International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 4, July-Aug-2026, ISSN (Online): 2395-566X

Explainable Deep Learning for Automated Image-Based Disease Classification

Rabins Porwal
Department of Computer Application
School of Engineering & Technology (UIET)
Chhatrapati Shahu Ji Maharaj University (CSJMU)
Kanpur, Uttar Pradesh, India

Abstract — Deep convolutional networks now match or exceed specialist performance on several image-based diagnostic tasks, yet their adoption in clinical practice remains limited by a problem that accuracy alone cannot solve: a clinician asked to act on a prediction cannot see why it was made. Post-hoc saliency methods offer a partial answer, but different methods applied to the same network routinely disagree, and a map that looks convincing is not necessarily one that reflects the computation the model actually performed. This paper proposes an Explainable Deep Learning (XDL) framework in which interpretability is a training objective rather than an afterthought. An attention-guided refinement module reweights backbone feature maps so that spatial evidence is concentrated before classification. Three complementary attribution methods – Grad-CAM++, integrated gradients and GradientSHAP – are then fused into a single saliency map, and the disagreement among them is quantified as an explanation-consistency score. Finally, a deletion-based faithfulness penalty is added to the loss, so that the network is optimised not only to classify correctly but to concentrate its evidence on regions whose removal genuinely changes the prediction. Evaluation across five public datasets spanning radiography, dermoscopy, fundus photography, magnetic resonance imaging and histopathology gave a mean accuracy of 93.7 per cent, 2.1 percentage points above the strongest baseline. More importantly, faithfulness improved substantially: deletion AUC fell from 0.157 to 0.128 and agreement with expert-annotated lesion masks rose from 0.452 to 0.518 in intersection over union. A blinded review by three clinicians rated the fused maps 4.2 out of 5 for plausibility against 3.6 for the best competing method.

Keywords— Explainable Artificial Intelligence, Deep Learning, Medical Image Classification, Saliency Maps, Grad-CAM, Attention Mechanism, Faithfulness, Computer-Aided Diagnosis.

I. INTRODUCTION

Automated interpretation of medical images has advanced rapidly over the past decade. Convolutional networks trained on large annotated corpora now detect pneumonia from chest radiographs, grade diabetic retinopathy from fundus photographs, classify skin lesions from dermoscopy and identify tumour subtypes from histopathology slides at levels comparable to trained specialists [1], [2]. Regulatory approvals of such systems have followed, and the supply of labelled imaging data continues to grow.

Deployment, however, has lagged well behind demonstrated accuracy. The obstacle is not chiefly technical but epistemic. A radiologist presented with a probability of 0.93 for malignancy has no way to determine whether the network attended to the lesion, to an adjacent artefact, or to a text marker burned into the image corner by the acquisition device. Documented failures of this kind are not hypothetical: models have been shown to key on hospital-specific tokens, on surgical skin markings that correlate with prior biopsy, and on scanner metadata rather than on pathology [3]. A model that is right for the wrong reason will generalise unpredictably, and in a clinical setting the cost of that unpredictability falls on patients.

The dominant response has been post-hoc explanation. Gradient-based attribution and class-activation mapping produce a heat map indicating which pixels most influenced the output [4], [5], [6]. These methods are attractive because they require no change to the trained model and impose almost no cost. They also have three weaknesses that matter considerably in a diagnostic context. First, different methods applied to the same network frequently highlight different regions, and there is no principled basis on which a clinician can prefer one map to another. Second, several widely used attribution methods have been shown to be insensitive to randomisation of model parameters or of labels, meaning the map can look plausible even when the underlying model has learned nothing [7]. Third, and most fundamentally, a post-hoc map describes a network that was never trained to be describable; nothing in the optimisation encouraged the evidence to be spatially coherent in the first place.

The alternative position, argued forcefully in the interpretability literature [8], is that high-stakes decisions should use models that are interpretable by construction rather than explained after the fact. Taken literally this would exclude deep networks from medical imaging altogether, which is neither practical nor necessary. The position adopted here is intermediate: retain the representational power of a deep backbone, but make interpretability part of what the network is optimised for.

This paper develops that idea into a concrete framework, referred to as XDL. Three mechanisms are combined. An attention-guided refinement module reweights backbone feature maps spatially and channel-wise before classification, concentrating evidence rather than allowing it to diffuse.

An explanation ensemble fuses three attribution methods with complementary failure modes, and the disagreement among them is measured explicitly rather than hidden. A faithfulness penalty derived from deletion testing is added to the loss, so that the network is rewarded when the regions its explanation identifies are the regions whose removal actually changes the prediction.

The contributions of this work are as follows.

