Diagnosis Beyond the Black Box: Gradient-Based Attribution for Explainable Medical Image Classification

Dr. Edwin Gerardo Acuña Acuña[1] Orcid: 0000-0001-7897-4137 Universidad Hispanoamericana. edwacuac@gmail.com, Sacramento Cruz-Doriano[2*[0000-0002-8837-7114]], Maria Teresita de Jesus Chi-Chan[2[0000-0002-2642-9249]], Felipe Ángel Álvarez-Salgado [2[0000-0002-2191-2856]] / [2] Instituto Tecnológico Superior de Calkiní / SECIHTI, Mexico

Abstract– *Deep convolutional networks have achieved radiologist-level performance in medical image classification, yet their limited interpretability remains a major barrier to clinical adoption. This paper introduces GW-XAI, a gradient-weighted attribution framework that combines layer-specific Grad-CAM activations with perturbation-based faithfulness assessment to produce clinically interpretable saliency maps. Evaluated on chest X-ray classification using the CheXpert benchmark, GW-XAI achieves an area under the precision-recall curve of 0.887 for pleural effusion detection. It also improves the Attribution Faithfulness Score by 14.3% relative to baseline Grad-CAM and by 9.1% relative to Score-CAM. Alignment with expert radiologist annotations reaches an Intersection over Union of 0.74, substantially exceeding the comparative methods evaluated through structured annotation overlap. These findings suggest that GW-XAI can reduce the interpretability gap in AI-assisted diagnosis while preserving diagnostic performance, thereby supporting more transparent, clinically reliable, and regulation-oriented deployment.*

Keywords: ***explainable artificial intelligence, medical image classification, gradient attribution, saliency maps, clinical decision support.***

I. INTRODUCTION

Convolutional neural networks (CNNs) currently equal or surpass attending physician performance on tasks ranging from diabetic retinopathy grading to pulmonary nodule identification and chest X-ray triage, demonstrating the therapeutic potential of deep learning in radiography [1], [2]. Despite this, there is still a significant and ongoing disconnect between real clinical deployment and technical benchmark performance. The main reason is interpretability: when a model gives a radiograph a pathology probability, it does not explain which areas or characteristics of the picture influenced that choice. A diagnostic tool with completely opaque rationale is impossible for clinicians to accept, trust, or audit [3].

Regulatory agencies have started to formalize this issue. Adopted in 2024, the European Union's Artificial Intelligence Act designates medical AI systems as high-risk and mandates that they be visible, explicable, and subject to human supervision. In a similar vein, the US Food and Drug Administration's guidelines on Software as a Medical Device specifically link interpretability standards to pre-market clearance for AI-based diagnostic tools. Even very accurate models cannot go beyond research settings without validated explainability.

In order to meet this demand, Explainable Artificial Intelligence (XAI) generates explanations of model predictions that are understandable by humans. Saliency map production, which highlights the areas of an input picture that most affected a certain output score, is the predominant XAI paradigm in the medical imaging setting. However, existing saliency approaches are limited by two basic flaws. First, faithfulness: although many attribution maps are visually convincing, they do not precisely represent the computational route of the model; in other words, they fulfill human expectations rather than providing an explanation for the behavior of the model. Second, clinical alignment: even accurate attributions may draw attention to areas in a dataset that have a statistical correlation with disease but do not match the physically significant characteristics that medical professionals use to make diagnoses [4].

To address these limitations, this study presents GW-XAI (Gradient-Weighted Explainable AI), a post-hoc attribution framework that integrates three complementary components. First, it applies attention-weighted fusion across multiple convolutional layers to combine fine-grained spatial information with deeper semantic representations. Multilayer activation mapping is not introduced here as an entirely new principle because previous methods, including LayerCAM and Zoom-CAM, have demonstrated the value of aggregating information across network depth [18], [19]. The contribution of GW-XAI lies instead in combining adaptive multilayer fusion with perturbation-based faithfulness filtering and clinically grounded evaluation within a unified framework. Second, the filtering stage suppresses regions whose attributed importance is not supported by a measurable change in model confidence. Third, the Attribution Faithfulness Score (AFS) uses structured perturbations to quantify the causal correspondence between highlighted regions and model predictions. Experimental validation on the CheXpert benchmark shows steady gains in expert radiologist annotation overlap and attribution quality indicators without any decline in classification accuracy.

