SHAP (Shapley Additive Explanations)

'SHAP' stands for 'Shapley Additive Explanations'. It is a method from cooperative game theory applied to machine learning to explain how much each input feature contributes to a model's output.

Originally derived from Shapley values in game theory, SHAP calculates the average **marginal contribution** of each feature to the prediction, across all possible combinations of features.

Key characteristics

- **Model-agnostic**: Can be applied to any machine learning model.
- **Additive**: The sum of SHAP values equals the model output.
- **Interpretable**: Assigns an importance score to each variable for a specific prediction.

Clinical relevance

SHAP is increasingly used in medical AI to:

- Understand which variables drive risk predictions.
- Improve transparency in black-box models.
- Support clinician trust in algorithmic decision tools.

Limitations

- SHAP explains the behavior of the **model**, not the **underlying physiology**.
- It does **not imply causality** — a high SHAP value does not mean a variable causes disease.
- Computationally expensive for complex models and large datasets.

'In summary:' SHAP helps interpret machine learning outputs, but must be used with caution in clinical settings to avoid overinterpreting spurious correlations.