1.5em 0pt

UNIVERSITÀ
DI TRENTO
Department of
Information Engineering and Computer Science

Masters Degree in
Computer Science

Project Course Report

Explainable Artificial Intelligence (XAI) and Ensemble Approaches in Image Classification

Supervisor

Prof. Giovanni Iacca

Student

Saad Raza Hussain Shafi

256453

Academic year 2024/2025

Contents

Abstract

The rapid advancement of deep learning, particularly in image classification tasks, has led to highly accurate models capable of achieving state-of-the-art performance. However, the "black-box" nature of these models raises significant concerns, especially in high-stakes domains like healthcare, where model interpretability is critical for trust and decision-making. Explainable Artificial Intelligence (XAI) aims to address these challenges by providing insights into the inner workings of deep learning models, enabling better understanding and validation of their predictions. This paper explores the integration of multiple XAI techniques—specifically Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM)—for enhancing the interpretability of image classifiers. The study begins by individually applying these techniques to a chosen image classification task, analyzing their strengths and limitations in terms of computational efficiency, localization effectiveness, and alignment with human expectations. The central aim of the research is to investigate how combining these techniques in an ensemble framework can generate more robust, reliable, and accurate explanations compared to using any single method. Different ensemble strategies are examined, including simple averaging, weighted averaging based on technique performance, and advanced logistic stacking, to assess their effectiveness in enhancing localization accuracy and overall explanation quality. The results demonstrate that individual XAI methods, such as LIME's high recall for relevant regions and Grad-CAM's computational efficiency, each provide valuable insights but also have notable shortcomings. The ensemble strategies, particularly the advanced logistic stacking approach, significantly outperform individual methods and simple ensembles by offering better balance in precision, recall, and localization metrics. These findings underline the importance of ensemble XAI approaches in improving the interpretability and trustworthiness of deep learning models, especially for complex and critical image classification tasks like cancer detection. By systematically combining multiple interpretability methods, this work offers a pathway to more transparent, reliable, and clinically applicable AI systems.

1 Introduction

Deep learning models, particularly Convolutional Neural Networks (CNNs), have revolutionized image classification, achieving state-of-the-art performance in various applications, from medical diagnostics to autonomous vehicles. However, these models' "black-box" nature often leaves users questioning how decisions are made, especially in critical fields like healthcare where interpretability is paramount. Explainable Artificial Intelligence (XAI) seeks to bridge this gap, providing transparency in machine learning models by offering explanations for their predictions.

In the context of image classification, various XAI techniques have been developed to highlight regions of an image that significantly influence a model's decision. Three of the most prominent techniques are Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM). Each of these techniques has its strengths and limitations in terms of accuracy, computational efficiency, and interpretability.

This study aims to explore how combining these techniques into an ensemble approach can yield more robust and reliable explanations than using each technique in isolation. The primary goal is to determine the best method for integrating different XAI techniques to improve the accuracy of localization and enhance the trustworthiness of the generated explanations. By applying this ensemble approach to a standard image classification task, we seek to demonstrate the potential of combining multiple XAI techniques to address the challenges of explaining complex models in critical real-world applications.

In this paper, we first evaluate each technique independently, analyzing their performance on a dataset and their ability to localize important image regions. We then propose several ensemble strategies, ranging from simple averaging to more complex methods such as logistic stacking, to investigate how different combinations of these techniques can enhance the overall explanation quality. Through rigorous evaluation using various metrics, including decision impact ratio, cosine similarity, and accordance recall, we demonstrate that ensemble methods significantly improve interpretability, providing more precise and actionable insights into the inner workings of deep learning models.

2 Literature Review

The advancement of deep learning, particularly in image classification, has led to highly accurate models, yet their inherent "black-box" nature poses significant challenges, especially in critical domains like healthcare[7]. Explainable Artificial Intelligence (XAI) has emerged as a crucial field addressing this issue by providing insights into how these complex models arrive at their decisions[13]. This literature review explores the landscape of XAI, distinguishing between model-agnostic and model-specific techniques, delving into prominent methods such as Grad-CAM, LIME, and SHAP, and finally examining the burgeoning field of ensemble XAI and its documented improvements in explanation quality and trustworthiness.

2.0.1 Understanding XAI and its Methodologies

Explainable AI (XAI) in deep learning typically employs saliency maps, heatmaps, or attention maps to highlight regions of an image that are deemed important for a classification decision [5][7]. These visualizations aim to make deep learning systems more transparent to developers, validators, and end-users. However, a significant limitation of many existing XAI methods is their failure to accurately identify regions that human experts consider meaningful, leading to a disparity in explanations between AI and human experts, for instance, in medical imaging [10, 13]. While heatmaps can be useful for developers in identifying irrelevant regions the classifier focuses on, they often do not align with the semantically meaningful, labeled explanations naturally produced and expected by human experts [5].

XAI methods can be broadly categorized into two main approaches:

- **Model-Agnostic Methods:** These techniques operate independently of the underlying deep learning model's architecture. They typically involve manipulating input data (e.g., pixels or superpixels) and observing how these perturbations affect the model's output[7]. If a change in an input region significantly impacts the classification decision, that region is considered important. **Local Interpretable Model-Agnostic Explanation (LIME)** and **Shapley Additive Explanations (SHAP)** are prominent examples within this category.
- **Model-Specific Methods (Activation/Weight-Based):** These methods directly examine the internal components of the deep neural network, such as activations or weights, to pinpoint regions of importance[5]. Examples include **Grad-CAM**, Integrated Gradients, Saliency, GradientShap, and Layerwise Relevance Propagation (LRP) [5, 13].

Key XAI Techniques: Grad-CAM, LIME, and SHAP

Grad-CAM

(Gradient-weighted Class Activation Mapping) is a model-specific technique that has gained popularity for making Convolutional Neural Networks (CNNs) more transparent by visualizing input regions critical for predictions[5]. It works by using the averaged gradient score as weights for the feature map from the last convolutional layer, thereby highlighting discriminative regions of the image. While effective, Grad-CAM has certain shortcomings; it may struggle to localize an object if there are multiple instances in an image or localize only a portion of an object due to the unweighted average of partial derivatives. Fundamentally, Grad-CAM tends to highlight general areas within an image[8].

