Evaluation of explainability tools and methods in medical diagnosis

Presented by: Asfa Jamil , Luca Reggiani , Daniele Marini

Alma Mater Studiorum - University of Bologna
University of Cyprus

What is eXplainable AI ?

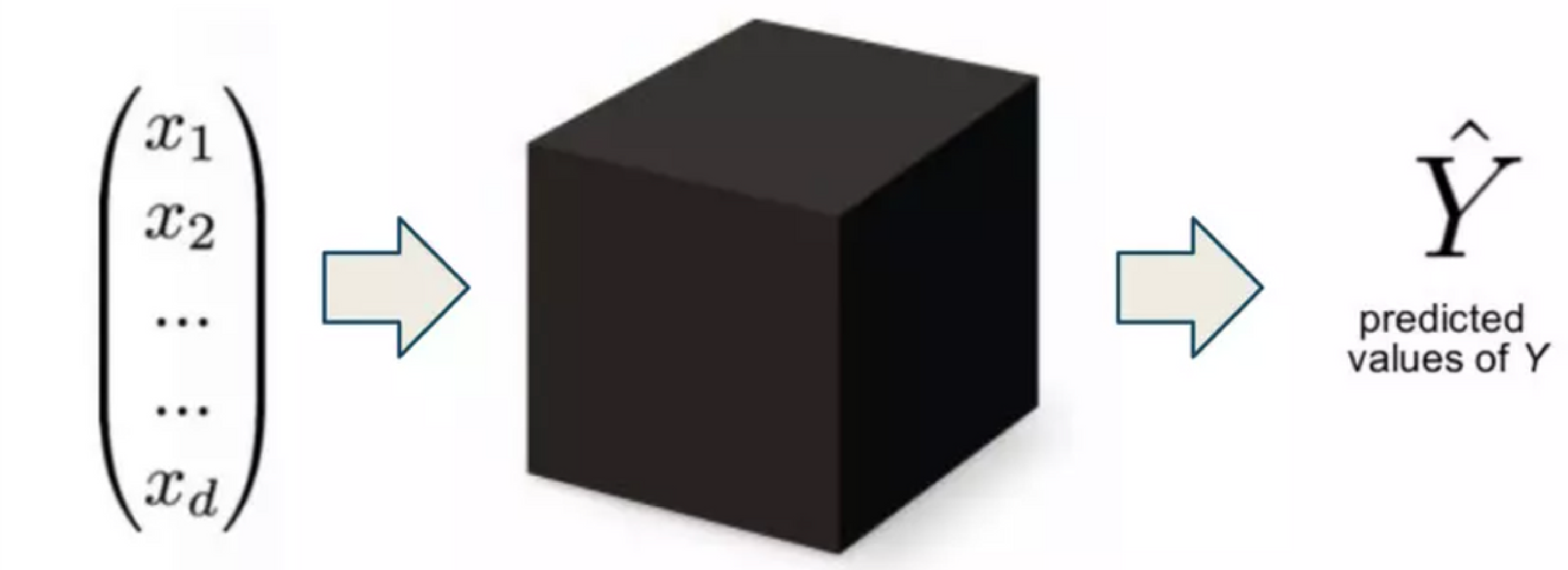
" Explainable AI refers to AI systems that provide transparent explanations for their decision-making, enabling humans to understand how and why the AI arrived at a particular output "

Why eXplainable AI ?

Understading the behaviour behind a machine learning model is extremely useful for many reasons:

- enhance the **trust** and the **confidence** of the users
- allow us to identify issues of the model
- identify biases in the dataset