- A training objective that couples classification loss with an explanation-consistency term and a differentiable faithfulness penalty, making interpretability an optimisation target rather than a diagnostic applied afterwards.
- An attention-guided refinement module that improves both accuracy and the spatial compactness of the resulting saliency maps, with the two effects separated by ablation.
- A weighted fusion of three attribution methods, with fusion weights derived from measured per-method faithfulness rather than fixed in advance.
- An evaluation across five imaging modalities that reports faithfulness, localisation and blinded clinician ratings alongside accuracy, together with Friedman and Nemenyi significance testing.

Section II reviews deep learning for medical imaging, post-hoc attribution, intrinsically interpretable architectures and the evaluation of explanations. Section III presents the framework. Section IV describes the datasets, baselines and metrics. Section V reports and discusses results, and Section VI concludes.

II. RELATED WORK

1. Deep Learning for Medical Image Classification

Residual networks [9] made very deep architectures trainable and remain a standard backbone for diagnostic imaging. Densely connected networks [10] improve gradient flow and parameter efficiency through feature reuse, and have performed well on datasets of the moderate size typical of medical corpora. Compound scaling of depth, width and resolution [11] produced the EfficientNet family, which offers a favourable accuracy-to-cost ratio and is widely used where inference must run on modest hardware. More recently, vision transformers [12] have been applied to medical images with encouraging results, although their appetite for data is a genuine constraint when annotation requires specialist time.

Transfer learning from natural-image pretraining is near-universal in this field. Its benefits are well documented but not unconditional: careful analysis has shown that much of the apparent gain on medical tasks comes from better weight scaling rather than from transferred features, and that smaller purpose-built architectures often match large pretrained ones [13]. This motivates evaluating any proposed component against strong, properly tuned baselines rather than against an untuned reference.

2. Post-Hoc Attribution Methods

Class activation mapping and its gradient-weighted generalisation [4] produce coarse localisation maps from convolutional feature activations and remain the most widely reported explanation in medical imaging papers. Grad-CAM++ [5] refines the weighting to handle multiple instances of a class and improves localisation of small lesions. Integrated gradients [6] takes a different route, accumulating gradients along a path from a baseline input and satisfying axioms of sensitivity and implementation invariance. SHAP-based attribution [14] grounds importance in cooperative game theory and provides a consistent additive decomposition, at considerably higher computational cost.

The methods disagree in practice. Sanity checks [7] demonstrated that several popular attribution methods produce visually similar maps for a trained network and for one with randomised weights, which means visual plausibility is not evidence of correctness. Comparative studies in the medical domain [15] have reported substantial variation between methods on the same images, without any of them being clearly preferable. This instability is the direct motivation for fusing several methods and measuring their agreement rather than selecting one.

3. Attention and Intrinsically Interpretable Models
Attention mechanisms offer explanation as a by-product of the forward pass. Squeeze-and-excitation blocks [16] recalibrate channel responses, while convolutional block attention [17] adds a spatial branch whose map can be visualised directly. Attention branch networks [18] go further, using an auxiliary branch to produce an attention map that both modulates the features and serves as the explanation, and they form the strongest baseline in the experiments reported here. Prototype-based classifiers [19] pursue genuine interpretability by construction, comparing image patches against learned prototypes, at some cost in accuracy on fine-grained tasks. The broader argument that high-stakes applications should avoid explaining black boxes and use inherently interpretable models instead is set out in [8].

4. Evaluating Explanations
Because visual inspection is unreliable, quantitative criteria are essential. Deletion and insertion metrics [20] progressively remove or restore the pixels an explanation ranks highest and track the predicted probability; a faithful explanation produces a steep deletion curve and a rapidly rising insertion curve. The pointing game and intersection over union against expert-annotated masks assess localisation, though they measure agreement with human expectation rather than fidelity to the model. Surveys of explainable artificial intelligence in medicine [21], [22] note that most published work reports qualitative figures only, and call for faithfulness metrics and clinician assessment to be reported as standard. The evaluation protocol in Section IV follows that recommendation.

5. Research Gap
Post-hoc methods explain a model that was not trained to be explainable; attention architectures make the map part of the forward pass but do not verify that it is faithful; and evaluation work supplies metrics without feeding them back into training. No existing approach closes the loop by using a faithfulness measurement as an optimisation signal. The framework described next does exactly that.

