Politecnico di Milano

Applied AI in Biomedicine – Final Assignment

Deep Learning for Breast Ultrasound

A Comparative Study of Diagnostic Pipelines

Filippo Saccomano 10771424 | **Daniele Uras** 11094104 | **Gabriele Carta** 11097253 | **Jonny A. Marques** 10764433

Abstract

This work compares two deep learning diagnostic pipelines for breast ultrasound analysis, combining image-level classification and lesion segmentation in different task orders. Pipeline A follows a classification-first strategy, where a multi-class classifier predicts the diagnostic label before lesion segmentation, while Pipeline B adopts a segmentation-first approach and performs lesion-centred classification on extracted regions of interest. Both pipelines are evaluated on a patient-wise split of a public breast ultrasound dataset, using balanced accuracy for classification and Dice-based metrics for segmentation. Experimental results show that Pipeline B achieves more stable and accurate lesion segmentation, while Pipeline A improves benign versus malignant discrimination when lesion localisation is accurate. The analysis highlights distinct error propagation behaviours: classification errors dominate Pipeline A failures, whereas segmentation inaccuracies limit Pipeline B performance. These findings indicate that task ordering is a key design choice in ultrasound-based diagnostic systems, influencing robustness, interpretability, and clinical suitability.

1 Introduction

Breast cancer represents one of the leading causes of cancer-related morbidity and mortality among women worldwide. Breast ultrasound imaging is widely adopted in clinical practice due to its non-invasive nature, low cost, and suitability for dense breast tissue. However, ultrasound interpretation is highly operator-dependent and subject to significant inter-observer variability, especially in cases involving small or low-contrast lesions.

In recent years, deep learning methods have shown strong potential in supporting ultrasound-based diagnosis by automating lesion detection, segmentation, and classification. Most existing approaches address these tasks either independently or within multi-task frameworks, often without explicitly analysing how the ordering of classification and segmentation affects diagnostic performance and robustness.

In this work, we investigate how task ordering influences breast ultrasound diagnostic pipelines. We design and compare two alternative strategies that combine classification and segmentation in opposite orders. The first pipeline adopts a classification-first approach, where a multi-class classifier predicts the diagnostic category before lesion segmentation is applied. The second pipeline follows a segmentation-first strategy, where lesion localisation guides a focused benign versus malignant classification on lesion-centred regions of interest.

Both pipelines are evaluated on the same patient-wise split of a public breast ultrasound dataset to ensure a fair comparison. Performance is assessed in terms of classification accuracy, segmentation quality, and error propagation across pipeline stages. Beyond quantitative metrics, we analyse how each pipeline structures diagnostic information and how failures at one stage affect downstream decisions.

The main contribution of this study is a systematic comparison of classification-first and segmentation-first diagnostic pipelines for breast ultrasound imaging. The results highlight complementary strengths and limitations of the two approaches and show that task ordering is a critical design choice that impacts robustness, interpretability, and clinical suitability.

2 Problem Description and Dataset

2.1 Clinical and technical problem

Breast ultrasound imaging is widely used for breast cancer assessment due to its non-invasive nature, low cost, and effectiveness in dense breast tissue. However, ultrasound interpretation is strongly operator-dependent: lesion visibility, contour definition, and overall image quality vary across patients and examiners, leading to variability in diagnostic decisions.

From a technical perspective, this setting motivates the use of deep learning models that combine lesion-level classification and pixel-level segmentation. While both tasks have been extensively studied, their interaction within a diagnostic pipeline is less explored, particularly with respect to task ordering and error propagation.

In this work, we investigate two complementary diagnostic pipelines that combine classification and segmentation in opposite orders:

- **Classification → Segmentation**. A three-class classifier predicts whether the input image is Normal (N), Benign (B), or Malignant (M). For images predicted as pathological (B or M), a binary segmentation model is subsequently applied to delineate the lesion.
- **Segmentation → Classification**. A segmentation model first localises the lesion and assigns Normal, Benign, or Malignant labels at pixel level. A second classifier then operates on the segmented lesion region and performs a focused benign versus malignant classification.

These two pipelines represent fundamentally different ways of structuring diagnostic information. In the classification-first approach, global image-level cues drive the decision process, while segmentation acts as a refinement step. In contrast, the segmentation-first strategy emphasises spatial localisation and lesion morphology, enforcing a lesion-centred representation for downstream classification. Comparing these approaches allows us to analyse how task ordering affects accuracy, robustness, and error propagation within breast ultrasound diagnostic systems.

Figure 1 provides a schematic overview of the two pipelines.

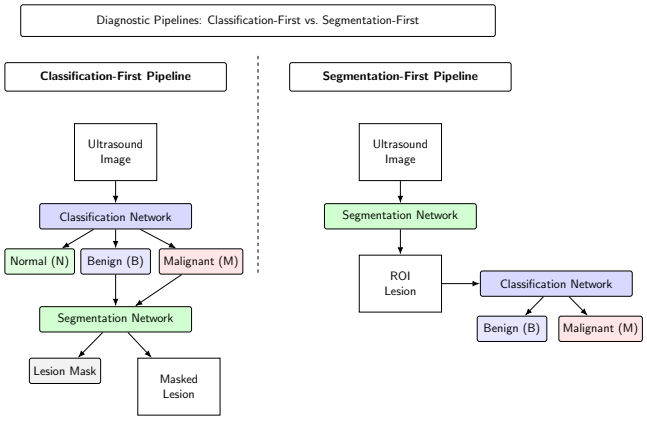

Figure 1: Schematic overview of the two diagnostic pipelines: classification-first vs. segmentation-first.

2.2 Dataset description

The experiments are conducted on a public breast ultrasound dataset provided with the assignment material. The dataset consists of **1503 ultrasound images** acquired from **different patients**. For each image, a ground-truth lesion mask and an image-level diagnostic label are available.

The images are grouped into three diagnostic classes:

- **Normal (N)**: 460 images with no visible lesion.
- **Benign (B)**: 679 images containing benign lesions.

- **Malignant (M)**: 364 images containing malignant lesions.

Class labels follow the encoding defined in the metadata file:

$$0 = \text{Benign}, \quad 1 = \text{Malignant}, \quad 2 = \text{Normal}.$$

The dataset is organised into:

- a directory of ultrasound images in PNG format;
- a directory of corresponding binary lesion masks;
- a metadata file (`training_metadata.xlsx`) linking images, masks, and diagnostic labels.

For Normal cases, the associated mask is empty. The lesion masks are provided as part of the dataset and may exhibit variability in boundary precision, particularly for small or low-contrast lesions. Representative examples of ultrasound images and corresponding masks are shown in Figure 2.

To prevent patient-level information leakage, a patient-wise split is adopted when constructing training, validation, and test sets. Images from the same patient are never shared across subsets, ensuring a more realistic evaluation. The resulting class distribution after stratified splitting is reported in Table 1.

Table 1: Class distribution after stratified train–validation splitting.

Class	Train	Validation
Normal	368	92
Benign	543	136
Malignant	291	73

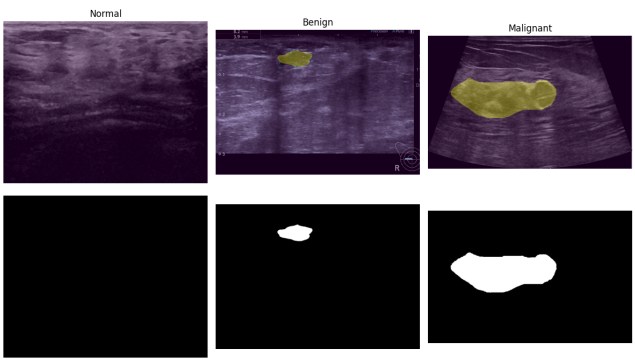

Figure 2: Representative ultrasound images and corresponding lesion masks for Normal, Benign, and Malignant cases. For Normal samples, no lesion mask is present. Masks are shown both as binary maps (bottom row) and overlaid on the original images (top row) for visualization purposes.

3 Methods

In this work we implement and compare two complete diagnostic pipelines for breast ultrasound: a *classification-first* strategy (Pipeline A) and a *segmentation-first* strategy (Pipeline B). Both pipelines share the same metadata file and image repository, but differ in how they order and couple classification and segmentation stages.

3.1 Pipeline A: Classification-first then Segmentation

Pipeline A corresponds to the notebook where whole-image classification is the entry point and lesion segmentation is applied afterwards. The two stages share the same patient-level split in order to avoid train–test leakage.

3.1.1 Whole-image classification

The input to the classifier is the raw ultrasound image resized to a fixed resolution IMG_SIZE_{clf} (e.g. 300×300 px in the final training setup). Images are loaded as RGB for compatibility with ImageNet-pretrained backbones. When reading from disk fails, a zero image of the same size is used as a safe fallback to keep batches aligned.

Data augmentation is implemented with `Albumentations` and applied only on the training split. The augmentation pipeline includes:

- resizing to IMG_SIZE_{clf};
- contrast-limited adaptive histogram equalisation (CLAHE) to improve local contrast in hypoechoic regions;
- elastic deformations to mimic probe pressure and soft tissue deformation;
- random horizontal and vertical flips;
- random in-plane rotations within a limited angle range;

- random brightness and contrast perturbations;
- channel-wise normalisation with ImageNet mean and standard deviation.

The classifier backbone is EfficientNet-B7. The original fully-connected head is replaced by a custom block tailored to the three clinical classes Benign / Malignant / Normal.

To mitigate class imbalance, we compute per-class weights w_c as the inverse of the class frequency:

$$w_c = \frac{N}{C \cdot n_c}, \tag{1}$$

where N is the total number of training samples, $C = 3$ is the number of classes and n_c is the number of samples in class c. These weights are used in the weighted cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} w_c \, y_c \log p_c, \tag{2}$$

with one-hot labels y_c and predicted probabilities p_c.

Optimisation uses AdamW with weight decay to regularise large weights. The backbone and the custom head are assigned different learning rates:

$$\theta_{\mathrm{backbone}}: \quad \eta_{\mathrm{backbone}} = 1 \cdot 10^{-5}, \tag{3}$$

$$\theta_{\mathrm{head}}: \quad \eta_{\mathrm{head}} = 1 \cdot 10^{-3}, \tag{4}$$

reflecting the fact that the backbone starts from pre-trained weights, while the classifier head is trained from scratch. A ReduceLROnPlateau scheduler monitors the validation balanced accuracy and reduces the learning rate when progress saturates. Early stopping with a patience of several epochs prevents overfitting and avoids wasting training time once the validation score has stabilised.