Grad-CAM computes the class-discriminative localization map for a class c as follows:

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

where * $A^k \in \mathbb{R}^{u \times v}$ is the activation map of the k-th channel in the last convolutional layer. * The weights α_k^c are obtained by global average pooling of the gradients of the class score y^c wṙ.t. the feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

with $Z = u \times v$ being the number of spatial locations. * The ReLU ensures that only the features positively correlated with the target class contribute to the localization

Grad-CAM++
is an extension designed to address the limitations of its predecessor. It offers an improved measure of importance for each pixel in a feature map, ensuring that all spatially relevant regions of the input image are equally emphasized. This enhancement allows Grad-CAM++ to localize the entire object more effectively, even when multiple occurrences of the same object are present [3]. Grad-CAM++ is noted for being a fast interpretation method.

LIME
(Local Interpretable Model-Agnostic Explanation) operates by locally approximating a complex "black-box" model with a simpler, interpretable model to explain individual predictions[7, 6]. For image classification, LIME uses superpixels—contiguous patches of similar pixels—as its interpretable representation. The process involves generating perturbed versions of the input image by graying out selected superpixels, obtaining predictions from the original deep learning model for these perturbed images, and then fitting a weighted, interpretable model to explain the original prediction[7, 6]. Despite its utility, LIME faces challenges: its computation can be time-consuming, and its superpixel-based explanations are sensitive to small amounts of noise, leading to instability[2]. Human evaluation studies have also shown that LIME's explanations can be inconsistent and less trusted by experts, often highlighting irrelevant areas outside the main object[7, 6].

SHAP
(Shapley Additive Explanations) is another model-agnostic method rooted in game theory, assigning an importance value to each feature (or region in an image) for a specific prediction[7, 4]. The SHAP gradient explainer extends the integrated gradients method, which attributes feature importance based on Aumann-Shapley values. It calculates feature contributions by integrating gradients along paths between a baseline (e.g., a black image) and the input image, effectively approximating Shapley values [9]. SHAP has demonstrated strong performance, often ranking as the second-best interpretation method in terms of localization effectiveness and radiologists' trust in certain studies[13].

Ensemble XAI: Enhancing Explanation Accuracy and Trust
Ensemble learning, a long-standing technique in machine learning, has traditionally been employed to reduce prediction error by mitigating bias and variance across multiple models[13, 5, 1]. The observation that XAI methods applied to deep learning models trained under slightly different conditions produce explanations that are not highly correlated suggests that combining these explanations could reduce error and improve robustness [5, 11]. This concept has led to the development of **Ensemble XAI**, which leverages the strengths of multiple explanations or models to generate more accurate, robust, and trustworthy interpretations [5, 1].

Several methods can be used to create diverse ensembles for generating explanations:

1. **Different Random Weights:** Training multiple identical base networks with distinct initial random weights on the same data leads to slightly different solutions, whose explanations can then be averaged [5].

2. **Leave Out One Bucket:** Dividing the training data into 'N' buckets and training 'N' identical architectures, each on 'N-1' buckets, creates variations in the learned models[5].

3. **Bootstrap Aggregation (Bagging):** This technique involves creating multiple training sets by sampling with replacement from the original data, resulting in diverse models[1].

Once individual explanations (e.g., heatmaps) are generated from the ensemble members, a common approach for forming an ensemble explanation is to average the relevance scores for each pixel[4].

The literature highlights several key ways in which ensembling has demonstrably improved XAI results:

- **Increased Explanation Accuracy and Robustness:** Averaging explanations from an ensemble of learners significantly improves explanation accuracy across various XAI algorithms. Ensemble methods are shown to be more robust, better aligned with human explanations, and capable of attributing relevance to a broader range of features, thereby increasing completeness[5, 7]. By considering areas of consensus across multiple networks, ensembles reduce irrelevant highlighted areas and enhance the focus on truly relevant regions. Quantitative metrics like **Intersection over Union (IoU)**, correlation, and center of mass distance consistently show improvements when explanations are averaged across an ensemble. For example, studies demonstrated a statistically significant increase in IoU for Grad-CAM and LIME when using ensemble explanations. Furthermore, averaging explanations can increase the "intersection" (finding more relevant areas) and decrease the "union" (finding fewer irrelevant areas) for methods like LIME, Integrated Gradients, and Saliency [12, 13, 5].

- **Human Preference and Trust:** Human experts, such as bird watchers and radiologists, consistently prefer explanations generated by ensembles over those from individual networks[5]. Studies showed that experts rated averaged LIME annotations significantly higher for emphasizing important areas and for recommendation compared to standard LIME [5]. This preference stems from the ensemble's ability to identify more relevant features and avoid emphasizing irrelevant patches. In the medical domain, a specific **Ensemble XAI** approach combining **SHAP and Grad-CAM++** using Kernel Ridge regression was developed [13]. This method achieved the **highest trust score (70.2% mean vote) from a panel of radiologists** when identifying critical areas in medical images. This ensemble also outperformed individual SHAP and Grad-CAM++ by effectively assigning low weight to distracting, non-diagnostic areas outside the regions of interest, such as text or catheters in X-ray images[13].

- **Improved Localization Effectiveness:** The aforementioned SHAP-Grad-CAM++ ensemble XAI method demonstrated superior localization effectiveness in medical image analysis. It achieved the **highest mean set F1 score (0.50), mean set accordance recall (0.57), mean set accordance precision (0.52), and mean set IoU (0.36)**, indicating that it accurately identified a larger fraction of annotated areas and maintained a higher consistency with expert annotations [13]. This performance confirms that the constituent methods, SHAP and Grad-CAM++, complement each other effectively within the ensemble.