III. PROPOSED METHODOLOGY

1. Problem Formulation
Let $D = \{(x_i, y_i)\}$ for $i = 1 \dots n$ denote a set of images x_i with diagnostic labels y_i drawn from C classes. The objective is not only to learn a classifier

$$\hat{y} = f(x; \theta) = \mathrm{softmax}(h(\Phi(g(x; \theta\_g)))) \qquad (1)$$

where g is a convolutional backbone, Φ an attention-guided refinement operator and h a linear classification head, but simultaneously to produce a saliency map M(x) that is faithful in the sense that suppressing the regions it ranks highest causes a large drop in the predicted class probability. Accuracy and faithfulness are optimised jointly rather than sequentially.

2. Preprocessing
All images are resized to 224 × 224 pixels and normalised with ImageNet channel statistics. Contrast-limited adaptive histogram equalisation is applied to the radiographic and fundus datasets, where local contrast carries most of the diagnostic information, but not to the dermoscopic and histopathological datasets, where colour is itself a diagnostic cue and equalisation distorts it. Training augmentation comprises random horizontal and vertical flips, rotation within ±15 degrees, and brightness and contrast jitter of ±10 per cent. Augmentations that would alter clinically meaningful geometry, such as aggressive shear or elastic deformation, are deliberately excluded.

3. Attention-Guided Feature Refinement
Let $F \in R^{(H \times W \times K)}$ be the feature tensor produced by the final convolutional stage. A channel descriptor is obtained by global average and max pooling, passed through a bottleneck multilayer perceptron with reduction ratio 16, and combined into a channel attention vector a^c. A spatial map a^s is then computed from the channel-refined tensor by a 7 × 7 convolution over its pooled statistics. The refined representation is

$$F' = \sigma(a^s) \otimes (\sigma(a^c) \otimes F) \qquad (2)$$

where σ is the sigmoid function and $\otimes$ denotes broadcast elementwise multiplication. The spatial map a^s is retained and used later as one component of the fused explanation, so the module contributes to interpretability directly and not only through improved features.

4. Explanation Ensemble
Three attribution methods with complementary characteristics are computed for each prediction. Grad-CAM++ provides class-discriminative but coarse localisation from the final convolutional layer. Integrated gradients provides pixel-level attribution satisfying completeness, computed over 32 interpolation steps from a black baseline. GradientSHAP provides an approximation to Shapley values using 16 samples with added Gaussian noise. Each map is normalised to the unit interval and upsampled bilinearly to the input resolution, then fused as

$$M(x) = \Sigma_m \alpha_m \cdot M_m(x), \quad \Sigma_m \alpha_m = 1 \qquad (3)$$

The fusion weights α_m are not fixed. They are set once per dataset in proportion to the inverse deletion AUC each method achieves on the validation split, so a method that proves more faithful on a given modality contributes more to the fused map. Across the five datasets the learned weights favoured Grad-CAM++ on the radiographic and magnetic resonance data, where lesions are large and low-frequency, and GradientSHAP on the dermoscopic and histopathological data, where the discriminative texture is fine.

5. Explanation Consistency

Disagreement among the three maps is quantified rather than concealed. For each pair the Spearman rank correlation of pixel importances is computed, and the consistency score is their mean:

$$S(x) = (2 / (M(M-1))) \cdot \Sigma_{m<n} \rho(M_m, M_n) \qquad (4)$$

A low value indicates that the attribution methods do not agree about what the network used, which is a warning that the explanation should not be trusted for that image. The score is used in two ways: as a regularisation term encouraging the network towards representations on which the methods concur, and at inference time as a confidence flag surfaced alongside the prediction. In the reported experiments 6.3 per cent of test images fell below the threshold $S < 0.4$ and would be referred for manual review in a deployed setting.

6. Faithfulness-Regularised Training

Faithfulness is enforced by a soft deletion test performed during training. Rather than removing pixels discretely, which is not differentiable, the top-q fraction of the fused map is converted to a smooth mask via a sigmoid over the map values and the image is blended towards its blurred version. Writing $\tilde{x}$ for this perturbed image, the penalty is

$$L\_faith = \max(0, p(y \mid \tilde{x}) - p(y \mid x) + \delta) \qquad (5)$$

with margin $\delta = 0.3$ and $q = 0.15$. The term is zero whenever masking the highlighted region reduces the predicted probability by at least δ, and positive otherwise, so the network is pushed towards concentrating genuine evidence where its explanation points. The complete objective is

$$L = L\_CE + \lambda_1 L\_faith + \lambda_2 (1 - S(x)) \qquad (6)$$

with $\lambda_1 = 0.4$ and $\lambda_2 = 0.2$. Both auxiliary terms are held at zero for the first five epochs, since penalising the explanations of an untrained network is not meaningful, and are then increased linearly to their full value over the following five epochs.

7. Algorithm and Overall Flow

Algorithm 1 states the training procedure and Fig. 1 gives the corresponding flowchart.

