values of their sum should equal the sum of their SHAP values.

$$\phi_i^{f+g} = \phi_i^f + \phi_i^g$$

Implication: Attributions are consistent across ensemble or additive models.