ExBale: Quantitative Evaluation of Explainability Alignment in Ovarian Ultrasound Imaging

Mina Kamel[1], Mahmoud Nour[1], Mina A. Awad[1], Mohamed A. Adelhamed[2], and Sahar Selim[1*]

[1] Center for Informatics Science (CIS),
School of Information Technology and Computer Science, Nile University,
26th of July Corridor, Sheikh Zayed City, Giza, 12588, Egypt
{mi.mamdouh, ma.nour, mi.adly, sselim}@nu.edu.eg
[2] School of Engineering, Nile University,
26th of July Corridor, Sheikh Zayed City, Giza, 12588, Egypt
m.alaa2155@nu.edu.eg
*Corresponding author

Abstract. The interpretation of deep learning decisions in ovarian tumour diagnosis is a fundamental requirement for clinical adoption and trust. While convolutional neural networks achieve high accuracy in ultrasound classification, their decision making processes remain opaque, necessitating post classification explainability. ExBale is proposed as a composite alignment metric that quantifies how well explainability heatmaps correspond to expert annotated tumour segmentation masks by combining segmentation coverage, CAM containment, and weighted activation intensity via the geometric mean. The framework is evaluated across five backbone architectures and six explainability methods using the MMOTU dataset. Activation based class activation mapping methods demonstrate statistical superiority over gradient based methods (Wilcoxon signed rank, $p < 10^{-42}$ across all backbones), with Swin Transformer and Eigen CAM achieving the highest global alignment score of 47.85%, placing it at 23.1% of the range between a random noise baseline (32.22%) and a perfect alignment ceiling (100%). Spatial alignment metrics serve as a secondary evaluation signal for explainability quality; their utility as a standalone triage criterion is limited by modest Youden J indices (range 0.001 to 0.138) and the absence of clinical outcome validation.

Keywords: ovarian tumour classification · explainability evaluation · CAM alignment · ultrasound imaging · ExBale · trustworthy medical AI

1 Introduction

Ovarian cancer is a highly lethal malignancy requiring early, accurate ultrasound classification for effective treatment planning. While deep learning architectures like DenseNet, ResNet, and YOLO achieve high diagnostic accuracy on datasets like MMOTU [3, 1, 2], their opaque decision-making restricts clinical adoption.

2D ultrasound suffers from noise and low contrast, meaning confident predictions are clinically questionable unless models demonstrably focus on actual tumour regions rather than irrelevant background. Recent medical imaging studies highlight the critical need for quantitative validation of model attention [19], especially in ultrasound where qualitative inspection is insufficient for trust calibration. Post-classification explainability methods, including activation-based CAM and gradient-based saliency maps [9, 13], visualize features driving model predictions. Although Grad-CAM is widely applied in clinical imaging, persistent gaps in quantitative validation remain [20]. Furthermore, recent CT classification analyses reveal Grad-CAM's localisation reliability degrades substantially in vision transformers due to non-local attention, proving that method selection is architecture-dependent and visual inspection cannot distinguish genuine alignment from incidental overlaps [21]. Quantitative alignment is typically evaluated using segmentation coverage, CAM containment, and weighted intensity inside the region of interest. While these metrics individually compare model focus against ground truth, no unified, scale-calibrated composite metric exists for ovarian ultrasound. Current literature predominantly relies on qualitative XAI inspection, usually evaluating a single architecture-method pair. Consequently, it remains unknown whether alignment quality stems from the architecture, the XAI method, or their interaction, and the statistical relationship between spatial alignment and prediction reliability across diverse architectures remains unexplored. This work addresses these gaps through three contributions. First, ExBale is proposed as a unified alignment metric quantifying spatial correspondence between model attention and expert masks by combining coverage, containment, and intensity via the geometric mean. Second, a systematic, statistically validated evaluation (using paired Wilcoxon and Friedman tests) compares five backbone architectures and six XAI methods to identify optimal techniques for ovarian ultrasound. Third, the statistical relationship between attention alignment and prediction reliability is analysed via Youden J optimisation, including an assessment of the limitations of alignment-based reliability screening. 2D ultrasound is particularly relevant to resource-constrained clinical settings, as it requires no advanced imaging infrastructure, is widely available in low-income and rural healthcare facilities, and carries low per-examination cost. Deploying trustworthy AI in these settings demands explainability frameworks that are both quantitatively rigorous and operationally practical, motivating the evaluation presented here.