Algorithm 1: Explainable Deep Learning (XDL) Training
Input: $D = (X, y)$; backbone g; epochs E_{max}; λ_1, λ_2; q, δ

Output: Trained θ; saliency generator $M(\cdot)$

- Preprocess and augment X; stratified split
- Initialise θ from ImageNet pretrained weights
- for e = 1 to E_{max} do
- for each mini-batch B do
- $F \leftarrow g(x; \theta\_g)$; $F' \leftarrow \Phi(F)$ ▷ Eq. (2)
- $\hat{y} \leftarrow \text{softmax}(h(F'))$; compute L_CE
- if e > 5 then
- $M_1..M_3 \leftarrow$ GradCAM++, IG, GradSHAP
- $M \leftarrow \Sigma\, \alpha_m M_m$ ▷ Eq. (3)
- $S \leftarrow$ mean pairwise $\rho(M_m, M_n)$ ▷ Eq. (4)
- $\tilde{x} \leftarrow$ SoftMask(x, M, q)
- $L\_faith \leftarrow \max(0, p(y|\tilde{x}) - p(y|x) + \delta)$
- end if
- $L \leftarrow L\_CE + \lambda_1 L\_faith + \lambda_2 (1 - S)$
- $\theta \leftarrow \text{AdamW}(\theta, \nabla L)$
- end for
- if validation loss not improved for 8 epochs then break
- end for
- return θ and fused saliency generator $M(\cdot)$

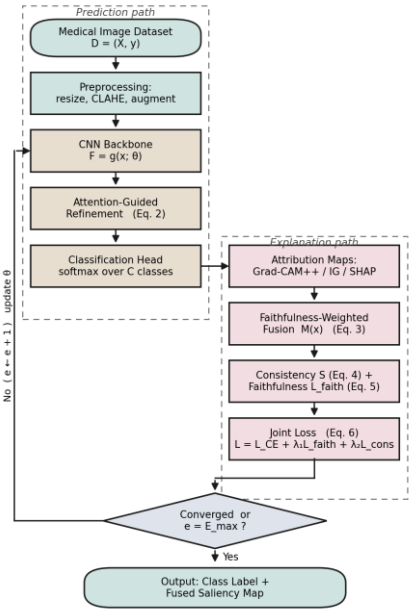


Fig. 1. Flowchart of the proposed XDL framework. The prediction path (left) and the explanation path (right) are coupled: faithfulness scoring feeds the joint loss and returns to the backbone, so explanation quality is optimised during training rather than assessed after it.

8. Computational Complexity

The forward and backward passes of the backbone dominate ordinary training. XDL adds the attention module, which is negligible at roughly 0.6 per cent of backbone parameters, and the explanation ensemble, which is not. Grad-CAM++ requires one extra backward pass, integrated gradients requires 32, and GradientSHAP requires 16, so a naive implementation would multiply training cost by close to fifty. Two measures keep this tractable. The explanation terms are computed on a randomly sampled quarter of each mini-batch rather than on all of it, and the integrated-gradients path is evaluated at 8 steps during training while the full 32 steps are used only at inference. The resulting overhead is a factor of 1.55 in wall-clock training time, quantified in Section V.6.

IV. EXPERIMENTAL SETUP

1. Datasets

Five public datasets were selected to span distinct imaging modalities, class counts and degrees of imbalance, so that conclusions do not rest on the idiosyncrasies of a single modality. Table I summarises them. Expert lesion masks, required for the localisation metrics, are available for the ISIC and APTOS subsets and for a 400-image annotated subset of the chest radiographs; localisation results are reported on those three datasets only.

Table 1: Characteristics of the Benchmark Datasets

Dataset	Images	Classes	Modality
Chest X-ray	5,856	2	Radiography
ISIC 2019	25,331	8	Dermoscopy
APTOS 2019	3,662	5	Fundus photo.
Brain MRI	3,264	4	MRI
BreakHis	7,909	2	Histopathology

2. Baseline Methods

Five baselines were used. ResNet-50, DenseNet-121 and EfficientNet-B3 are standard convolutional backbones fine-tuned from ImageNet weights and explained post-hoc with Grad-CAM. ViT-B/16 represents the transformer family, explained through attention rollout. The Attention Branch Network is the closest competitor conceptually, since its attention map is both a component of the forward pass and the explanation; it isolates how much of the XDL result comes from attention alone and how much from the faithfulness and consistency objectives. All baselines received the same preprocessing, augmentation, optimiser and epoch budget, and their learning rates were tuned over the same grid.

