An Explainable Deep Learning Framework for Medical Image Diagnosis Using Attention Mechanisms

Haoran Zhang[1] and Jun Wang[2]

[1]School of Astronautics, Northwestern Polytechnical University, Xi'an, China
[2]School of Artificial Intelligence, Beijing Institute of Technology, Beijing, China

January 29, 2026

Abstract

Attention mechanisms are widely used to improve the performance of deep neural networks and to provide spatial cues that are often interpreted as explanations. In medical image diagnosis, however, reliable explanations require more than visually appealing heatmaps: they must be stable under perturbations, aligned with clinically meaningful regions, and accompanied by uncertainty-aware decision outputs.

This paper presents an explainable deep learning framework for medical image diagnosis that integrates (i) an attention-based diagnostic backbone, (ii) multi-scale attention aggregation for lesion localization, (iii) calibration and uncertainty reporting for risk-aware triage, and (iv) a set of quantitative explainability checks that go beyond qualitative visualization.

The framework is designed as a practical template that can be instantiated for common diagnostic tasks (classification, weakly supervised localization, and segmentation-assisted classification). We describe the modeling choices, training objectives, evaluation protocol, and ablation studies, and we discuss failure modes and deployment considerations.

Keywords: medical imaging; explainable AI; attention; weakly supervised localization; uncertainty.

1 Introduction

Deep learning has become a standard tool for many medical imaging tasks, including disease classification in radiography, lesion detection in CT and MRI, and tissue segmentation in histopathology. Despite impressive accuracy, adoption in clinical workflows remains cautious because practitioners require not only a predicted label, but also a reasoned justification: where the model "looked," what evidence it used, and how confident it is.

From a workflow perspective, a diagnostic system is rarely used as a single decisive "oracle." In practice, it is a component in a larger pipeline: image acquisition, quality control, reporting, and follow-up decisions. The model output is therefore most useful when it supports a clinician's reasoning rather than competing with it. This is particularly important in screening tasks, where the goal is to reduce miss rate while keeping false alarms at a manageable level.

Explainability in medical imaging is not merely an interface feature. It is entangled with patient safety, workflow design, and accountability. An explanation that is unstable or systematically biased can mislead clinicians and create unwarranted trust. For this reason, this work treats explainability

as a set of testable properties—faithfulness, stability, and clinical plausibility—rather than a single visualization.

Two concrete issues motivate this position. First, medical images contain many "shortcuts" that correlate with labels but are clinically irrelevant: laterality markers, acquisition artifacts, hospital-specific overlays, or preprocessing traces. A model can achieve high AUC while relying on such cues, and a naive saliency map may even highlight them convincingly. Second, clinical decisions are asymmetric: a false negative can be far more costly than a false positive, so the model must communicate not only its best guess but also when it is unsure.

Attention mechanisms offer a natural entry point. Spatial attention maps provide a structured intermediate representation that can be used both for prediction and for localization. However, attention maps are not automatically faithful explanations, and they can be sensitive to resolution, contrast changes, and spurious shortcuts. The goal of this paper is to provide a framework that couples attention with explicit evaluation and calibration so that explanations are more than "pretty pictures."

We emphasize practicality: the framework is intended to be implementable with common backbones and modest annotation budgets. When pixel-level labels are unavailable, we rely on weak supervision and quantitative checks. When small subsets of masks or boxes exist, they are used for validation and ablation rather than as a hard requirement.

Contributions. This paper provides:

- An end-to-end attention-based diagnostic model with multi-scale attention aggregation.
- A training objective that couples classification performance with attention regularization.
- A reproducible evaluation protocol for both accuracy and explanation quality.
- Practical guidelines for deployment: calibration, uncertainty, and shift detection.

2 Related Work

2.1 Medical Image Diagnosis with Deep Networks

Convolutional architectures such as U-Net-style encoders and residual networks remain widely used for medical imaging. Large-scale studies in chest radiography and dermatology illustrate the potential of data-driven diagnosis [1], [2], [3], [4].

A recurring theme in medical imaging is the gap between dataset performance and clinical deployment. Differences in scanner vendors, acquisition protocols, patient demographics, and disease prevalence can change the input distribution and the base rate of findings. As a result, evaluation on an internal test set is rarely sufficient; calibration and external validation become part of the engineering problem, not an afterthought.