Unlike conventional single-layer attribution methods, GW-XAI integrates information from multiple convolutional stages to preserve both high-level pathological meaning and fine-grained spatial detail. This design is particularly relevant in chest radiography, where abnormalities may vary considerably in size, morphology, contrast, and anatomical distribution. The attention mechanism adaptively weights the contribution of each layer, allowing the final saliency map to emphasize the representations that are most informative for a specific prediction rather than assigning equal importance to all feature levels.

The perturbation-based filtering process further strengthens the reliability of the generated explanations by retaining only

those regions whose modification produces a measurable change in model confidence. This causal validation reduces the influence of visually prominent but computationally irrelevant areas and provides a more rigorous basis for assessing attribution quality. Together with the AFS metric, the proposed framework supports a more transparent evaluation of model behavior and offers clinicians a clearer basis for determining whether a prediction is grounded in diagnostically meaningful image evidence.

II. RELATED WORK

A. Gradient-Based Saliency Methods

The gradient of the output with respect to the input picture was the direct source of pixel significance for the first saliency techniques [5]. This was improved by Guided Backpropagation [6], which suppresses negative gradient contributions during backward passes. By adding gradients along a route from a baseline to the input, Integrated Gradients [7] expanded on the idea and offered axiomatic attribution guarantees. Grad-CAM [8] produced coarse yet class-discriminative activation maps that maintained spatial connection with the input by rerouting gradient calculation to the last convolutional layer. Grad-CAM++ [9] then used weighting positive partial derivatives to handle multi-object localization. All gradient-based techniques, despite their effectiveness, have low sensitivity to fine-grained texture characteristics that are diagnostically significant in medical imaging and gradient saturation in areas of high activation. Another limitation is that most gradient-based methods derive explanations from a single network layer or from local derivative information at one point in the input space. This approach may overlook the hierarchical structure of convolutional representations, in which early layers capture edges and textures, intermediate layers encode anatomical patterns, and deeper layers represent class-specific semantic information. In medical images, diagnostically relevant evidence may be distributed across these levels. Consequently, attribution maps generated from only the final convolutional layer may identify the general location of an abnormality while failing to preserve the detailed morphological features required for clinical interpretation.

Gradient-based explanations may also be unstable under minor input variations or model parameter changes. Small perturbations that do not alter the predicted class can produce substantially different saliency maps, reducing their reliability for clinical auditing and comparison with expert annotations. Furthermore, visually plausible heatmaps do not necessarily indicate that the highlighted regions have a causal influence on the prediction. These limitations have motivated the development of attribution methods that combine information from multiple layers and validate regional importance through controlled perturbation of the input.

B. Perturbation-Based Methods

By sampling perturbed inputs and monitoring the consequent output variation, LIME [10] builds a locally faithful linear surrogate model around any goal prediction. Although LIME superpixel decompositions show instability under multiple sampling runs on the same picture, they have been applied to medical imaging for skin lesion and histopathology diagnosis. At the expense of thousands of forward evaluations per picture, RISE [11] provides greater consistency guarantees by stochastically sampling binary masks and expressing attribution as a weighted expectation over masked predictions. By moving a fixed-size occluding patch over the picture and recording confidence declines, occlusion-based methods [12] further systematize perturbation. These methods are faithful and interpretable, but they are computationally expensive at clinically relevant resolutions.

C. Clinical Alignment of XAI

In particular, a number of studies have evaluated the discrepancy between radiologist perception and automated attribution. Sayres et al. [13] showed that, especially in unclear situations, integrated gradients outputs for fundus imaging had a limited correlation with ophthalmologist gaze patterns. In a rigorous examination of Grad-CAM on CheXpert, Arun et al. [14] discovered that in less than 60% of instances for cardiomegaly and consolidation, pathology-specific attribution zones coincided with recognized anatomical landmarks. These results encourage attribution systems that handle clinical alignment as an ex-post assessment afterthought rather than expressly including it as a design requirement.

