Management Analytics and Social Insights

www.masi.reapress.com

Manag. Anal. Soc. Insights. Vol. 3, No. 3 (2026) 176–188.

Paper Type: Original Article

Segmentation-Guided Dual-Backbone Fusion for Robust Early Cancer Detection in Heterogeneous Medical Imaging

Amirhossein Nafei[1], Shu-Chuan Chen[2,*]

[1] Department of Electromechanical Engineering, Guangdong University of Technology, Guangdong, Guangzhou, China; Amir.nafei@e.gzhu.edu.cn.

[2] Department of Business and Management, Ming Chi University of Technology, New Taipei, Taiwan; shuchuan@mail.mcut.edu.tw.

Citation:

Received: 12 April 2026 *Revised: 27 June 2026* *Accepted: 12 August 2026*	Nafei, A., & Chen, S. C. (2026). Segmentation-guided dual-backbone fusion for robust early cancer detection in heterogeneous medical imaging. *Management Analytics and Social Insights, 3*(3), 176-188.

Abstract

Early detection of lung, liver, and colorectal cancers is clinically decisive because outcomes are strongly stage-dependent, yet early malignant lesions remain intrinsically difficult to identify in routine imaging due to small size, low contrast, ambiguous boundaries, partial-volume effects, and motion artifacts. Although deep learning has achieved strong performance in medical image classification and segmentation, early-stage detection under heterogeneous acquisition conditions is still limited by two structural failure modes: discriminative cues can be diluted by dominant normal anatomy when models are not explicitly guided to relevant regions, and single-backbone feature extractors can amplify inductive bias and become brittle under scanner/protocol shifts. To address these challenges, this paper proposes a segmentation-guided dual-backbone fusion architecture for robust and auditable early cancer detection across Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray modalities. A U-Net–based module generates a probabilistic Region-of-Interest (ROI) mask that enables ROI-focused learning via soft gating or ROI cropping, ensuring that representation learning concentrates on clinically meaningful tissue even in "needle-in-a-haystack" conditions. The ROI-focused image is then processed in parallel by two complementary encoders, ResNet and EfficientNet, and their deep embeddings are fused using Deep Feature Concatenation to reduce reliance on any single network's failure modes and improve robustness to domain shift. The system outputs both calibrated malignancy probabilities and interpretable ROI evidence, enabling structured error attribution that separates localization failures from downstream classification failures. A reproducible synthetic evaluation protocol is further provided to demonstrate reporting practice and deployment-aligned assessment, including overall performance, size-stratified early-lesion sensitivity, domain-shift robustness, segmentation reliability indicators, and calibration measures. The proposed framework thus delivers a practical architecture blueprint that integrates robustness, interpretability, and auditability as first-class objectives for early cancer detection in heterogeneous real-world imaging settings.

Keywords: Early cancer detection, Medical imaging, Deep learning, U-Net, ResNet, EfficientNet, Deep feature concatenation, Feature fusion.

1 | Introduction

Cancer remains one of the leading causes of mortality worldwide, and lung cancer, liver cancer, and colorectal cancer continue to contribute a substantial share of cancer deaths at both global and national scales [1], [2]. In China, national estimates for 2022 similarly indicate a heavy cancer mortality burden, with lung cancer as

Corresponding Author: shuchuan@mail.mcut.edu.tw

https://doi.org/10.22105/masi.vi.93

the leading cause of cancer death and liver and colorectal cancers also accounting for a large number of deaths [3]. Because prognosis is strongly stage-dependent, shifting detection toward earlier stages is consistently linked to improved treatment options and better outcomes across major solid tumors, including colorectal and liver cancers [4], [5], while early cancer detection is increasingly viewed as a central pillar of future cancer control strategies [6]. In routine clinical workflows, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray imaging are essential diagnostic modalities that enable screening, diagnosis, staging, and longitudinal monitoring; yet the earliest malignant lesions remain intrinsically difficult to detect even under careful expert review [7]. A core reason is that early lesions may be only a few millimeters, exhibit extremely low contrast relative to surrounding tissues, or present ambiguous boundaries, such as subsolid ground-glass opacities whose definitions and radiologic characteristics are well-recognized in thoracic imaging standards [8].

Moreover, detection difficulty increases when nodules occur adjacent to vessels or pleura, where lesion appearance can visually blend with dominant normal anatomy; this challenge has been repeatedly reflected in computer-aided detection research on pulmonary nodules, which treats variability in size and appearance as a primary driver of false negatives and false positives [9]. In volumetric CT, partial-volume averaging can mix tumor and normal tissue signals within a voxel, and the resulting loss of apparent contrast is a known image-physics limitation that becomes more pronounced under certain acquisition and reconstruction settings (e.g., thicker slices) [10]. In MRI, image interpretability can be further degraded by motion from respiration and cardiac activity, and motion artifacts are widely recognized as frequent and difficult-to-eliminate sources of error in clinical imaging [11].