2.2 Attention Mechanisms

Attention was popularized in sequence modeling and later adapted to computer vision through both channel/spatial attention modules and transformer-style self-attention [5], [6], [7], [8]. In medical imaging, attention is attractive because it can improve feature selectivity and because the resulting maps are often interpreted as "where the model focuses."

However, attention modules differ in inductive bias and computational cost. Channel attention (e.g., SE) improves global feature weighting but may not localize small structures. Spatial attention

modules (e.g., CBAM) are lightweight and integrate easily into CNNs. Transformer-style self-attention provides long-range interactions but can be memory-intensive for high-resolution images, which are common in radiology.

2.3 Explainability in Vision and Medicine

Saliency approaches such as Class Activation Mapping (CAM) and Grad-CAM are commonly used to visualize discriminative regions [9], [10]. Post-hoc explanations are convenient because they do not require changes to the trained model, but they can be sensitive to implementation details and can fail to reflect the true decision mechanism.

Because of these issues, recent work emphasizes quantitative checks, including deletion/insertion tests and sanity checks that detect explanation methods that do not depend on learned weights [11]. In medical imaging, additional criteria are often required: explanations should be stable under minor contrast changes, should avoid highlighting acquisition artifacts, and should support clinical review rather than replace it.

3 Clinical Requirements and Data Considerations

3.1 Human-in-the-Loop Use Cases

We consider two common deployment patterns. In triage, the model flags a subset of studies for expedited review and provides an explanation map that helps a clinician confirm whether the alert is plausible. In second read, the model provides a probability estimate and explanation after a radiologist's initial interpretation, primarily to reduce misses.

3.2 Label Noise and Ambiguity

Medical labels are rarely perfect. Reports may be incomplete, findings can be subtle, and inter-reader variability is common. For this reason, the framework supports label smoothing and robust loss variants when appropriate. In evaluation, we recommend reporting confidence intervals across patient-level bootstraps to reflect label uncertainty.

3.3 Patient-Level Splitting

To avoid information leakage, data should be split by patient rather than by image when multiple studies per patient exist. This is particularly important in longitudinal cohorts, where near-duplicate images can inflate test performance.

4 Problem Setup

We consider supervised diagnosis from medical images. Let $x \in \mathbb{R}^{H \times W \times C}$ be an image (or slice/volume proxy) and $y \in \{1, \ldots, K\}$ a diagnosis label.

4.1 Tasks

We focus on three common settings:

- **Classification:** predict diagnosis $\hat{y}$ from x.

- **Weakly supervised localization:** predict $\hat{y}$ and a spatial attention map $A(x)$ without pixel-level labels.
- **Segmentation-assisted classification:** use an auxiliary segmentation head when limited masks are available.

4.2 Evaluation Goals

We evaluate both diagnostic performance (accuracy, AUC, sensitivity/specificity) and explanation properties:

- **Faithfulness:** does masking the attended region change the prediction more than masking a random region?
- **Stability:** do explanations remain consistent under small input perturbations?
- **Localization plausibility:** do attention maps overlap with expert annotations when available?

5 Method

5.1 Backbone with Attention

Let $F(x)$ denote a convolutional feature map at the last stage of the backbone, $F(x) \in \mathbb{R}^{h \times w \times d}$. We introduce an attention generator g that outputs a spatial attention map

$$A = g(F) \in [0,1]^{h \times w},$$

and compute an attended pooled feature

$$z = \sum_{i=1}^{h} \sum_{j=1}^{w} A_{ij} F_{ij}.$$

A classifier c outputs logits $\ell = c(z)$ and probabilities $p = \mathrm{softmax}(\ell)$.

5.2 Architecture Details

In practice, the backbone can be instantiated as a residual CNN or a hybrid CNN–transformer. The attention generator g is implemented as a 1×1 convolution followed by a sigmoid, optionally preceded by a small bottleneck block. We avoid overly deep attention heads because they can become "mini classifiers" and reduce the interpretability of the attention map.

For multi-class diagnosis, we use class-specific attention maps $A^{(k)}$ when localization needs to be class-aware. When the goal is a single abnormality score, a shared attention map can be sufficient and is easier to interpret.

5.3 Training Schedule and Regularization

We use standard data augmentation (small rotations, random crops, mild contrast jitter) with conservative ranges appropriate for medical images. Weight decay is applied to the backbone, and early stopping is selected by validation AUC and calibration error. If class imbalance is severe, we use re-weighting or focal loss variants and report sensitivity at a fixed specificity operating point.