Black Box



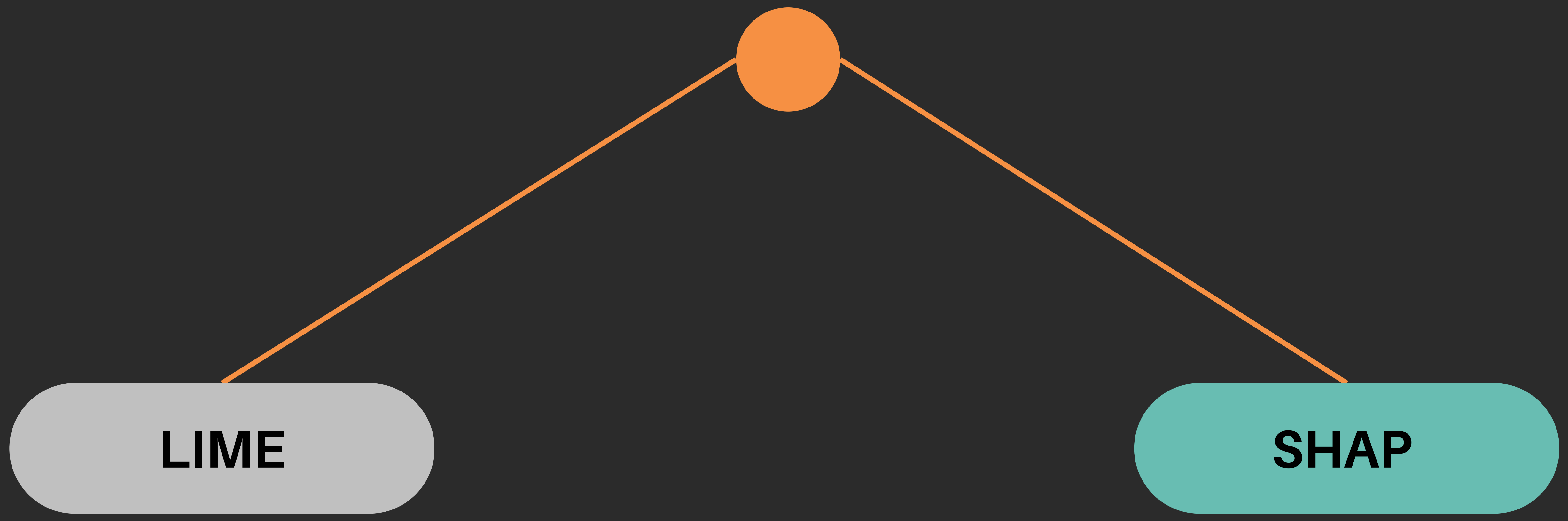
Input

Output

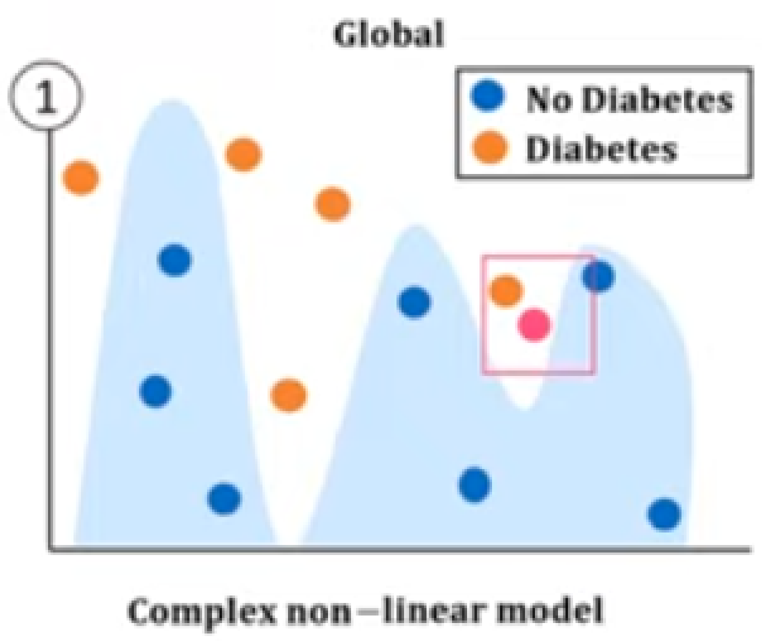
Model Agnostic techniques

Global interpretations

Local explaination

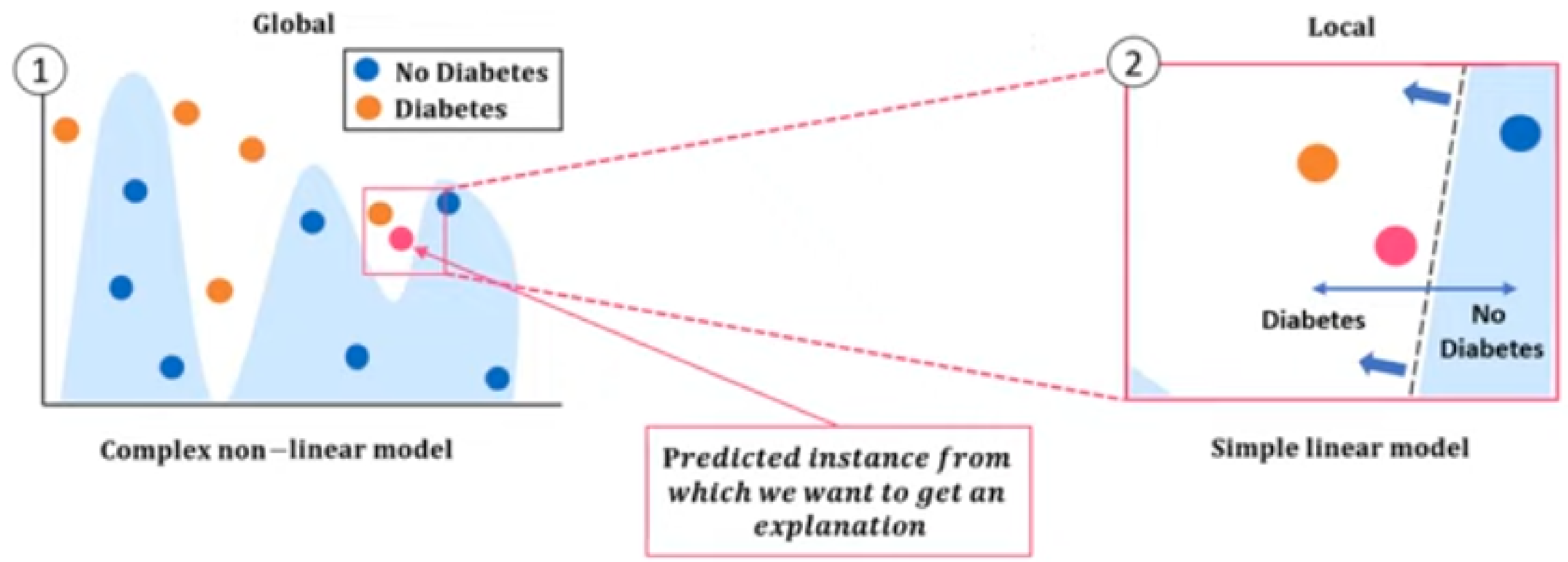


LIME
Local Interpretable Model agnostic Explanation

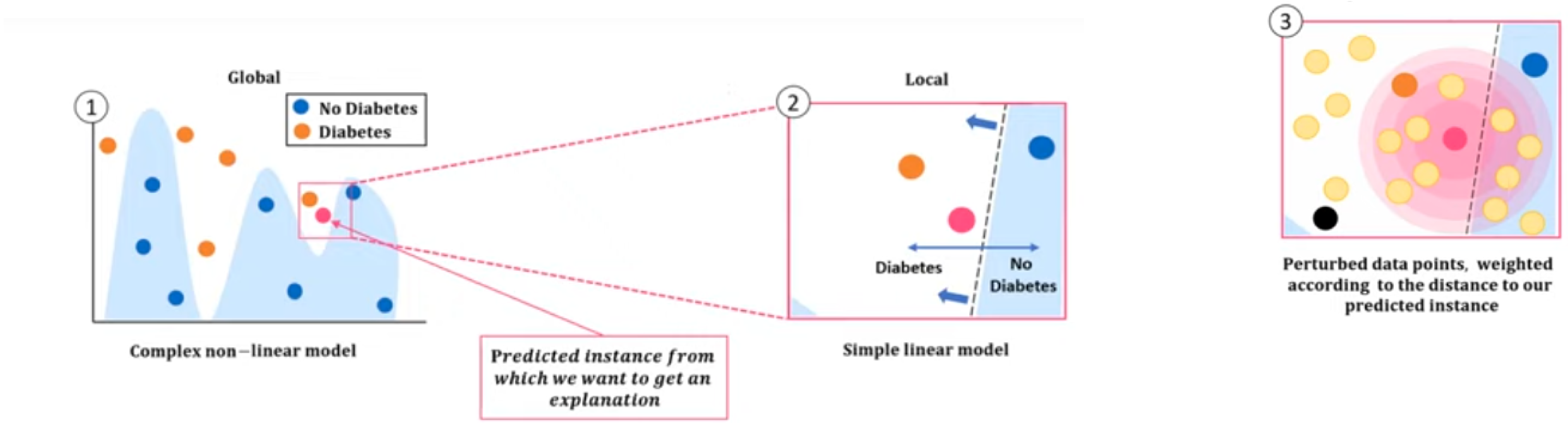


LIME

Local Interpretable Model agnostic Explanation

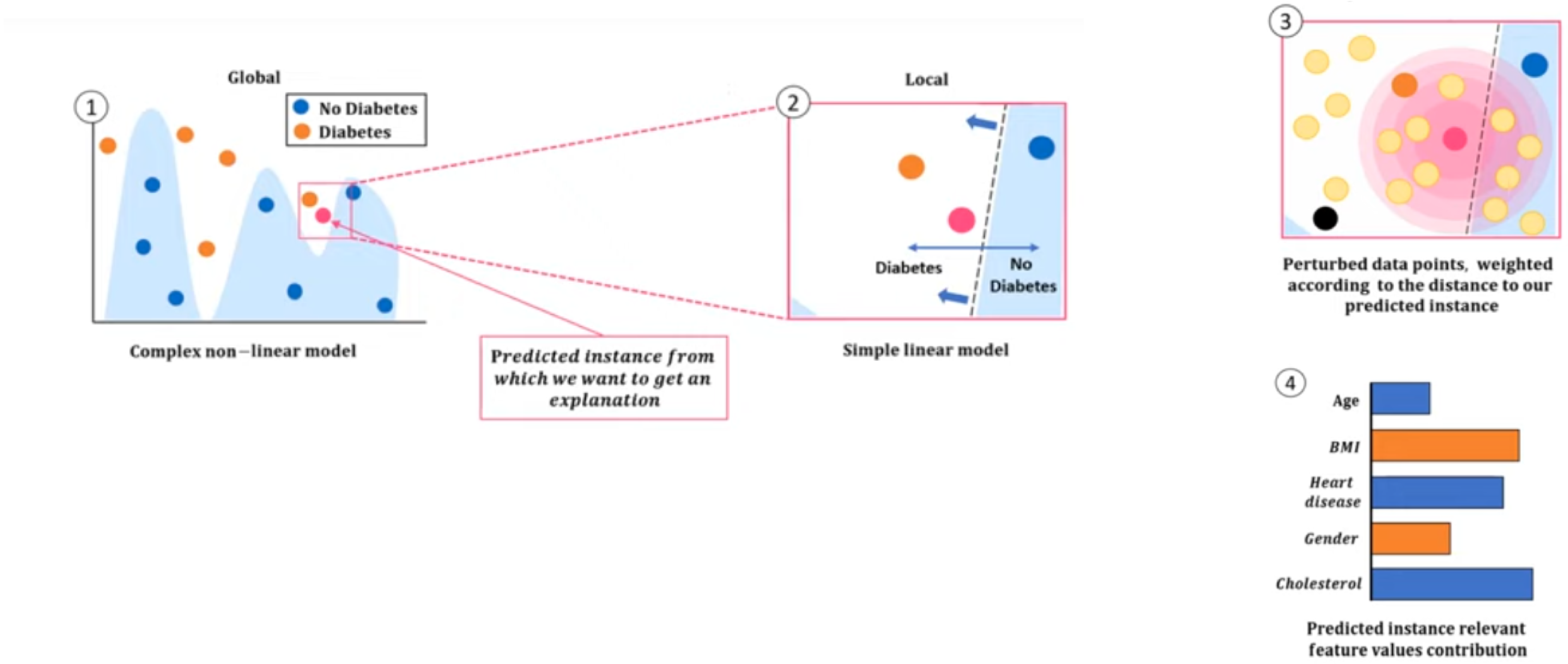


LIME
Local Interpretable Model agnostic Explanation



LIME

Local Interpretable Model agnostic Explanation



SHAP

SHapley Additive ExPlainations

Model-Agnostic technique to compute features contribution to a model output.

SHAP

SHapley Additive ExPlainations

Model-Agnostic technique to compute features contribution to a model output.

Accurate and consistent features importance values.

SHAP

SHapley Additive ExPlainations

Model-Agnostic technique to compute features contribution to a model output.

Accurate and consistent features importance values.

Based on the concept of the Shapely Values from cooperative game theory.

SHAP

SHapley Additive ExPlainations

Model-Agnostic technique to compute features contribution to a model output.

Accurate and consistent features importance values.

Based on the concept of the Shapely Values from cooperative game theory.