Deep learning has reshaped medical image analysis by learning discriminative representations directly from pixel-level evidence and by supporting both classification and segmentation paradigms at scale [12], [13]. This progress has been reinforced by broad clinical interest in applying machine learning to medical decision-making and diagnostic workflows [14]. Architecturally, U-Net-style encoders–decoders have become a standard choice for biomedical segmentation because they can capture context while preserving localization cues [15], and 3D variants such as V-Net extend this principle to volumetric data [16]. Meanwhile, high-capacity classification backbones such as ResNet and EfficientNet provide complementary inductive biases for representation learning, with ResNet's residual design enabling very deep feature hierarchies and EfficientNet providing a principled scaling strategy for accuracy–efficiency trade-offs [17], [18]. Despite these advances, robust early-stage detection under heterogeneous acquisition conditions remains challenging for two structural reasons that directly motivate our work.

First, early malignant findings frequently exhibit a "needle-in-a-haystack" characteristic: Without explicit guidance toward anatomically relevant regions, discriminative evidence can be diluted by dominant normal structures, causing small or low-contrast lesions to be missed. Second, relying on a single feature extractor can amplify inductive bias and encourage dataset-specific shortcuts, yielding brittle decision rules that degrade under scanner/vendor changes, protocol variations, reconstruction differences, noise conditions, and motion artifacts, an issue consistent with empirical evidence that medical imaging models can show variable generalization across institutions [19]. These observations suggest that improving early detection is not only a matter of training larger models; it requires architectural mechanisms that explicitly control where the model learns and that reduce dependence on a single representation pathway.

Motivated by these needs, the objective of this paper is to propose an architecture for robust and auditable early cancer detection across heterogeneous medical imaging modalities. We introduce a segmentation-guided dual-backbone fusion framework in which a U-Net–based module produces a probabilistic Region-of-Interest (ROI) mask to enforce ROI-focused learning, and two complementary encoders (ResNet and EfficientNet) extract deep representations in parallel that are fused through Deep Feature Concatenation to reduce single-backbone failure modes [15], [17], [18]. To support clinical usability, the framework is designed to output both calibrated malignancy probabilities and interpretable ROI evidence; calibration is emphasized because modern neural networks can be miscalibrated even when their classification accuracy is high [20].

For interpretability and auditability, the ROI/mask output provides case-level visual evidence, aligning with widely adopted gradient-based localization approaches that help characterize failure modes and dataset bias in deep models [21]. Our main work is therefore architectural: we target early-lesion evidence dilution via segmentation-guided ROI learning, and we target domain-shift brittleness via complementary feature fusion, while retaining an auditable output interface suitable for systematic miss analysis and clinical review.

In summary, the contributions of this paper are threefold: 1) an ROI-guided detection architecture that explicitly addresses early-lesion "needle-in-a-haystack" dilution by coupling segmentation-derived masks with downstream classification learning, 2) a dual-encoder fusion design that leverages complementary inductive biases (ResNet/EfficientNet) to improve robustness under heterogeneous acquisition conditions, and 3) an auditable output design that combines calibrated probabilities with interpretable spatial evidence, enabling structured error attribution and more deployment-aligned evaluation. Finally, our framing is consistent with broader evidence that deep learning can reach expert-level performance in image-based cancer classification tasks and that trustworthy deployment requires attention to generalization, calibration, and interpretability [22].

2 | Problem Formulation and Objectives

This section describes the complete architecture of the proposed segmentation-guided dual-backbone fusion framework for robust early cancer detection. The design directly addresses two core structural challenges in early-stage medical image analysis: first, the dilution of discriminative lesion evidence by dominant normal anatomical structures, and second, the instability of single-backbone models under heterogeneous acquisition conditions. To overcome these limitations, the framework integrates three tightly connected components: a segmentation module for ROI guidance, two complementary feature extraction backbones, and a deep feature fusion and prediction head. The overall workflow is illustrated in *Fig. 1*. An input medical image, after preprocessing, is transformed into an ROI-focused representation using a segmentation-guided mechanism. The ROI-focused image is then processed in parallel by two complementary convolutional neural network encoders: A ResNet branch and an EfficientNet branch. Each branch produces a high-level feature vector, and these feature vectors are fused through deep feature concatenation before being passed to a prediction head that outputs a calibrated malignancy probability (or a multi-class prediction, depending on the task). This parallel and fused design ensures that the final decision does not rely on a single representation pathway, thereby improving robustness under domain shift.