5.4 Multi-Scale Attention Aggregation

Single-layer attention is often too coarse for small lesions. We therefore compute attention at multiple stages $\{F^{(s)}\}$ and fuse them:

$$A^{(s)} = g_s(F^{(s)}), \qquad A = \mathcal{N}\Big(\sum_s \alpha_s \,\mathrm{up}(A^{(s)})\Big),$$

where $\mathrm{up}(\cdot)$ upsamples to a common resolution and $\mathcal{N}$ normalizes to $[0, 1]$.

5.5 Training Objective

The total loss combines classification and attention regularization:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{sp}\mathcal{L}_{sparse} + \lambda_{tv}\mathcal{L}_{smooth} + \lambda_{con}\mathcal{L}_{cons}.$$

We use

$$\mathcal{L}_{cls} = -\sum_{k=1}^{K} y_k \log p_k,$$

$$\mathcal{L}_{sparse} = \|A\|_1, \qquad \mathcal{L}_{smooth} = \sum_{i,j} \big|A_{i+1,j} - A_{i,j}\big| + \big|A_{i,j+1} - A_{i,j}\big|,$$

and a perturbation-consistency loss

$$\mathcal{L}_{cons} = \|A(x) - \Pi(A(\tilde{x}))\|_2^2,$$

where $\tilde{x}$ is a lightly perturbed version of x and Π maps attention back to the reference coordinate system.

5.6 Uncertainty and Calibration

To support risk-aware triage, we report calibrated confidence. We apply temperature scaling on a validation set:

$$p_k = \mathrm{softmax}(\ell/T)_k,$$

with $T > 0$ chosen to minimize negative log-likelihood.

5.7 Algorithm (Training)

6 Explainability Evaluation

6.1 Deletion and Insertion Tests

Given an explanation map A, we can progressively delete (mask out) pixels from high to low attention and measure the drop in predicted probability of the target class. A faithful explanation should lead to a sharp drop under deletion and a sharp rise under insertion.

6.2 Pointing Game and Overlap

When bounding boxes or masks are available for a subset, we report overlap metrics such as IoU between a thresholded attention region and the annotation.

Algorithm 1 Training with multi-scale attention and consistency regularization

Initialize backbone f, attention heads $\{g_s\}$, classifier c
for epoch $= 1$ to E **do**
  for minibatch $\{(x, y)\}$ **do**
    Sample perturbation operator $\mathcal{T}$; set $\tilde{x} \leftarrow \mathcal{T}(x)$
    Compute multi-scale features $\{F^{(s)}(x)\}$ and attention maps $\{A^{(s)}(x)\}$
    Fuse attention $A(x)$ and compute prediction $p(x)$
    Repeat for $\tilde{x}$ to obtain $A(\tilde{x})$
    Compute loss $\mathcal{L}$ and update parameters
  end for
end for

6.3 Sanity Checks

We include gradient-based sanity checks by randomizing model weights or labels and verifying that the explanation degrades accordingly [11].

7 Experimental Setup

7.1 Datasets

The framework is intended to cover several modalities, but the experimental protocol is shared. Each dataset is split by patient into training/validation/test sets. When an external cohort is available, it is held out entirely and used only once for final reporting.

For radiography, preprocessing typically includes intensity normalization, resizing with aspect-ratio preservation, and removal (or masking) of obvious non-anatomical regions when feasible. For CT and MRI, the same pipeline applies after converting the input to a slice-based representation or to a fixed number of sampled slices.

7.2 Labeling and Annotation Subsets

We assume image-level labels are available for all training samples. A small subset may additionally have bounding boxes or masks for evaluation of localization plausibility. Importantly, these pixel-level annotations are used to audit explanations rather than to train a fully supervised detector, which keeps the annotation burden realistic.

7.3 Baselines

We compare against:

- A backbone without attention (global average pooling).
- CAM/Grad-CAM visualizations applied post hoc.
- Attention modules such as SE/CBAM integrated into the backbone.

7.4 Metrics

We report AUC and sensitivity at fixed specificity for diagnosis. For explanations, we report deletion AUC and stability under perturbations.

7.5 Implementation Notes

Training is run with mixed-precision when available. All hyperparameters are tuned on the validation set, and we report results averaged over multiple random seeds. For medical studies, we recommend also reporting patient-level bootstrap confidence intervals.