Two kind of explainations available

Local Explainations

Global Explainations

SHAP

SHapley Additive ExPlainations

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$

SHAP

SHapley Additive ExPlainations

Shapley value for feature i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$

SHAP

SHapley Additive ExPlainations

Shapley value for
feature i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \underbrace{\left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].}_{\text{Marginal Contribution}}$$

SHAP

SHapley Additive ExPlainations

Shapley value for feature i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \underbrace{\left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)\right]}_{\text{Marginal Contribution}}.$$

Subset of features

Total number of features (without i)

Marginal Contribution

SHAP

SHapley Additive ExPlainations

Shapley value for feature i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \overbrace{\frac{|S|!(|F| - |S| - 1)!}{|F|!}}^{\text{Weighting}} \underbrace{\left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]}_{\text{Marginal Contribution}} .$$

Subset of features

Total number of features (without i)

SHAP

SHapley Additive ExPlainations

Weighting

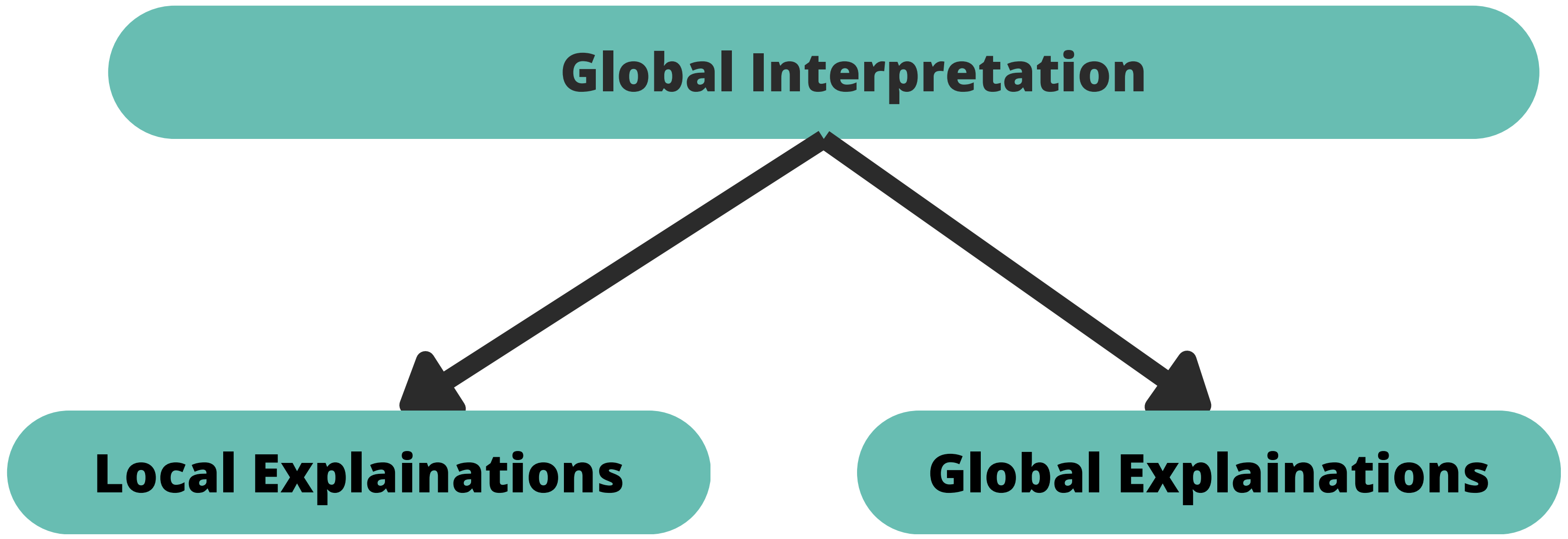
Shapley value for feature i

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right].$$

Subset of features

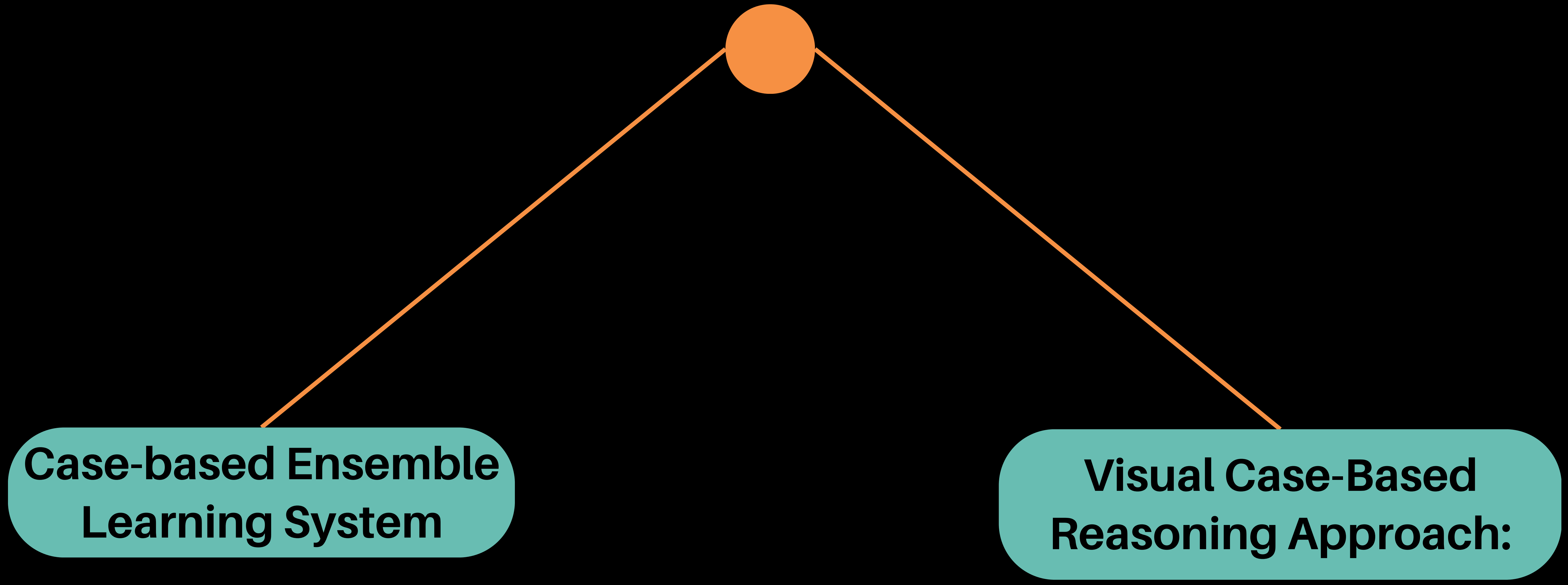
Total number of features (without i)

Marginal Contribution



Case based Reasoning Techniques

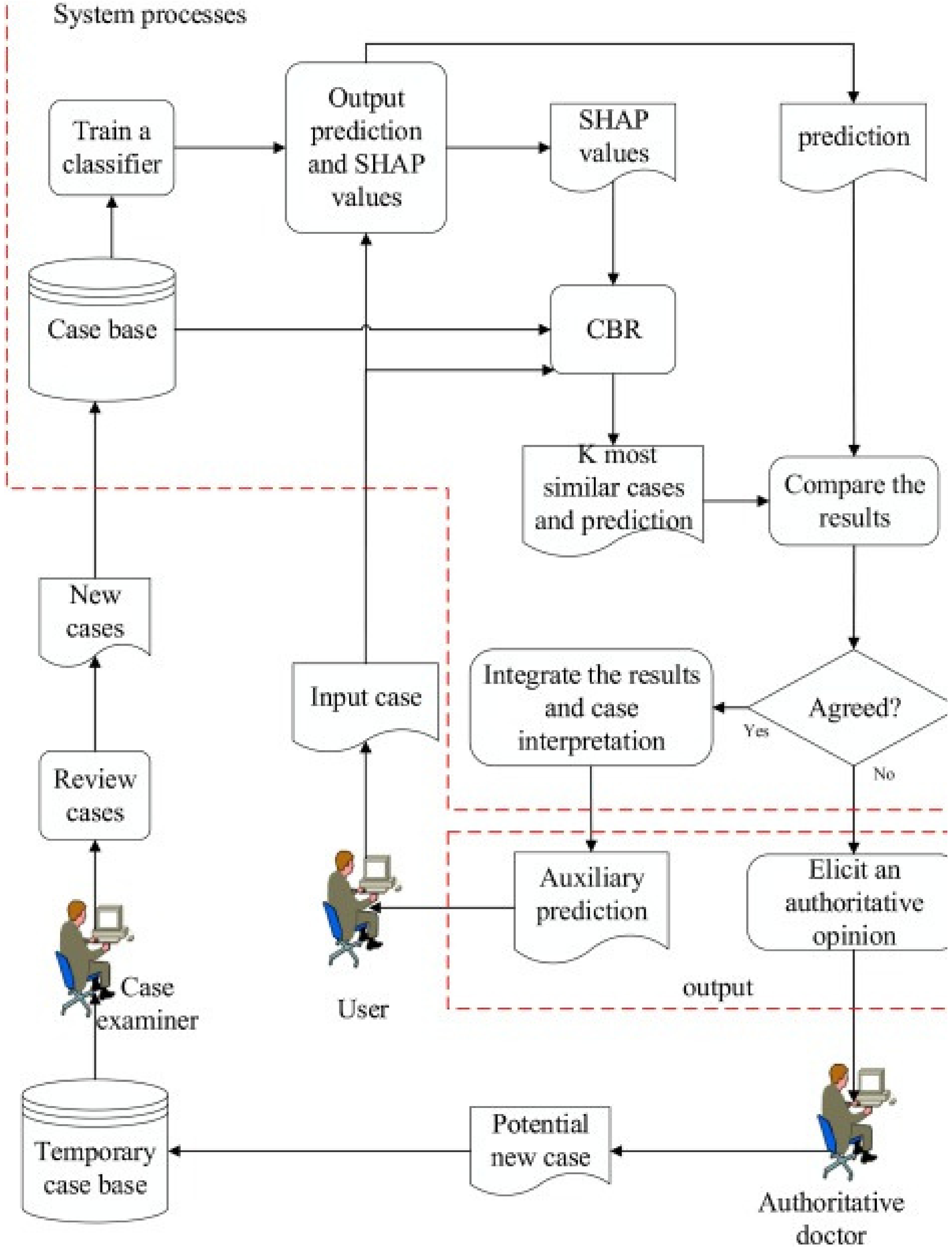
Case-based reasoning (CBR) is a problem-solving methodology that relies on past experiences, or "cases," to solve new problems.