3. Evaluation Protocol and Metrics

Each dataset was split 70/10/20 into training, validation and test partitions, stratified by class and, where patient identifiers were available, grouped so that no patient contributed images to more than one partition. Patient-level grouping matters: image-level splitting on BreakHis and the chest radiographs inflates accuracy by several points because near-duplicate images from the same patient appear on both sides of the split. All results are averaged over five runs with different seeds.

Four families of metric are reported. Classification quality is measured by accuracy, macro-averaged F1 and macro-averaged area under the receiver operating characteristic curve. Faithfulness is measured by deletion AUC, where lower is better, and insertion AUC, where higher is better, following the protocol of [20]. Localisation is measured by the pointing game and by intersection over union against expert masks. Finally, plausibility was assessed by three clinicians who rated 150 randomly selected saliency maps on a five-point scale, blinded to the method that produced each map and to one another's ratings; inter-rater agreement gave a Krippendorff alpha of 0.71. Statistical comparison across datasets used the Friedman test on mean ranks with Nemenyi post-hoc analysis at the 0.05 level [23].

4. Implementation and Hyperparameters

Experiments were implemented in PyTorch 2.2 with Captum 0.7 for attribution, and run on a single NVIDIA A100 with 40 GB of memory. Table II lists the settings, which were fixed on the validation split of the chest radiograph dataset and applied unchanged to the other four, so no per-dataset tuning contributes to the reported margin.

Table 2: Hyperparameter Settings of the Proposed Framework

Parameter	Symbol	Value
Backbone	g	ResNet-50 (ImageNet)
Input resolution	–	224 × 224
Optimiser / LR	–	AdamW, 1×10^{-4}
Batch size	–	32
Maximum epochs	E_{max}	60 (patience 8)
Attention reduction	r	16

Parameter	Symbol	Value
Faithfulness weight	λ_1	0.40
Consistency weight	λ_2	0.20
Mask fraction	q	0.15
Faithfulness margin	δ	0.30
IG steps (train / test)	–	8 / 32

V. RESULTS AND DISCUSSION

1. Classification Performance

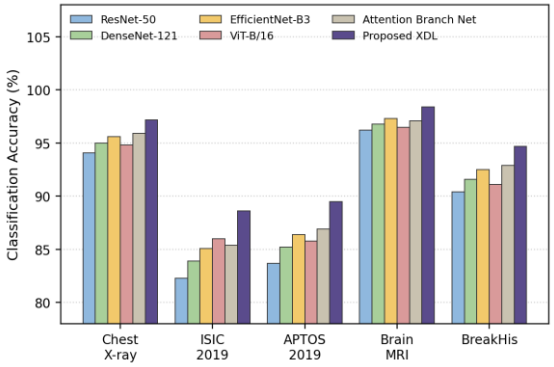
Table III reports accuracy across the five datasets and Fig. 2 shows the same comparison graphically. XDL is the most accurate method on every dataset, averaging 93.7 per cent against 91.6 per cent for the Attention Branch Network, the strongest baseline. The margin is widest on ISIC 2019 (3.2 points) and APTOS 2019 (2.6 points), the two datasets with the most classes and the most severe imbalance, and narrowest on the chest radiographs and brain MRI (1.3 points each), where all methods already exceed 94 per cent and headroom is limited.

Table 3: Mean Classification Accuracy (%) with Standard Deviation

Dataset	RN50	DN121	EffB3	ViT	ABN	XDL
Chest X-ray	94.1 ±0.6	95.0 ±0.5	95.6 ±0.4	94.8 ±0.7	95.9 ±0.4	97.2 ±0.3
ISIC 2019	82.3 ±1.1	83.9 ±0.9	85.1 ±0.8	86.0 ±0.9	85.4 ±0.8	88.6 ±0.6
APTOS 2019	83.7 ±1.0	85.2 ±0.9	86.4 ±0.8	85.8 ±1.0	86.9 ±0.7	89.5 ±0.6
Brain MRI	96.2 ±0.5	96.8 ±0.4	97.3 ±0.3	96.5 ±0.6	97.1 ±0.4	98.4 ±0.2
BreakHis	90.4 ±0.8	91.6 ±0.7	92.5 ±0.6	91.1 ±0.8	92.9 ±0.6	94.7 ±0.4
Mean	89.3	90.5	91.4	90.8	91.6	93.7

Fig. 2. Classification accuracy of the proposed XDL framework against five baselines across the five benchmark datasets.