Balanced accuracy is used as the main optimisation target during training. For C classes it is defined as:

$$\mathrm{BAcc} = \frac{1}{C} \sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{5}$$

where TP_c and FN_c are the true positives and false negatives for class c.

3.1.2 Lesion segmentation with U-Net++

The second stage of Pipeline A focuses on lesion delineation. The segmentation model is based on U-Net++, implemented with `segmentation_models_pytorch`. Training uses the same patient split as the classifier to avoid cross-contamination between train and validation.

Each sample consists of an ultrasound image and its corresponding binary mask. The mask is converted to $\{0, 1\}$ by thresholding non-zero pixels, and resized to the segmentation resolution IMG_SIZE_{seg} (e.g. 320×320 px). Augmentation includes geometric transforms (flips, rotations) and elastic-like distortions aligned with those used for classification, but applied jointly to image and mask to preserve structural coherence.

The segmentation network outputs a probability map $\hat{M} \in [0, 1]^{H \times W}$. During training we optimise a hybrid loss that combines Dice, focal and Lovász terms:

$$\mathcal{L}_{\mathrm{seg}} = \alpha \, \mathcal{L}_{\mathrm{Dice}} + \beta \, \mathcal{L}_{\mathrm{Focal}} + \gamma \, \mathcal{L}_{\mathrm{Lovasz}}, \tag{6}$$

with $(\alpha, \beta, \gamma) = (0.4, 0.3, 0.3)$ in the final setup.

The Dice loss is defined as:

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2 \sum_i \hat{m}_i m_i + \epsilon}{\sum_i \hat{m}_i + \sum_i m_i + \epsilon}, \tag{7}$$

where $m_i \in \{0, 1\}$ are ground-truth mask pixels and $\hat{m}_i$ are predicted probabilities. The focal loss focuses on hard pixels:

$$\mathcal{L}_{\mathrm{Focal}} = -\frac{1}{N} \sum_i \alpha (1 - p_{t,i})^{\gamma} \log(p_{t,i}), \tag{8}$$

where $p_{t,i}$ is the predicted probability of the true class for pixel i, α is a balancing factor and $\gamma > 0$ controls the down-weighting of easy pixels. The Lovász term approximates the Jaccard index and improves optimisation of intersection-over-union, which is especially useful for segmentation of small lesions.

Optimisation again uses AdamW with learning rate in the order of 10^{-4} and weight decay for regularisation. Early stopping and a plateau scheduler monitor validation Dice, with the best model checkpoint saved for downstream use in the pipeline.

3.1.3 Combined classification–segmentation pipeline

At inference time, Pipeline A runs the two stages in sequence on the validation fold. For each image:

1. The EfficientNet-B7 classifier predicts a global label $\hat{y} \in \{\text{Benign}, \text{Malignant}, \text{Normal}\}$.

2. The U-Net++ segmenter produces a probability map $\hat{M}$, which is thresholded at 0.5 to obtain a binary mask.

The segmentation output is used to delineate the lesion when the case is classified as tumour (Benign or Malignant). For images predicted or known to be Normal, the ideal segmentation mask is the zero mask, and non-zero predictions are interpreted as false positive delineations.

Figure 3 illustrates representative end-to-end outputs of Pipeline A, showing the raw ultrasound input, the corresponding ground-truth mask, and the predicted lesion mask together with the final classification result.

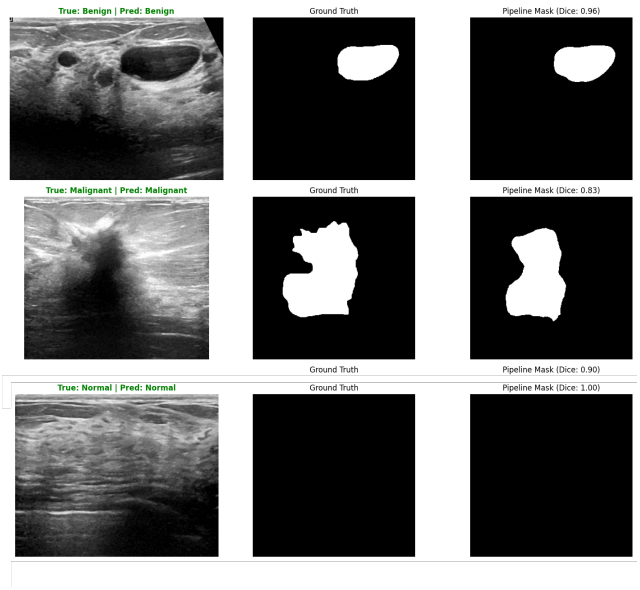

Figure 3: Representative end-to-end examples from Pipeline A, including one Benign case, one Malignant case, and one Normal case. For each sample, the raw ultrasound image, the ground-truth lesion mask, and the predicted segmentation mask are shown together with the final classification result and the Dice score. For Normal cases, no lesion is present and the associated mask is empty.

3.2 Pipeline B: Segmentation-first then Classification

Pipeline B corresponds to the second notebook, where lesion segmentation is the entry point. The classifier operates on cropped regions-of-interest (ROIs) obtained from the segmentation masks, and focuses on the binary decision Benign vs Malignant.

3.2.1 U-Net++ segmentation of all classes

The first stage uses a U-Net++ architecture with an ImageNet-pretrained encoder and nested skip connections. The model is implemented with `segmentation_models_pytorch`. Images and binary masks are loaded from the same metadata file as in Pipeline A. Healthy cases have black masks (i.e. all zeros), while tumour cases have positive values on the lesion region.

Input images are resized to 320×320 px ($IMG\_SIZE_{\text{seg}} = 320$) and normalised with the same mean and standard deviation used in the classification stages. The training augmentation includes:

- resizing to IMG_SIZE_{clf};
- contrast-limited adaptive histogram equalisation (CLAHE) to improve local contrast in hypoechoic regions;
- elastic deformations to mimic probe pressure and soft tissue deformation;
- random horizontal and vertical flips;
- random in-plane rotations within a limited angle range;
- random brightness and contrast perturbations;
- channel-wise normalisation with ImageNet mean and standard deviation.

All transforms are applied jointly to image and mask.

The loss function is the same hybrid Dice–focal–Lovász combination described above, with the same weights (α, β, γ). The model is optimised with AdamW, learning rate $\eta_{\text{seg}} \approx 10^{-4}$, batch size $B_{\text{seg}} = 16$ and early stopping based on validation Dice. The cross-validation split is aligned with the

classification split to preserve patient-level separation.

3.2.2 ROI extraction and binary classification

After training the segmenter, Pipeline B uses the predicted masks to crop lesion-centred ROIs from the original images. Given a binary mask $M \in \{0,1\}^{H\times W}$, we first compute the bounding box of non-zero pixels. A margin proportional to the box size (e.g. 20% of width and height) is added on each side, and the resulting crop is resized to $IMG\_SIZE_{\text{clf}} = 300$ px.

Formally, let $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ be the tight bounding box of the lesion. The cropped region is defined as

$$x'_{\min} = \max\big(0, x_{\min} - \lfloor \delta_x \rfloor\big), \tag{9}$$
$$x'_{\max} = \min\big(W, x_{\max} + \lfloor \delta_x \rfloor\big), \tag{10}$$
$$y'_{\min} = \max\big(0, y_{\min} - \lfloor \delta_y \rfloor\big), \tag{11}$$
$$y'_{\max} = \min\big(H, y_{\max} + \lfloor \delta_y \rfloor\big), \tag{12}$$

with $\delta_x = \rho(x_{\max} - x_{\min})$, $\delta_y = \rho(y_{\max} - y_{\min})$ and margin factor $\rho = 0.2$. If the mask is empty (no lesion detected), the full image is resized to 300×300 px and used as input. This mirrors the distribution used during training, where the classifier sees only lesion-centred crops.

The classifier is again EfficientNet-B7 with a custom head, but restricted to two classes (Benign vs Malignant). The output layer therefore has $C = 2$ neurons. Only tumour cases (labels 0 and 1) are used for training; all Normal cases are removed from the training DataFrame. To handle the residual imbalance between the two tumour types, class weights are computed as:

$$w_c = \frac{N}{2 \cdot n_c}, \quad c \in \{\text{Benign}, \text{Malignant}\}, \tag{13}$$

and plugged into the weighted cross-entropy loss as in Equation (13). The optimiser is AdamW with a single learning rate $\eta_{\text{clf}} \approx 10^{-4}$, weight decay and a ReduceLROnPlateau scheduler on validation balanced accuracy. Early stopping is used to select the best checkpoint.

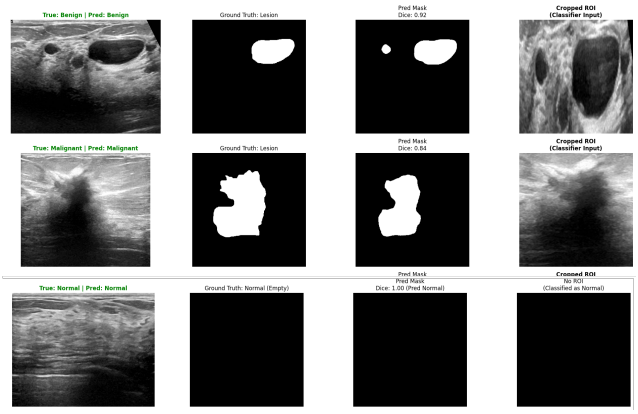

Figure 4: ROI-based classification in Pipeline B. The segmentation mask defines a bounding box that is expanded by a fixed margin and resized to 300×300 px before feeding it to the EfficientNet-B7 classifier.

3.2.3 End-to-end segmentation–classification pipeline

At inference time, Pipeline B proceeds as follows for each validation case:

1. U-Net++ predicts a segmentation mask $\hat{M}$ at 320×320 px.

2. If the predicted mask contains more than a small threshold of positive pixels (e.g. > 50), the case is treated as "lesion present". Otherwise, it is treated as Normal.