Approach 1: Case-based Ensemble Learning System

Combining case-based reasoning (CBR) and ensemble learning

Qualitative Explanations and prediction of breast cancer recurrence


System processes
Train a classifier
Output prediction and SHAP values
SHAP values
prediction
Case base
CBR
K most similar cases and prediction
Compare the results
New cases
Input case
Integrate the results and case interpretation
Agreed?
Yes
No
Review cases
Auxiliary prediction
Elicit an authoritative opinion
output
Case examiner
User
Temporary case base
Potential new case
Authoritative doctor

Case-based Ensemble Learning System Methodology Overview

Data preprocessing: 1,286 breast cancer patient data

Ensemble learning: XGBoost implementation
Case-based reasoning: Justification of prediction reasoning

10-fold cross-validation and user survey

Case-based Ensemble Learning System Experimentation and Results

Outperforms logistic regression, SVM, random forest, and deep learning

Superior performance in accuracy, sensitivity, specificity, and AUC-ROC

Survey among oncologists: found to be useful and easy to use

Case-based Ensemble Learning System Advantages

- High accuracy
- Interpretable
- User-friendly
- Enhances clinical decision-making

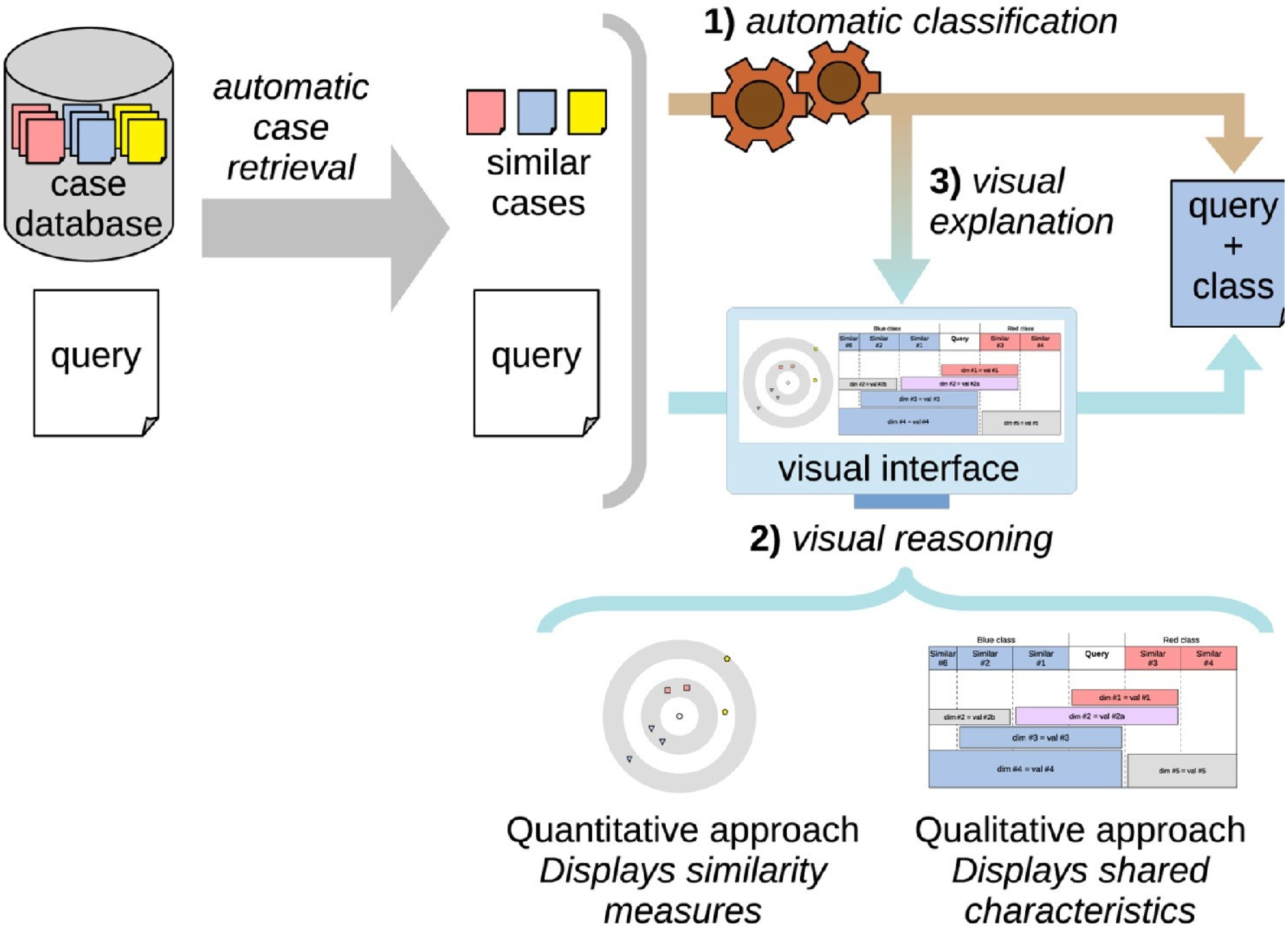
Case-based Ensemble Learning System Limitations

- Limited data
- Limited features
- Limited evaluation
- Limited scalability

Approach 2: Visual Case-Based Reasoning

User-friendly visual interface for exploring similarities

Qualitative and quantitative explanations


1) automatic classification
case database
query
automatic case retrieval
similar cases
query
3) visual explanation
query + class
visual interface
2) visual reasoning
Blue class
Red class
Similar #6
Similar #2
Similar #1
Query
Similar #3
Similar #4
dim #1 = val #1
dim #2 = val #2b
dim #2 = val #2a
dim #3 = val #3
dim #4 = val #4
dim #5 = val #5
Quantitative approach
Displays similarity measures
Qualitative approach
Displays shared characteristics

Approach 2: Visual Case-Based Reasoning Methodology

- Data gathering
- Feature identification
- Similarity calculation
- Case retrieval
- Visual interface
- Automatic algorithm
- Explanation generation

Approach 2: Visual Case-Based Reasoning Datasets

- The Breast Cancer Wisconsin (BCW) dataset
- The Mammographic Mass (MM) dataset
- The Breast Cancer (BC) dataset

Approach 2: Visual Case-Based Reasoning Experiments and Results

Visual CBR outperforms conventional CBR: 85% vs. 75% accuracy rate

Superior precision and recall measures for visual CBR

Positive user feedback on interface and decision explanations

Approach 2: Visual Case-Based Reasoning Advantages

- Accuracy
- Explainability
- Usability
- Adaptability

Approach 2: Visual Case-Based Reasoning Disadvantages

- Limited applicability
- Data availability
- Technical proficiency

Post Hoc Approach: Explaining Individual Classification Decisions

Approximating the classifier by using simple classifer

Quantitative explainability with limitation in qualitative insights

Explaining Individual Classification Decisions Methodology

- local explanation vectors as class probability gradients
- Gaussian Process Classification (GPC)
- Approximation of classifier
- Selection of an appropriate classifier
- Estimation of local explanations

Explaining Individual Classification Decisions
Results and Experimentation

- Application to SVM classifier
- Outperforms other methods (LIME, SHAP) in terms of accuracy and computational efficiency