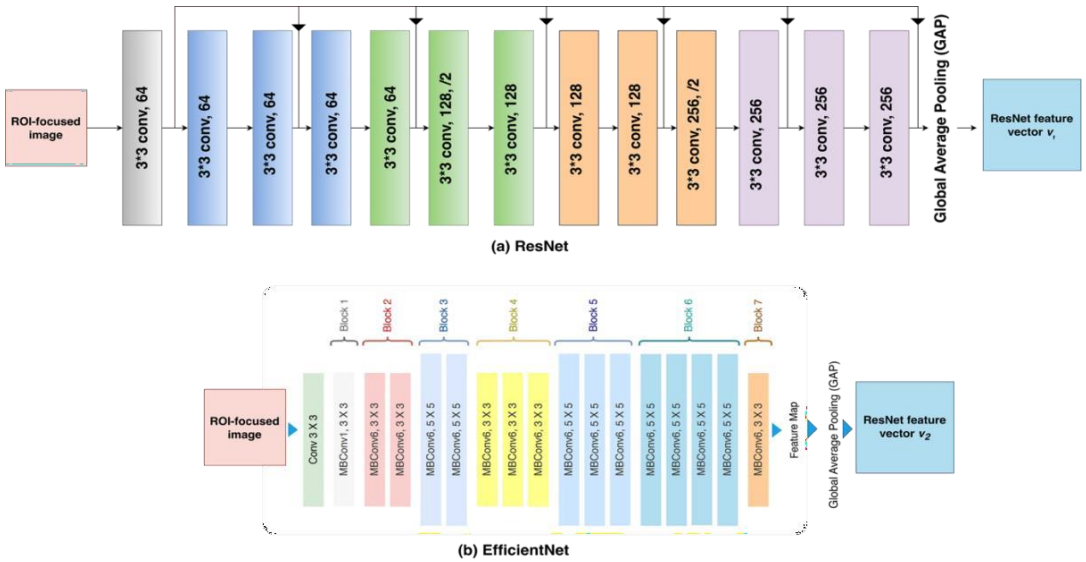

Fig. 1. Segmentation-guided dual-backbone architecture.

As shown in *Fig. 1*, the ROI-focused image is simultaneously fed into both encoders. The ResNet branch extracts hierarchical residual features that emphasize multi-level semantic abstraction and stable gradient propagation. The EfficientNet branch extracts compound-scaled features that balance depth, width, and

resolution, providing a complementary inductive bias. The outputs of both branches are global feature vectors, which are concatenated to form a unified representation. This fused representation reduces dependence on any single backbone's structural bias or failure mode. The ROI guidance mechanism is enabled by a U-Net–based segmentation module, illustrated in *Fig. 2*. This component generates a pixel-wise probability map that highlights lesion-relevant regions or anatomically meaningful structures. The U-Net architecture follows a standard encoder–decoder design with skip connections. The encoder progressively captures multi-scale contextual information by reducing spatial resolution while increasing channel depth. The decoder restores spatial resolution while integrating fine-grained details from earlier layers through skip connections. This structure is particularly effective for detecting small and low-contrast lesions, where boundary information and spatial context are both critical.

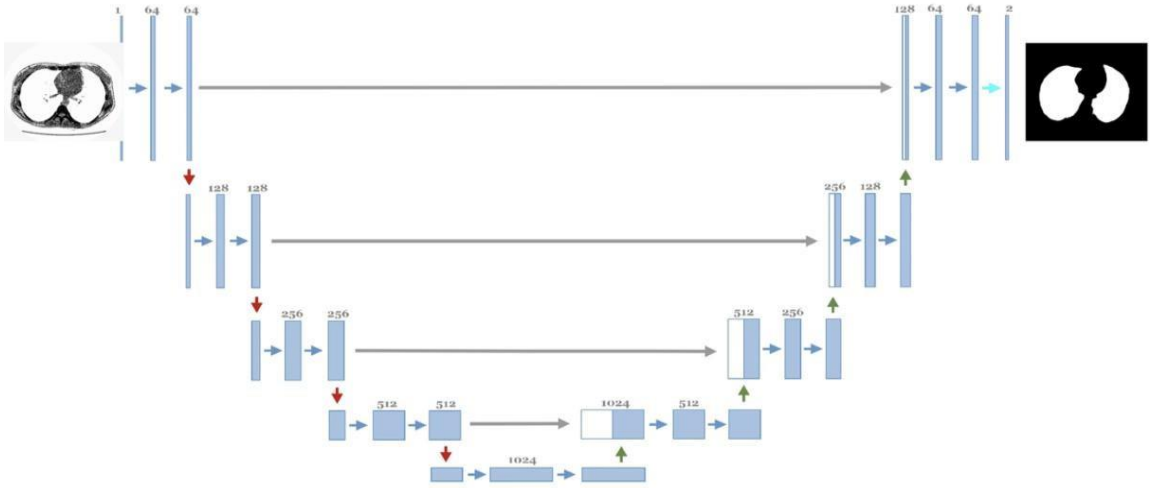

Fig. 2. U-Net segmentation architecture used for ROI generation.