2 Methodology

2.1 Dataset and Pipeline Overview

The OTU 2D subset of the Multi Modality Ovarian Tumour Ultrasound (MMOTU) dataset is utilised, containing 1469 two dimensional images across eight ovarian tumour classes [3]. Each image is paired with a binary pixel level segmentation mask delineating the tumour region of interest. Patient level stratified

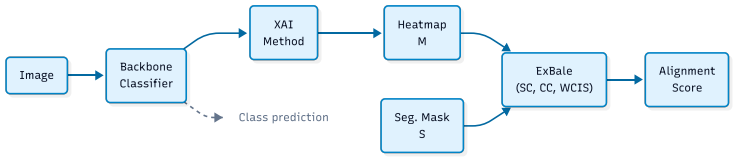

Fig. 1: ExBale evaluation pipeline: A backbone classifier and an XAI method process an ultrasound image. ExBale quantifies spatial alignment between the generated heatmap and the expert segmentation mask using Segmentation Coverage (SC), CAM Containment (CC), and Weighted CAM Intensity in Segmentation (WCIS).

splitting is implemented using `StratifiedGroupKFold` to prevent data leakage, yielding 940 training images, 235 validation images, and 294 test images. This approach discards the predefined dataset splits, which lack a held-out test partition and do not guarantee patient level separation, in favour of principled patient-independent evaluation. Figure 1 illustrates the ExBale evaluation pipeline, where explainability heatmaps are compared with expert segmentation masks to compute alignment scores.

The pipeline evaluates five backbones (DenseNet 121, ResNet 50, EfficientNet B3, MobileNet V3 Large, Swin Transformer) connected to a unified classification head containing a 512 unit projection layer, batch normalisation, ReLU activation, 50% dropout, and an 8 class output logit layer.

2.2 Training Protocol

Training utilises the AdamW optimiser with a backbone learning rate of 1×10^{-4} and head learning rate of 1×10^{-3} over a 60 epoch schedule, governed by a 5 epoch linear warmup and cosine annealing. The backbone is initially frozen for five epochs. The loss function is class weighted cross entropy with 10% label smoothing to address class imbalance. Optimisation incorporates mixed precision training, gradient clipping (max norm 1.0), and Stochastic Weight Averaging activated at epoch 45. Data augmentation comprises RandAugment, spatial flips, and post freeze MixUp and CutMix. Spatial flips and rotations are included during training as data regularisation to reduce overfitting on the small dataset; however, these transforms are anatomically inappropriate at inference time for ultrasound imaging, where probe orientation is fixed. This explains the performance degradation observed during test time augmentation evaluation, in which these same transforms produced unrealistic acquisition geometries and degraded ensemble accuracy.

2.3 Explainability Methods

Explainability heatmaps are generated using six distinct methods comprising Grad CAM [9], Grad CAM plus plus [10], Score CAM [11], Eigen CAM [12], Saliency [13], and Integrated Gradients [14]. Target layers for CAM computation are selected as the final convolutional or feature blocks for each architecture.

For Swin Transformer, the target layer is the layer normalisation module (`norm2`) of the final block within the last feature stage (`features[-1][-1].norm2`). Because Swin Transformer produces token sequences of shape (B, L, C) rather than spatial feature maps, a reshape transform is applied: the sequence is reshaped to (B, H, W, C) where $H{=}W{=}\sqrt{L}$, then permuted to (B, C, H, W) before CAM computation. Integrated gradients are computed using a zero tensor baseline and 50 integration steps. Faithfulness is evaluated through insertion and deletion protocols using the RISE framework [15].