- **Enhanced Classification Accuracy:** Beyond improving explanations, ensembling models can also lead to an increase in the classification accuracy of the deep learning system itself. For instance, on the ISIC-2018 melanoma detection dataset, ensembling models improved the AUC score from 0.82 to 0.86, a statistically significant increase [5].

- **Computational Efficiency:** While requiring more computation than a single network, the linear increase in computation for ensembles can be mitigated by parallel training. Critically, the ensemble XAI method (SHAP-Grad-CAM++) has been shown to require significantly less computation time compared to LIME.

In summary, ensemble XAI addresses fundamental limitations of individual XAI methods by mitigating variance, reducing bias, and producing more complete and robust explanations. The combination of diverse explanation methods or models, especially when averaged, results in explanations that are not only quantitatively more accurate (as measured by IoU, correlation, and center of mass distance) but also qualitatively preferred and more trusted by human experts, particularly in complex domains like medical imaging. This approach signifies a vital step towards creating more trustworthy and clinically useful AI systems.

3 Methodology and Implementation

3.1 Dataset

For the data set, the pancreatic tumor detection data set has been used. This dataset contains 281 positive class images and 81 negative class images. Each image is 336 x 338 pixels. The dataset also contains annotation of positive class images in the form of .csv with image index and matrix size of (224,224). we have divided both of these positive and negative images into three subsets: training with 259 images, validation set with 56 images and test set with 56 images with batch size of 16. Training and validation set has been used for fine tuning the model. While test dataset has been used to for final evaluation metrics and XAI techniques.

3.2 Pretrained Model Selection

Different pretrained deep learning models are available on Keras. This include VGGNet, ResNet, InceptionV3 and Xception. InceptionV3 was initially selected for this experiment. The model introduces a network within a network to improve accuracy and efficiency. This model adopts multi-level feature extraction adapting well to different input sizes and complexities. There is an Xception model further increasing the capabilities of InceptionV3. Finally, Xception has been selected, however, its performance is better on larger datasets; 350 million images. This model required the images to be 299 by 299 hence the initial images have been transformed to this size.

3.2.1 Finetuning the Pretrained Model

The model has been finetuned using the training dataset. For this , the model is loaded in the inference mode without the top classification layer. An extra dense layer is added and fitted to the input data and validated in the validation test. Finally, the model is evaluated on the test dataset using binary accuracy. The resulting model achieved an accuracy of 91.47% on training data, 92.12% on the validation data, and 85.365% on the test data.

Dataset	**Xception Accuracy**
Training	0.9147
Validation	0.9219
Test	0.8536

Table 3.1: Fine-Tuned Model Accuracy

3.3 Evaluation Criteria

The evaluation criteria for the performance of the interpretation methods are as follows:

Decision Impact Ratio:

The Decision Impact Ratio quantifies the effect of omitting critical regions identified by an interpretable model method on the decision-making process. Specifically, it measures the percentage change in decision accuracy when important regions (as determined by the interpretation method) are excluded from the input data. The identification of critical area is carried out using a threshold where values greater than threshold are identified as true and thus critical.

$$\text{Decision Impact Ratio} = \frac{D_{\text{without critical regions}} - D_{\text{with critical regions}}}{D_{\text{with critical regions}}} \times 100$$

Where:

- $D_{\text{without critical regions}}$ is the decision or output after excluding the critical regions.
- $D_{\text{with critical regions}}$ is the decision or output with the critical regions included.

3.3.1 Localization Effectiveness with Annotations

The following evaluation metrics are used for identifying the localization effectiveness of the XAI methods when compared to the normalized heatmaps of annotated ground truth.

Cosine Similarity:

Cosine Similarity measures the cosine of the angle between two vectors. It is used to determine how similar two vectors are in an n-dimensional space, regardless of their magnitude.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Where:

- $\mathbf{A}$ and $\mathbf{B}$ are the vectors being compared.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (norms) of the vectors.

Accordance Recall:

Accordance Recall is a metric used to evaluate how well a model's predictions align with the ground truth, specifically focusing on the recall of accurately identified positive samples. This is given by the fraction of annotated area correctly identified by the XAI method.

$$\text{Accordance Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Adaptive Accordance Recall:

Adaptive Accordance Recall is a variation of Accordance Recall, where the recall metric is adjusted dynamically based on changing conditions, such as the difficulty of detecting certain regions or classes over time.

$$\text{Adaptive Accordance Recall} = \frac{\sum_i \text{True Positives}_i}{\sum_i (\text{True Positives}_i + \text{False Negatives}_i)}$$

Mean Recall:

Mean Recall is the average recall across all classes or categories within a dataset. It reflects the model's ability to correctly identify positive samples.

$$\text{Mean Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i}$$

Where N is the total number of classes.

Mean Intersection over Union (IoU):

Mean IoU is the average of the Intersection over Union (IoU) metric across all classes. It measures the overlap between predicted and true regions, with higher values indicating better model performance.

$$\text{Mean IoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{Intersection}_i}{\text{Union}_i}$$

Where:

- Intersection_i is the area of overlap between the predicted and ground truth for class i.
- Union_i is the total area covered by both the predicted and ground truth for class i.

Mean F1 Score:
The Mean F1 Score is the average of the F1 Scores for each class. The F1 Score is the harmonic mean of Precision and Recall, balancing the trade-off between these two metrics.

$$\text{Mean F1 Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Mean Precision:
Mean Precision is the average precision across all classes, reflecting the model's ability to avoid false positives. Precision is the ratio of true positives to the total predicted positives.

$$\text{Mean Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i}$$

3.4 General Workflow

The general workflow has been applied to all the XAI intrepretation. Until this point, the model finetuning has already been carried out. It can be summarized in four phases:

1. **Dataset and Annotation:** The test dataset containing 56 images are loaded, with 44 positive images. The negative images are ignored as they are not relevant for the analysis below. These images are transformed to (299, 299) as required by the Xception model in the next step. As for the annotation data, the attention matrices .csv file is loaded and the column entries are converted to (224,224) matrix. Since it contains annotation for the whole positive images, it is filtered using the filenames from test dataset. In the end, we have 44 images and list with 44 corresponding attention matrices (ground truth).