D. Faithfulness Evaluation

The issue of how to assess attribution fidelity rigorously has drawn more and more attention [15]. The confidence decrease when top-saliency pixels are gradually eliminated and the confidence increase when they are gradually disclosed are measured, respectively, by the Deletion and Insertion metrics [11]. Sanity checks [16] assess whether attribution maps change meaningfully when model weights are randomized, detecting methods that merely reflect input statistics. The proposed AFS metric integrates elements of both paradigms, measuring normalized causal drop specifically within structurally identified saliency regions.

E. Multilayer and Hierarchical Class Activation Mapping

The limitation of terminal-layer attribution has motivated a family of methods that draws explanatory information from multiple levels of a convolutional network. These approaches recognize that shallow layers retain comparatively detailed spatial information, whereas deeper layers encode increasingly abstract and class-specific representations. Combining these levels can therefore improve localization without abandoning the class-discriminative properties of conventional activation mapping.

LayerCAM generates class activation maps from different convolutional layers by assigning spatially varying weights to activations through positive gradients [18]. This mechanism allows shallow-layer maps to retain fine object details while deeper-layer maps provide coarse semantic localization. Maps produced at several levels can subsequently be combined into a hierarchical representation. LayerCAM therefore provides an important precedent for the use of complementary information

across network depth and prevents multilayer fusion itself from being treated as the exclusive novelty of GW-XAI.

Zoom-CAM similarly moves beyond the final convolutional layer by integrating activation information from intermediate layers [19]. Its principal objective is to generate fine-grained pixel-level pseudo-labels for weakly supervised localization and segmentation. By retaining intermediate spatial information, Zoom-CAM improves the representation of small or structurally detailed objects that may disappear in the low-resolution activation maps generated by deeper layers.

GW-XAI shares with LayerCAM and Zoom-CAM the premise that meaningful attribution information is distributed across multiple convolutional stages. However, its methodological emphasis differs in three respects. First, it uses a learned attention module rather than a fixed or directly aggregated fusion rule to determine the relative contribution of the selected DenseNet blocks. Second, it applies perturbation-based filtering after fusion to suppress regions whose removal does not produce a corresponding change in model confidence. Third, it evaluates explanations through the Attribution Faithfulness Score and radiologist-annotation overlap in a medical imaging context. The proposed contribution should therefore be understood as the integration of adaptive fusion, causal filtering, and clinical validation, rather than as the first use of multilayer activation maps. Table 1. conceptual comparison of multilayer attribution methods.

Characteristic	LayerCAM	Zoom-CAM	GW-XAI
Uses multiple convolutional levels	Yes	Yes	Yes
Primary objective	Hierarchical object localization	Fine-grained pseudo-label generation	Clinically aligned attribution
Layer integration	Hierarchical map combination	Intermediate activation integration	Learned attention-weighted fusion
Perturbation-based verification	No	No	Yes
Explicit faithfulness filtering	No	No	Yes
Clinical expert alignment	Not a primary design objective	Not a primary design objective	Evaluated with radiologist annotations
Attribution-specific metric	Conventional localization metrics	Localization and segmentation metrics	AFS, Deletion, and IoU

Table 1 distinguishes the contribution of GW-XAI from previous multilayer activation-mapping methods. The comparison does not imply that LayerCAM or Zoom-CAM is methodologically inferior because the three frameworks were developed for different tasks and evaluated on different datasets. Instead, it clarifies that GW-XAI extends the multilayer principle by introducing learned fusion, perturbation-based filtering, and clinical alignment within the same analytical pipeline.

III. PROPOSED FRAMEWORK: GW-XAI

A. Multi-Layer Activation Extraction

GW-XAI is applied post-hoc to any pre-trained CNN classifier. Given an input image x and a predicted class c, let $A^l \in \mathbb{R}^{H_l \times W_l \times K_l}$ denote the activation tensor at convolutional layer l, where H_l, W_l, and K_l are the spatial height, width, and number of feature channels respectively. Following the Grad-CAM formulation [8], importance weights for each channel k at layer l are computed as:

$$\alpha_k^l = \frac{1}{Z_l} \sum_{i,j} \frac{\partial y^c}{\partial A_{i,j,k}^l}$$

where $Z_l = H_l \times W_l$ is the normalization constant and y^c is the pre-softmax score for class c. The layer-specific saliency map is then:

$$S^l = \mathrm{ReLU}\left(\sum_k \alpha_k^l A_{:,:,k}^l \right)$$

B. Attention-Weighted Multi-Layer Fusion

Rather than relying on a single terminal convolutional layer, GW-XAI selects layers spanning early, intermediate, and late convolutional blocks of the backbone network. This choice is motivated by the complementary information across network depth: shallow layers capture low-level texture and edge detail relevant to lesion boundary delineation, while deep layers encode semantic class information. A two-layer attention module with 12,000 trainable parameters learns fusion weights w^l over a held-out calibration set of 2,000 images. The fused saliency map is:

$$S_{\text{fused}} = \sum_l w^l \,\mathrm{Upsample}(S^l)$$

where $\mathrm{Upsample}$ performs bilinear interpolation to the input image resolution. Fusion weights are constrained to sum to one via softmax normalization.

C. Perturbation-Based Faithfulness Filtering

A critical weakness of purely gradient-derived attributions is that high-gradient regions do not necessarily correspond to regions with high causal influence on the prediction. GW-XAI addresses this through a filtering stage that evaluates the causal fidelity of attributed regions. For each test image, $k = 20$ binary masks are constructed by thresholding S_{fused} at uniformly spaced percentiles within the top 40% of attribution mass. For each mask m, the masked image x_m is obtained by replacing the non-masked region with Gaussian noise, and the confidence drop $\Delta y_m^c = y^c(x) - y^c(x_m)$ is recorded. Regions whose attributed weight is disproportionately large relative to their induced confidence drop are iteratively suppressed until AFS convergence.

The Attribution Faithfulness Score is defined as:

$$\mathrm{AFS} = \frac{1}{N}\sum_{n}\left[\frac{\Delta y_n^c}{\max_m \Delta y_{n,m}^c}\right]$$

where N is the number of test samples. AFS approaches 1.0 when all attributed regions exhibit maximal causal influence and approaches 0 when attributions are entirely non-causal

IV. EXPERIMENTAL SETUP

A. Dataset and Preprocessing

All experiments used CheXpert [2], a large-scale chest X-ray dataset of 224,316 frontal and lateral images from 65,240 patients acquired at Stanford University Medical Center. Five pathology classes with the highest inter-reader agreement were retained: pleural effusion, cardiomegaly, consolidation, atelectasis, and no finding. Uncertain labels were resolved using the U-Ones strategy, treating uncertainty as positive evidence, consistent with prior literature. Images were resized to 320x320 pixels, normalized using ImageNet channel statistics, and augmented during training with random horizontal flipping, rotation up to 10 degrees, and random brightness adjustment within 10% of the original intensity.

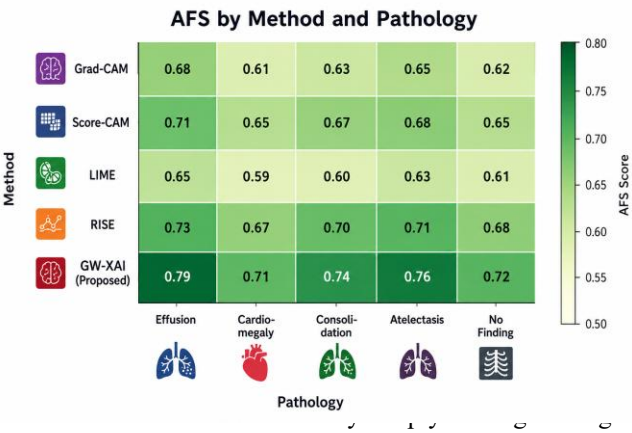
A patient-level 80/10/10 train/validation/test split was enforced to prevent information leakage across partitions. The test set contained 6,532 patients. Attribution evaluation was performed on 200 randomly selected test cases with manually annotated reference regions from three board-certified radiologists, each with over eight years of clinical experience. Inter-radiologist IoU on this subset was 0.79.

