# LIME vs SHAP for Explainable AI (XAI)

**Motivation:** Many modern ML/AI models are “black boxes” (complex, non-linear, deep, or ensemble-based) — hard for humans to understand.

In high-stakes domains (medicine, lending, justice, hiring, etc.), stakeholders need to trust and validate model decisions — explanations build trust.

Explanations help detect errors, overfitting, dataset bias or unjust dependencies (e.g. model relying on spurious correlations).

# LIME vs SHAP for Explainable AI (XAI)

- Breast Cancer dataset (scikit-learn), Breast Cancer Wisconsin Diagnostic Dataset
  Target classes: 0 = malignant tumor, 1 = benign tumor

Class distribution (typical for this dataset):

212 malignant

357 benign

→ This means the dataset is mildly imbalanced toward benign cases.

- Focus: interpreting a single prediction + global behavior

# LIME vs SHAP for Explainable AI (XAI)

- Breast Cancer dataset (scikit-learn), Breast Cancer Wisconsin Diagnostic Dataset

Target classes: 0 = malignant tumor, 1 = benign tumor

Class distribution (typical for this dataset):

212 malignant

357 benign

→ This means the dataset is mildly imbalanced toward benign cases.

- Focus: interpreting a single prediction + global behavior

# LIME vs SHAP for Explainable AI (XAI)

## The features include

Radius

Texture

Perimeter

Area

Smoothness

Compactness

Concavity

Symmetry

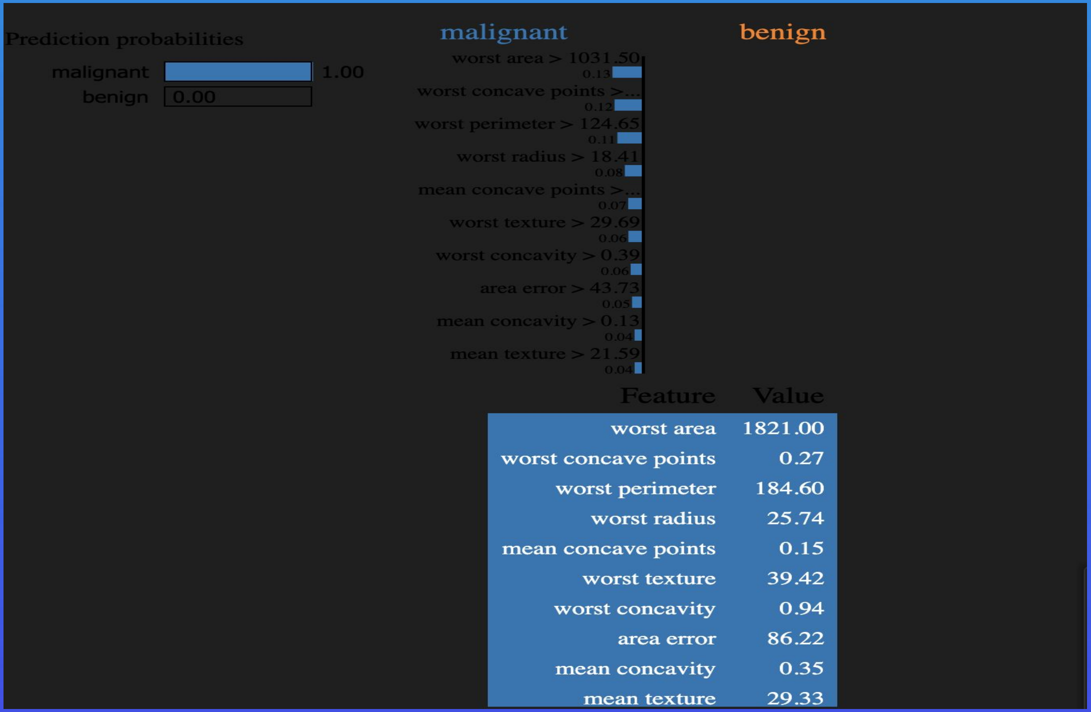
Fractal dimension

# TL;DR

- LIME ⟶ simple, human-readable rules around one prediction
- SHAP ⟶ Shapley-value attributions with global + local views
- For our model, worst area / concavity / radius dominate decisions
- SHAP is more stable; LIME is more story-like for clinicians