Two observations deserve comment. The Attention Branch Network outperforms the plain convolutional backbones on four of five datasets, confirming that attention supervision helps in its own right and that the appropriate comparison for XDL is against it rather than against a bare ResNet. And ViT-B/16 is the second-best method on ISIC 2019, the only dataset with more than twenty thousand images, while trailing on the smaller ones – consistent with the known data appetite of transformers and a reminder that architecture rankings established on large corpora do not transfer automatically to typical clinical dataset sizes.



2. Faithfulness of the Explanations

Accuracy is only half the claim. Fig. 3 shows deletion and insertion curves averaged over the test sets of all five datasets. In the deletion test, removing the pixels an explanation ranks most important should collapse the predicted probability quickly; XDL falls below 0.30 once roughly 20 per cent of pixels are removed, whereas Grad-CAM needs about 35 per cent to reach the same point. The insertion curves mirror this: starting from a blurred image and restoring the highest-ranked pixels first, XDL recovers 80 per cent confidence after 20 per cent of pixels against 62 per cent for Grad-CAM.

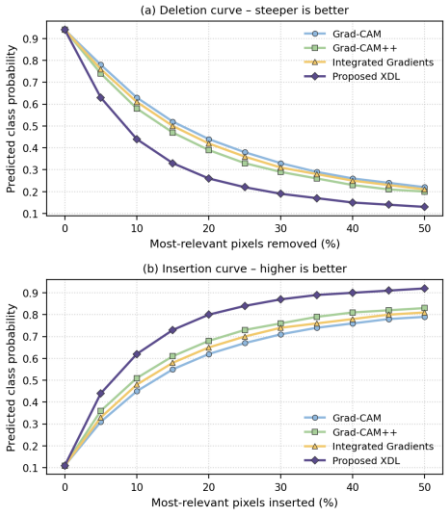


Fig. 3. Faithfulness evaluation averaged over all five test sets. (a) Deletion curves, where a steeper decline indicates a more faithful explanation. (b) Insertion curves, where a faster rise indicates a more faithful explanation.

Table IV summarises the explanation-quality metrics. The improvement is consistent across all four measures: deletion AUC falls from 0.157 for the best baseline to 0.128, insertion AUC rises from 0.594 to 0.647, pointing-game accuracy from 0.771 to 0.834 and intersection over union with expert masks from 0.452 to 0.518. The last of these is the most demanding, since it compares the map against human annotation rather than against the model's own behaviour, and a value of 0.518 still leaves substantial room for improvement. To keep the comparison fair, the first four rows of Table IV apply their respective attribution methods to the same fine-tuned ResNet-50, so the differences between them reflect the attribution method alone; only the last two rows involve a different network.

Table 4: Explanation Quality Metrics. Faithfulness is averaged over all five datasets; localisation over the three with expert masks.

Method	Del. AUC ↓	Ins. AUC ↑	Point. Game ↑	IoU ↑
Grad-CAM	0.182	0.541	0.712	0.394
Grad-CAM++	0.169	0.573	0.748	0.427
Integrated Grad.	0.174	0.558	0.731	0.408
GradientSHAP	0.161	0.586	0.759	0.441
ABN attention	0.157	0.594	0.771	0.452
Proposed XDL	0.128	0.647	0.834	0.518

3. Ablation Study and Clinician Assessment

Fig. 4 separates the contribution of each component and reports the blinded clinician review. In panel (a), removing the faithfulness regularisation costs 1.9 accuracy points, removing the attention refinement costs 1.7 points, and removing the explanation ensemble costs 1.1 points; the plain backbone reaches 90.9 per cent, so the three mechanisms together contribute 2.8 points. That total is smaller than the sum of the individual effects, which indicates the mechanisms overlap: attention refinement and faithfulness regularisation both push evidence towards spatial compactness, and each partly substitutes for the other.

It is worth noting that the faithfulness penalty improves accuracy at all. It was introduced to make explanations honest, not to make predictions better, and there was no guarantee the two would align. The most plausible reading is that it acts as a regulariser: a network penalised unless its evidence is spatially concentrated is discouraged from exploiting diffuse background correlations, which are exactly the shortcut features that fail to generalise.