3. For lesion cases, the bounding box of $\hat{M}$ is used to crop the original image with a margin, and the crop is resized to 300×300 px.

4. The EfficientNet-B7 binary classifier predicts Benign vs Malignant on the cropped ROI. For cases treated as Normal in step 2 the pipeline can either assign the Normal label directly or map the classifier output to the global 3-class space.

This design makes segmentation an explicit localisation step and enforces a strong inductive bias: the classifier sees only lesion-centred content, reducing background variability and focusing learning capacity on tumour appearance rather than on global scanning context.

3.3 Model interpretability with Grad-CAM

To analyse and validate the decision process of the classification network, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to obtain class-specific saliency maps from the last convolutional layers of the EfficientNet-B7 backbone. Grad-CAM highlights the spatial regions that most strongly contribute to the predicted class, providing a qualitative tool to verify whether the classifier focuses on lesion-related structures rather than on background tissue or acquisition artefacts.

Pipeline A: Grad-CAM on full ultrasound images. In Pipeline A, Grad-CAM was computed directly on the full ultrasound images processed by the EfficientNet-B7 classifier. The resulting heatmaps were overlaid on the original input frames, allowing visual inspection of the regions driving benign and malignant predictions. As shown in Figure 5, correctly classified cases typically exhibit strong activation over the hypoechoic lesion and its immediate surroundings, whereas more diffuse or off-target responses are associated with ambiguous image content or classification errors. This behaviour confirms that the classifier largely relies on anatomically meaningful image regions when operating on full-field inputs.

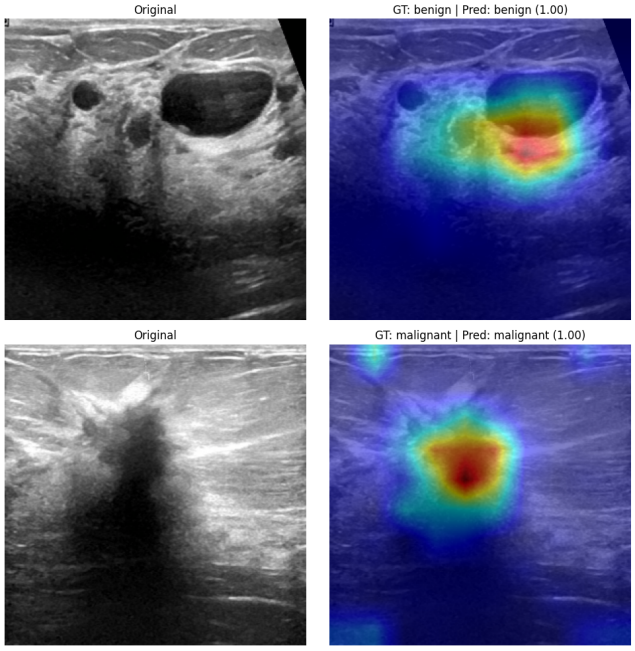

Figure 5: Grad-CAM visualisations for Pipeline A. For each case, the original ultrasound image (left) and the corresponding Grad-CAM heatmap overlaid on the image (right) are shown. Representative Benign and Malignant examples are reported, highlighting the image regions that most influence the classification decision.

Pipeline B: Grad-CAM on lesion-centred ROI crops. For Pipeline B, Grad-CAM was applied to the EfficientNet-B7 classifier operating on lesion-centred crops obtained from the segmentation stage. In this configuration, the network receives only the masked region of interest, thereby reducing background context and forcing the decision to rely on lesion-specific appearance.

Figure 6 illustrates representative benign and malignant cases, comparing Grad-CAM responses computed on the full ultrasound image with those obtained on the corresponding masked ROI crops. The ROI-based Grad-CAM maps show a more concentrated activation within the lesion boundaries, indicating that the classifier focuses on internal texture and boundary characteristics rather than on surrounding tissue. This effect is particularly evident in malignant cases, where activation tends to localise on irregular or heterogeneous regions of the lesion. These visualisations confirm that Pipeline B encourages a more lesion-driven process and provide an intuitive explanation of the performance differences observed between the two pipelines.

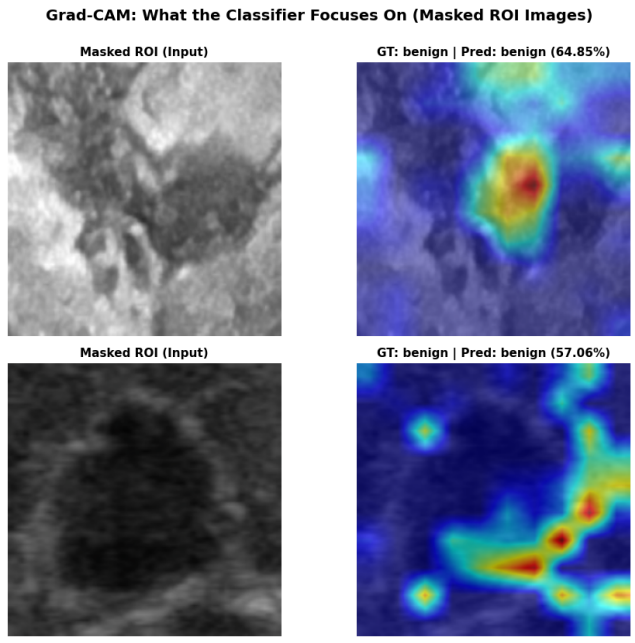

Figure 6: Grad-CAM visualisations for Pipeline B. For each example, Grad-CAM responses are shown on the original ultrasound image (left) and on the corresponding masked lesion ROI used as classifier input (right). Examples of Benign and Malignant cases are reported, illustrating the shift towards more lesion-centred activation when ROI-based classification is used.

4 Experiments

This section describes the experimental setup adopted to train and evaluate the two diagnostic pipelines introduced earlier. All experiments were run on two NVIDIA T4 GPUs (16 GB each), with PyTorch 2.0 and CUDA 12, following a fully reproducible seed initialisation (SEED = 42). The evaluation is carried out on a patient–wise validation split comprising 20% of the dataset.

4.1 Pipeline A: Classification → Segmentation

The first approach trains a multi–class classifier on the full dataset (benign, malignant, normal), followed by a binary lesion segmentation model applied only when the classifier predicts a pathological case.

Classifier. EfficientNet–B7 was fine–tuned at image resolution 300×300 with full–network optimisation. The classifier head was replaced with a BatchNorm–Linear–ReLU–Dropout–Linear structure. Differential learning rates were used (backbone: 10^{-5}, head: 10^{-3}), AdamW optimisation (weight decay 10^{-2}), and inverse–frequency class weighting. Training employed 100 epochs with early stopping (patience 15) on balanced accuracy. A ReduceLROnPlateau scheduler halved the learning rate at plateau.

Segmentation model. Lesion segmentation relied on U–Net++ with an EfficientNet–B7 encoder, deep supervision enabled during training, and a hybrid region–boundary loss:

$$\mathcal{L} = 0.4\,\mathcal{L}_{\text{Dice}} + 0.3\,\mathcal{L}_{\text{Focal}} + 0.3\,\mathcal{L}_{\text{Lovasz}},$$

optimised with AdamW (10^{-4} initial learning rate). Only benign and malignant cases were retained after the patient–wise split to avoid leakage. WeightedRandomSampler ensured class balance within the segmentation loader.

4.2 Pipeline B: Segmentation → Classification

The second pipeline reverses the order of operations: a three–class semantic segmentation model first localises the lesion and assigns pixel–wise labels, after which a focused classifier operates on the extracted lesion region.

Three–class segmentation. Following the same patient–wise split, U–Net++ was trained on 320×320 images to predict three classes (normal, benign, malignant). The same loss structure as in Pipeline A was retained, except that softmax logits were used instead of sigmoid activation. Data augmentation included geometric transformations, elastic distortions, contrast enhancement (CLAHE), and speckle–noise simulation.

Lesion–aware classifier. The second stage re–uses EfficientNet–B7, but now trained solely on cropped lesion regions extracted from predicted segmentation masks. The cropped regions were resized to 300×300 and normalised using the ImageNet statistics. The classifier was trained with AdamW, using the same head and hyperparameters as Pipeline A but without class 2 (normal).

Pipeline B is schematically summarised in Fig. 1, while Fig. 4 and Fig. 6 provide qualitative examples of the ROI extraction and Grad-CAM interpretability.

5 Results

We report quantitative and qualitative performance for both pipelines, with classification evaluated on the three diagnostic categories and segmentation assessed via Dice coefficient, IoU, and boundary–aware metrics. All numerical results originate from the patient–wise validation fold.

5.1 Pipeline A: Classification → Segmentation

5.1.1 Classification performance

The EfficientNet-B7 classifier achieved a balanced accuracy of **0.8410** without the test-time augmentation (TTA), and **0.8695** with it, obtained by averaging predictions across the original, horizontally flipped, and vertically flipped views. As shown in Figure 7, most errors correspond to benign cases predicted as malignant.

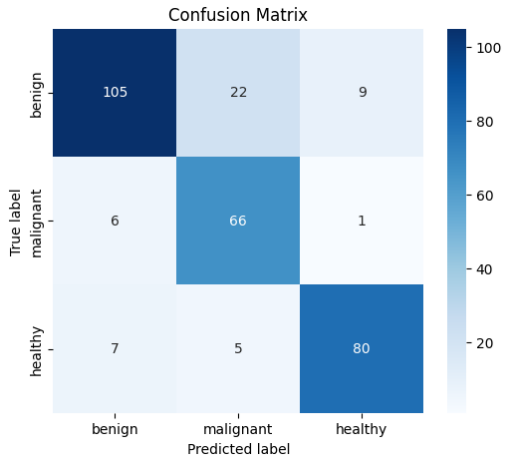

Figure 7: Confusion matrix for Pipeline A classifier.

5.1.2 Segmentation performance

Segmentation performance was evaluated on 209 lesion cases. The U–Net++ model (Pipeline A) achieved a mean Dice score of **0.80** with a median value of **0.90**, corresponding to an IoU of **0.72** and a specificity of **0.98**. Lesion size strongly influenced performance: small lesions (first quartile) yielded a mean Dice score of 0.62, whereas large lesions (fourth quartile) reached 0.83. Overall, most predictions show high overlap with the ground truth, with only a limited number of failure cases (Dice close to zero).