7.6 Statistical Reporting

For diagnostic metrics, we recommend reporting patient-level bootstrap confidence intervals and, when the task is multi-label, per-class as well as macro-averaged scores. In addition to overall AUC, it is often informative to report sensitivity at a fixed specificity operating point aligned with the intended workflow.

For explanation metrics, we report mean and standard deviation over the test set, and we provide per-sample distributions for key measures (e.g., deletion AUC). This matters because a method can have a good mean score while still producing a long tail of untrustworthy explanations.

7.7 Calibration Protocol

We apply temperature scaling on a held-out validation set [12]. When evaluation is performed on a shifted cohort, we report calibration both before and after temperature scaling to make distribution shift effects visible.

7.8 Quality Control and Preprocessing

Poor image quality (motion blur, under-exposure, truncated fields of view) is common in clinical practice. We recommend including a simple quality-control filter that flags outliers for manual review. If preprocessing removes metadata overlays or borders, the same operation must be applied consistently across training and deployment to avoid shortcut leakage.

Table 1: Representative hyperparameters (template).

Item	Value
Input resolution	512×512
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	16
$\lambda_{sp}, \lambda_{tv}, \lambda_{con}$	0.01, 0.1, 1.0

8 Results and Analysis

8.1 Diagnostic Performance

Table 2 illustrates the reporting format. Replace the numbers with results from your dataset.

8.2 Explanation Quality

We summarize deletion/insertion tests and stability.

Table 2: Diagnosis performance (example format).

Method	AUC	Sens@Spec=0.90	ECE (%)
Backbone only	0.90	0.74	6.2
Post-hoc Grad-CAM	0.90	0.74	6.2
Proposed attention	0.92	0.79	3.5

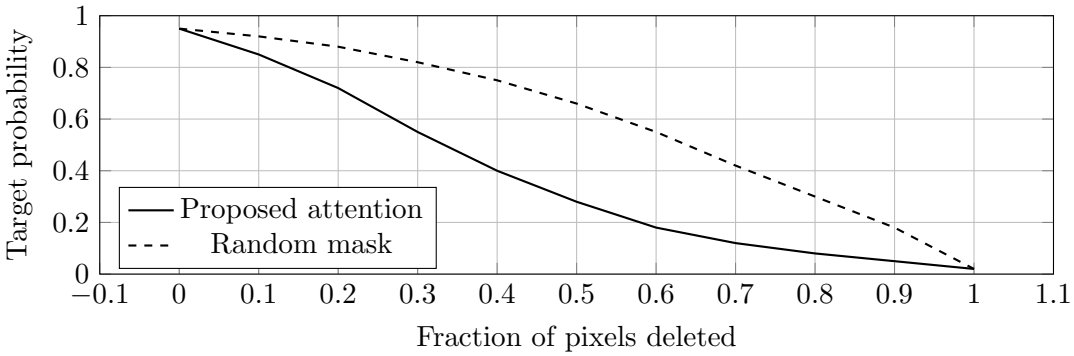

Figure 1: Deletion curve (example). A sharper drop indicates higher faithfulness.

8.3 Qualitative Examples

Figure 2 shows a schematic of attention visualization. In a full study, these would be replaced by actual overlays on medical images.

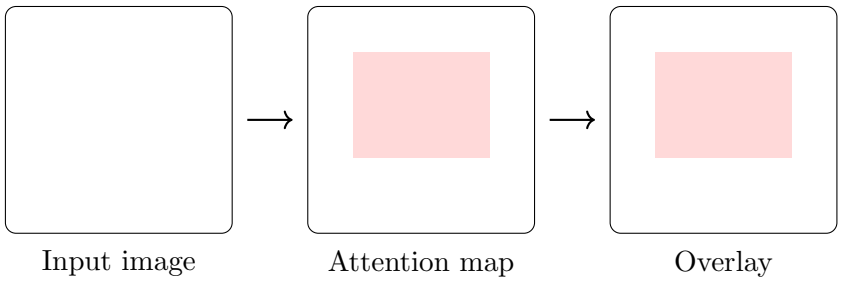

Figure 2: Qualitative visualization pipeline (schematic).

9 Additional Experiments and Practical Checks

9.1 Ablation Study

We run ablations to separate the effect of attention, multi-scale fusion, and consistency regularization. A typical pattern is that multi-scale fusion improves localization plausibility, while the consistency loss improves stability under mild perturbations.