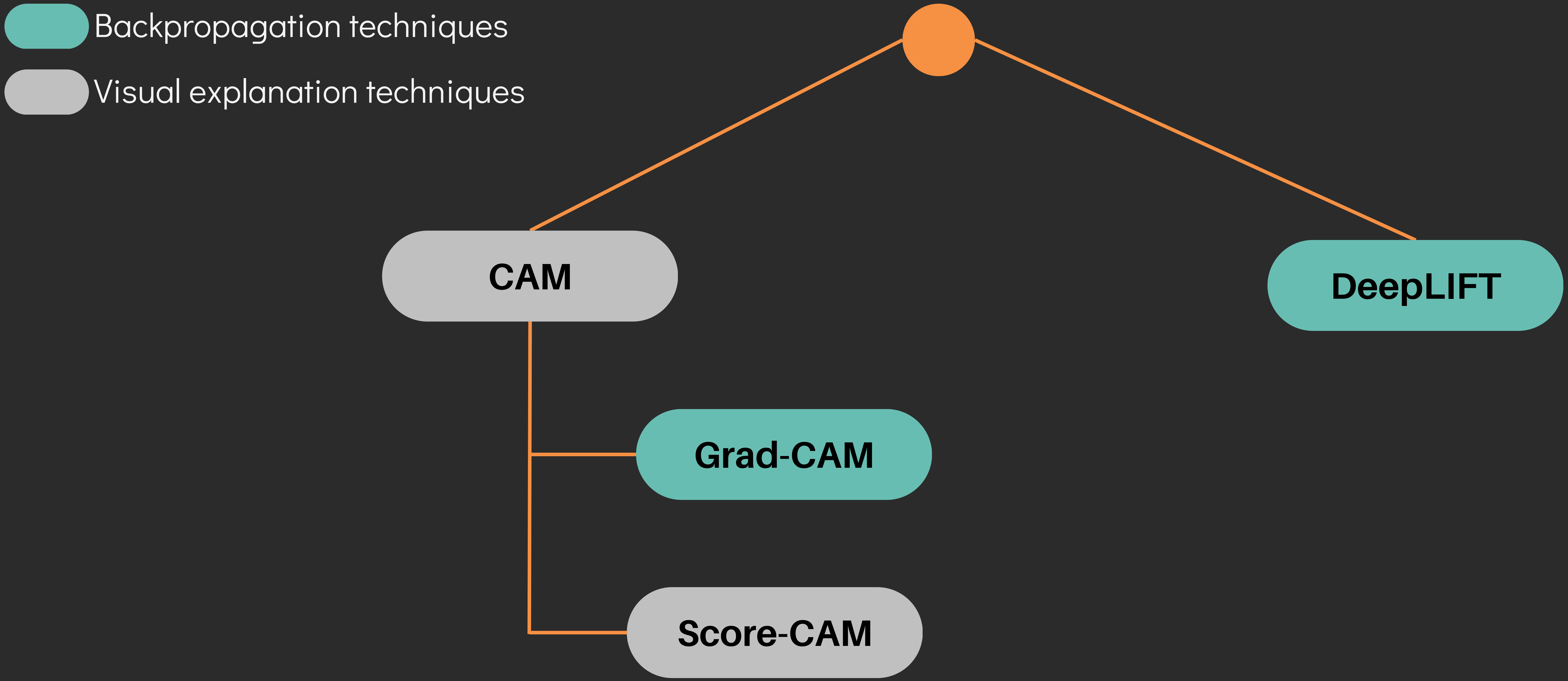
Explaining Individual Classification Decisions Advantages

- Quantitative measure of feature importance through local explanation vectors
- Capability to handle complex models and high-dimensional data
- Flexibility to apply to different types of classifiers

Explaining Individual Classification Decisions Disadvantages

- Dependency on accurate classifier approximation
- Focus on local data properties rather than global properties
- Potential high computation time for large datasets