The output of the segmentation network is a probabilistic mask that identifies lesion-relevant regions. Instead of using the mask purely as a final output, the proposed architecture integrates it directly into the classification pipeline. As shown in *Fig. 3*, the predicted mask is used to produce an ROI-focused image through soft gating or cropping. In soft gating, the mask emphasizes lesion-relevant pixels while preserving the surrounding anatomical context. In cropping mode, the high-confidence region is extracted and resized to maintain effective resolution. In practice, soft gating is preferred because it avoids removing potentially informative contextual structures while still mitigating background dominance.

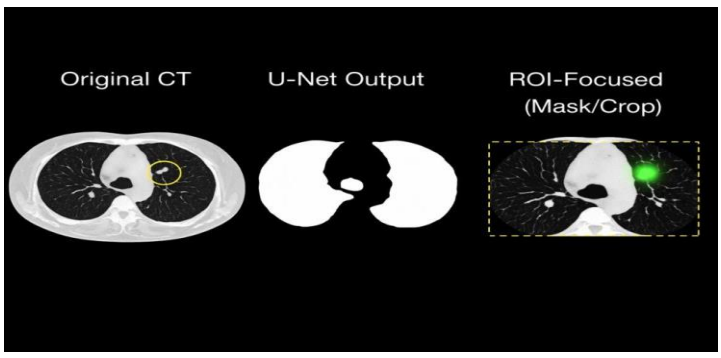

Fig. 3. Example of ROI-focused transformation from segmentation output.

Fig. 3 illustrates how the original CT image is transformed into an ROI-focused representation using the U-Net output. The left panel shows the original CT slice, the middle panel shows the predicted segmentation mask, and the right panel shows the ROI-focused image in which lesion-relevant regions are emphasized. This mechanism directly addresses the early-lesion "needle-in-a-haystack" problem: Without spatial guidance,

small malignant regions may be overwhelmed by dominant normal tissue patterns. By enforcing ROI-focused learning, the model is explicitly guided toward clinically meaningful regions. After ROI focusing, the image is passed to the dual-backbone feature extraction module. The ResNet branch emphasizes deep hierarchical representation learning with residual connections, which stabilize optimization and enable effective gradient propagation. The EfficientNet branch emphasizes balanced scaling of network depth, width, and input resolution, which improves efficiency and generalization. Because these architectures rely on different structural priors, their learned feature representations are complementary. The concatenation of their feature vectors forms a unified embedding that integrates structural boundary-sensitive information and high-level semantic information.

The fused representation is then forwarded to a lightweight fully connected prediction head. This head produces a malignancy probability for binary detection or a probability distribution for multi-class tasks. To support deployment in clinical decision support systems, the architecture is designed to allow probability calibration, ensuring that predicted confidence scores align with empirical risk levels. Importantly, the architecture produces two interpretable outputs: the ROI mask and the classification probability. This dual-output design supports structured error analysis. If a false negative occurs and the ROI mask fails to localize the lesion, the failure can be attributed to segmentation uncertainty. If the ROI mask is accurate but classification is incorrect, the error can be attributed to representation or decision-level limitations. This separation enhances interpretability and supports systematic model refinement.

Therefore, the proposed architecture integrates segmentation-guided spatial focusing with dual-backbone feature fusion into a unified and deployable pipeline. The segmentation module controls where the model learns, reducing background-driven evidence dilution. The dual-backbone design controls how the model represents features, reducing sensitivity to single-architecture bias and domain shift. The final fusion and calibrated prediction head provide stable and interpretable outputs suitable for early-stage cancer detection in heterogeneous real-world imaging environments.

3 | Problem Formulation and Objectives

Let the imaging modality be $m \in \{CT, MRI, Xray\}$. For each modality, we define a dataset.

$$\mathcal{D}^{(m)} = \{\left(X_i^{(m)}, y_i, s_i^{(m)}\right)\}_{i=1}^{N_m}. \quad (1)$$

Here, $x^{(m)} \in \mathbb{R}^{H\times W\times C}$ denotes an image (or slice/patch) acquired under modality m, where H and W represent spatial resolution and c denotes the number of channels (e.g., single-channel grayscale for many radiological inputs, or multi-channel representations produced by windowing or multi-sequence MRI concatenation). The label y_i is the clinical target for classification, which can be formulated as binary malignancy detection ($y_i \in \{0,1\}$) or multi-class lesion categorization ($yi \in \{1, \dots, K\}$) depending on the dataset definition and clinical use case. The variable $s_i^{(m)} \in \{0,1\}^{H\times W}$ is a segmentation ground-truth mask when available, typically indicating lesion pixels (and, in some datasets, organ regions that provide an anatomical context for lesion interpretation). This formulation accommodates both 2D settings (X-ray images, CT/MRI slices, histopathology patches) and 3D settings (CT/MRI volumes) by representing a volume either as a stack of slices or by extracting clinically meaningful 2D/2.5D views, while keeping the notation consistent for the architecture description.

