# Comparison of XAI Methods

## Team Name- Atoja

By Atul Santhosh & Pooja Negi
https://github.com/atulsan/Atoja

# Datasets used for comparison

1. Cats and Dogs Classification (Image Dataset)
2. Face Glasses Recognition (Image Dataset)

# What XAI Methods we chose?

1. **Grad-CAM (Gradient-weighted Class Activation Mapping):** Provides class-specific visual heat maps.

2. **LIME (Local Interpretable Model-Agnostic Explanations):** Offers localised explanations for image classification models.

3. **Integrated Gradients:** Provide pixel-wise attributions based on gradients.

# Applicability of XAI Methods

# 1. Grad-CAM (Gradient-weighted Class Activation Mapping)

## Why it makes sense?

- Grad-CAM visualises the regions influencing a model's decision, offering intuitive heatmaps with good faithfulness, as it directly reflects gradient-based activations.
- Sparsity is moderate as it highlights regions but does not generate as sparse an explanation as LIME.

## Considerations

Complexity is relatively low but is constrained to CNN- based architectures.

# 2. LIME (Local Interpretable Model-agnostic Explanations)

## Why it makes sense:

- LIME explains predictions by creating locally interpretable linear approximations, making it suitable for assessing faithfulness (how well explanations align with the model's behavior) in specific input instances.

- It generates sparse explanations by design, ensuring only a subset of features is considered important.

## Considerations:

The perturbation approach might lead to noisy results, and complexity increases with high-dimensional inputs like images.

# 3. Integrated Gradients (IG)

## Why it makes sense:

- IG directly attributes feature importance by integrating gradients along a path from a baseline to the input, ensuring faithful explanations that align well with the model's internal workings.
- Complexity is relatively low since IG works directly with the model's gradients, avoiding external approximation.

## Considerations:

Sparsity may not be as high as LIME because it calculates dense importance scores across all input features.

# Outputs of Dataset 1

https://github.com/atulsan/Atoja/tree/main/cat_and_dog_classification(Cat_Dog-XAI)

Original Image

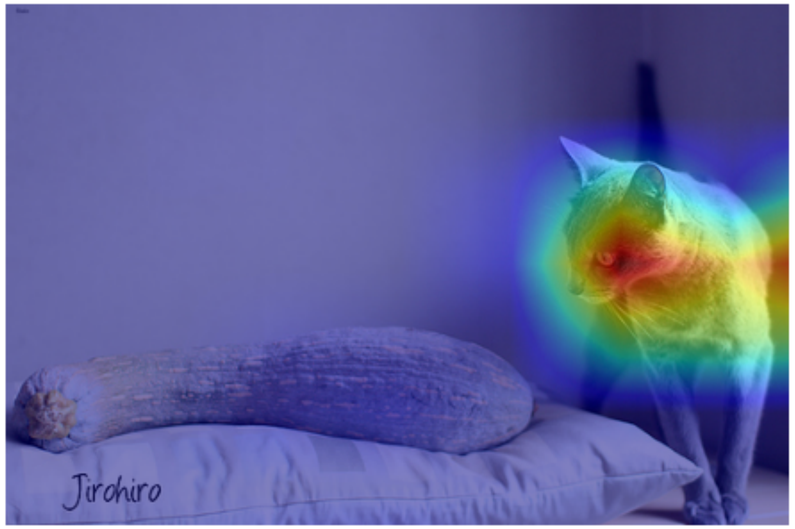
Grad-CAM
Faithfulness: 0.43488850794605766
Sparsity: 0.0000
Complexity: 7022

Integrated Gradients
Faithfulness: 0.49755407126294704
Sparsity: 0.0000
Complexity: 8

LIME
Faithfulness: -0.043038074698375214
Sparsity: 0.9051
Complexity: 3523

Original Image


Grad-CAM
Faithfulness: -0.49304201825858324
Sparsity: 0.0721
Complexity: 3171




Integrated Gradients
Faithfulness: -0.256477139366769
Sparsity: 0.0000
Complexity: 47


LIME
Faithfulness: -0.3181937294888557
Sparsity: 0.8494
Complexity: 6591

Original Image

Grad-CAM
Faithfulness: 0.190797780898683
Sparsity: 0.4655
Complexity: 4419

Integrated Gradients
Faithfulness: 0.1185047393727681
Sparsity: 0.0000
Complexity: 23