Backpropagation-based techniques

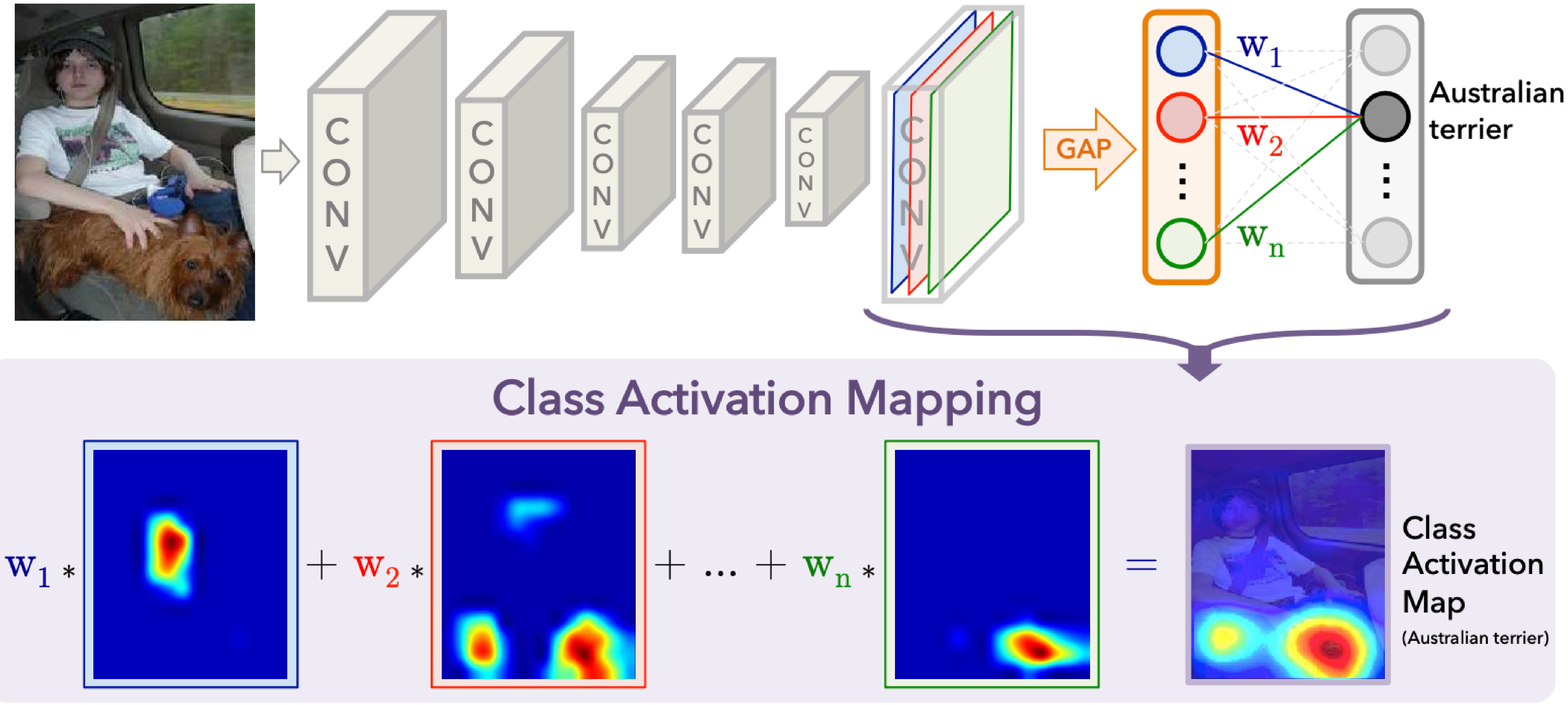


CAM
Class Activation Mapping

$$L^c_{CAM} = \sum_k \alpha^c_k A^k_{l-1}$$

where

$$\alpha^c_k = w^c_{l,l+1}[k]$$



Bolei Zhou et al, *Learning Deep Features for Discriminative Localization*, 2016

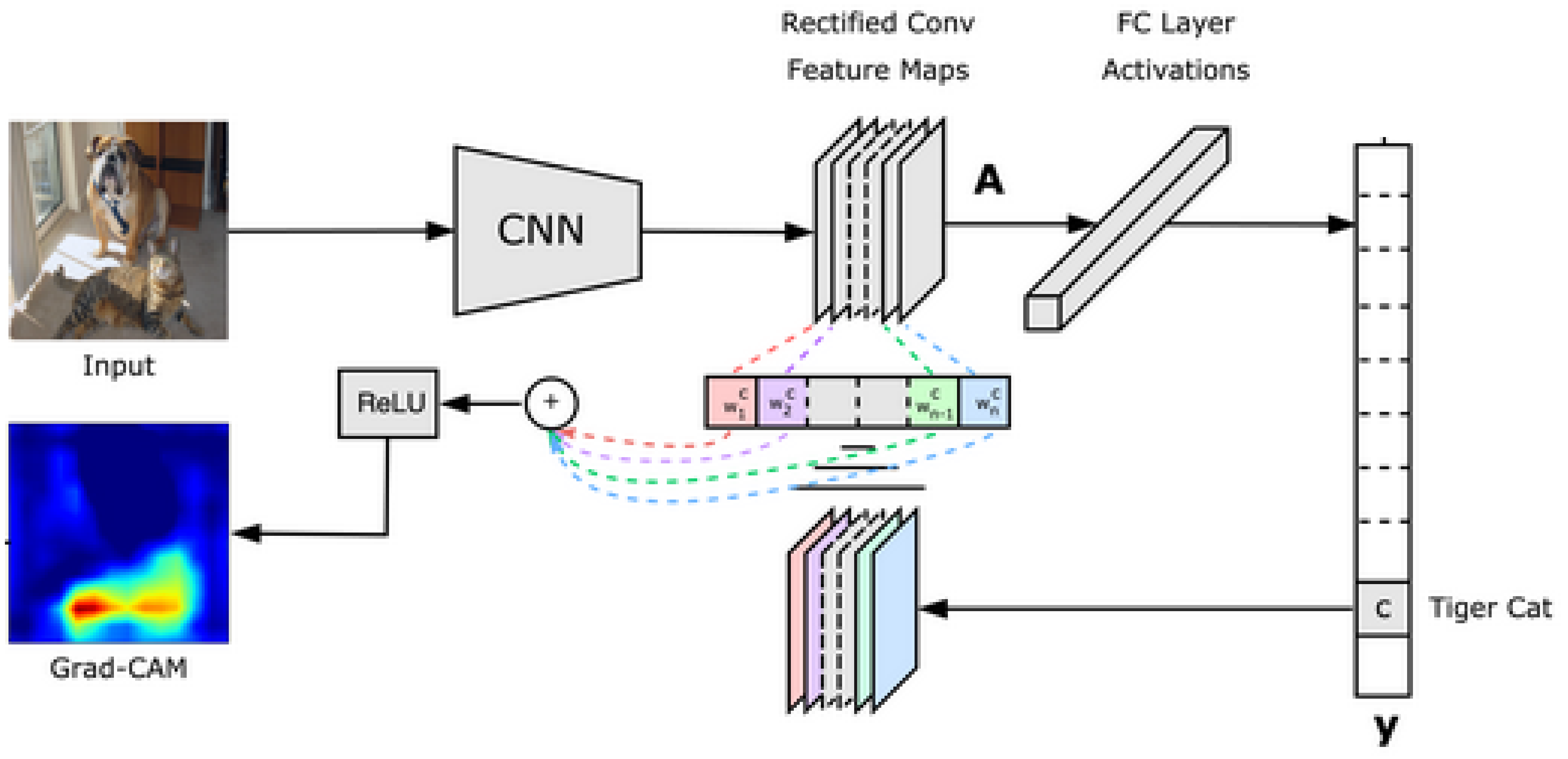
Grad-CAM

Gradient weighted Class Activation Mapping

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k_l\right)$$

$$where$$

$$\alpha^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{GAP}} \frac{\partial Y^c}{\partial A^k_{ij}}$$



Michael Cogswell et al, Grad-cam: Visual explanations from deep networks via gradient- based localization, 2017

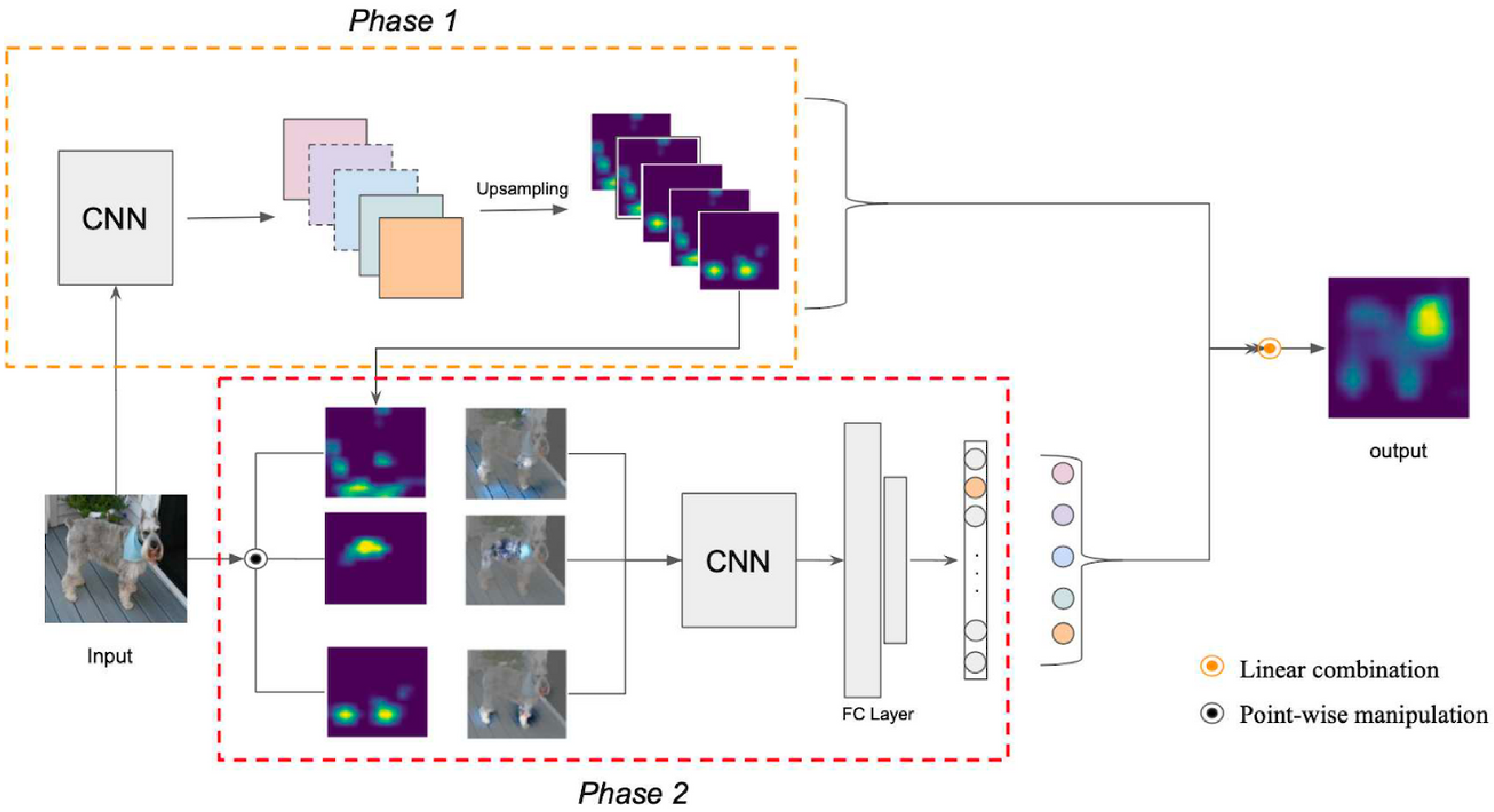
Score-CAM

Score weighted Class Activation Mapping

$$L^c_{Score-CAM} = ReLU\left(\sum_k \alpha^c_k A^k_l\right)$$

where

$$\alpha^c_k = C(A^k_l)$$



Haofan Wang et al, Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020

Score-CAM

Score weighted Class Activation Mapping

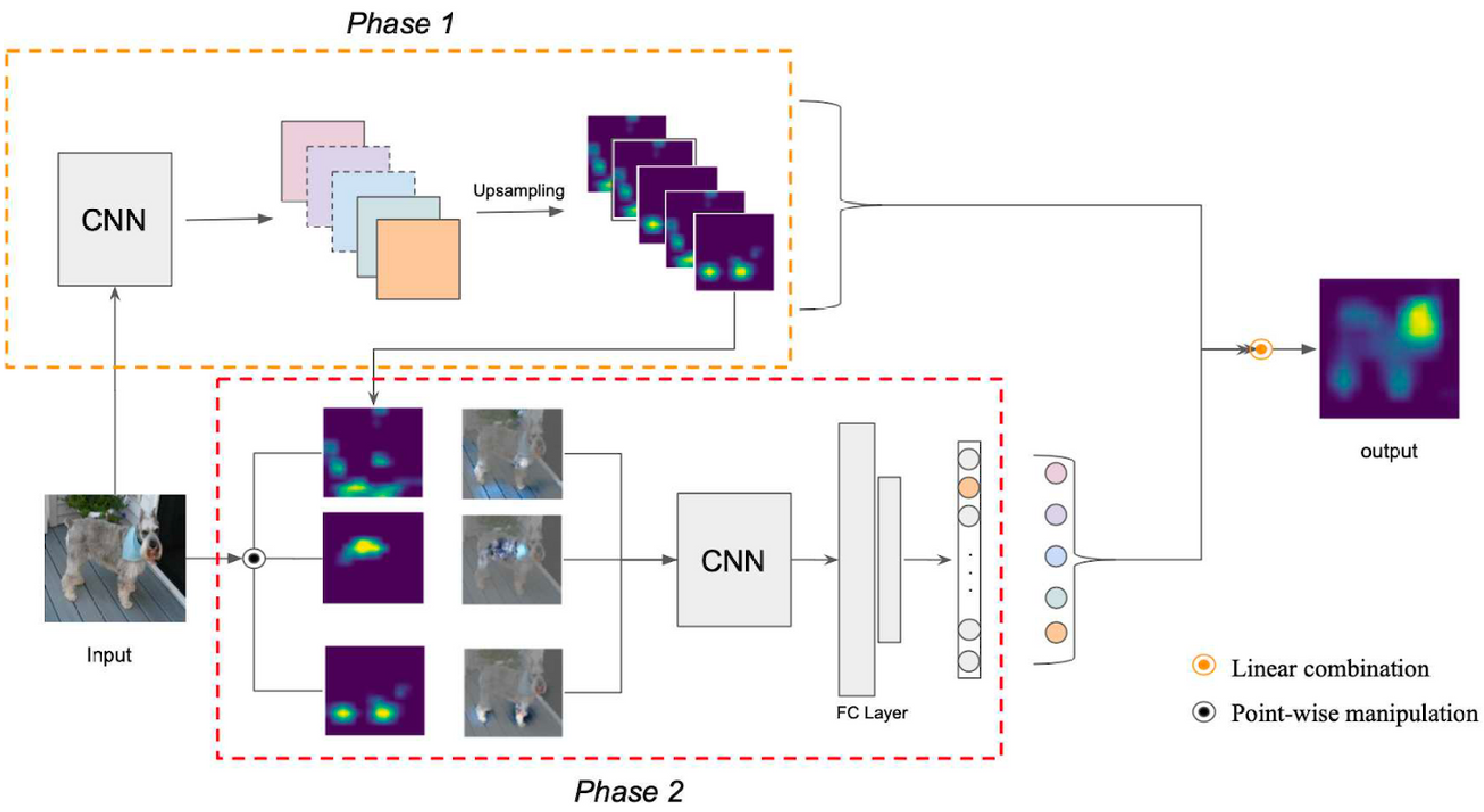
$$L^c_{Score-CAM} = ReLU\left(\sum_k \alpha^c_k A^k_l\right)$$