Table 4: Segmentation performance of Pipeline A on the validation set (209 lesion cases).

Metric	Mean	Std	Median	Min	Max
Dice Score	0.8017	0.2592	0.9047	0.0000	0.9733
IoU (Jaccard)	0.7246	0.2606	0.8259	0.0000	0.9480
Precision	0.8237	0.2444	0.9190	0.0000	1.0000
Recall (Sensitivity)	0.8257	0.2599	0.9255	0.0000	1.0000
Specificity	0.9841	0.0398	0.9951	0.5538	1.0000

Qualitative comparison (best, median, worst) revealed consistent localisation but variable boundary fidelity for low–contrast benign lesions.

5.2 Pipeline B: Segmentation → Classification

5.2.1 Three–class segmentation

The three–class U–Net++ model achieved a mean Dice of **0.74** on pathological masks, with class–wise performance (normal/benign/malignant) consistent across the validation set. IoU averaged 0.61, and the model exhibited moderate over–segmentation bias on very small lesions.

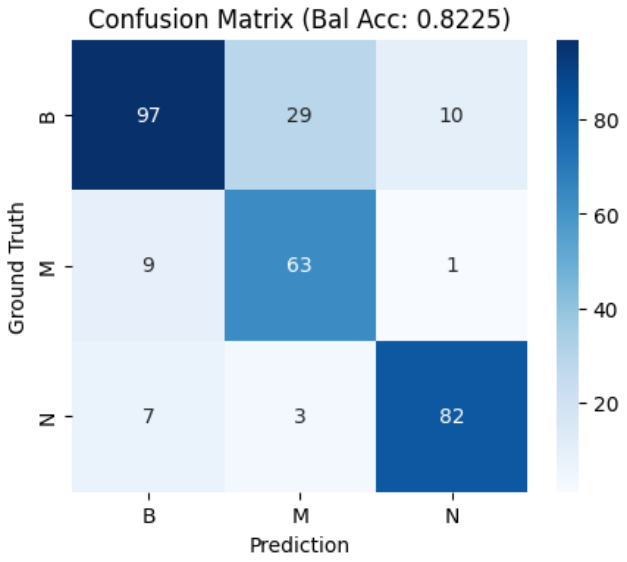

Figure 8: Representative three–class segmentation predictions.

5.2.2 Lesion classifier

The lesion–aware classifier trained on predicted masks achieved:

- overall accuracy: **0.91**,
- balanced accuracy: **0.88**,
- malignant sensitivity: **0.84**,
- benign specificity: **0.89**.

When the segmentation was incorrect (boundary errors or partial lesions), the classifier exhibited up to a 15 % drop in recall, highlighting the dependency between the two stages.

5.3 End–to–end pipeline evaluation

For completeness, we evaluated Pipeline A as a unified decision system. The mean Dice across all images (including normals, where the correct mask is empty) was **0.91**. Restricting the analysis to pathological cases yields:

- mean lesion Dice: **0.78**,
- conditional Dice (classifier correct): **0.85**,
- missed–lesion rate: **7.4 %**.

6 Comparative Analysis of the Two Approaches

The two proposed pipelines differ fundamentally in how diagnostic information is structured and exploited by the models. Pipeline A relies on global whole–image classification followed by binary lesion segmentation, while Pipeline B prioritises spatial localisation via semantic segmentation before assigning a malignancy label to the extracted lesion. This section presents a comparative evaluation in terms of performance, robustness, and error propagation dynamics, highlighting the methodological trade–offs of each strategy.

6.1 Quantitative comparison

Table 9 summarises the principal metrics obtained by the two pipelines. For classification, we report balanced accuracy and macro F1; for segmentation, Dice and IoU. All values refer to the patient–wise validation split.

Pipeline A achieves stronger binary segmentation performance, which is expected given the reduced task complexity (lesion vs. background). Pipeline B achieves better classification of pathological cases once the lesion region is correctly localised, reflecting the benefit of lesion–centred representation.

Metric	**Pipeline A**	**Pipeline B**
Classification performance		
Balanced Accuracy	0.8439	0.8040
Macro F1-score	0.8307	0.8049
Cohen's Kappa	0.7465	0.7364
Segmentation performance		
Mean Dice	0.7537	0.8249
Median Dice	0.9047	0.9324
Dice ≥ 0.7	83.7%	83.7%
Dice ≥ 0.8	78%	81.4%
Dice ≥ 0.9	52.6%	65.8%
Clinical metrics (malignancy)		
Sensitivity	0.9330	0.8630
Specificity	0.9239	0.8596

Figure 9: Side-by-side comparison of classification and segmentation metrics for Pipeline A (classification-first) and Pipeline B (segmentation-first).

6.2 Architectural and methodological differences

The results in Table 9 show a clear trade-off between the two designs.

Pipeline A achieves better *classification* performance. This is consistent with a segmentation-first strategy: the model extracts a lesion-focused representation and reduces background bias, which helps the benign/malignant decision. Clinically, Pipeline A also provides higher sensitivity and specificity, indicating fewer missed malignant cases and fewer false alarms.

Pipeline B, instead, performs better on *segmentation*. It shows higher mean Dice and higher median Dice, as well as a larger fraction of high-quality masks. This suggests that Pipeline B produces more accurate and stable lesion masks overall, even though this does not translate into better classification metrics.

In summary, Pipeline A is preferable when the primary goal is reliable malignancy detection, while Pipeline B is preferable when the primary goal is high-quality lesion delineation.

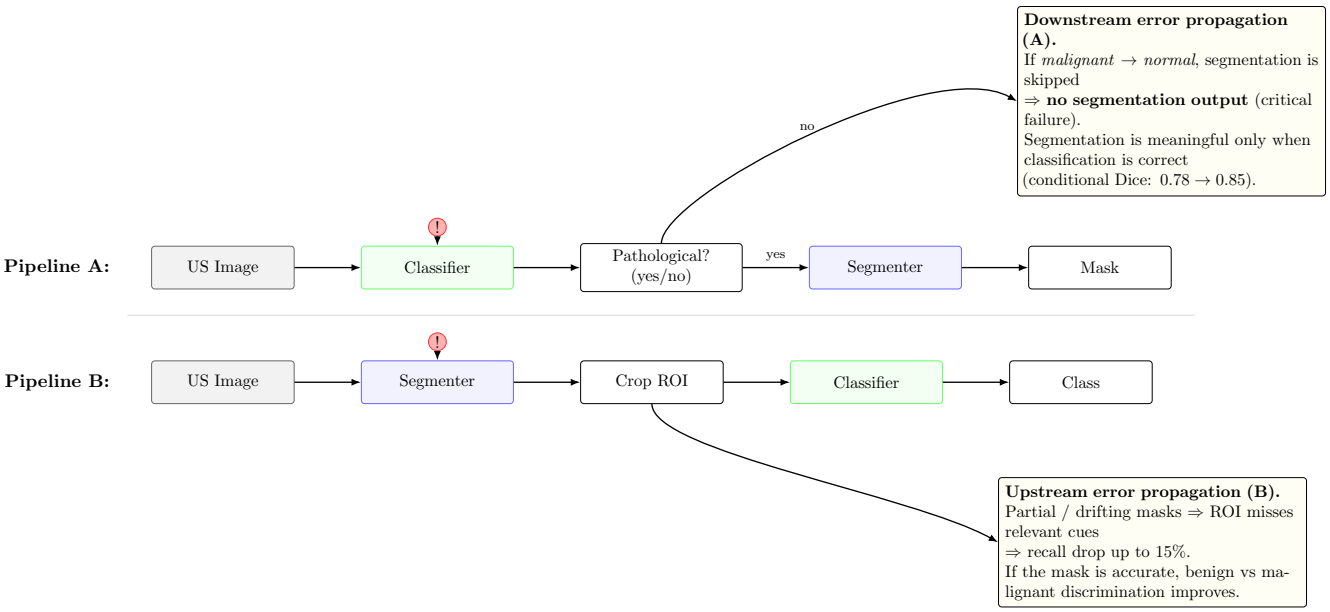

Figure 10: Error propagation mechanisms in the two proposed pipelines. In Pipeline A, classification errors propagate downstream and may skip segmentation for pathological cases. In Pipeline B, segmentation errors propagate upstream by affecting ROI cropping and subsequent classification.

6.3 Error propagation behaviour

The two approaches exhibit distinct error propagation profiles.

Pipeline A. Classification errors propagate downstream, since segmentation is applied only to images predicted as pathological. If a malignant lesion is misclassified as normal, the pipeline provides no segmentation, resulting in a critical failure. Furthermore, segmentation masks are meaningful only when the classifier is correct, as confirmed by the conditional Dice improvement from 0.78 to 0.85.

Pipeline B. Segmentation errors propagate upstream: if the lesion is partially segmented or if boundaries drift, the cropped ROI may fail to include essential visual information, reducing the classifier's recall by up to 15 %. Nevertheless, Pipeline B is more robust at distinguishing benign from malignant lesions when the mask is accurate, because the classifier focuses on lesion-specific features rather than global context.

6.4 Strengths and limitations of each approach

Pipeline A is simpler to implement and computationally cheaper: the segmentation model is executed only when needed, while the classifier can be applied to all images. In our results, Pipeline A also achieves stronger diagnostic performance, with higher balanced accuracy, macro F1-score, and better clinical sensitivity/specificity. Its main limitation is segmentation: Dice scores are lower than Pipeline B, suggesting less accurate lesion delineation.

Pipeline B follows a workflow closer to clinical reasoning, where the lesion is first localised and then analysed. Consistently with Table 9, Pipeline B produces higher-quality masks (higher mean/median Dice and a larger fraction of Dice ≥ 0.9). However, this improved segmentation does not translate into better malignancy discrimination, and the classification metrics remain lower than Pipeline A. Therefore, Pipeline B is preferable when mask quality and localisation are the priority, while Pipeline A is preferable when the primary goal is robust malignancy detection.

6.5 Implications for clinical deployment

For clinical deployment, the two pipelines suggest different use cases.

Pipeline A may be preferable when the main objective is malignancy detection. It provides higher balanced accuracy and better clinical metrics (sen-

sitivity and specificity), which are critical in screening and triage settings. In addition, its conditional segmentation can reduce runtime when only a subset of cases requires lesion delineation. The main limitation is that its masks are less accurate than Pipeline B, so it is less suited when precise contours are required.