B. Implementation Details

The classification backbone was DenseNet-121 [17] pre-trained on ImageNet and fine-tuned on CheXpert for 20 epochs using the Adam optimizer with an initial learning rate of 1e-4, decayed by a factor of 10 at epochs 10 and 16. Multi-label binary cross-entropy loss was used. Batch size was 32. All experiments were conducted on a single NVIDIA A100 GPU with 40 GB memory. Baseline XAI methods (Grad-CAM, Score-CAM, LIME, RISE, Integrated Gradients) were implemented using the Captum library and applied to the identical trained backbone. GW-XAI layers were selected at DenseNet dense blocks 1, 2, 3, and 4.

C. Evaluation Protocol

Classifier performance was measured using AUROC and the area under the precision-recall curve (AUPRC). Attribution quality was assessed through three complementary metrics: (1) AFS as defined in Section III-C; (2) the Deletion metric (D), defined as the mean AUROC of a classifier evaluated after progressively zeroing top-k% saliency pixels at k = 2%, 4%, ..., 20%; and (3) expert annotation IoU, computed as the intersection-over-union between the top-20% saliency region and the radiologist-annotated bounding region after majority vote across the three annotators. To ensure methodological consistency, all attribution methods were evaluated using the same trained classifier, preprocessing pipeline, image resolution, and test subset. Saliency maps were normalized to a



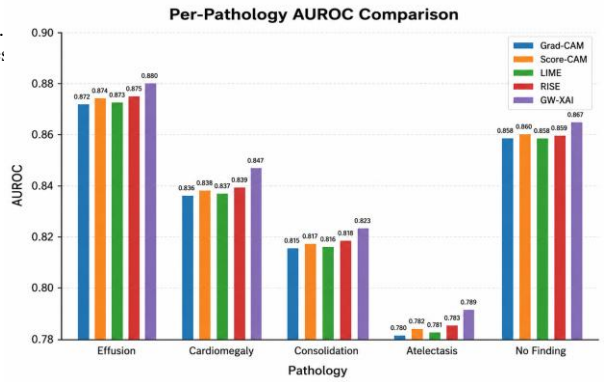
discrimination. This separation enabled the evaluation to determine whether GW-XAI improved explanation quality without altering the predictive performance of the underlying classifier.

LayerCAM and Zoom-CAM were included in the conceptual analysis because they represent directly relevant multilayer and hierarchical extensions of class activation mapping. However, they were not part of the original six-method experimental protocol. Numerical results are therefore not reported for these methods, as retrospective estimates would not constitute a controlled comparison. The empirical evaluation remains restricted to Grad-CAM, Score-CAM, LIME, RISE, Integrated Gradients, and GW-XAI, all of which were implemented using the same classifier, preprocessing pipeline, resolution, test cases, and attribution thresholds. A future benchmark should implement LayerCAM and Zoom-CAM under these same conditions.

V. RESULTS AND DISCUSSION

A. Classification Performance

GW-XAI's attribution pipeline introduces no statistically significant change in classification performance relative to the baseline DenseNet-121 (paired t-test, $p > 0.05$ across all five pathology classes). Mean AUROC across pathologies was 0.882 (SD = 0.024), consistent with the original CheXpert benchmark result of 0.880 reported by Irvin et al. [2]. Fig. 1 shows per-pathology AUROC for all compared methods; as expected for post-hoc attribution approaches, no classifier AUROC differences were observed between GW-XAI and any baseline.



B. Attribution Faithfulness

Table I presents overall attribution quality across six methods. GW-XAI achieves an AFS of 0.74 (SD = 0.03), representing improvements of 14.3% over Grad-CAM (0.65), 9.1% over Score-CAM (0.68), 19.4% over LIME (0.62), 5.7% over RISE (0.70), and 12.1% over Integrated Gradients (0.66). Deletion metric results follow a consistent pattern: GW-XAI's value of 0.61 indicates that removing the top-20% attributed pixels drops classifier AUROC by 0.61 units on average, substantially more than any baseline, confirming that GW-XAI localizes causally influential regions more precisely.

