Unit V – Research Directions & Future Trends

Evaluation Metrics for Explaination

Evaluation metrics ensure the explanation is faithful and not misleading.

Metrics allow objective comparison to select the most reliable method.

Evaluated explanations increase confidence and acceptance of AI outputs.

Metrics detect unstable or noisy explanations.

Metrics provide evidence that explanations are meaningful and reliable.

Metrics help identify flawed reasoning patterns in models.

Evaluation ensures explanations are useful, concise, and actionable.

Metrics act as quality control against false reassurance.

Unit V – Research Directions & Future Trends

Fidelity (Faithfulness)

Measures how accurately the explanation reflects the true behavior of the model.

An explanation is high-fidelity if changes to important features significantly alter the model’s predictions.
Without fidelity, explanations become post-hoc justifications, not real insights.

Local / Global FIdelity

LIME	Local fidelity (depends on sampling)
SHAP	High fidelity (theoretically grounded)
Integrated Gradients	High fidelity for differentiable models
Grad-CAM	Approximate, visualization-focused

Unit V – Research Directions & Future Trends

Consistency (Stability)

Consistency evaluates whether an explanation method gives similar explanations for similar inputs.

Consistency measures the stability of explanations under small input perturbations.

Inconsistent explanations reduce trust, even if the model itself is accurate.

Local / Temporal/ Model Consistency

Input Perturbation Test : Add a small noise to Input and Check for stability

Generate explanations for multiple samples in a small neighborhood.

Unit V – Research Directions & Future Trends

Completeness: Measures whether the explanation accounts for the full prediction.

Sensitivity: Tests whether explanations react appropriately to meaningful input changes.

Sparsity (Concise): Evaluates how compact and simple the explanation is.

Interpretability (Human Understandability): Measures how easily humans can understand the explanation.

Robustness: Assesses whether explanations remain reliable under noise or adversarial perturbations.

Actionability: Measures whether explanations suggest meaningful actions.

Computational Efficiency: Measures time and resources required to generate explanations.

Unit V – Research Directions & Future Trends

Case study on Robustness

A hospital deploys a **deep learning model** to predict **diabetes risk** using tabular patient data:

- Age
- BMI
- Blood glucose level
- Blood pressure
- Family history

To gain trust, clinicians use **SHAP** and **LIME** explanations for individual predictions.

Doctors observe that **small measurement variations** (e.g., glucose ±2 mg/dL) produce **significantly different explanations**, even though:

1. The predicted risk score remains almost the same.
2. The patient’s condition is clinically unchanged.

This raises concerns about **robustness of explanations**.

Unit V – Research Directions & Future Trends

Case study on Robustness

Step 1: Baseline Explanation

Generate explanation for original patient data using SHAP and LIME.

Record feature importance ranking.

Step 2: Perturb Input

Add small Gaussian noise to numeric features (within clinical tolerance).

Example:

Glucose: 140 → 142 mg/dL

BMI: 27.5 → 27.6

Step 3: Re-generate Explanations

Generate explanations for perturbed inputs.

Step 4: Measure Robustness

Compare explanations using:

Spearman rank correlation

Cosine similarity of attribution vectors