2. **Classification Model and XAI:** For each XAI technique, a heatmap is generated using Xception model and 44 transformed images. This acts as the prediction from the XAI methods.

3. **Comparing the Techniques:** The heatmaps generated from the previous step are transformed to (224,224), similar to ground truth. These heatmaps are normalized between [0,1]. Similarly, attention matrices are also normalized. At the end these heatmaps are evaluated for the evaluation metrics mentioned above. For the absent impact, only heatmaps of XAI techniques are used. The thresholds of 0.17 and 0.75 have been used to check for the changes in the signifigance of selected area. For localization and comparison using logistic model, both heatmaps and ground truth matrices are used. For the threshold in accordance recall, 0.2 is used. Adaptive recall used 80th percentile to select the threshold automatically. For the logistic model, we have used a stride of 1 and threshold of 0.1737.

4. **Updating the Technique:** The results are noted and further changes in the ensembling techniques are performed. This includes changing weights, devising methods and so on.

3.5 Workflow for Implementation of GradCAM

After loading the images and transforming them into (299,299) size, convert them into arrays.

- Create a keras sub-model that maps the input image to the output predictions and the last convolutional layer of Xception model.
- The last convolutional layer identified as 'block14_sepconv2_act'. This layer is used to compute the gradient of top predicted class of the input image with respect to the layer.
- The gradient is filtered for values over zero and pooled together using mean to gain weighted importance. A heatmap is generated similar in size to the last layer, thus (10,10).
- This heatmap is then upscale to 224 by 224 for further analysis and comparison.

3.6 Workflow for Implementation of Lime

Initially, the dataset is loaded and preprocessed using custom classes to ensure compatibility with the deep learning model, specifically the Xception architecture.

- Once the model is loaded with pretrained weights, the LIME (Local Interpretable Model-agnostic Explanations) methodology is employed to interpret the model's predictions on the test dataset.
- The custom 'LimeExplainer' class, as implemented in the 'custom_lime' module, wraps the standard LIME image explainer and adapts it for Keras models and TensorFlow datasets. It extracts individual images from the test set, preprocesses them, and then generates explanations by perturbing the input image and observing the model's output changes.
- The explainer highlights the most influential superpixels that drive the model's decision, producing both visual overlays and heatmaps that localize important regions.
- These explanations are collected for all test images and stored for further analysis and benchmarking. The approach ensures that the interpretability process is tightly coupled with the model's data pipeline and leverages LIME's strengths in providing local, human-interpretable insights into complex neural network predictions.

3.7 Workflow for Implementation of SHAP

In the context of deep learning and image classification, SHAP provides pixel-wise attributions, highlighting regions of an image that most influence the model's decision. The workflow for applying SHAP to image models typically involves several steps.

- First, the trained model is loaded, and a masker is defined to handle the inpainting or perturbation of image regions.
- The SHAP explainer is then initialized with the model and masker. For each image, the explainer computes SHAP values, which represent the impact of each pixel (or superpixel) on the model's prediction for a specific class. These values are aggregated and normalized to produce a heatmap, visually indicating the most influential areas of the image.
- The resulting heatmaps can be saved and overlaid on the original images to facilitate interpretation.

3.8 Workflow for Implementation of Ensemble XAI

For the various derived ensembling methods, the individual heatmaps and ground truth have been imported, transformed to grayscale, (224, 224) and combined in the various ways 3.1. Same goes for the ground truth.

3.8.1 Averaging the XAI

For this the heatmaps have been simply averaged with equal weights. The resulting heatmap is then normalized and compared using evaluation metrics.

3.8.2 Averaging the XAI using Weights

For this the heatmaps have been averaged using weights. The determination of weights is determined based on the evaluation metrics scores of individual XAI techniques. The resulting heatmap is then normalized and compared using evaluation metrics.

3.8.3 Ensembling the XAI using Maximum Value

The max-fusion strategy adopts a deterministic rule in which, at each spatial location, the attribution score is assigned as the maximum of the candidate methods:

$$H_{\max}(i,j) = \max\{L(i,j),\, G(i,j)\},$$

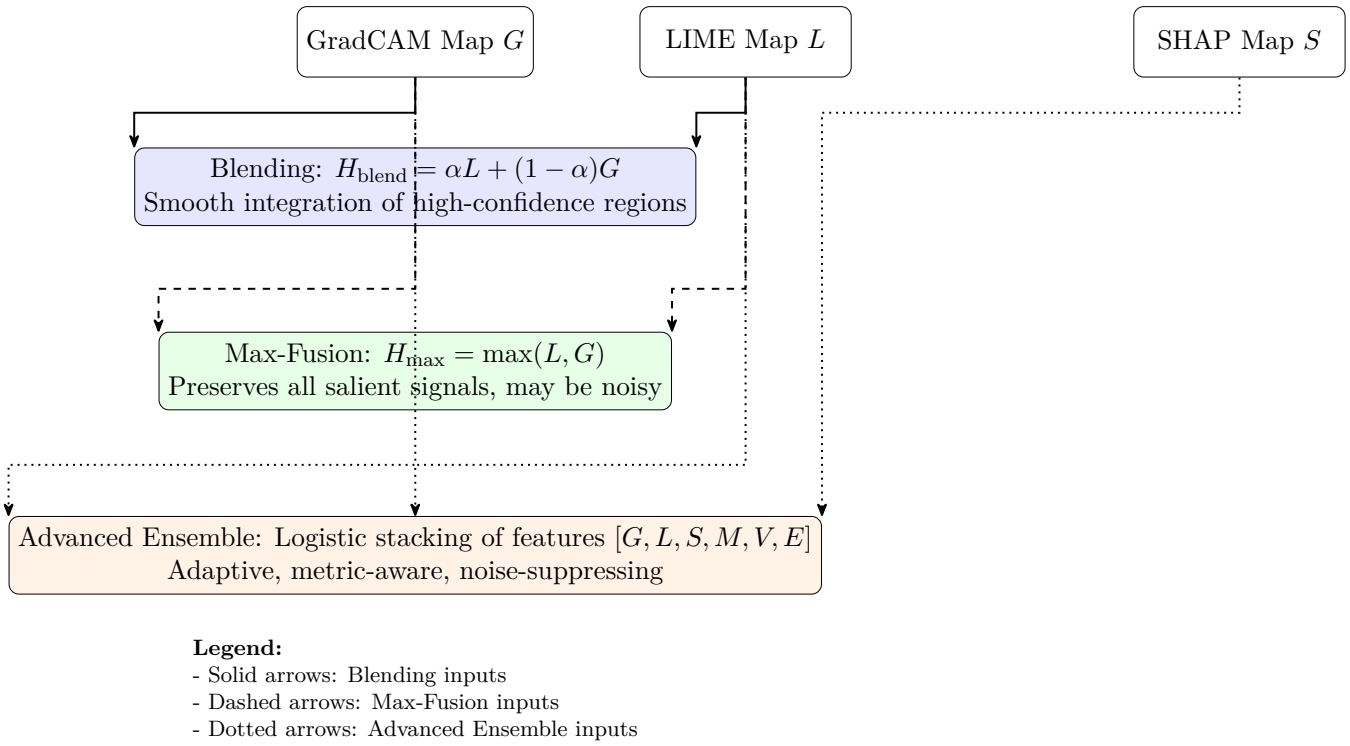

Figure 3.1: Comparison of ensemble strategies for combining interpretability maps

where L and G denote the normalized saliency maps from LIME and GradCAM, respectively. This ensures that any salient signal identified by either method is preserved. However, max-fusion tends to over-emphasize isolated high responses, even if they result from noise or spurious perturbations, leading to attribution maps that may lack spatial smoothness and interpretive stability. In practice, this can inflate recall at the expense of precision, generating explanations that highlight larger but less discriminative regions.

The The resulting heatmap is then normalized and compared using evaluation metrics.

3.8.4 Ensembling the XAI using Blended Value

The blending strategy integrates complementary properties of two attribution methods-LIME and GradCAM-by performing a weighted linear combination of their saliency values within regions of interest. Specifically, GradCAM highlights coarse, spatially consistent regions corresponding to class-discriminative structures, whereas LIME provides fine-grained, instance-specific attributions derived from perturbation analysis. Formally, in high-activation regions identified by GradCAM (i.e., regions where GradCAM scores exceed a predefined threshold), the blended attribution map is computed as:

$$H_{\text{blend}} = \alpha\, L + (1-\alpha)\, G,$$

where L and G represent normalized saliency maps from LIME and GradCAM respectively, and $\alpha \in [0, 1]$ is the blending coefficient controlling the relative contribution of each method. This mechanism achieves two goals:

- Reinforcement of salient regions: When both LIME and GradCAM agree on the importance of a region, blending amplifies its attribution while smoothing out noise from either method individually.
- Complementary correction: In cases where GradCAM highlights a broader region but lacks boundary precision, LIME's localized perturbation-based scores refine the attribution. Conversely, when LIME introduces spurious noise in irrelevant regions, GradCAM's spatial consistency suppresses such errors.

By combining local sensitivity (LIME) with global discriminative localization (GradCAM), blending produces attribution maps that are both spatially coherent and class-specific, thus offering a more stable and interpretable explanation in medical imaging tasks such as cancer detection.

3.8.5 Advanced Ensemble

The **Advanced Ensemble** is a feature-driven logistic stacking method designed to integrate multiple interpretability maps (GradCAM G, LIME L, and SHAP S) into a single high-fidelity saliency map for cancer image interpretation. This approach goes beyond simple averaging, weighted averaging, or max-fusion by constructing per-pixel features and learning an adaptive classifier to optimally combine the signals.

Feature Construction

For each pixel (i, j), a six-dimensional feature vector is constructed:

$$X(i,j) = \big[G(i,j),\, L(i,j),\, S(i,j),\, M(i,j),\, V(i,j),\, E(i,j)\big],$$

where:

- G, L, S are normalized GradCAM, LIME, and SHAP saliency maps.
- $M(i,j) = \frac{1}{3}(G + L + S)$ is the mean attribution across methods.
- $V(i,j) = \frac{1}{3}\sum_k (A^k - M)^2$ is the variance, capturing disagreement.
- $E(i,j) = \frac{M}{\sqrt{V+\lambda}}$ is a consensus score emphasizing regions with consistent high attribution, with λ a small regularization constant.

The features are reshaped into a two-dimensional array of shape $[N_{\text{pixels}}, 6]$ for model input.

Logistic Stacking

A logistic regression classifier is trained on these features to predict the probability that a pixel belongs to a positive region (e.g., cancerous tissue):

$$\hat{P}_{i,j} = \sigma\big(w^T X(i,j) + b\big),$$

where σ is the logistic sigmoid, and w and b are learned via cross-validated logistic regression. Key aspects include:

- **Class balancing:** Positive pixels are sampled with corresponding negatives up to a predefined ratio to mitigate class imbalance.
- **Cross-validation:** GroupKFold ensures training and validation splits respect image boundaries, preventing leakage.
- **Hyperparameter search:** Multiple values of regularization strength C and l_1 ratio are evaluated to maximize the target evaluation metric (e.g., F-beta or IoU).
- **Threshold optimization:** A per-metric threshold τ is determined to convert probabilities into binary predictions:

$$\hat{M}_{i,j} = \begin{cases} 1, & \text{if } \hat{P}_{i,j} \geq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Upsampling and Map Construction

Predictions are made at a downsampled stride (e.g., stride = 4) to reduce computational cost. Bilinear upsampling is then applied to reconstruct the full-resolution probability map:

$$\hat{P}^{\text{full}} = \text{BilinearUpsample}(\hat{P}, H, W),$$

where H and W are the original image dimensions.