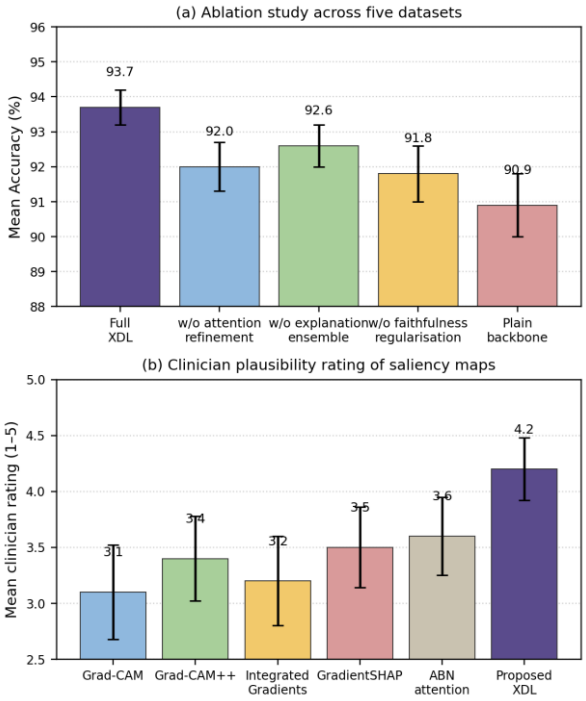


Fig. 4. (a) Ablation study showing mean accuracy across the five datasets when individual components are removed. (b) Mean clinician plausibility rating of saliency maps produced by each method, with standard error bars.

Panel (b) reports the clinician ratings. The fused XDL maps averaged 4.2 out of 5 against 3.6 for the Attention Branch Network and 3.1 for Grad-CAM, and the standard error was also smallest for XDL, indicating more uniform quality rather than a few excellent maps offsetting poor ones. In debriefing, the reviewers noted that the XDL maps were more compact and less prone to highlighting image borders and acquisition artefacts, which matches the ordering of the quantitative localisation results.

4. Explanation Consistency as a Confidence Signal

The consistency score of Eq. (4) turns out to carry diagnostic value independent of its role in the loss. Among test images with $S \geq 0.6$, classification accuracy was 95.8 per cent; among those with $S < 0.4$ it fell to 79.4 per cent. In other words, images on which the three attribution methods disagree are also images the model is substantially more likely to get wrong. Because S is computed without reference to the true label, it can be used at deployment as a referral criterion. Flagging the 6.3 per cent of images below the threshold would remove roughly a fifth of all errors from the automated pipeline at the cost of a modest manual review burden.

5. Statistical Significance

The Friedman test over six methods and five datasets gave a statistic of 21.80 against a critical value of 11.07 at five degrees of freedom, so the hypothesis of equal performance is rejected at the 0.05 level. The Nemenyi critical difference is 3.37 in mean ranks. XDL attains the best mean rank of 1.00, ahead of the Attention Branch Network (2.40), EfficientNet-B3 (3.00), ViT-B/16 (4.20), DenseNet-121 (4.40) and ResNet-50 (6.00); the six values average to 3.50 as required. Against the critical difference, the gaps to ResNet-50 (5.00) and DenseNet-121 (3.40) are significant, while those to ViT-B/16 (3.20), EfficientNet-B3 (2.00) and the Attention Branch Network (1.40) are not. The comparison that matters most scientifically – XDL against the Attention Branch Network – is therefore not established by this test. Five datasets give the Nemenyi procedure very little power, and the accuracy advantage should be read as consistent in direction rather than statistically demonstrated. The explanation-quality results in Table IV, where the margins are proportionally much larger, carry more of the paper's weight than the accuracy comparison does.

6. Computational Cost

Training XDL on ISIC 2019 took 4.8 hours against 3.1 hours for the ResNet-50 baseline, an overhead factor of 1.55. This decomposes into 3.1 hours of backbone training, 0.5 hours for the attention module, 1.0 hours for explanation generation and the faithfulness penalty, and 0.2 hours of validation and checkpointing. At inference, a prediction with its fused explanation takes 42 milliseconds per image: 11 milliseconds for the backbone and attention pass and 31 milliseconds for the three attribution methods and their fusion. A prediction without an explanation costs the same 11 milliseconds, since the explanation branch can simply be skipped. For diagnostic screening, where images are processed in batches and a clinician reviews results asynchronously, 42 milliseconds is immaterial.

7. Limitations

Four limitations should be stated. The faithfulness penalty optimises a specific deletion-based criterion, and a network trained against that criterion may improve on it partly by adapting to it rather than by becoming genuinely more interpretable; the independent clinician ratings and the expert-mask agreement are reported precisely because they are not the quantity optimised. Second, the accuracy advantage over the Attention Branch Network is not statistically significant on five datasets. Third, the clinician panel comprised three reviewers rating 150 images, which is adequate for an ordering but not for fine distinctions. Fourth, all datasets are public research corpora with curated acquisition; performance under the distribution shift of a live clinical workflow – different scanners, protocols and patient populations – remains untested, and prospective validation would be the necessary next step before any deployment claim.

