Understanding SHAP Values

Explaining model predictions – not patient outcomes

What does a SHAP graph show ?

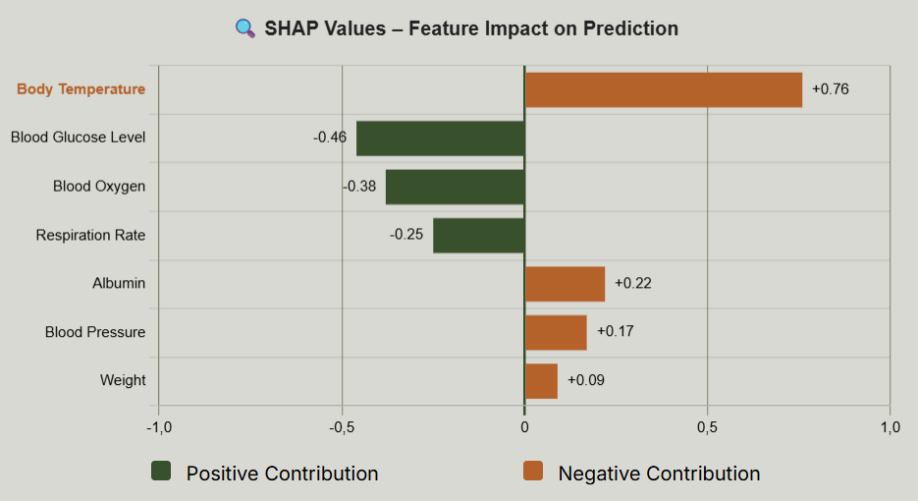
A SHAP graph explains how each input feature influenced the model's prediction for one specific case.

- Each bar shows the SHAP value of a feature: how much that feature pushed the prediction up or down compared to the average.
- Bars to the right increased the prediction; bars to the left decreased it.
- The longer the bar, the stronger the impact of the feature on the prediction – in the context of all other features.

> ⚠️ SHAP values do not represent clinical importance or causality.
> They describe how the model behaves, not necessarily how the body works.

Example: Body Temperature

Let's say the feature "Body Temperature" has a positive SHAP value and its bar points to the right.



- This means that in this specific case, the patient's temperature made the model predict a higher risk.
- If the temperature had been different, the SHAP value might have changed – but not necessary in the opposite direction.

👉 SHAP reflects the model's internal reasoning, not the clinical impact interventions.
It does not say what would happen if you changed the temperature – only how the current value contributed to the prediction.