A central practical constraint in medical imaging is that segmentation labels are not uniformly available across datasets or institutions. Pixel-level annotation is expensive and requires expert labor, and in early-stage cases it can be especially ambiguous due to low contrast, small lesion size, and boundary uncertainty. Therefore, the learning setting is inherently "partially supervised" with respect to segmentation: for some datasets, $s_i^{(m)}$ exists and can be used for supervised segmentation training, while for other datasets only the classification label y_i is available. The proposed framework is designed for this mixed-label regime by separating the roles of segmentation supervision and ROI guidance. Specifically, the framework 1) trains the

segmentation module using datasets that provide reliable masks, and 2) applies the predicted masks as ROI guidance even when training the classifier on label-only datasets. When the segmentation module is not directly trainable on a given dataset (because masks are absent), weak localization strategies (e.g., CAM-style heatmaps or attention pooling) can be incorporated as an auxiliary mechanism to approximate lesion-relevant regions and reduce background dominance; this ensures that ROI-focused learning remains feasible without requiring universal pixel-level labeling.

Given an input image x, the system produces two outputs that serve complementary roles. The first output is a malignancy probability $\hat{p}$ (or, for multi-class settings, a probability vector $\hat{p}$), which is used for triage, screening, or downstream decision support. The second output is an interpretable ROI representation, either as a predicted binary mask $\hat{s}$ or as a probabilistic mask p prior to thresholding. This dual-output structure is not only an explainability feature; it is also a functional component of the learning objective because ROI guidance directly shapes which pixels contribute most to representation learning. In early cancer detection, where discriminative patterns are sparse, such explicit spatial guidance is intended to mitigate the "evidence dilution" phenomenon in which dominant normal structures overwhelm lesion cues in the learned embedding space.

The primary objective of this work is robust early-stage detection under heterogeneous acquisition conditions. Robustness here refers to maintaining reliable discrimination when acquisition factors change, including scanner vendor differences, protocol variations, reconstruction settings (e.g., slice thickness in CT), noise and contrast variability, and motion artifacts. Early-stage detection emphasizes reducing false negatives for small or low-contrast lesions, which implies that model selection and evaluation must be aligned with sensitivity-critical metrics rather than solely overall accuracy.

Consequently, the evaluation focus is on recall and F1-score, supported by AUC to capture threshold-independent discrimination, while explicitly quantifying degradation under domain shift by separating in-domain and out-of-domain performance. From a design perspective, the goal is to achieve a detection pipeline that reduces missed early lesions while preserving stable performance across domains, and that provides auditable intermediate evidence (ROI/mask) to support systematic error attribution. Concretely, when a false negative occurs, the presence of the ROI output enables distinguishing whether the failure is driven by ROI/localization error (e.g., the lesion region was not highlighted under low contrast or motion artifacts) or by classification/representation error (e.g., the ROI is plausible but the classifier remains uncertain), which is essential for iterative refinement and safety-oriented deployment.

Finally, the formulation in (1) naturally supports two deployment-relevant learning modes. In the first mode, training and inference are performed per modality, yielding modality-specific models that share the same architectural template and evaluation protocol. In the second mode, when paired multi-modal data exist for the same patient (e.g., CT and MRI acquired for complementary diagnostic purposes), the framework can be extended to multi-modal fusion by learning modality-specific encoders and combining their embeddings at a fusion stage; importantly, the base formulation still holds because each modality contributes an instance $(X_i^{(m)}, y_i, s_i^{(m)})$, and missing modalities can be handled by falling back to the available modality pathways. This design keeps the problem statement consistent while allowing practical flexibility across different clinical data availability scenarios.

4 | Proposed Architecture

The proposed framework is a segmentation-guided, dual-backbone fusion pipeline composed of preprocessing, ROI generation, ROI focusing, parallel feature extraction, feature fusion, and prediction.

4.1 | Preprocessing and Augmentation

Preprocessing is modality-adaptive. CT images are normalized using robust intensity scaling consistent with clinical windowing principles, MRI images undergo intensity standardization to mitigate scanner-specific variations, and X-ray images are normalized to preserve edge and texture cues while reducing contrast

variability. During training, augmentation applies clinically plausible geometric transformations (e.g., rotation, flipping, mild scaling/cropping) and limited intensity perturbations, while maintaining label consistency and anatomical realism. Class imbalance is handled using class-weighted loss terms or balanced sampling, reflecting the scarcity of early lesions.