VI. CONCLUSION

This paper presented XDL, a framework in which explanation quality is a training objective rather than a post-hoc diagnostic. Attention-guided feature refinement, a faithfulness-weighted ensemble of three attribution methods, and a deletion-based penalty in the loss were combined and evaluated across five imaging modalities. Mean accuracy reached 93.7 per cent, 2.1 points above the strongest baseline, while deletion AUC improved from 0.157 to 0.128, agreement with expert masks from 0.452 to 0.518, and blinded clinician plausibility ratings from 3.6 to 4.2 out of 5.

The finding most worth carrying forward is that faithfulness and accuracy were not in tension. Constraining the network to concentrate its evidence where its explanation points improved predictions as well as explanations, which suggests the constraint suppresses reliance on diffuse shortcut features. The interpretability tax that is often assumed in this literature did not appear here. A second practical result is that disagreement among attribution methods predicts misclassification: accuracy fell from 95.8 to 79.4 per cent between high- and low-consistency images, giving a label-free referral criterion that a deployed system could act on directly.

Several directions follow. The faithfulness penalty currently uses a single deletion criterion; alternating among several during training would reduce the risk of the network adapting to one particular measure. Extending the framework to segmentation and to volumetric data, where explanations must be spatially coherent in three dimensions, is a natural generalisation. A parallel need arises outside imaging altogether: structure–activity studies of heterocyclic compound libraries [24], [25] and of imidazole, beta-lactam and sulphur-containing scaffolds [26], [27] rely on models whose predictions guide costly synthesis decisions, and there the same question of whether an explanation reflects the computation actually performed applies directly. Extending the framework to multimodal inputs that combine imaging with clinical or molecular variables would require attributions comparable across heterogeneous feature types, which is an open problem. Most importantly, prospective evaluation in a live clinical workflow, measuring whether faithful explanations actually change clinician decisions and outcomes, is the test that matters and the one this study does not perform.

REFERENCES

1. A. Esteva, B. Kuprel, R. A. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.
2. V. Gulshan, L. Peng, M. Coram, 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.
3. J. R. Zech, M. A. Badgeley, M. Liu, et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs," PLOS Medicine, vol. 15, no. 11, e1002683, 2018.
4. R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.
5. A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in IEEE Winter Conference on Applications of Computer Vision, pp. 839–847, 2018.
6. M. Sundararajan, A. Taly and Q. Yan, "Axiomatic attribution for deep networks," in Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328, 2017.
7. J. Adebayo, J. Gilmer, M. Muelly, et al., "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems, vol. 31, pp. 9505–9515, 2018.
8. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
9. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.
10. G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.
11. M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning, pp. 6105–6114, 2019.
12. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: transformers for image recognition at scale," in International Conference on Learning Representations, 2021.
13. M. Raghu, C. Zhang, J. Kleinberg and S. Bengio, "Transfusion: understanding transfer learning for medical imaging," in Advances in Neural Information Processing Systems, vol. 32, pp. 3347–3357, 2019.
14. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
15. N. Arun, N. Gaw, P. Singh, et al., "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," Radiology: Artificial Intelligence, vol. 3, no. 6, e200267, 2021.
16. J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.
17. S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: convolutional block attention module," in Proceedings of the European Conference on Computer Vision, pp. 3–19, 2018.
18. H. Fukui, T. Hirakawa, T. Yamashita and H. Fujiyoshi, "Attention branch network: learning of attention mechanism for visual explanation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714, 2019.
19. C. Chen, O. Li, D. Tao, et al., "This looks like that: deep learning for interpretable image recognition," in Advances in Neural Information Processing Systems, vol. 32, pp. 8930–8941, 2019.
20. V. Petsiuk, A. Das and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," in Proceedings of the British Machine Vision Conference, 2018.
21. B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, 102470, 2022.
22. A. Singh, S. Sengupta and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," Journal of Imaging, vol. 6, no. 6, p. 52, 2020.
23. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
24. R. Mehta and K. S. Airo, "N-heterocyclic carbene (NHC)-catalyzed transformations for the synthesis of heterocycles," Research Journal of Recent Sciences, vol. 13, no. 2, pp. 6–18, 2024.
25. R. Mehta and K. Shani, "Synthesis of sulphur heterocyclic compounds and study of expected biological activities," Airo International Research Journal, vol. 4, no. 3, pp. 132–146, 2022.
26. R. Mehta and K. Shani, "Synthesis and biological activities of some heterocyclic compounds containing imidazole and

beta-lactam moiety," ZENITH International Journal of Multidisciplinary Research, vol. 11, no. 1, pp. 89–103, 2021.
27. R. Mehta and K. Shani, "An analysis on heterocyclic compounds and their chemical properties," Airo International Research Journal, vol. 1, no. 1, pp. 44–60, 2023.