2.4 ExBale Metric Formulation

ExBale is introduced as a composite alignment metric that quantifies the spatial correspondence between an explainability heatmap and the expert annotated tumour segmentation mask. Unlike a simple average of the three components, the geometric mean formulation ensures that all three dimensions of alignment must be simultaneously satisfied: a near zero value in any single component collapses the overall score, preventing an explanation that achieves broad coverage but poor spatial precision from receiving an inflated score. The arithmetic mean does not satisfy this property: a method achieving SC =1.0, CC =0, WCIS =1.0 would receive an arithmetic mean of 0.67, falsely suggesting useful alignment despite zero containment. The harmonic mean is more conservative but collapses to zero under the same condition without distinguishing near-zero from exactly-zero components. The geometric mean uniquely balances sensitivity to low-scoring components against robustness to component-level noise, making it appropriate for a multi-criterion spatial alignment task where each component captures a non-redundant property of heatmap quality. The metric is defined in Equation 1.

$$\mathrm{ExBale}(M, S, \tau) = \Big(\mathrm{SC}(M_\tau, S) \cdot \mathrm{CC}(M_\tau, S) \cdot \widetilde{\mathrm{WCIS}}(M, S)\Big)^{1/3} \tag{1}$$

In this formulation, M represents the continuous heatmap in $[0, 1]$, S denotes the binary segmentation mask, and M_τ is the binarised heatmap at threshold $\tau{=}0.5$. The segmentation coverage is defined in Equation 2.

$$\mathrm{SC}(M_\tau, S) = \frac{|M_\tau \cap S|}{|S|} \tag{2}$$

A high segmentation coverage indicates that the explanation highlights most of the tumour region. The CAM containment is defined in Equation 3.

$$\mathrm{CC}(M_\tau, S) = \frac{|M_\tau \cap S|}{|M_\tau|} \tag{3}$$

A high CAM containment indicates that the explanation is focused on the tumour rather than background tissue. The weighted CAM intensity in segmentation is defined in Equation 4.

$$\mathrm{WCIS}(M, S) = \frac{1}{|S|} \sum_{(i,j):S_{ij}=1} M_{ij} \tag{4}$$

Table 1: Classification performance on the held out test set (n=294).

Model	Top 1 Acc	Top 2 Acc	Macro F1	Bal Acc
DenseNet 121	74.47%	84.68%	70.18%	70.94%
ResNet 50	77.87%	87.23%	70.18%	69.60%
EfficientNet B3	74.04%	83.83%	68.60%	69.29%
MobileNet V3 Large	77.45%	88.09%	71.18%	72.40%
Swin T	77.87%	88.09%	70.10%	69.76%
Ensemble	78.57%	—	74.36%	75.59%

Table 2: Per class performance of the ensemble on the held out test set.

Class	Precision	Recall	F1	AUC ROC
Chocolate	83.58%	82.35%	82.96%	95.96%
Serous	71.43%	83.33%	76.92%	96.00%
Teratoma	94.55%	77.61%	85.25%	97.23%
Theca cell	78.95%	68.18%	73.17%	92.83%
Clear cell	50.00%	43.75%	46.67%	92.38%
Dermoid	75.47%	86.96%	80.81%	95.51%
Simple cyst	66.67%	62.50%	64.52%	85.86%
Normal ovary	73.33%	96.42%	83.30%	99.33%

The WCIS component is globally normalised using dataset level statistics computed across all test images and all XAI methods prior to evaluation: $\widetilde{\text{WCIS}} = (\text{WCIS} - \text{WCIS}_{\text{min}})/(\text{WCIS}_{\text{max}} - \text{WCIS}_{\text{min}} + \varepsilon)$, where ε=10^{-8} prevents division by zero, WCIS_{min}=0.0, and WCIS_{max}=1.0, as heatmap values are normalised to $[0, 1]$ per image before metric computation. The normalised value is clipped to $[0, 1]$. To provide an interpretable scale, calibration anchors are established across 100 randomly sampled test masks: a uniform random noise heatmap yields a mean ExBale of 32.22% (chance baseline) and a perfect mask heatmap yields 100% (perfect ceiling). Scores above 32.22% represent meaningful spatial correspondence with the expert annotated tumour region.