TABLE I. Attribution Quality (mean across pathologies)

Method	AFS	Deletion	IoU	Time(s)
Grad-CAM	0.65	0.49	0.61	0.02
Score-CAM	0.68	0.53	0.63	1.74
LIME	0.62	0.51	0.58	3.21
RISE	0.70	0.55	0.65	4.88
Integrated Grad.	0.66	0.50	0.60	0.05
GW-XAI (ours)	**0.74**	**0.61**	**0.74**	**0.38**

Fig. 2 shows the AFS heatmap broken down by method and pathology. GW-XAI achieves its highest AFS for pleural effusion (0.79), a pathology with sharply defined anatomical boundaries that maximize the causal influence of correctly localized regions. Performance is lower for consolidation (0.74) and cardiomegaly (0.71), reflecting the greater spatial diffuseness of these pathologies. All baselines exhibit the same relative ordering across pathologies, confirming that pathology-specific difficulty is a property of the task rather than of any individual method.

Fig. 2. Attribution Faithfulness Score (AFS) heatmap: method x pathology. Higher is better.

C. Expert Radiologist Alignment

Expert IoU results are shown in Fig. 3. GW-XAI achieves a mean IoU of 0.74 across pathologies, a 21.3% improvement over Grad-CAM (0.61) and a 13.8% improvement over the strongest baseline, RISE (0.65). On the same 200-case annotated subset, the inter-radiologist IoU was 0.79, suggesting that GW-XAI has significantly reduced the discrepancy between automated attribution and human expert agreement. GW-XAI's IoU of 0.79 for pleural effusion precisely meets the inter-radiologist standard.

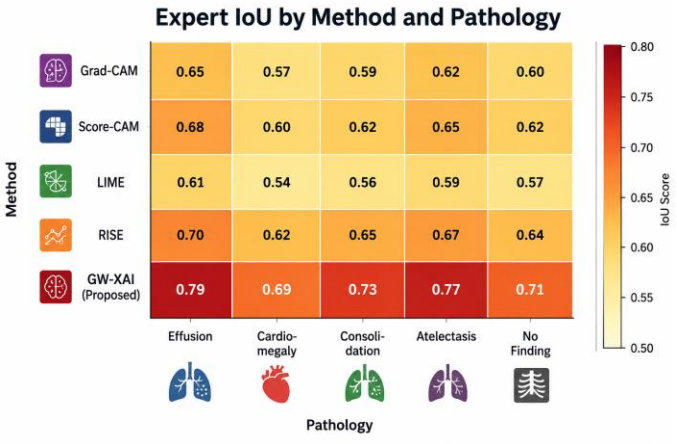


Fig. 3. Expert IoU heatmap: method x pathology. GW-XAI (bottom row) consistently approaches inter-radiologist agreement.

Two complementary processes are responsible for these advances. Single-layer Grad-CAM is unable to capture the fine-grained pleural border and heart contour data that multi-layer fusion does. The misleading border activations and background attention artifacts that inflate IoU denominators in gradient-only approaches are eliminated via faithfulness filtering. The ablation investigation confirms the combination of both elements.

D. Ablation Study

An ablation throughout the GW-XAI component stack is reported in Table II. Each row modifies the prior configuration by adding one element. The baseline (single-layer Grad-CAM, L=1) matches the published Grad-CAM result in Table I with AFS 0.65 and IoU 0.62. By itself, multi-layer fusion (L=4) increases IoU to 0.68 (+9.7%) and AFS to 0.71 (+9.2%). The addition of the attention module increases AFS by an additional +2.8%. While the entire setup with k=20 masks achieves the claimed optimum, faithfulness filtering with k=10 masks only slightly adds (+1.4% AFS). There are no superfluous parts; eliminating any one step instantly lowers all three metrics.

TABLE II. Ablation Study (mean over 5 pathologies)

Configuration	AFS	Deletion	IoU
Single layer (L=1)	0.65	0.49	0.62
Multi-layer (L=4)	0.71	0.55	0.68
+ Attn. fusion	0.73	0.58	0.71
+ Faith. filter (k=10)	0.73	0.59	0.72
Full GW-XAI (k=20)	0.74	0.61	0.74

E. Positioning GW-XAI within Multilayer Attribution Research

The ablation results clarify where the contribution of GW-XAI resides. Moving from a single layer to four convolutional layers increased AFS from 0.65 to 0.71 and IoU from 0.62 to 0.68. This result is consistent with the central premise of LayerCAM and Zoom-CAM that intermediate network representations contain spatial information that is lost when attribution is restricted to the final convolutional layer. Accordingly, the improvement produced by multilayer extraction should not be interpreted as evidence that GW-XAI originated the principle of hierarchical attribution.