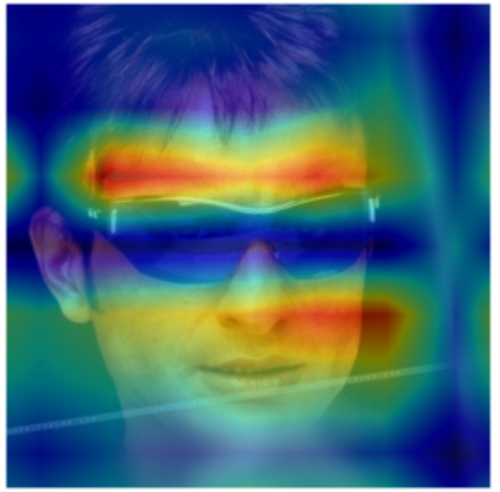
LIME
Faithfulness: 0.1887613114611376
Sparsity: 0.7452
Complexity: 11269

# Outputs of Dataset 2

https://github.com/atulsan/Atoja/tree/main/face_glasses_recongition(Glasses-XAI)

Original Image

Grad-CAM
Faithfulness: 0.0019
Sparsity: 0.0000
Complexity: 203185.0000

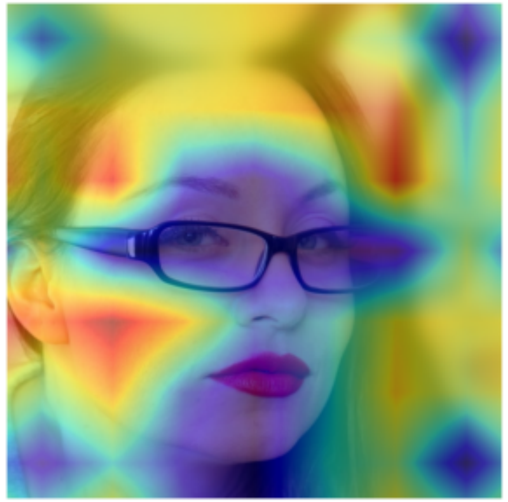


Integrated Gradients
Faithfulness: 0.0282
Sparsity: 0.0000
Complexity: 476.0000

LIME
Faithfulness: 0.0000
Sparsity: 0.9593
Complexity: 42644.0000

Original Image

Grad-CAM
Faithfulness: 0.0086
Sparsity: 0.0000
Complexity: 471131.0000



Integrated Gradients
Faithfulness: 0.0738
Sparsity: 0.0000
Complexity: 185.0000

LIME
Faithfulness: 0.0000
Sparsity: 0.9851
Complexity: 15627.0000

Original Image

Grad-CAM
Faithfulness: 0.0003
Sparsity: 0.0203
Complexity: 186886.0000

Integrated Gradients
Faithfulness: 0.0224
Sparsity: 0.0000
Complexity: 159.0000

LIME
Faithfulness: 0.0000
Sparsity: 0.9859
Complexity: 14775.0000

# Evaluation of XAI Methods

# Metrics for Evaluation

**Faithfulness:**

- Measures how accurately the explanation reflects the model's decision process.
- Test by removing the most important regions/features (according to the explanation) and measuring the drop in prediction confidence.
- Higher faithfulness indicates a better alignment between the model and explanation.

.