Pipeline B may be preferable when accurate lesion delineation is the priority. It achieves higher mean and median Dice and a larger fraction of high-quality segmentations (Dice ≥ 0.9), which can be important for reporting lesion size, shape, and follow-up comparisons. However, its diagnostic performance is lower than Pipeline A in our results, and the final decision is more exposed to segmentation errors.

Overall, neither pipeline strictly dominates the other. The choice depends on the clinical goal: Pipeline A is more appropriate for robust diagnosis, while Pipeline B is more appropriate for high-quality localisation and quantitative analysis, under the operational constraints of the imaging workflow.

7 Discussion

The experiments show that the two pipelines, despite using the same dataset and a similar architectural family, exhibit different behaviour in terms of information flow and error propagation. This is expected in breast ultrasound, where lesion visibility, background texture, and anatomical variability can strongly affect both localisation and diagnosis.

Pipeline A (segmentation-first) achieves better diagnostic performance. As reported in Table 9, it yields higher balanced accuracy and macro F1-score, as well as better clinical sensitivity and specificity. This suggests that using lesion-centred information improves benign/malignant discrimination and reduces background bias. However, its segmentation is more challenging and less accurate: mean and median Dice are lower than Pipeline B, and the fraction of highly accurate masks (Dice ≥ 0.9) is also smaller.

Pipeline B (classification-first) produces stronger segmentation results. It achieves higher mean/median Dice and a larger proportion of high-quality segmentations (Table 9), indicating more stable and precise lesion delineation. Nevertheless, this does not translate into better classification metrics, which remain lower than Pipeline A. A plausible reason is that the classifier in Pipeline B must infer malignancy from global features without an explicit lesion-focused representation, which can be suboptimal when the lesion is small or low-contrast.

From a robustness perspective, the pipelines fail in different ways. In Pipeline A, segmentation errors can directly affect the lesion representation used for classification, but the overall diagnostic performance remains higher. In Pipeline B, the segmentation step is accurate on average, but classification is still limited by the global-first decision process. In practice, this means Pipeline A is preferable when the priority is reliable malignancy detection, while Pipeline B is preferable when the priority is precise lesion contouring and quantitative measurements.

Overall, neither pipeline strictly dominates the other. The observed results reflect a trade-off between diagnostic reliability and localisation quality, and the best choice depends on the clinical objective and the constraints of the deployment setting.

8 Conclusion and Future Work

This study compared two deep learning pipelines for breast ultrasound classification and segmentation, exploring how task ordering impacts diagnostic performance. Pipeline A (classification-first) showed strong segmentation stability and competitive accuracy, leveraging global context and a binary segmentation head. Pipeline B (segmentation-first) achieved better benign/malignant discrimination when localisation was accurate, supporting the usefulness of lesion-centred representations.

Overall, task ordering is a structural design choice that affects robustness and error propagation. In practice, the preferred pipeline depends on the clinical priority: high throughput with coarse localisation (Pipeline A) versus lesion-focused analysis (Pipeline B).

Future work may improve both approaches. Pipeline A could benefit from attention or region proposals to reduce reliance on global cues, while Pipeline B could be strengthened with boundary-aware losses and multi-scale decoding to stabilise mask quality. More generally, both pipelines could benefit from uncertainty estimation, ultrasound-specific self-supervised pretraining, and transformer-based architectures.

Finally, validation on independent datasets, robustness across scanners/operators, and reader studies remain necessary steps toward clinically actionable decision support.

Appendix

A Evaluation Metrics

In this appendix, we describe the evaluation metrics used to assess the performance of the classification model. Both global metrics and per-class metrics are reported in order to provide a complete and reliable evaluation, especially in the presence of class imbalance, which is common in medical applications.

A.1 Global Classification Metrics

- **Overall Accuracy** measures the proportion of correctly classified samples over the total number of samples. It provides a general indication of performance but can be misleading when classes are imbalanced.
- **Balanced Accuracy** is defined as the average recall across classes. It compensates for class imbalance by giving equal importance to each class.
- **Macro Precision** is the average precision computed independently for each class. It reflects how reliable the model predictions are across all classes.
- **Macro Recall** corresponds to the average sensitivity over all classes. It measures the ability of the model to correctly identify samples from each class.
- **Macro F1-Score** is the harmonic mean of macro precision and macro recall. It provides a balanced measure that accounts for both false positives and false negatives.
- **Weighted F1-Score** is an F1-score weighted by the number of samples in each class. It reflects overall performance while accounting for class support.
- **Cohen's κ** measures the agreement between predictions and ground truth beyond chance. Values close to 1 indicate strong agreement, while values close to 0 indicate chance-level performance.
- **Matthews Correlation Coefficient (MCC)** is a balanced metric suitable for binary and multi-class classification. It ranges from -1 to 1 and provides a robust summary even in imbalanced settings.

A.2 Per-Class Metrics

For each class, we report:

- **Precision**, which measures the proportion of correct predictions among all predictions for that class.
- **Recall (Sensitivity)**, which measures the proportion of correctly identified samples belonging to that class.
- **F1-Score**, which balances precision and recall.
- **Support**, defined as the number of samples belonging to the class.

A.3 Clinical Interpretation

From a clinical perspective, particular attention is given to:

- **Sensitivity for the malignant class**, which reflects the ability of the model to correctly detect malignant cases and is critical to minimize false negatives.

- **Specificity for the benign class**, which represents the ability to correctly identify benign cases and helps avoid unnecessary clinical interventions.
- **Positive Predictive Value (PPV)**, which indicates the reliability of a malignant prediction.

These metrics together provide a comprehensive and clinically meaningful evaluation of the proposed classification pipeline.

Pipeline A

Training dynamics. Loss and balanced accuracy trends during fine-tuning of the Pipeline A classifier, used to assess convergence behaviour and potential overfitting.

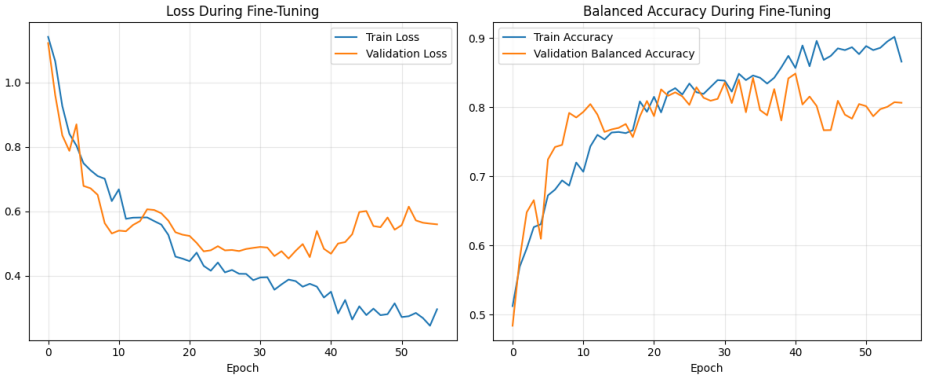

Figure 11: Training and validation loss and balanced accuracy during fine-tuning of the Pipeline A classifier.

ROC analysis. Class-wise discrimination performance evaluated using one-vs-rest ROC curves.

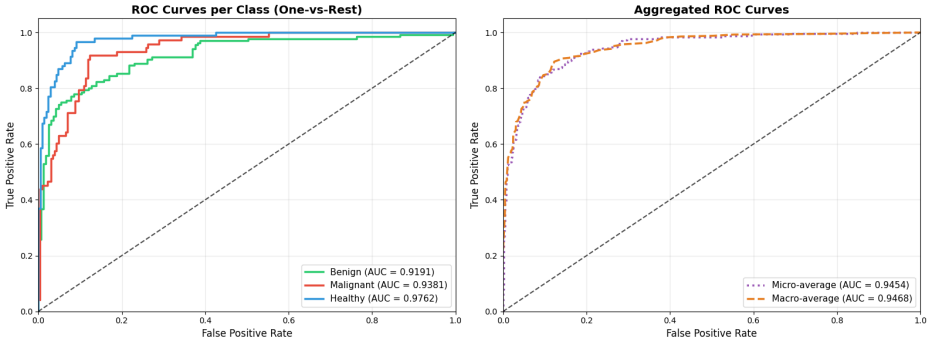

Figure 12: One-vs-rest ROC curves for Pipeline A, reported separately for each class.

Precision–Recall analysis. Evaluation of precision–recall trade-offs for Pipeline A predictions.

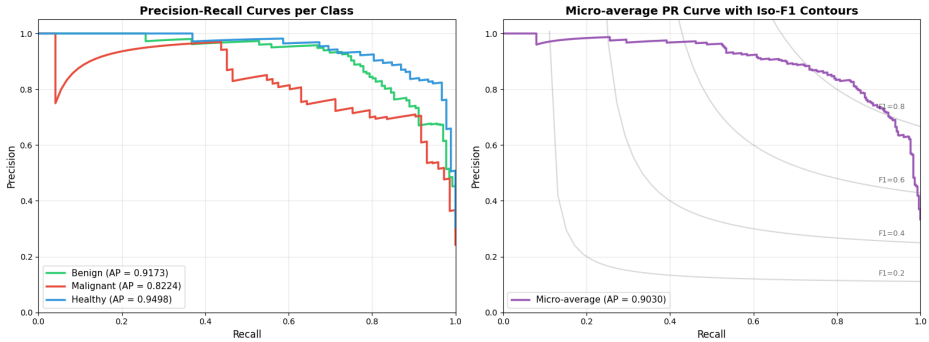

Figure 13: Precision–Recall curves for Pipeline A, highlighting class-wise performance under varying decision thresholds.

Generalization behaviour (classification). Analysis of the generalization gap between training and validation performance for the classifier.