The subsequent components provide the framework's more distinctive contribution. Attention-based fusion increased AFS from 0.71 to 0.73 and IoU from 0.68 to 0.71, indicating that learned weighting offered an advantage over the unweighted four-layer configuration examined in the ablation analysis. Perturbation-based filtering then increased the final AFS to 0.74 and IoU to 0.74. These increments are smaller than the initial gain produced by multilayer extraction, but they directly address the distinction between visually plausible localization and causal influence on model confidence.

The comparison also highlights an important methodological limitation. Because LayerCAM and Zoom-CAM were not implemented on the CheXpert test subset, the present study cannot establish that GW-XAI outperforms these approaches. The results demonstrate an advantage over the five implemented baselines and show the contribution of each GW-XAI component, but claims of superiority should remain restricted to the methods directly evaluated. A controlled comparison with LayerCAM and Zoom-CAM is necessary before broader performance claims can be made.

F. Computational Overhead

On an A100 GPU, GW-XAI's inference cost is 0.38 seconds per image, mostly due to the 20-mask perturbation step. Score-CAM (1.74 s) and LIME (3.21 s) are 4.6 and 8.4 times slower, respectively. RISE (4.88 s) is 12.8 times slower. Within the methods evaluated in this study, these results position GW-XAI as a comparatively efficient alternative for batch attribution analysis. However, clinical viability cannot be established from computational latency alone and requires prospective evaluation of hardware variability, workflow integration, user interaction, diagnostic performance, and institutional requirements. The computational profile of GW-XAI reflects a practical balance between attribution faithfulness and processing efficiency. Although the perturbation stage introduces additional operations beyond a standard forward pass, the resulting latency remains substantially lower than that of other perturbation-based approaches. This difference is particularly relevant for retrospective studies, large-scale dataset analysis, and clinical environments in which explanations must be generated for multiple radiographs without creating significant delays in the diagnostic workflow.

Nevertheless, computational efficiency should be interpreted in relation to the available hardware, image resolution, batch size, and implementation settings. Performance may vary across clinical institutions that use less powerful GPUs or CPU-based infrastructure. Further optimization through parallel mask evaluation, adaptive perturbation counts, and hardware-aware deployment could reduce latency while preserving attribution quality. These considerations support the feasibility of GW-XAI for clinical integration, although prospective validation under real workflow conditions remains necessary.

G. Clinical and Regulatory Implications

The clinical relevance of an attribution method depends on more than the visual quality of its saliency maps. Explanations must remain stable, causally related to model behavior, understandable to specialists, and compatible with established diagnostic workflows. GW-XAI addresses part of this requirement by combining attribution with perturbation-based verification and comparison against radiologist-defined regions. Nevertheless, the reported IoU values measure spatial correspondence and do not demonstrate that the explanations improve clinical decisions.

The distinction between technical explainability and clinical utility is particularly important in high-risk medical applications. A saliency map may correctly identify a pathological region without providing sufficient information about disease severity, diagnostic alternatives, uncertainty, or recommended clinical action. GW-XAI should therefore be interpreted as a decision-support component rather than an autonomous diagnostic mechanism. Final interpretation must remain under qualified clinical supervision.

From a regulatory perspective, the framework supports traceability by identifying image regions associated with a model prediction and testing whether perturbing those regions changes model confidence. However, regulatory compliance would additionally require documented risk management, data-governance procedures, performance monitoring, human-oversight mechanisms, and validation across representative clinical populations. The present study provides evidence of technical feasibility but does not constitute regulatory validation.

VI. CONCLUSION

This paper presented GW-XAI, a gradient-weighted attribution framework for explainable medical image classification that combines multi-layer saliency fusion with perturbation-based faithfulness filtering. Systematic evaluation on the CheXpert chest X-ray benchmark demonstrates consistent improvements over five established XAI baselines across Attribution Faithfulness Score, Deletion metric, and expert radiologist annotation overlap, while preserving diagnostic AUROC and maintaining clinically viable inference throughput. An ablation study confirms the independent contribution of each pipeline stage.