# LIME Local Explanation – How to Read This

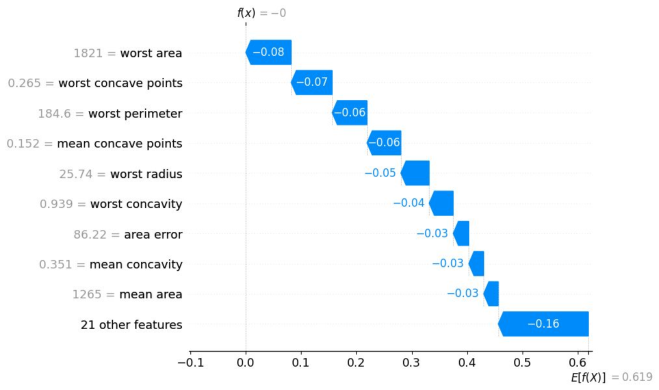
- Left: model says malignant with probability 1.00 (blue bar)
- Middle: blue bars = features pushing toward malignant; orange would be toward benign
- Each rule (e.g., "worst area > 1031.50") is a local decision boundary
- The larger the bar, the larger that feature’s impact on this patient’s malignancy score
- Table at bottom lists the actual values (e.g., worst area = 1821, worst concave points = 0.27)



| Feature | Value |
|---|---|
| worst area | 1821.00 |
| worst concave points | 0.27 |
| worst perimeter | 184.60 |
| worst radius | 25.74 |
| mean concave points | 0.15 |
| worst texture | 39.42 |
| worst concavity | 0.94 |
| area error | 86.22 |
| mean concavity | 0.35 |
| mean texture | 29.33 |

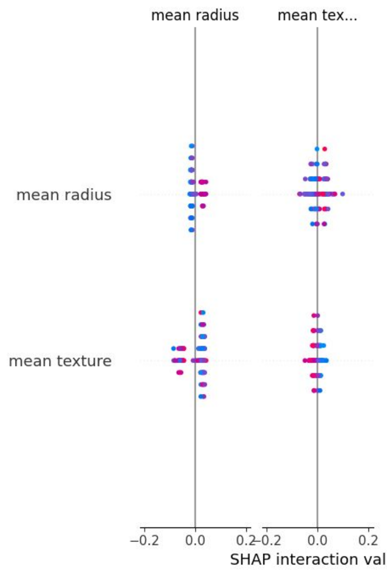
# SHAP Waterfall Plot – Local Explanation

- Baseline on the right: E[f(X)] = average model output over all patients
- Blue blocks show how each feature moves us from baseline to this patient’s prediction
- Rightward contributions increase malignancy probability; leftward decrease it
- In this example, high worst area and concavity-related features strongly push toward malignant



# SHAP Interaction / Additional View

- SHAP can visualize interactions between features (e.g., mean radius × mean texture)
- Each point is a patient; color and position reflect joint influence on prediction
- Helps reveal nonlinear patterns that simple feature importance misses
- Complements LIME by showing how groups of patients behave, not only one case



# LIME vs SHAP – Comparison

LIME:

- Pros: very interpretable (if/then style rules), easy to show to clinicians
- Cons: explanations can change with different sampling; no direct global picture

SHAP:

- Pros: theoretically grounded, consistent global rankings, rich plots (summary, waterfall, interaction)
- Cons: slower to compute; visualizations slightly harder to explain to non-technical audiences

# References

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD.
Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.
SHAP documentation: https://shap.readthedocs.io
LIME documentation: https://github.com/marcotcr/lime