3 Results

3.1 Classification Accuracy and Ensemble Performance

Classification performance is evaluated on the held out test set containing 294 images. Single backbone results are summarised in Table 1. The large performance differences between backbones are consistent with known receptive field and attention mechanism differences: Swin Transformer's global window attention captures long range dependencies, while EfficientNet B3 and DenseNet 121 rely on local convolutions less effective on noisy ultrasound backgrounds. A uniform ensemble achieves macro F1 of 74.36% and balanced accuracy of 75.59%, exceeding all individual backbones and reflecting improved generalisation. Per class analysis (Table 2) shows the highest precision for teratoma (94.55%) and the lowest F1 for clear cell carcinoma (46.67%), consistent with its low prevalence (16 images).

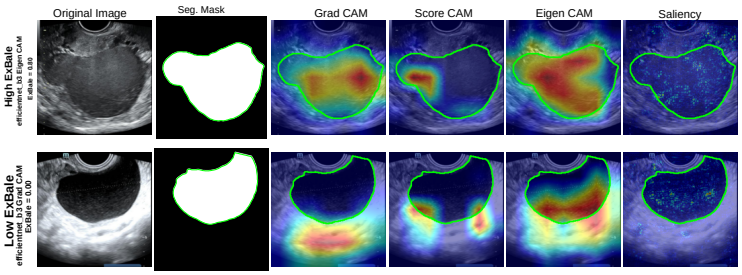

Fig. 2: Qualitative examples of high (top) and low (bottom) ExBale alignment cases. Activation-based CAM methods successfully localize within the tumour boundary (green contour), whereas Saliency yields diffuse, non-localised responses.

Table 3: Mean ExBale scores (%) at τ=0.5. All CAM vs gradient comparisons are statistically significant (paired Wilcoxon signed rank test per test image, n=294, $p<10^{-42}$ for all backbones). Random baseline: 32.22%. Perfect ceiling: 100%.

XAI Method	DenseNet 121	ResNet 50	EfficientNet B3	MobileNet V3	Swin T
Grad CAM	24.22%	28.65%	30.63%	30.07%	23.60%
Grad CAM++	35.85%	24.12%	23.56%	19.17%	35.53%
Score CAM	**42.96%**	10.94%	8.59%	7.35%	33.81%
Eigen CAM	25.55%	15.68%	13.90%	11.61%	**47.85%**
Saliency	4.13%	3.82%	3.73%	4.61%	5.34%
Integr Grads	3.96%	3.97%	3.44%	4.48%	4.12%
Random baseline			32.22%		
Perfect ceiling			100%		

3.2 ExBale Alignment Results

Explainability alignment is quantified via ExBale (τ=0.5). Per-image paired Wilcoxon signed-rank tests (n=294) confirm CAM-based methods significantly outperform gradient-based methods across all backbones (all $p<10^{-42}$). As Table 3 shows, gradient methods (Saliency, Integrated Gradients) fall below the 32.22% random baseline, showing negligible spatial overlap with tumour regions. Conversely, activation-based methods successfully concentrate within tumour boundaries (Figure 2). A Friedman test confirmed significant alignment variance across backbones (χ^2=46.07, p=2.38$\times10^{-9}$). Bonferroni-corrected post-hoc analysis shows Swin Transformer and DenseNet 121 achieve the highest alignment (p=1.0 between them), significantly outperforming EfficientNet B3 and MobileNet V3 Large ($p<0.001$). The global best alignment is achieved by Swin Transformer with Eigen CAM (47.85% ExBale, covering 23.1% of the range above chance), while DenseNet 121 peaks with Score CAM (42.96%). This variation occurs because CAM based methods operate directly on internal feature maps, so alignment quality depends on the spatial resolution, channel count, and activation distribution of the chosen target layer, properties that differ substantially across the five backbones evaluated.

Table 4: Faithfulness evaluation across backbone architectures.