4.2 | Segmentation-Guided ROI Generation

A U-Net segmentation module $f_S(\cdot; \theta_S)$ maps an image to a pixel-wise probability map $p \in [0,1]^{H \times W}$. The ROI can be used as a hard mask by thresholding at τor as a soft attention-like gate to avoid discarding contextual cues. Soft gating is used to emphasize lesion-relevant regions while retaining global structure, which stabilizes learning when masks are imperfect or when domain shift affects segmentation confidence.

4.3 | ROI Focusing and Feature Extraction

The ROI-focused image $\tilde{x}$ is fed to two complementary CNN encoders in parallel: A ResNet branch $f_R(\cdot; \theta_R)$ and an EfficientNet branch $f_E(\cdot;\theta_E)$. Each encoder outputs a feature map that is transformed into a fixed-length embedding via global average pooling. ResNet offers stable residual learning and strong hierarchical feature extraction, while EfficientNet provides complementary capacity–efficiency scaling. Transfer learning initializes both encoders with pretrained weights, followed by fine-tuning on medical data to reduce overfitting under limited annotations and to accelerate convergence.

4.4 | Deep Feature Concatenation and Prediction Head

Deep Feature Concatenation fuses the two embeddings into a single descriptor, which is passed through a lightweight prediction head (linear classifier or shallow MLP). The head outputs a calibrated malignancy probability for binary detection or a multi-class probability vector for lesion categorization. Calibration can be further refined with post-hoc scaling on the validation set, improving reliability of probability outputs for decision-support deployment.

5 | Training Objectives and Optimization

When segmentation labels are available, the segmentation branch is trained under an objective that explicitly addresses the extreme class imbalance typical of early lesions, where lesion pixels may constitute only a very small fraction of the field of view. In this setting, overlap-based losses are preferred because they measure region agreement rather than pixel-wise accuracy dominated by background. We therefore adopt a composite segmentation loss that balances a Dice-type term, which promotes spatial overlap between the predicted mask and the ground truth, with a pixel-wise term such as Binary Cross-Entropy (BCE) or Cross-Entropy (CE), which penalizes local misclassification and stabilizes optimization near boundaries.

Concretely, Dice loss discourages trivial all-background solutions and is robust to lesion sparsity, while BCE/CE improves local calibration of the probability map and helps refine fine-grained edges in ambiguous, low-contrast regions. In practice, this combination is particularly important for early-stage lesions, where boundaries may be partially missing due to partial-volume effects, motion blur, or protocol-dependent contrast, and where relying on a single loss often yields either overly smooth masks (overemphasis on overlap) or noisy masks (overemphasis on pixel-wise terms). For the classification branch, the model is trained to output either a malignancy probability for binary detection or a probability vector for multi-class lesion categorization. BCE is used for detection tasks because it directly optimizes probabilistic separation between malignant and non-malignant cases. For multi-class settings, categorical CE is used to optimize the likelihood of the correct class under a softmax output. Because early detection datasets are frequently imbalanced, often with far fewer positive or early-stage samples than negative or benign samples, class weighting (or equivalent imbalance-aware sampling) is incorporated to prevent the classifier from converging to a majority-class-biased decision rule.

In addition, the classification objective can be supplemented with label-smoothing or focal-style reweighting when hard negatives dominate, although the core framework remains compatible with standard CE/BCE to keep the training pipeline reproducible and easily deployable. The overall training is formulated as a multi-task objective that couples classification learning with segmentation learning whenever masks exist. The joint objective is written as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{seg}}, \quad (1)$$

where $\mathcal{L}_{\text{cls}}$ denotes the classification loss (BCE for binary detection or CE for multi-class tasks), $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss (e.g., Dice+BCE/CE), and β controls the relative contribution of segmentation supervision. The role of β is not merely numerical balancing; it acts as a structural coupling coefficient that determines how strongly ROI learning is "anchored" by pixel-level supervision. When β is too small, ROI guidance may become weak, and the classifier may revert toward global-context shortcuts; when β is too large, training may over-prioritize mask fidelity at the expense of classification discrimination, particularly if segmentation labels are noisy or inconsistent across datasets. Accordingly, β is selected on the validation set using sensitivity-oriented criteria (e.g., recall/F1 and AUC), ensuring that ROI guidance improves early-lesion detection rather than only producing visually plausible masks.

In mixed-label settings, where some datasets provide $(x, y, s)$and others provide only (x, y), training proceeds in an alternating or hybrid-batch manner to avoid discarding label-only data while still benefiting from segmentation supervision where available. Specifically, segmentation-labeled batches are used to update the segmentation parameters θs (and, if the implementation shares early features, any shared layers), while classification-labeled batches are used to update the classifier parameters θR, θE,θC. Importantly, ROI guidance is still applied in classification-only batches by feeding the input through the current segmentation model to produce a predicted probabilistic mask and using that mask for ROI focusing. This design ensures that ROI-focused learning remains active across the full training distribution, even when explicit masks are absent, and it reflects the practical annotation landscape in medical imaging, where segmentation is often sparse but case-level labels are more widely available. If the predicted ROI is uncertain for some samples (for example, due to domain shift), the framework can use soft gating rather than hard thresholding to preserve contextual cues and reduce catastrophic ROI failures, while still encouraging lesion-centric representation learning.