$where$

$$\alpha^c_k = C(A^k_l)$$

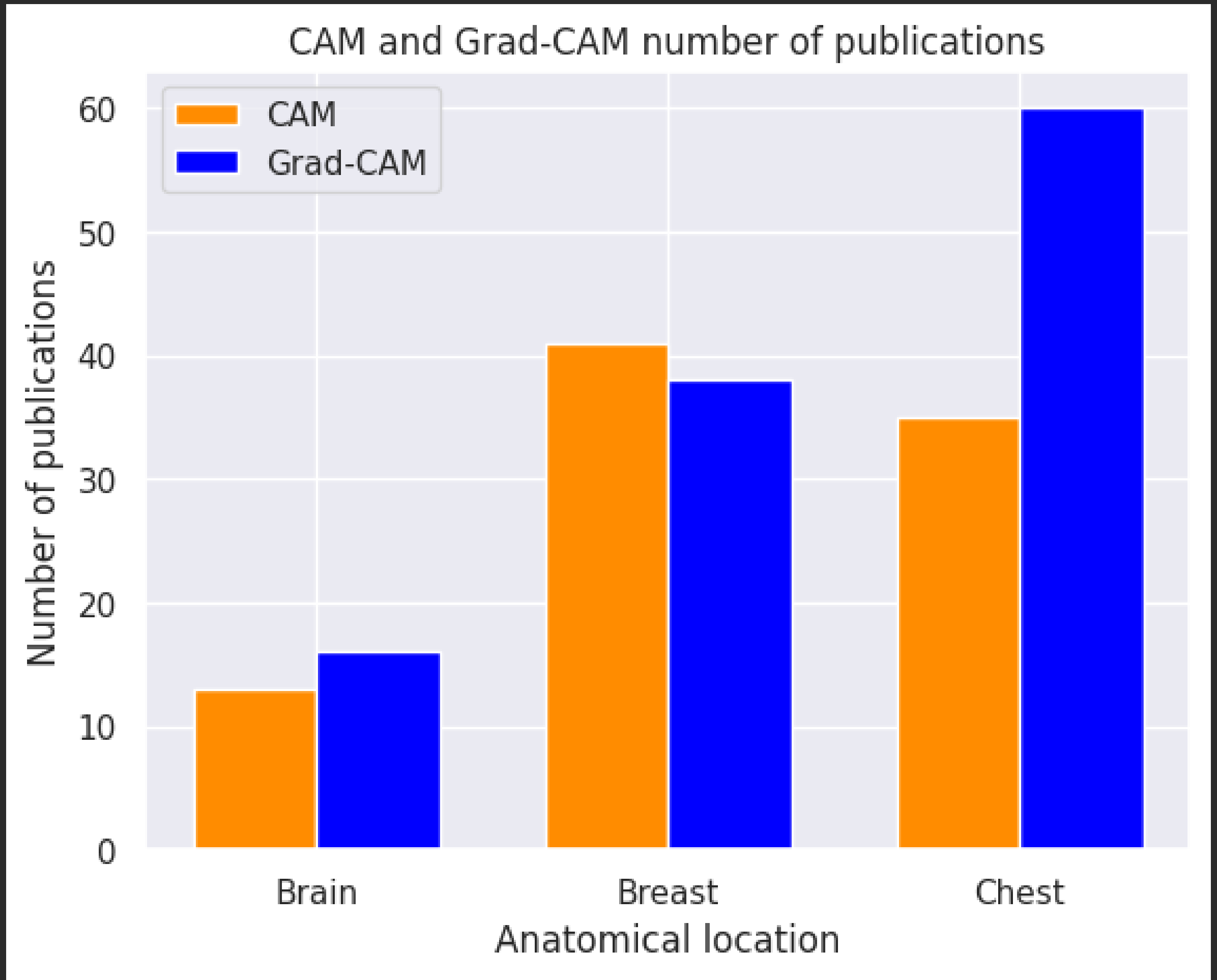
$$C(A^k_l) = f(X \circ H^k_l) - f(X_b)$$

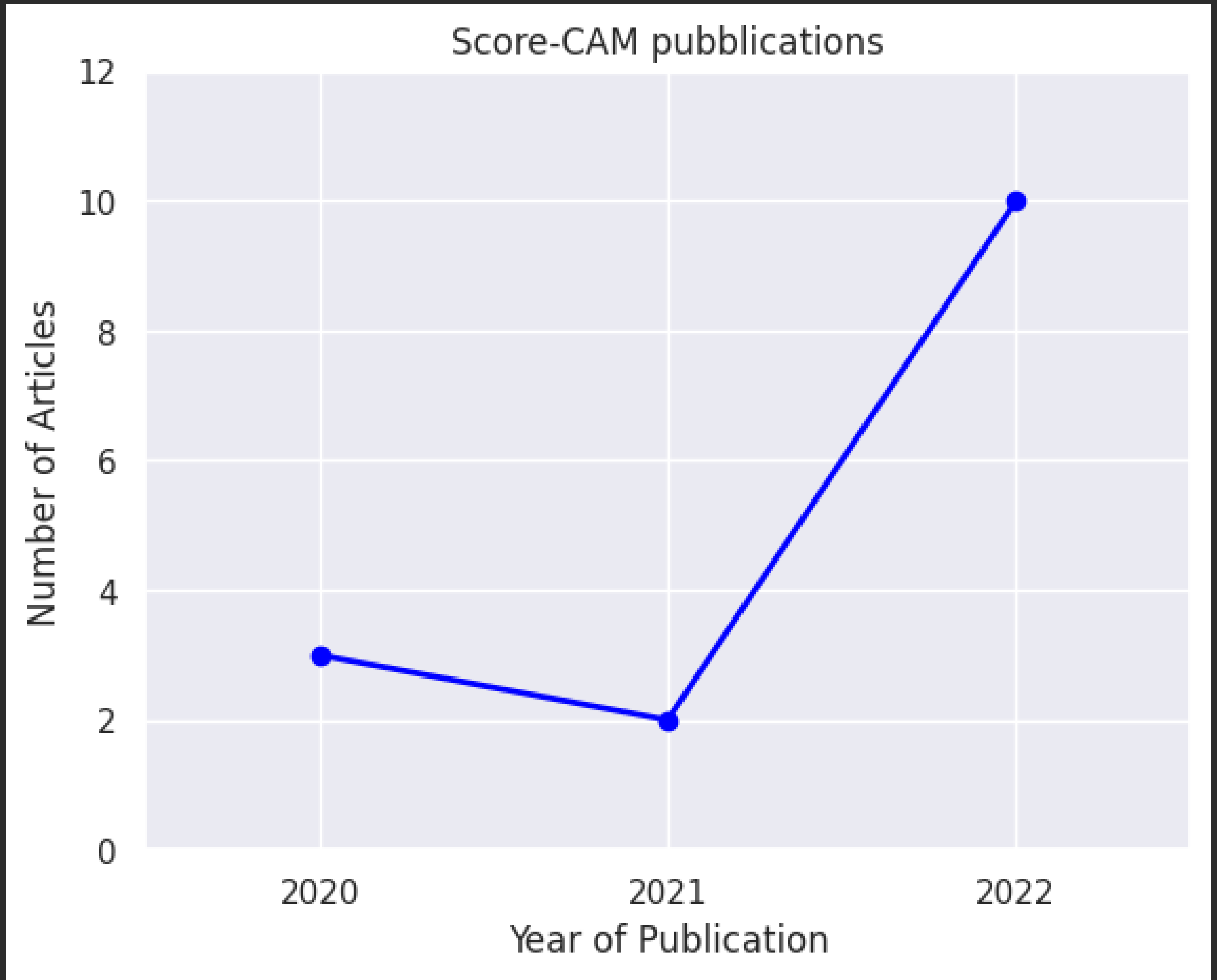
$$H^k_l = s(Up(A^k_l))$$



Haofan Wang et al, Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020

Healthcare Applications





DeepLIFT

Deep Learning Important FeaTures

1 **Define a reference value:**

- Select a reference value for each feature or variable in the input

2 **Compute the baselines:**

- Propagate the reference values through the neural network
- Calculate the expected activation of each neuron