Advantages

The Advanced Ensemble offers several key benefits:

1. **Adaptive fusion:** The logistic classifier learns the optimal combination of LIME, GradCAM, and SHAP per pixel.
2. **Noise suppression:** Pixels with low agreement across methods are automatically down-weighted via the variance and consensus features.
3. **Metric-aware optimization:** Thresholding and model selection explicitly optimize for evaluation metrics such as F1-score or IoU.
4. **Balanced performance:** Empirically, this method achieves high F1, IoU, and cosine similarity while maintaining reasonable Decision Impact Ratios, outperforming max-fusion and simple blending approaches.

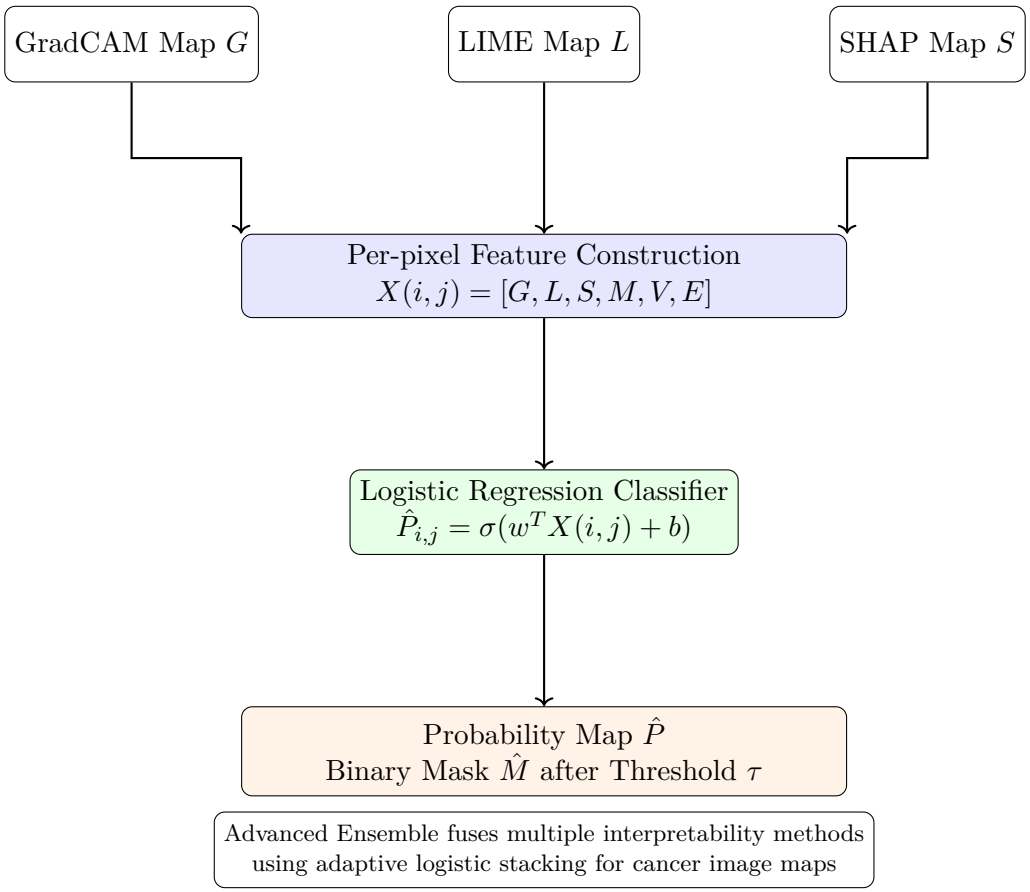

Figure 3.2: Schematic of the Advanced Ensemble pipeline. GradCAM, LIME, and SHAP maps are combined via per-pixel feature construction, followed by logistic stacking to generate probability maps and binary masks.

4 Results

In this chapter, we compare the interpretability performance of LIME, GradCAM, SHAP, and multiple ensemble strategies on cancer image classification tasks. The ensembles include simple averaging, weighted averaging, max-fusion, blending, and an advanced hybrid strategy. Performance is assessed across three categories: **Absent Impact**, **Localization**, and **Logistic Model behavior**, with specific evaluation metrics reported in Table 4.1.

4.1 Results for Individual Techniques

4.1.1 Impact of Absence of Regions

In this section, the significant regions identified by different XAI interpretations have been considered for importance. The heatmaps are normalized and decision impact ratios are calculated. The importance of significant areas is controlled using a threshold. For a higher threshold of 0.75, the mean decision impact ratio ranges from 0.011 for SHAP to 0.0415 for GradCAM. This means only selecting regions with a value of 0.75 or above as significant regions. Reducing the threshold leads to a generous selection mechanism. When the threshold is reduced to 0.17, the decision impact ratio for GradCAM rises to 0.52, 0.96 for lime, and 0.49 for SHAP. This threshold has been used in analysis next as well.

4.1.2 Localization Metrics

The cosine similarity defines the correlation among the features and the ground truth using the dot product. Higher values indicate that two features are similar, while lower values suggest that these two are independent. Average cosine similarity for the three methods ranges from 0.088 for GradCAM, 0.110 for Shap, and 0.139 for Lime. The same trend has been further validated by accordance recall, which is a localization effectiveness benchmark. For the 80^{th} percentile, high scores have been observed for lime with 0.9624, followed by Shap with 0.458, and followed by GradCAM with 0.3851. The mean precision is low at 0.024 for SHAP and LIME while 0.03129 for GradCAM. These results indicate that Lime performs better than other methods across all benchmark scores, giving it a higher weight than the other methods.

4.2 Results for Averages

In the first attempt to ensemble the feature maps, a simple unweighted average has been taken. The results indicate a worst performance compared to Lime when used individually. These benchmarks fall in the middle, equivalent to Grad-CAM's performance metrics. Moreover, there is no advantage in relation to ensemble calculation time as all three XAI methods have to be used in conjunctions in this method.