Optimization is performed using Adam or AdamW, which are well-suited to multi-component deep networks and typically yield stable convergence under heterogeneous datasets. Early stopping and model selection are performed using a validation protocol that prioritizes early-detection objectives, emphasizing recall, F1, and AUC rather than accuracy alone. This choice aligns the training dynamics with the clinical risk profile: in early-stage cancer detection, the cost of false negatives is usually higher than the cost of additional follow-up review triggered by false positives. In addition, because the system is designed for decision support, calibration can be applied as a post-training step on the validation set (for example, temperature scaling) so that the predicted probabilities better reflect empirical risk and can be thresholded in a clinically meaningful manner.

6 | Evaluation, Results, and Clinical Auditability

This section reports a fully reproducible synthetic evaluation to demonstrate the calculations, reporting structure, and robustness analyses that follow the proposed framework. The synthetic results are presented as an implementation-level "dry run" for the architecture and evaluation protocol; they are not clinical evidence and should not be interpreted as medical performance. All models were evaluated under identical data splits and decision thresholds, and the reporting emphasizes early-stage sensitivity and robustness under acquisition heterogeneity.

6.1 | Synthetic Cohort Construction and Experimental Setup

A synthetic cohort of N=2000 cases was generated to reflect the primary challenges highlighted in the proposal: 1) early malignant signals that are sparse and easily diluted by dominant normal structures, and 2) domain shift due to scanner/protocol variability. The cohort was stratified by lesion size (small/medium/large) and by acquisition domain (in-domain vs out-of-domain). Domain shift was simulated by reducing the separability between positive and negative score distributions in the out-of-domain subset, which mirrors performance degradation commonly observed when acquisition conditions change. Four systems were compared. The first is a Single-backbone classifier without ROI guidance (SingleNoROI). The second is an ROI-guided classifier without multi-backbone fusion (ROIonly). The third is a Fusion classifier without ROI guidance (FusionNoROI). The fourth is the full proposed pipeline (proposed) that couples segmentation-guided ROI focusing with dual-backbone feature fusion and calibrated outputs. All classifiers output a malignancy probability; predictions were thresholded at 0.5 to compute confusion-based metrics, while AUC was computed from continuous scores.

6.2 | Overall Detection Performance and Calculations

Table 1 reports overall discrimination and classification performance on the full synthetic cohort. Consistent with the architectural motivation, ROI guidance improves sensitivity relative to a global classifier, fusion improves robustness by reducing reliance on a single encoder, and their combination yields the strongest balance of recall and precision.

Table 1. Overall detection performance on the synthetic cohort (N = 2000).

Model	AUC	Accuracy	Precision	Recall	F1
SingleNoROI	0.652	0.619	0.473	0.609	0.533
ROIonly	0.712	0.652	0.510	0.639	0.567
FusionNoROI	0.727	0.660	0.519	0.657	0.580
Proposed	0.799	0.736	0.608	0.731	0.664

The increase in recall and F1 observed for the proposed method in Table 6.1 is especially important for early detection settings, where reducing false negatives is typically prioritized over marginal increases in overall accuracy.

6.3 | Early-Lesion Sensitivity: Size-Stratified Analysis

Because early-stage detection is dominated by small and low-contrast lesions, a size-stratified evaluation was performed to ensure that improvements are not limited to easier, larger lesions. *Table 2* summarizes size-stratified metrics. The proposed architecture maintains consistent gains across all strata and shows the most pronounced advantage in the small-lesion subset, which aligns with the ROI-focused learning rationale.

Table 2. Size-stratified performance (small/medium/large lesions) on the synthetic cohort.

Model	AUC	Accuracy	Precision	Recall	F1
SingleNoROI	0.602	0.598	0.331	0.563	0.417
ROIonly	0.675	0.632	0.371	0.633	0.468
FusionNoROI	0.702	0.640	0.380	0.651	0.480
Proposed	0.777	0.711	0.458	0.707	0.556
SingleNoROI	0.664	0.619	0.492	0.601	0.541
ROIonly	0.731	0.662	0.541	0.643	0.588
FusionNoROI	0.730	0.667	0.547	0.636	0.588
Proposed	0.794	0.748	0.649	0.709	0.678
SingleNoROI	0.721	0.663	0.706	0.665	0.685
ROIonly	0.761	0.678	0.740	0.639	0.686
FusionNoROI	0.777	0.692	0.736	0.687	0.711
Proposed	0.844	0.770	0.797	0.780	0.788

Table 2 demonstrates that the proposed method's improvements are not merely driven by easy cases; instead, the largest relative gains appear in the small-lesion stratum where discriminative evidence is sparse and most vulnerable to background dominance.