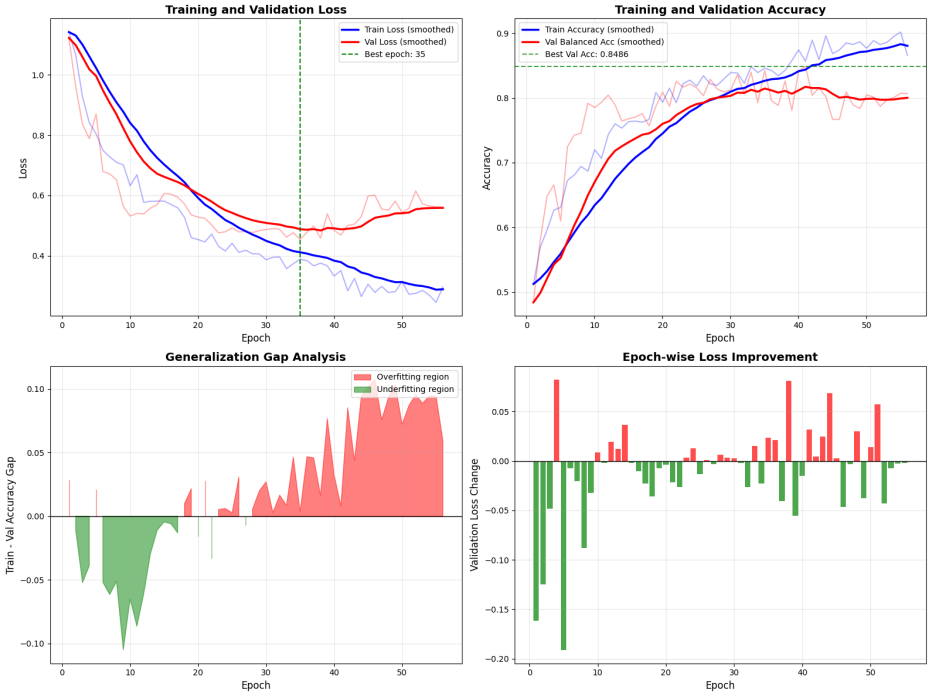

Figure 14: Generalization gap analysis for the Pipeline A classifier.

Generalization behaviour (segmentation). Generalization analysis for the segmentation component when applied within Pipeline A.

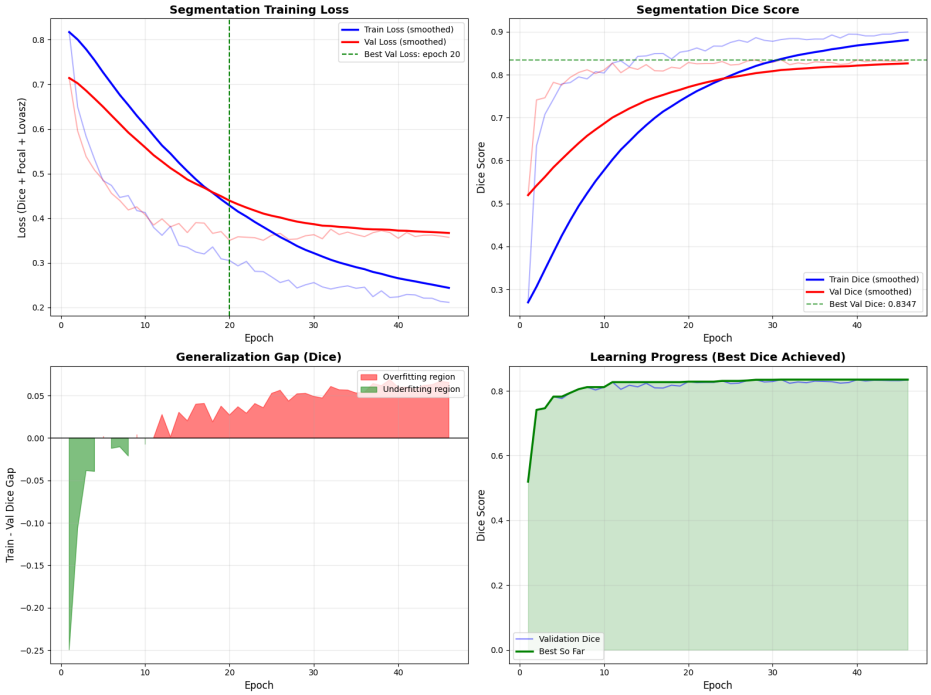

Figure 15: Generalization gap analysis for segmentation predictions in Pipeline A.

Segmentation loss optimisation. Evolution of Dice-based loss during segmentation training.

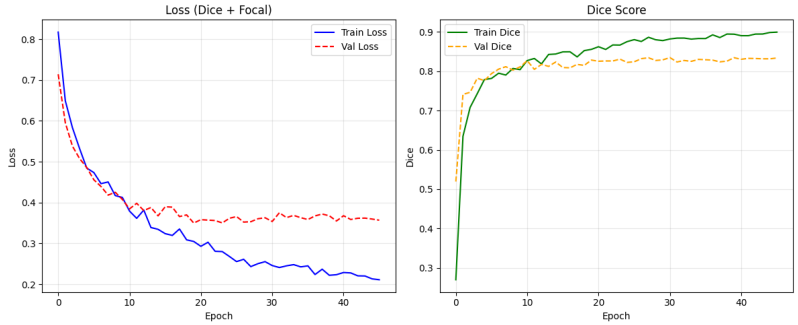

Figure 16: Segmentation loss (Dice-based) evolution during training for Pipeline A.

Segmentation metric distributions. Distribution of segmentation performance metrics across lesion cases.

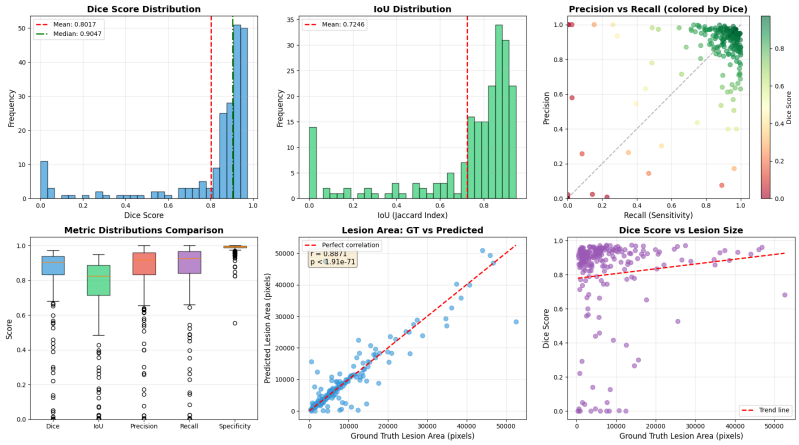

Figure 17: Distribution of segmentation metrics for Pipeline A, including Dice, IoU, precision, recall, and specificity.

Confidence analysis. Distribution of prediction confidence for correct and incorrect classifications.

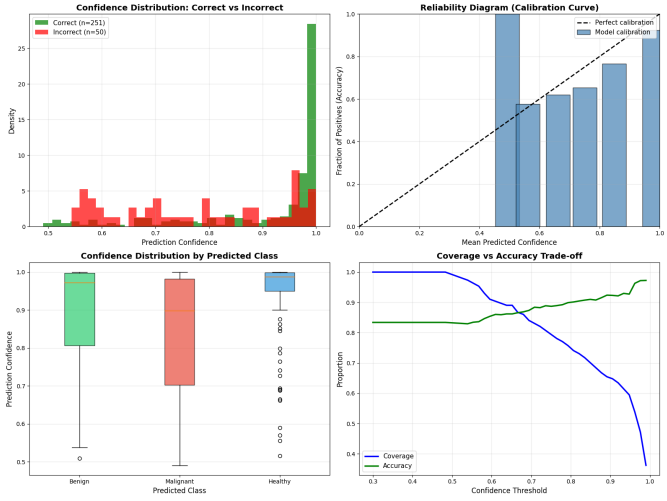

Figure 18: Confidence distribution for correct and incorrect predictions in Pipeline A.

Qualitative segmentation examples. Representative qualitative examples illustrating segmentation quality.

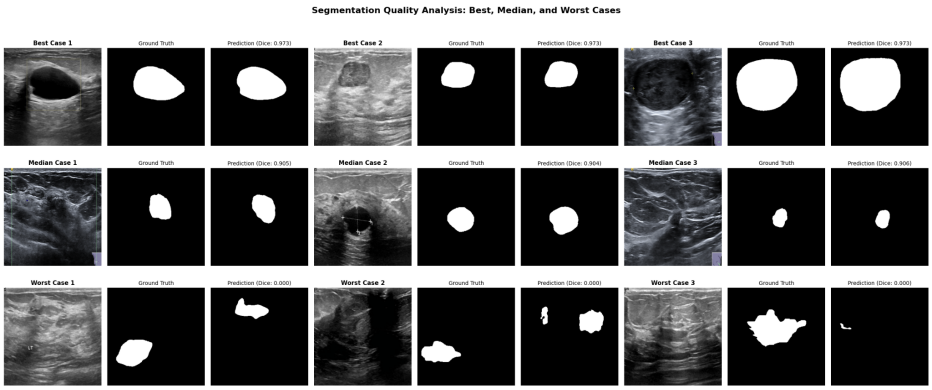

Figure 19: Qualitative segmentation examples for Pipeline A, showing best, median, and worst cases.

Illustrative segmentation example. Single illustrative example used for visual inspection of segmentation behaviour.