Backbone	Insertion AUC	Deletion AUC
DenseNet 121	0.3632 ± 0.1768	**0.2374 ± 0.1316**
ResNet 50	0.3557 ± 0.1564	0.2540 ± 0.1505
EfficientNet B3	0.3797 ± 0.1771	0.2725 ± 0.1645
MobileNet V3	0.3377 ± 0.1751	0.2614 ± 0.1535
Swin T	**0.4982 ± 0.1788**	0.2381 ± 0.1588

3.3 Faithfulness and Clinical Screening

Faithfulness evaluation results are presented in Table 4. Swin Transformer achieves the highest average insertion AUC of 0.4982 (± 0.1788), indicating that model confidence strongly recovers when salient features are introduced. DenseNet 121 achieves the lowest average deletion AUC of 0.2374 (± 0.1316), closely followed by Swin Transformer at 0.2381, demonstrating rapid confidence degradation when salient regions are removed. Screening analysis determines optimal thresholds for flagging unreliable predictions using Youden J optimisation [17] over a systematic sweep of SC thresholds ($\{0.10, 0.15, 0.20, 0.25, 0.30\}$) and CC thresholds ($\{0.30, 0.40, 0.50, 0.60, 0.70\}$). Swin Transformer provides a specificity of 72.97% at a low flag rate of 28.4% using the optimal pair (SC<0.20, CC<0.60). The systematic sweep revealed a clear gradient: screening reliability strictly improves as the CC threshold increases and the SC threshold decreases, converging on the optimal pair.

4 Discussion

The results establish the statistical superiority of activation-based CAM methods over gradient-based techniques for ovarian ultrasound explainability ($p<10^{-42}$). Gradient methods fall completely below the random alignment baseline (32.22%) across all backbones, indicating poor spatial correspondence with expert-annotated tumour regions. Interestingly, while Grad-CAM alignment acts as a statistically significant reliability indicator for two backbones (EfficientNet-B3 and MobileNet-V3 Large, $p<0.01$), Eigen CAM despite achieving the highest absolute alignment score does not discriminate between correct and incorrect predictions ($p=1.000$ for Swin Transformer). This divergence demonstrates that localisation quality and prediction reliability signalling serve distinct clinical purposes and should be evaluated independently. An important architectural insight emerges from the Swin Transformer results. Although Grad-CAM yields poor spatial alignment for this model (23.60% ExBale), its high insertion AUC (0.4982) confirms the functional relevance of these regions. Conversely, Eigen CAM optimally balances spatial alignment (47.85% ExBale) and functional faithfulness. We hypothesize that the shifted-window self-attention mechanism diffuses gradient flow across non-contiguous tokens, making gradient-weighted maps spatially diffuse. Eigen CAM bypasses this limitation by extracting the principal component of the activation tensor directly. This aligns with prior interpretability research [22,

21] and suggests that gradient-free CAM variants are better suited for evaluating spatial alignment in transformer-based medical imaging models.

4.1 Limitations and Future Work

The modest Youden J values across all backbones (range 0.001 to 0.138) indicate that ExBale based thresholding produces a weak and variable triage signal. Spatial alignment is shown here to be a statistically significant but practically limited predictor of classification correctness, and ExBale based flagging is appropriate only as a secondary evaluation aid rather than a primary diagnostic criterion. All assertions of clinical utility in this work are therefore bounded by this empirical observation. All experiments are conducted on the single MMOTU dataset, and external validation on independent ultrasound cohorts is required before clinical deployment. ExBale computation requires expert pixel level segmentation masks, which may not be available in routine clinical settings, limiting applicability to annotated research datasets. Interpretability and trust are ultimately human-centric concepts that require validation through clinician engagement. The present work does not include a user study and this constitutes a key limitation. Future work should evaluate sonographer error rates and trust calibration under ExBale based flags through a prospective study in which sonographers classify ultrasound images with and without ExBale alignment feedback, measuring diagnostic agreement and confidence calibration as primary outcomes. ExBale stability under annotation disagreement has not been evaluated in this work. Future studies should assess ExBale score variance across multiple annotators for the same image to determine the metric's sensitivity to boundary definition uncertainty, particularly for morphologically heterogeneous classes such as Dermoid cyst.