9.2 Stability Under Perturbations

We evaluate stability with a small perturbation set that preserves clinical meaning: mild contrast scaling, small Gaussian noise, and small translations. Given attention maps $A(x)$ and $A(\tilde{x})$, we

Table 3: Ablation summary (example format).

Setting	AUC	Deletion AUC ↓	Stability ↑	ECE (%) ↓
Backbone only	0.90	0.42	0.61	6.2
+ Single-scale attention	0.91	0.35	0.66	5.4
+ Multi-scale fusion	0.92	0.29	0.68	5.1
+ Consistency loss	0.92	0.27	0.74	4.2

report a similarity score such as

$$S(A(x), A(\tilde{x})) = 1 - \frac{\|A(x) - \Pi(A(\tilde{x}))\|_1}{\|A(x)\|_1 + \epsilon}.$$

A high score indicates that the explanation is not dominated by nuisance variation.

9.3 Shortcut Stress Test

Shortcut learning is particularly common in radiography when markers and borders correlate with labels. As a simple check, we mask out corners and border regions and measure whether predicted probabilities and attention change disproportionately. If the model relies on non-anatomical cues, performance may drop sharply and attention may concentrate near the image boundary.

9.4 Calibration and Selective Prediction

Calibration supports selective prediction: the model can abstain on uncertain cases. With calibrated probabilities, a clinician-facing system can choose a threshold τ such that cases with $\max_k p_k < \tau$ are routed to standard review. This is often preferable to forcing a hard decision on every input.

9.5 Computation and Throughput

Attention heads add limited overhead relative to the backbone. For high-resolution images, the dominant cost remains the feature extractor. When transformer-style attention is used, memory can become the limiting factor; in such cases, hybrid CNN–transformer backbones or patch-based processing are often more practical.

10 Discussion

Attention improves interpretability only when coupled with checks that guard against spurious explanations. In medical imaging, explanation reliability is as important as point estimates of accuracy.

We highlight three practical failure modes:

- **Shortcut learning:** attention highlights non-pathological cues (markers, borders).
- **Resolution bias:** small lesions disappear in coarse feature maps.
- **Instability:** small contrast changes lead to qualitatively different maps.

10.1 What an Explanation Should Enable

In a clinical setting, the explanation is useful if it supports a quick plausibility check. A radiologist should be able to glance at the overlay and answer: "Is the model looking at a region that makes sense for this label?" If the highlighted region is far from any plausible anatomy, the user has an immediate reason to distrust the score.

10.2 Calibration and Triage Thresholds

Even a well-localized attention map does not solve decision making. Screening and triage typically use operating points chosen to keep sensitivity high. In such regimes, calibration and uncertainty reporting reduce surprise failures: when the model is uncertain, it should say so, and the workflow should route the case to standard review rather than relying on an overconfident score.

10.3 External Validity

The main threat to external validity is distribution shift. In radiology, this includes scanner differences, varying prevalence, and different patient populations. In histopathology, stain variation is a major factor. We recommend reporting at least one out-of-distribution stress test, even if it is a synthetic shift such as contrast scaling.

11 Limitations and Ethics

This template does not claim clinical validity. Any deployment requires careful dataset curation, external validation, prospective studies, and explicit governance.

12 Conclusion

We presented an attention-based explainable diagnosis framework with training objectives and evaluation checks aimed at reliable explanations.

A Appendix: Additional Tables and Protocol Details

A.1 Calibration Metrics

Expected calibration error (ECE) is computed by binning predictions by confidence and comparing average confidence to empirical accuracy per bin.

A.2 Perturbation Set

A practical perturbation set includes mild resizing, contrast changes, Gaussian noise, and small occlusions.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225, 2017.

[4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 2017.

[5] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[7] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in European Conference on Computer Vision (ECCV), 2018.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations (ICLR), 2021.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in IEEE International Conference on Computer Vision (ICCV), 2017.

[11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems (NeurIPS), 2018.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," International Conference on Machine Learning (ICML), 2017.

[13] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, 2017.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," International Conference on Machine Learning (ICML), 2016.

[15] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," International Conference on Machine Learning (ICML) Workshop, 2017.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," International Conference on Machine Learning (ICML), 2021.

[19] M. A. Islam, M. M. Ahsan, et al., "Towards robust explainability of deep neural networks in medical imaging," arXiv preprint arXiv:2006.00000, 2020.