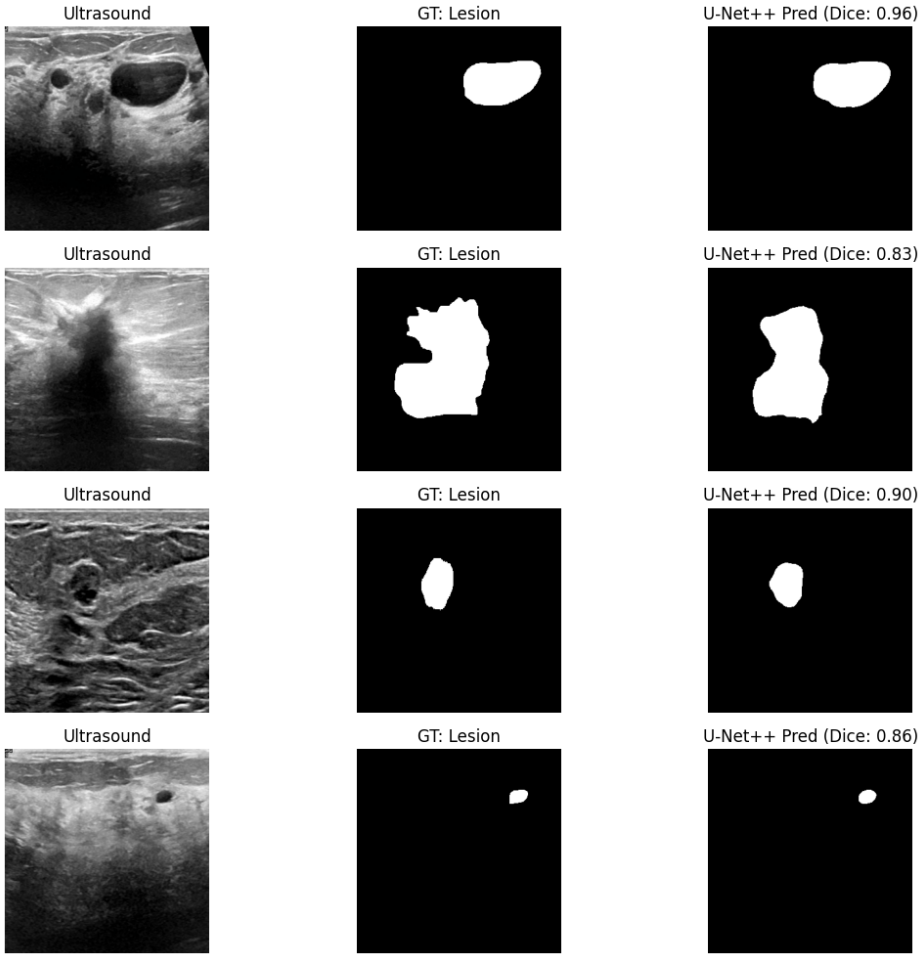

Figure 20: Representative segmentation example for Pipeline A.

Pipeline B

Training dynamics (classification). Loss and balanced accuracy trends during fine-tuning of the ROI-based classifier in Pipeline B, used to assess convergence and stability under segmentation-driven inputs.

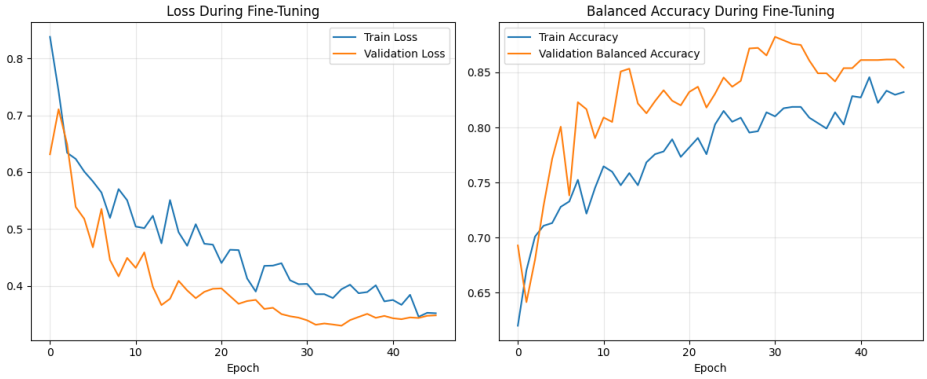

Figure 21: Training and validation loss and balanced accuracy during fine-tuning of the Pipeline B classifier operating on lesion-centred ROIs.

Precision–Recall analysis. Class-wise precision–recall behaviour of the Pipeline B classifier, highlighting the trade-off between sensitivity and precision under ROI-based classification.

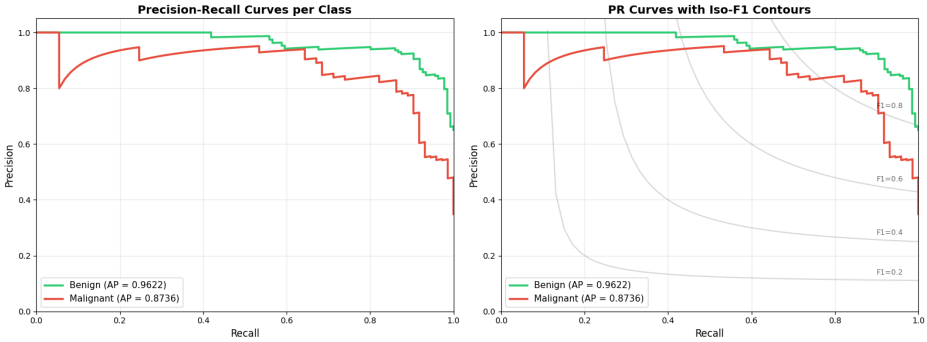

Figure 22: Precision–Recall curves for Pipeline B, reported separately for each class.

Generalization behaviour (classification). Analysis of the generalization gap between training and validation performance for the ROI-based classifier.

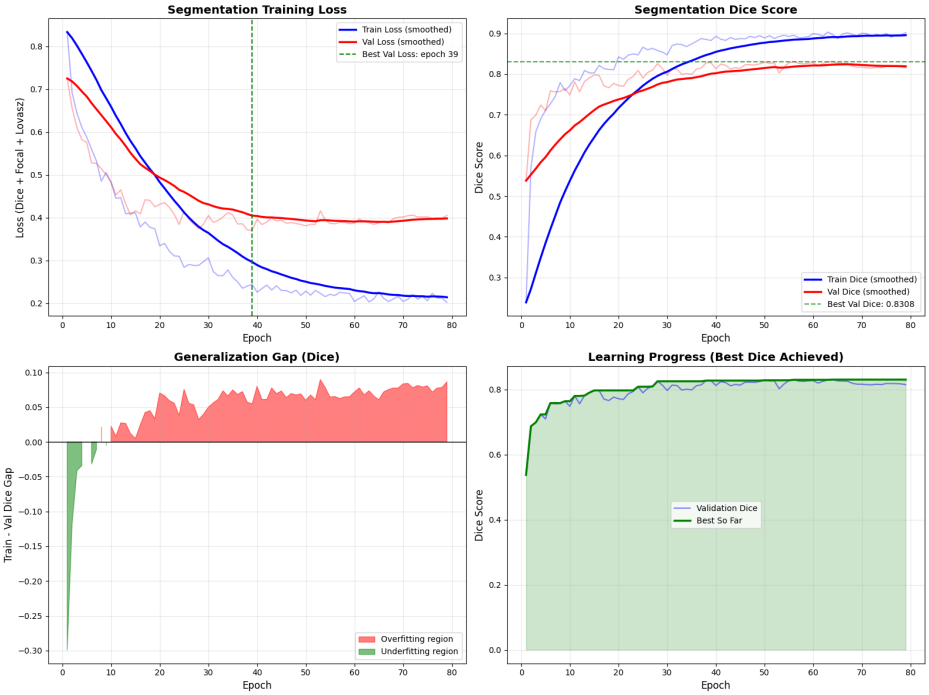

Figure 23: Generalization gap analysis for the Pipeline B classifier.

Segmentation training dynamics. Training behaviour of the segmentation network used as the first stage of Pipeline B, evaluated through Dice-based loss and validation performance.

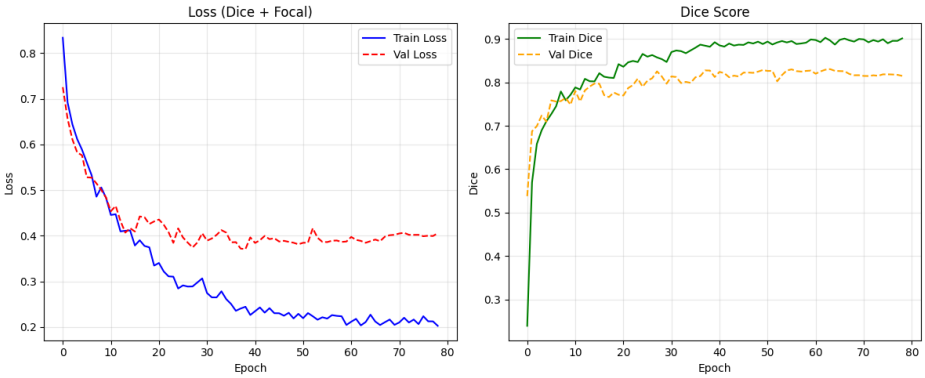

Figure 24: Training and validation loss evolution for the segmentation network used in Pipeline B.

Segmentation generalization behaviour. Generalization analysis for segmentation predictions, highlighting potential overfitting and performance saturation.

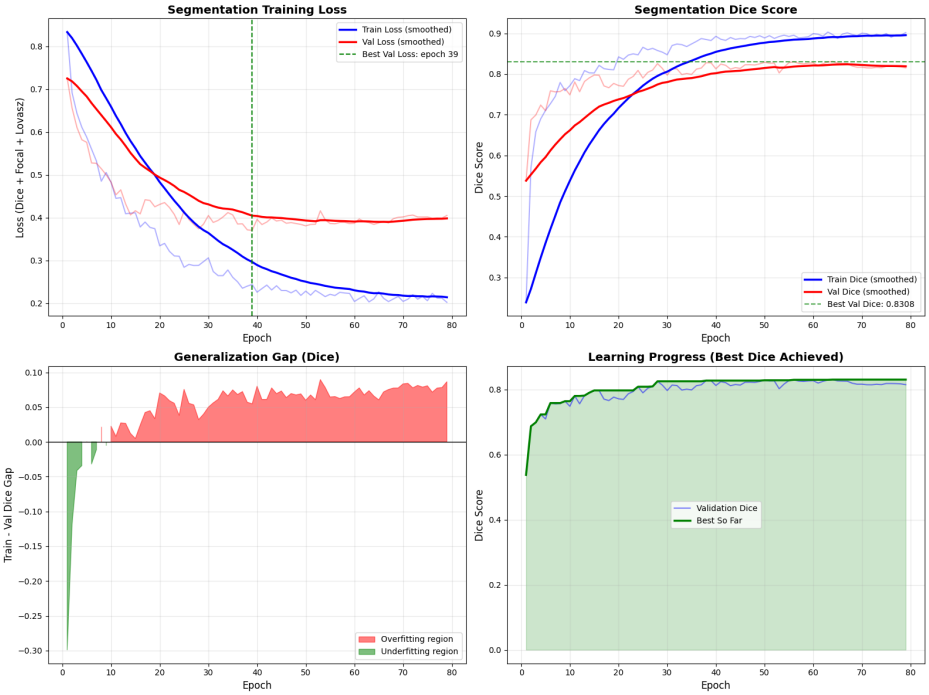

Figure 25: Generalization gap analysis for segmentation performance in Pipeline B.