5 Conclusion

ExBale is introduced as a composite alignment metric for quantitative evaluation of post classification explainability in ovarian ultrasound imaging. Activation based CAM methods are demonstrated to be statistically superior to gradient based methods across five backbone architectures (paired Wilcoxon signed rank test per test image, n=294, $p<10^{-42}$). Swin Transformer combined with Eigen CAM achieves the highest alignment with expert annotated tumour masks (47.85%, placed at 23.1% of the achievable range above the random baseline of 32.22%). The relationship between attention alignment and prediction correctness is established for Grad CAM, although absolute alignment quality does not universally correlate with reliability. The Youden J analysis confirms that ExBale based flagging serves as a structured secondary evaluation signal rather than a standalone clinical triage criterion, and that clinical deployment requires a prospective user study and external dataset validation. These findings provide a structured framework for validating deep learning decisions in trust critical medical applications and highlight the importance of selecting architecture specific explainability methods for reliable AI assisted diagnosis.

References

1. F. Christiansen, E.L. Epstein, E. Smedberg, M. Akerlund, K. Smith, E. Epstein, Ultrasound image analysis using deep neural networks for discriminating between benign and malignant ovarian tumors: Comparison with expert subjective assessment, Ultrasound in Obstetrics and Gynecology, vol. 57, no. 1, pp. 155 163, 2021.
2. W. Xie, W. Lin, P. Li, H. Lai, Z. Wang, P. Liu, Y. Huang, Y. Liu, L. Tang, G. Lyu, Developing a deep learning model for predicting ovarian cancer in Ovarian Adnexal Reporting and Data System Ultrasound (O RADS US) category 4 lesions: A multicenter study, Journal of Cancer Research and Clinical Oncology, vol. 150, no. 346, 2024.
3. Q. Zhao, S. Lyu, W. Bai, L. Cai, B. Liu, G. Cheng, M. Wu, X. Sang, M. Yang, L. Chen, MMOTU: A multi modality ovarian tumour ultrasound image dataset for unsupervised cross domain semantic segmentation, arXiv:2207.06799, 2022.
4. G. Huang et al., Densely connected convolutional networks, CVPR, 2017.
5. K. He et al., Deep residual learning for image recognition, CVPR, 2016.
6. M. Tan et al., EfficientNet: Rethinking model scaling for convolutional neural networks, ICML, 2019.
7. A. Howard et al., Searching for MobileNetV3, ICCV, 2019.
8. Z. Liu et al., Swin Transformer: Hierarchical vision transformer using shifted windows, ICCV, 2021.
9. R. Selvaraju et al., Grad-CAM: Visual explanations from deep networks via gradient based localization, ICCV, 2017.
10. A. Chattopadhyay et al., Grad-CAM++: Generalized gradient based visual explanations for deep convolutional networks, WACV, 2018.
11. H. Wang et al., Score-CAM: Score weighted visual explanations for convolutional neural networks, CVPR Workshops, 2020.
12. M.B. Muhammad, M. Yeasin, Eigen-CAM: Class activation map using principal components, arXiv:2008.00299, 2020.
13. K. Simonyan et al., Deep inside convolutional networks: Visualizing image classification models and saliency maps, ICLR, 2014.
14. M. Sundararajan et al., Axiomatic attribution for deep networks, ICML, 2017.
15. V. Petsiuk et al., RISE: Randomized input sampling for explanation of black box models, BMVC, 2018.
16. F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin, 1945.
17. W. Youden, Index for rating diagnostic tests, Cancer, 1950.
18. J. Tukey, Comparing individual means in the analysis of variance, Biometrics, 1949.
19. J.M. Metsch, A.C. Hauschild, BenchXAI: Comprehensive benchmarking of post hoc explainable AI methods on multi modal biomedical data, Computers in Biology and Medicine, vol. 191, 110124, 2025.
20. M. Haupt, M.H. Maurer, R.P. Thomas, Explainable artificial intelligence in radiological cardiovascular imaging: A systematic review, Diagnostics, vol. 15, no. 11, 1399, 2025.
21. T. Panboonyuen, Seeing isn't always believing: Analysis of Grad-CAM faithfulness and localization reliability in lung cancer CT classification, arXiv:2601.12826, 2026.
22. H. Chefer, S. Gur, L. Wolf, Transformer Interpretability Beyond Attention Visualization, Advances in Neural Information Processing Systems (NeurIPS), 2021.