**Sparsity:**

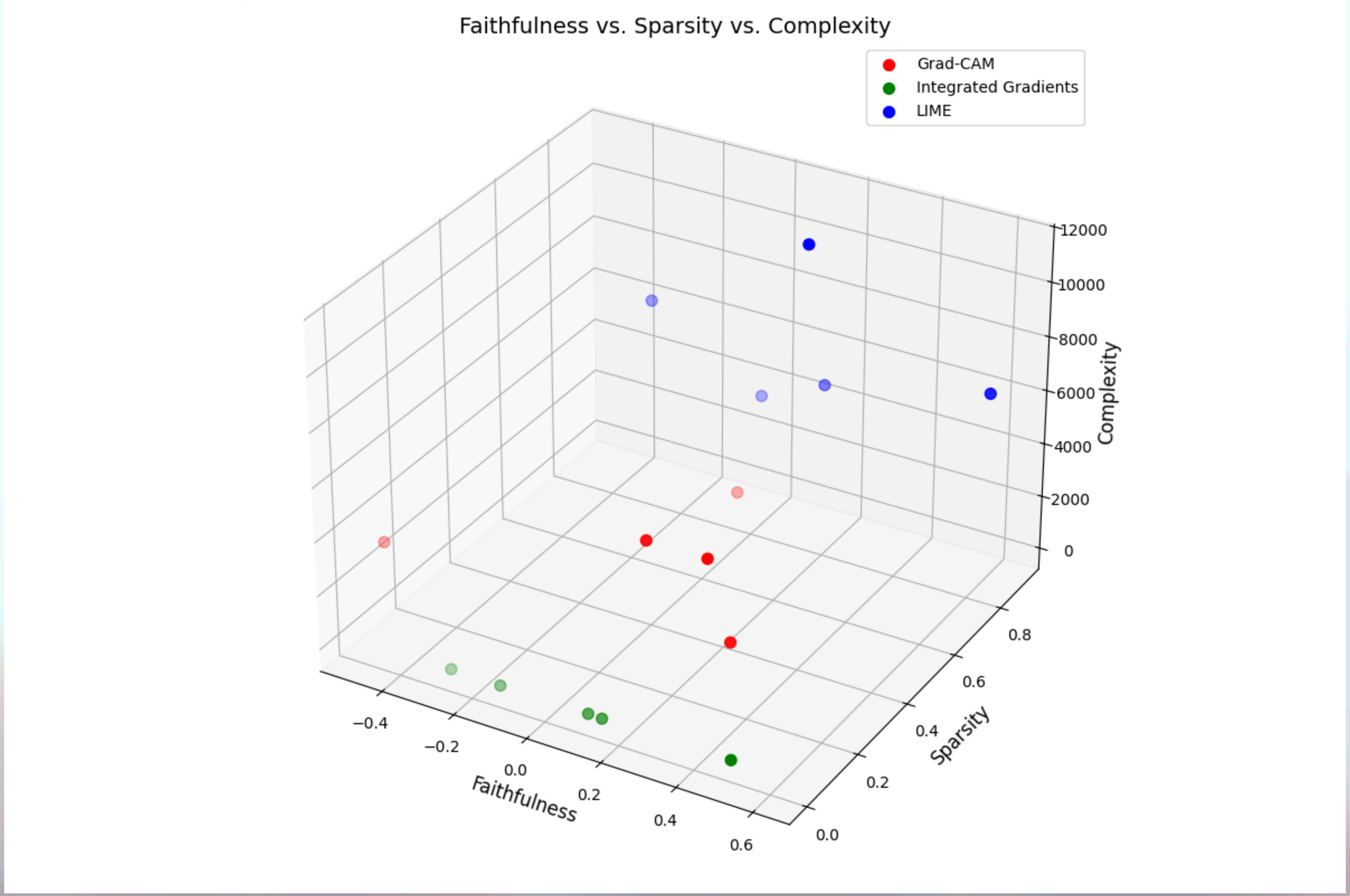
- Reflects the conciseness of the explanation—whether only the most critical features are highlighted.
- Sparse explanations are often more interpretable, especially for tasks involving humans (e.g., determining if the model focuses solely on glasses for face recognition).
- Compare the percentage of input features or pixels deemed "important" by each method.

.