Category	Metric	LIME	GradCAM	SHAP	Average	Weighted Avg.	Fused_max	Fused_blend	Advanced
Absent Impact	Decision Impact Ratio (0.75)	0.0887	0.0415	0.0111	0.0422	0.0457	0.0118	0.0000	0.5400
	Decision Impact Ratio (0.17)	0.9676	0.5256	0.4963	0.5584	0.8023	0.1996	0.8250	0.5430
Localization	Cosine Similarity	0.1390	0.0880	0.1100	0.0900	0.1150	0.0550	0.1210	0.1270
	Accordance Recall (0.2)	0.9625	0.3851	0.4582	0.3940	0.7144	0.1445	0.7923	0.7900
	Adaptive Accordance Recall	0.2000	0.2000	1.0000	0.2003	0.2002	0.2046	0.2003	0.2000
Logistic Model	Mean Recall (stride=1, at 0.1737)	0.9990	0.5876	0.9900	0.4587	0.9900	0.8608	0.8947	0.7967
	Mean IoU	0.0244	0.0309	0.0244	0.0285	0.0244	0.0270	0.0257	0.0302
	Mean F1	0.0465	0.0555	0.0465	0.0530	0.0465	0.0513	0.0489	0.0564
	Mean Precision	0.0240	0.0313	0.0240	0.0305	0.0244	0.0293	0.0258	0.0303

Table 4.1: Evaluation metrics grouped by impact category: absent impact, localization, and logistic model performance.

4.3 Results for Weighted Average XAI

From the results of individual methods and averages, we have concluded that LIME consistently performs better; however, the method is computationally expensive. Although, GradCAM does not perform as well as Lime, but its computational power is lesser as it is calculated by the last convolutional layer of the model, a 10 by 10 in size. Hence, it makes sense to devise a method reaching metrics as good as individually. Given the performance of Lime, it has been assigned a weight of 0.802, followed by Shap with 0.149 and lastly, GradCAM with 0.05. The **Weighted Average Ensemble** incorporates predefined weights ($w = [0.0491, 0.8018, 0.1491]$) favoring LIME.

Absent Impact Decision Impact Ratio (0.75) improves to 0.0457, slightly above simple averaging. Importantly, at a stricter threshold (0.17), performance increases to 0.8023, showing this weighting effectively leverages LIME's strengths.

Localization Cosine similarity (0.1150) improves over simple averaging and all individual methods, demonstrating better spatial alignment. Accordance Recall (0.7144) provides a middle ground between LIME's high recall (0.9625) and GradCAM/SHAP's weaker performance, indicating stability in region detection.

Logistic Model Mean Recall (0.9900) remains competitive with LIME and SHAP. IoU (0.0244) mirrors individual weaknesses, while F1 (0.0465) stagnates. Mean Precision (0.0244) shows no gain, indicating weighting alone does not resolve precision deficiencies.

4.3.1 Fused_max

The **Fused_max strategy** selects the maximum importance between LIME and GradCAM per region.

Absent Impact This approach sharply reduces the Decision Impact Ratio (0.75) to 0.0118, highlighting a conservative behavior that may reduce false positives. Conversely, performance at the 0.17 threshold (0.1996) is poor compared to weighted averaging, suggesting instability under strict decision conditions.

Localization Cosine similarity drops to 0.0550, underperforming compared to all baselines. Accordance Recall (0.1445) is also the lowest among all tested methods, suggesting max fusion discards important contextual signals in cancer localization.

Logistic Model Logistic metrics show modest gains: Mean Recall (0.8608) is better than GradCAM (0.5876) but far below LIME. Mean IoU (0.0270) and F1 (0.0513) outperform LIME and SHAP individually, showing selective fusion can balance sensitivity and overlap, though with reduced precision.

4.3.2 Fused_blend

The **Fused_blend strategy** linearly combines GradCAM and LIME in high-activation regions.

Absent Impact Decision Impact Ratio (0.75) collapses to 0.0000, eliminating overconfidence at high thresholds but potentially at the cost of interpretability. At 0.17, the score rises sharply to 0.8250, rivaling the weighted ensemble and demonstrating adaptive strength at moderate thresholds.

Localization Cosine similarity improves to 0.1210, second only to the advanced strategy. Accordance Recall (0.7923) nearly matches weighted averaging, suggesting blended fusion maintains both GradCAM's localization strengths and LIME's regional recall.

Logistic Model Performance is consistent: Mean Recall (0.8947) is higher than GradCAM and closer to ensemble averages. IoU (0.0257) is stable, while Mean F1 (0.0489) remains modest. Importantly, precision (0.0258) is slightly better than Fused_max, indicating blending improves reliability.

4.3.3 Advanced Ensemble

The **Advanced Ensemble** integrates multiple strategies adaptively.

Absent Impact Decision Impact Ratio (0.75) reaches 0.5400, significantly outperforming all other methods. At 0.17, performance (0.5430) is moderate, balancing conservatism and flexibility.

Localization Cosine similarity peaks at 0.1270, surpassing all baselines. Accordance Recall (0.7900) remains competitive with weighted and blended ensembles, confirming stable localization across cancerous regions. Adaptive Accordance Recall remains near constant (≈ 0.200), suggesting robustness.

Logistic Model Mean Recall (0.7967) is lower than blended strategies but still higher than GradCAM. IoU (0.0302) and F1 (0.0564) are the highest across all strategies, marking the advanced ensemble as the most balanced in overlap and harmonic precision-recall trade-offs.