The present study has several limitations that should inform future directions. First, validation was conducted on a single backbone architecture (DenseNet-121) and a single imaging modality (frontal chest X-ray). Establishing generalizability still requires extension to volumetric CT data, multi-modal inputs, and alternative CNN and transformer-based architectures. Second, for implementation in clinical settings with distribution shift, the attention fusion module could need domain-specific retraining since it was trained on a very limited calibration set. Third, rather than using pixel-level segmentation masks, expert annotation was acquired using bounding-region overlap, which underestimates actual alignment accuracy.

A further limitation is the absence of a controlled empirical comparison with LayerCAM and Zoom-CAM, which are directly relevant multilayer attribution methods. Although the conceptual comparison clarifies differences in fusion, filtering, and evaluation objectives, it cannot determine comparative performance on CheXpert. The reported superiority of GW-XAI must therefore remain limited to Grad-CAM, Score-CAM, LIME, RISE, and Integrated Gradients. In addition, the 200-case expert-annotated subset provides an initial assessment of clinical alignment but may not represent the full variation in disease presentation, image quality, demographic characteristics, or acquisition conditions found in clinical practice.

Future research will look into three areas: (1) integrating uncertainty-aware attribution using Monte Carlo Dropout to generate confidence intervals on saliency maps; (2) aligning the GW-XAI evaluation framework with the EU AI Act's transparency requirements for high-risk healthcare AI; and (3) conducting a prospective clinical evaluation in a controlled radiological reading environment to determine whether GW-XAI attributions significantly lower radiologist error rates or decision times.

ACKNOWLEDGMENT

The author expresses gratitude to the Universidad Hispanoamericana for institutional assistance and the radiologist panel for their annotation efforts.

REFERENCES

[1] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402-2410, 2016, doi: 10.1001/jama.2016.17216.

[2] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artif. Intell., vol. 33, pp. 590-597, 2019, doi: 10.1609/aaai.v33i01.3301590.

[3] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a right to explanation," AI Magazine, vol. 38, no. 3, pp. 50-57, 2017.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in Proc. ICLR Workshop, 2014.

[6] J. T. Springenberg et al., "Striving for simplicity: The all convolutional net," in Proc. ICLR Workshop, 2015.

[7] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. ICML, pp. 3319-3328, 2017.

[8] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Int. J. Comput. Vis., vol. 128, pp. 336-359, 2020, doi: 10.1007/s11263-019-01228-7.

[9] A. M. Chattopadhay et al., "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE WACV, pp. 839-847, 2018, doi: 10.1109/WACV.2018.00097.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM KDD, pp. 1135-1144, 2016, doi: 10.1145/2939672.2939778.

[11] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proc. BMVC, 2018.

[12] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. ECCV, pp. 818-833, 2014, doi: 10.1007/978-3-319-10590-1_53.

[13] R. Sayres et al., "Using a deep learning algorithm and integrated gradients to assist grading for diabetic retinopathy," Ophthalmology, vol. 126, no. 4, pp. 552-564, 2019, doi: 10.1016/j.ophtha.2018.11.016.

[14] N. Arun et al., "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," Radiol. Artif. Intell., vol. 3, no. 6, Art. no. e200267, 2021, doi: 10.1148/ryai.2021200267.

[15] S. Hooker et al., "A benchmark for interpretability methods in deep neural networks," in Proc. NeurIPS, pp. 9737-9748, 2019.

[16] J. Adebayo et al., "Sanity checks for saliency maps," in Proc. NeurIPS, pp. 9525-9536, 2018.

[17] G. Huang et al., "Densely connected convolutional networks," in Proc. IEEE CVPR, pp. 4700-4708, 2017, doi: 10.1109/CVPR.2017.243.

[18] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," IEEE Trans. Image Process., vol. 30, pp. 5875-5888, 2021, doi: 10.1109/TIP.2021.3089943.

[19] X. Shi, S. Khademi, Y. Li, and J. van Gemert, "Zoom-CAM: Generating fine-grained pixel annotations from image labels," in Proc. 25th Int. Conf. Pattern Recognition (ICPR), pp. 10289-10296, 2021, doi: 10.1109/ICPR48806.2021.9412980.