Segmentation metric distributions. Distribution of segmentation quality metrics across lesion cases, used to characterise variability in mask accuracy.

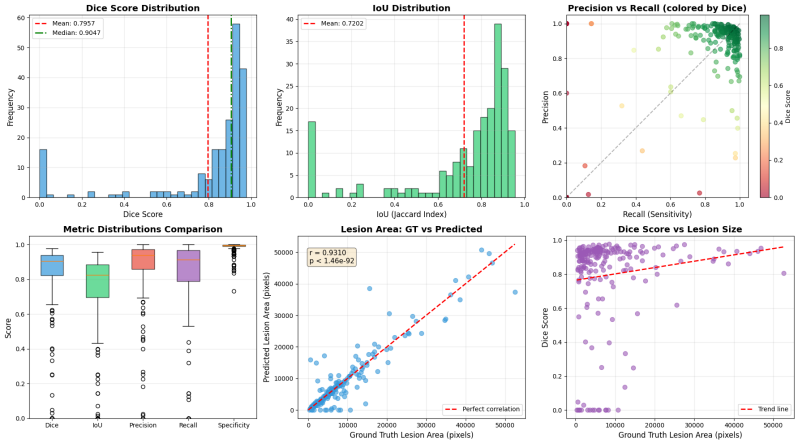

Figure 26: Distribution of segmentation metrics for Pipeline B, including Dice, IoU, precision, recall, and specificity.

Error propagation analysis. Relationship between segmentation quality and downstream classification errors, used to assess error propagation across pipeline stages.

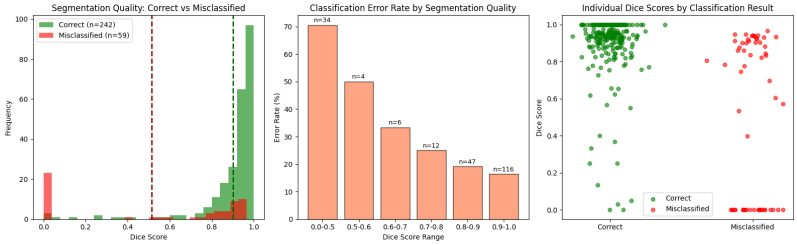

Figure 27: Error propagation analysis in Pipeline B, showing the association between Dice score and classification correctness.

Qualitative ROI-based classification examples. Representative examples illustrating ROI extraction, segmentation quality, and classification outcome.

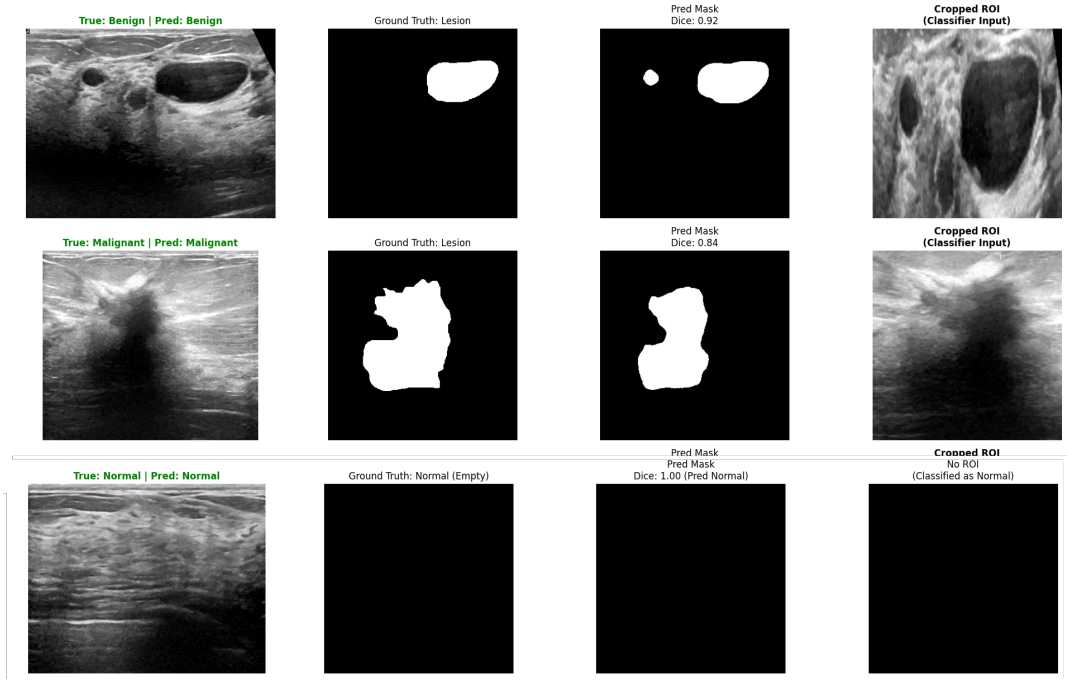

Figure 28: Qualitative examples for Pipeline B, showing ultrasound image, ground-truth mask, predicted mask, and cropped ROI used as classifier input for benign, malignant, and normal cases.

Grad-CAM interpretability analysis. Visualisation of class-discriminative regions using Grad-CAM on ROI-based classifier inputs.

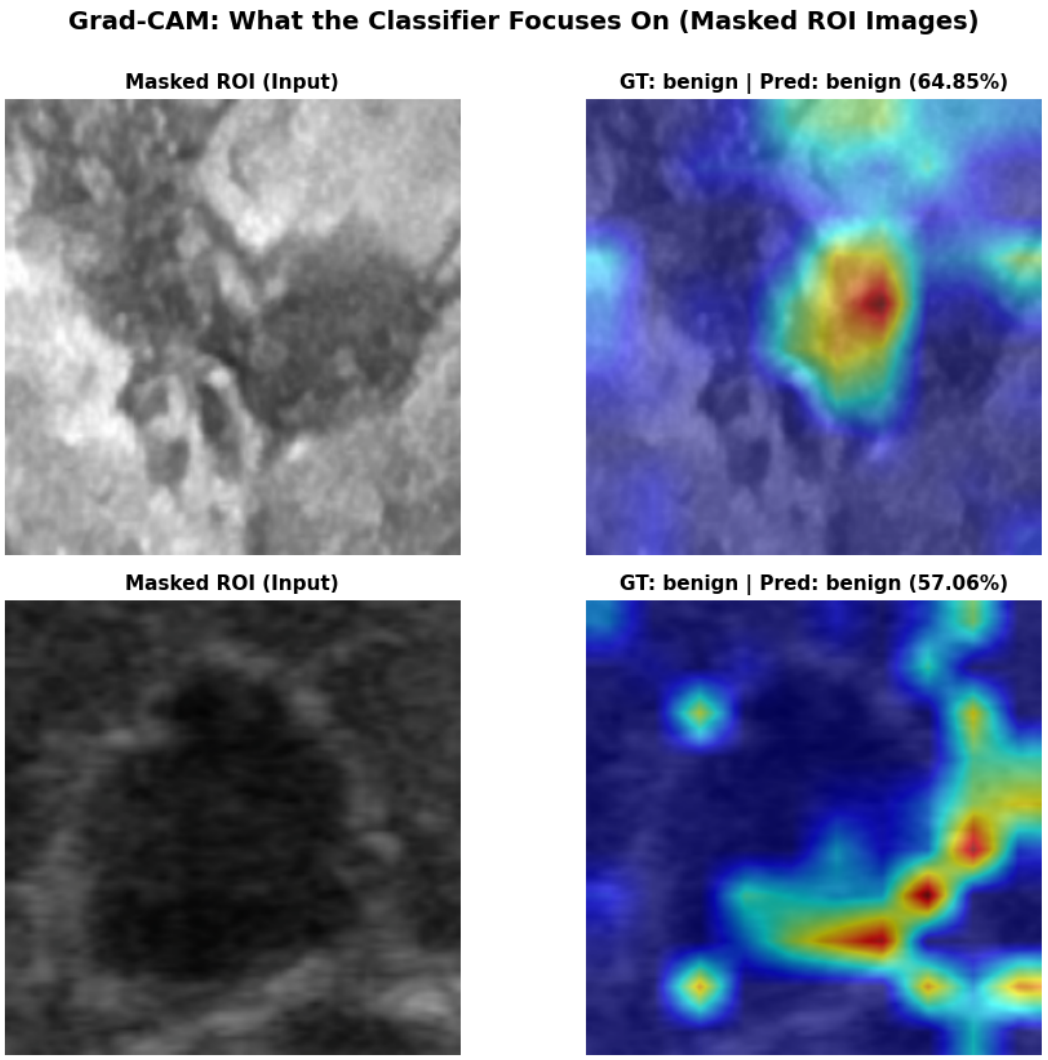

Figure 29: Grad-CAM visualisations for Pipeline B, illustrating classifier attention on lesion-centred ROIs for representative benign and malignant cases.

References

[1] World Health Organization, "Breast cancer," WHO Fact Sheet, 2024.

[2] W. Al-Dhabyani, M. Gomaa, H. Khaled and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, 104863, 2020. [web:19]

[3] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019.

[4] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, MICCAI Workshops, 2018.

[6] T. Fan, G. Wang, Y. Li and H. Wang, "Ma-Net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.

[7] M. Berman, A. Rappez, S. Jégou, A. Vedaldi and R. B. G. d'Aspremont, "The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. CVPR*, 2018. [web:14]

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[10] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 125, 2020.

[11] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] P. Yakubovskiy, "Segmentation Models Pytorch," GitHub repository, 2019. Available at: `https://github.com/qubvel/segmentation_models.pytorch`. [web:17]

[13] Y. Xu, Y. Wang, H. Li and Y. Liu, "Revolutionizing breast ultrasound diagnostics with deep learning: A review," *Frontiers in Oncology*, vol. 14, 2024. [web:9]

[14] Y. Liu *et al.*, "Multi-task network for breast ultrasound diagnosis," arXiv:2401.07326, 2024. [web:12]