3 **Propagate the actual input:**

- Perform a forward pass with the actual input values
- Compute the activations of each neuron

4 **Compute the contribution:**

- Compare the activations obtained with the actual input and the baseline activations obtained from the reference values

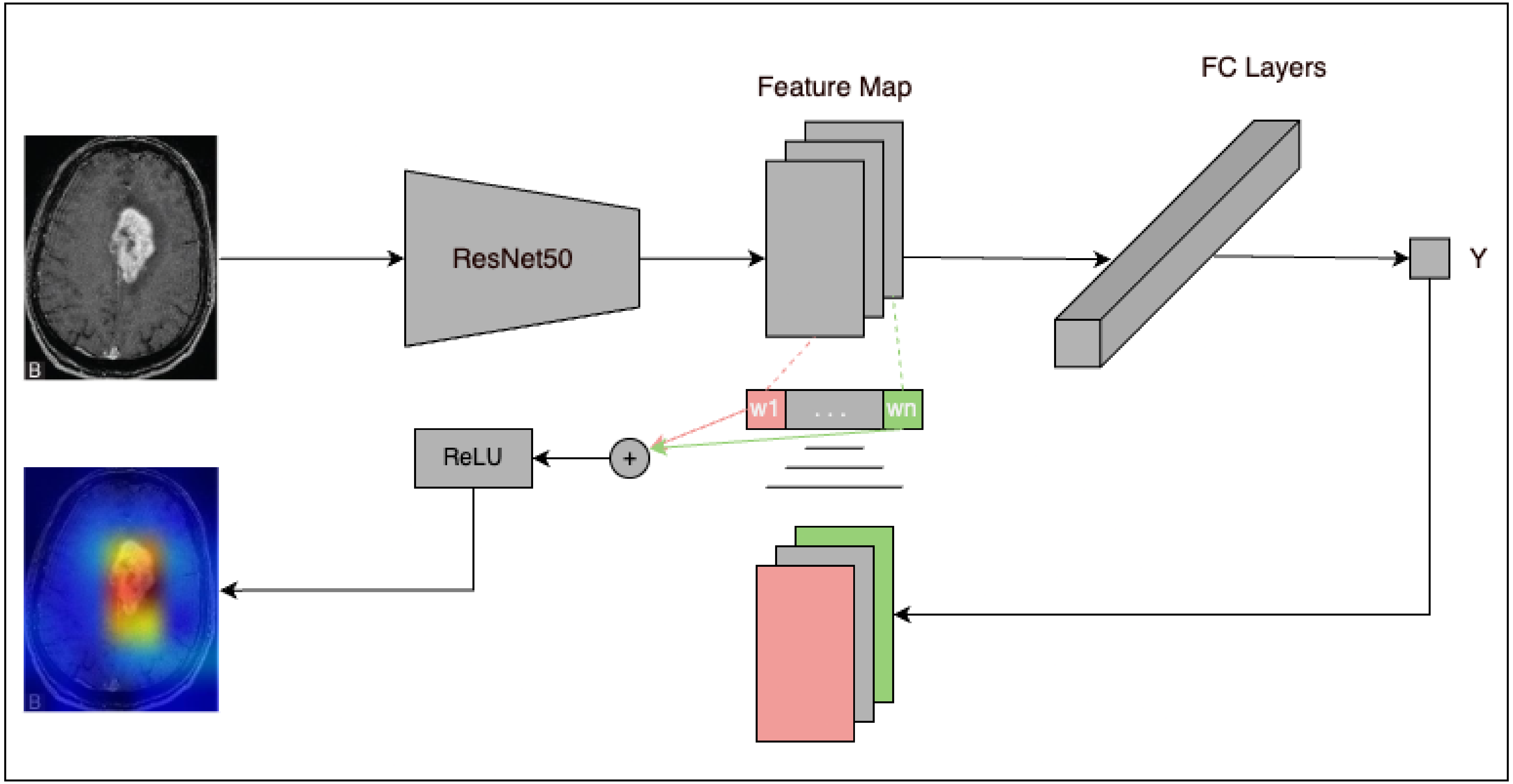
5 **Assign the importance scores:**

- Scale the contribution values to assign importance scores to each input feature

Experimental Analysis

Setup

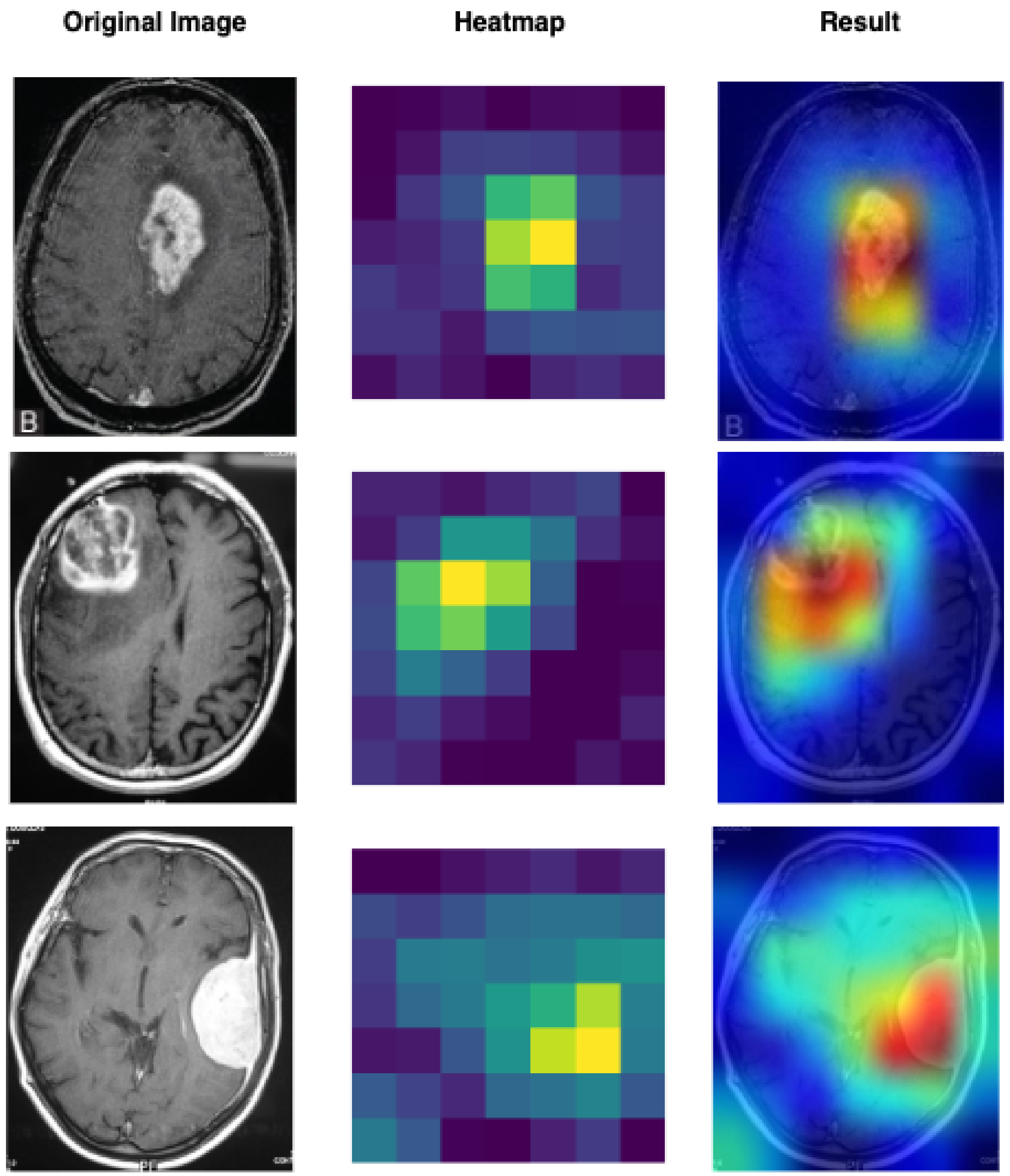
- Fine-tuned ResNet50
- Applied Grad-CAM to visualize the visual explanation



Experimental Analysis

Results

- Satisfactory localization ability
- Strong dependence from the classification model



https://github.com/daniele-marini/Grad-CAM

Thank you