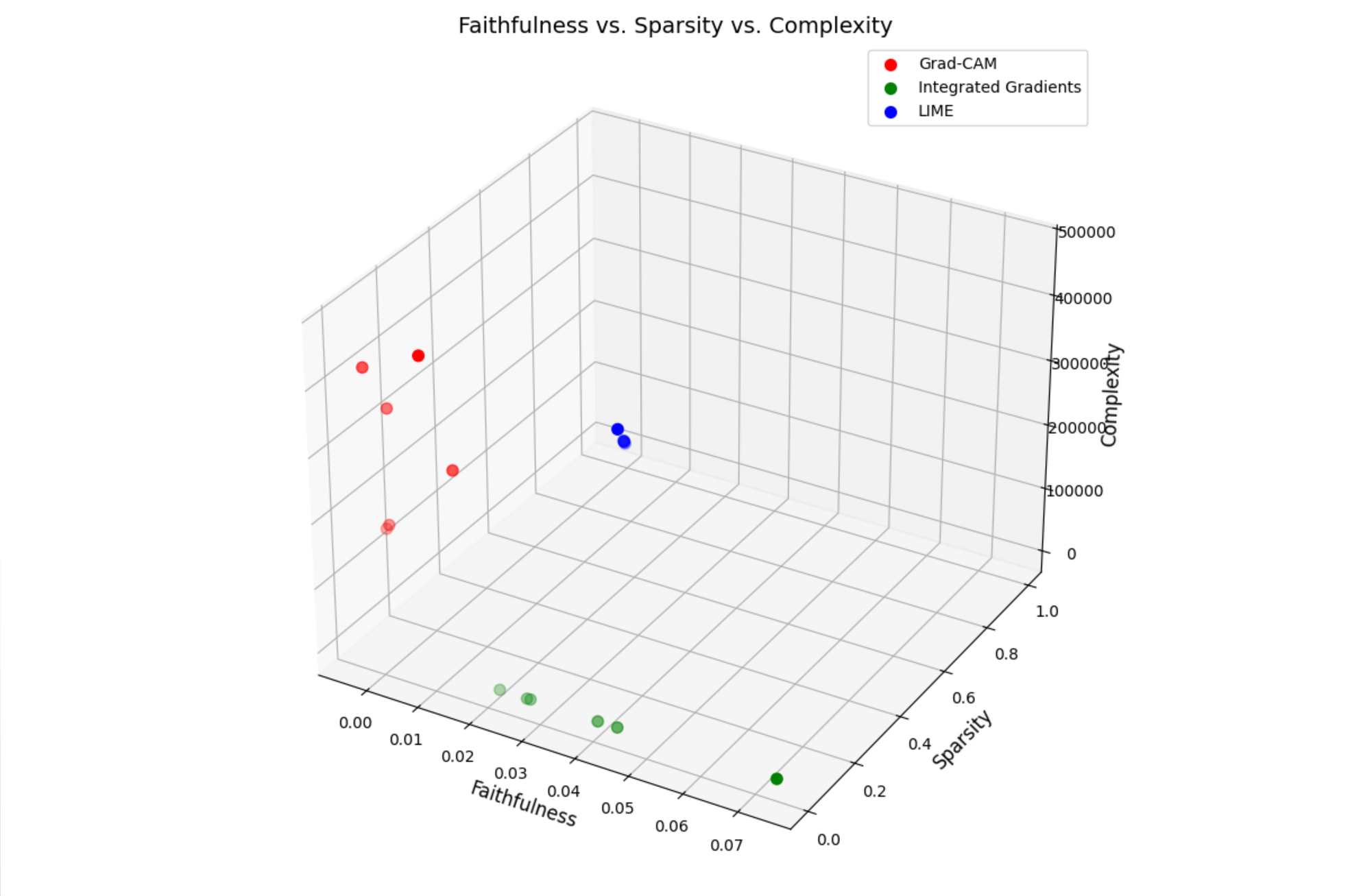
**Complexity:**

- Evaluates the computational overhead of generating explanations.
- Time taken to compute explanations and memory usage can be tracked for each method, especially relevant when scaling to larger datasets or real-time applications.

# Dataset 1 (Cat & Dog Classification)



# Dataset 2 (Face Glasses Recognition)



# Comparison of Methods:

| Metric | LIME | Integrated Gradients (IG) | Grad-CAM |
|---|---|---|---|
| **Faithfulness** | Moderate-High (depends on perturbation quality) | High (directly tied to model gradients) | High (gradient-based with activation focus) |
| **Sparsity** | High (selects a subset of features) | Moderate (dense feature attribution) | Moderate (regional focus, less sparse than LIME) |
| **Complexity** | High (perturbation and linear model fitting) | Moderate (computes gradients efficiently) | Low (relatively simple gradient computation) |

# Insights:

- **LIME** excels in sparsity but can suffer from lower faithfulness if the perturbations do not reflect real-world variations. Computational complexity is also higher, especially for large-scale image tasks.
- **Integrated Gradients** offers high faithfulness but is less sparse, as it assigns importance to all input features. Its computational complexity is moderate, making i balanced choice.
- **Grad-CAM** provides intuitive and faithful visualizations with low computational cost. However, it lacks the sparsity of LIME, making it less interpretable for tasks requiring concise explanations.

# Which one performs better?

- For **faithfulness**, both IG and Grad-CAM perform well, with IG offering more precise attributions.
- For **sparsity**, LIME is the best choice as it explicitly targets sparse explanations.
- For **complexity**, Grad-CAM is the most efficient method, followed by IG.

# Observations and Model Improvements

# Observations on based Metrics-

### 1. Faithfulness Observations:

- If explanations show poor faithfulness (e.g., removing "important" regions does not impact predictions), the model might rely on spurious correlations.
- **Action:** Refine the training dataset to reduce biases and include edge cases & Regularise the model to encourage reliance on relevant features.

### 2. Sparsity Observations:

- Sparse explanations (LIME) might reveal that the model bases predictions on limited, distinct regions (e.g., glasses area for face recognition).
- **Action:** Validate whether these regions are sufficient for accurate predictions. If not, encourage the model to consider broader context through data augmentation or attention mechanisms.

### 3. Complexity Observations:

- High complexity in LIME may hinder scalability for large datasets or real-time applications.
- **Action:** Prefer Grad-CAM or IG for production scenarios, as they are computationally efficient while maintaining high faithfulness.

# Model Improvements

## Dog/Cat Classification:

Use observations from IG and Grad-CAM to verify whether the model focuses on salient features like ears, tails, or faces. If not, retrain with additional targeted augmentations.

## Glasses Recognition:

Use sparse explanations (LIME) to ensure the model does not focus on irrelevant features like facial contours or hair. Fine-tune the model if necessary.

# Thank you !