4.4 Key Takeaway

The evaluation summarized in Table 4.1 demonstrates that individual interpretability methods exhibit complementary strengths: LIME achieves high absent impact sensitivity (Decision Impact Ratio 0.0887 at 0.75), SHAP attains perfect adaptive accordance recall (1.0), and GradCAM provides computational efficiency albeit moderate localization performance. Simple ensemble strategies, including averaging and weighted averaging, improve accordance recall and cosine similarity modestly, with weighted averaging leveraging method-specific reliability to achieve an accordance recall of 0.7144. Max- and blend-fusion approaches further explore nonlinear integration, with blend-fusion notably improving localization (accordance recall 0.7923) by combining high- and low-confidence regions. The advanced logistic stacking ensemble surpasses these strategies, achieving balanced performance across absent impact (Decision Impact Ratio 0.5400), localization (cosine similarity 0.1270, accordance recall 0.7900), and logistic model metrics (mean IoU 0.0302, mean F1 0.0564). By constructing per-pixel features from normalized saliency maps, mean, variance, and consensus, and adaptively learning optimal combination weights with threshold optimization, this approach effectively suppresses noise, emphasizes consistent attributions, and provides clinically relevant saliency maps. These findings highlight the efficacy of adaptive ensemble strategies in producing robust, interpretable, and actionable explanations for cancer image analysis.

Metric	Best Method	Value	Notes
Decision Impact Ratio (0.75)	Advanced Ensemble	0.5400	Highest absent impact sensitivity
Decision Impact Ratio (0.17)	Fused_blend	0.8250	Best low-threshold decision coverage
Cosine Similarity	Advanced Ensemble	0.1270	Strongest agreement with reference maps
Accordance Recall (0.2)	Fused_blend	0.7923	Best localization recall
Adaptive Accordance Recall	SHAP	1.0000	Perfect consistent region highlighting
Mean Recall (stride=1)	LIME / Weighted Avg.	0.9990 / 0.9900	Highest true positive detection
Mean IoU	Advanced Ensemble	0.0302	Best overlap between prediction and ground truth
Mean F1	Advanced Ensemble	0.0564	Optimal balance between precision and recall
Mean Precision	GradCAM	0.0313	Highest precision among methods

Table 4.2: Summary of top-performing methods per evaluation metric across absent impact, localization, and logistic model categories.

5 Conclusion

In this study, we systematically investigated the integration of multiple post-hoc interpretability methods like LIME, GradCAM, and SHAP, through a variety of ensemble strategies, culminating in an advanced logistic stacking framework. The evaluation, summarized in Table 4.1, highlights several important insights regarding both individual methods and ensemble approaches in the context of cancer image interpretation.

Individual Methods

Among the singular interpretability techniques, LIME exhibited strong performance in capturing absent impact regions, with a Decision Impact Ratio of 0.0887 at the 0.75 threshold, indicating sensitivity to non-informative regions. SHAP achieved perfect adaptive accordance recall (1.0), reflecting its capacity to highlight highly consistent regions across images. GradCAM, while computationally efficient due to its utilization of the final convolutional layers, demonstrated moderate performance across metrics, particularly in accordance recall (0.3851) and cosine similarity (0.0880), highlighting limitations in fine-grained localization without complementary methods.

Simple Ensemble Approaches

Average and weighted average fusion methods provided modest improvements over single-method baselines. Weighted averaging, which assigns empirically derived weights favoring LIME, achieved higher accordance recall (0.7144) compared to simple averaging (0.3940), demonstrating the utility of incorporating method-specific reliability into ensemble weighting. Max-fusion and blend-fusion further explored nonlinear integration strategies. Fused max preserved salient regions effectively but underperformed in absent impact and logistic metrics, while fused blend improved localization metrics (accordance recall 0.7923) by combining information from complementary high- and low-confidence regions.

Advanced Ensemble

The advanced logistic stacking approach integrates normalized maps from LIME, GradCAM, and SHAP into a six-dimensional per-pixel feature space (G, L, S, mean, variance, and consensus). Logistic regression adaptively learns the combination of features and optimizes threshold selection with respect to evaluation metrics. This method demonstrated balanced performance across categories: high absent impact sensitivity (Decision Impact Ratio 0.5400 at 0.75), strong localization metrics (cosine similarity 0.1270, accordance recall 0.7900), and competitive logistic model outputs (mean IoU 0.0302, mean F1 0.0564). Importantly, it achieved these results while mitigating noise and enhancing agreement across methods, outperforming simpler averaging or max-based fusion strategies in scenarios requiring adaptive weighting of heterogeneous attribution maps.

Overall Implications

The results indicate that while individual methods provide valuable insights, ensemble strategies, particularly the advanced logistic stacking, offer superior robustness and interpretability for cancer image explanation tasks. By learning context-sensitive combinations of interpretability signals, the advanced ensemble effectively balances sensitivity to relevant features with suppression of spurious activations, thereby enhancing the reliability and clinical relevance of AI-driven saliency maps.

List of Figures

List of Tables

Bibliography

[1] K. M. Ali and M. J. Pazzani. Error reduction through learning multiple descriptions. *Machine learning*, 24(3):173–202, 1996.

[2] D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods, 2018.

[3] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*, 2018.

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Michael Pazzani, Severine Soltani, Sateesh Kumar, Kamran Alipour, and Aadil Ahamed. Improving explanations of image classifiers: Ensembles and multitask learning. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 13(6):51–72, November 2022.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[7] Laura Rieger and Lars Kai Hansen. Aggregating explainability methods for neural networks stabilizes explanations. *arXiv preprint arXiv:1903.00519*, 2019.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017.

[9] A. Shrikumar et al. Learning important features through propagating activation differences, 2017.

[10] M. Sundararajan et al. Axiomatic attribution for deep networks. In *Proc. 34th Int. Conf. Mach. Learn.*, page 3319–3328, 2017.

[11] M. Watson, B. A. S. Hasan, and N. Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 875–884, 2022.

[12] J. Zhang et al. Dense gan and multi-layer attention-based lesion segmentation method for covid-19 ct images. *Biomed. Signal Process. Control*, 69, 2021.

[13] Lin Zou, Han Leong Goh, Charlene Jin Yee Liew, Jessica Lishan Quah, Gary Tianyu Gu, Jun Jie Chew, Mukundaram Prem Kumar, Christine Gia Lee Ang, and Andy Wee An Ta. Ensemble image explainable ai (xai) algorithm for severe community-acquired pneumonia and covid-19 respiratory infections. *IEEE Transactions on Artificial Intelligence*, February 2022.