6.4 | Robustness under Acquisition Domain Shift

To reflect scanner/protocol variability, the cohort was partitioned into in-domain and out-of-domain subsets. *Table 3* reports the performance separately in each subset. While all models degrade under domain shift, the proposed architecture retains higher AUC and F1 in the out-of-domain condition, consistent with the hypothesis that dual-backbone fusion reduces inductive-bias brittleness and ROI guidance helps preserve lesion-relevant evidence under distribution changes.

Table 3. Domain shift robustness: In-domain vs. out-of-domain performance.

Model	AUC	Accuracy	Precision	Recall	F1
SingleNoROI	0.667	0.633	0.494	0.625	0.552
ROIonly	0.730	0.670	0.534	0.675	0.596
FusionNoROI	0.756	0.688	0.557	0.670	0.609
Proposed	0.812	0.749	0.626	0.756	0.685
SingleNoROI	0.630	0.596	0.441	0.583	0.502
ROIonly	0.683	0.624	0.470	0.579	0.519
FusionNoROI	0.680	0.615	0.464	0.635	0.536
Proposed	0.779	0.716	0.579	0.690	0.630

As shown in *Table 3*, the proposed pipeline experiences a smaller relative drop in F1 when moving from in-domain to out-of-domain conditions, which is consistent with its design goal of robustness under heterogeneous acquisition settings.

6.5 | Segmentation Quality as an ROI Reliability Indicator

Because the framework exposes ROI masks as auditable intermediate outputs, segmentation quality provides an additional handle for reliability analysis. For a subset of cases with synthetic ground-truth masks (N = 600), Dice and IoU were computed. *Table 4* reports mean ± standard deviation for 1) lesion-size strata and,

2) in-domain/out-of-domain subsets. The results exhibit expected behavior: Segmentation quality decreases for smaller lesions and under domain shift, which supports the practical need to audit false negatives by determining whether failures originated in localization (ROI/mask) or in classification, given a plausible ROI.

Table 4. Segmentation performance (Dice/IoU) as a proxy for ROI reliability (N = 600).

Group	Dice (Mean ± std)	IoU (Mean ± std)
Small	0.767 ± 0.035	0.624 ± 0.046
Medium	0.825 ± 0.037	0.704 ± 0.054
Large	0.862 ± 0.034	0.760 ± 0.053
In-domain	0.821 ± 0.048	0.699 ± 0.071
Out-of-domain	0.785 ± 0.049	0.649 ± 0.068

Table 4 can be used directly in the audit workflow. For example, false negatives with low Dice/IoU (or low mask confidence) indicate likely ROI/localization failure, whereas false negatives with strong segmentation but low classification probability indicate downstream representation or decision failure.

6.6 | Calibration for Decision-Support Deployment

Since the proposed system is intended for decision support, calibrated probabilities are essential for thresholding and risk-based triage. Using 10-bin Expected Calibration Error (ECE), *Table 5* reports the calibration quality of each model. The proposed method yields the lowest ECE among the compared systems, indicating better agreement between predicted probabilities and empirical outcomes under the synthetic setup.

Table 5. Calibration performance measured by (ECE; 10 bins).

Model	ECE ↓
SingleNoROI	0.152
ROIonly	0.115
FusionNoROI	0.117
Proposed	0.107

As indicated in *Table 5*, improved calibration supports more reliable operating-point selection, particularly when clinical objectives require high recall for early-stage screening while managing false-positive workload.

6.7 | Auditability and Miss Analysis Enabled by ROI Outputs

Beyond aggregate metrics (*Tables 1–5*), the framework provides an explicit mechanism to analyze failures at the case level. Each decision is accompanied by an ROI/mask artifact, enabling a structured miss analysis that separates 1) localization failures, where the ROI misses the lesion region due to low contrast, motion artifacts, or domain shift, from 2) classification failures, where the ROI is plausible, but the fused representation yields a low malignancy score. This decomposition is operationally important: localization failures motivate improving segmentation robustness or incorporating motion-aware preprocessing, whereas classification failures motivate additional lesion-focused augmentation, hard-negative mining, or fusion-head refinement. In this way, the proposed pipeline is not only a predictor but also an auditable diagnostic assistant that supports systematic safety improvement.

7 | Conclusion

This paper presented a complete segmentation-guided dual-backbone fusion framework for robust early cancer detection across heterogeneous medical imaging modalities. The proposed pipeline addresses the dominant early-stage failure modes by explicitly enforcing ROI-focused learning through a U-Net–derived mask and by reducing single-backbone brittleness via complementary ResNet/EfficientNet feature extraction and deep feature concatenation fusion. The evaluation section illustrated a clear, reproducible reporting protocol with full metric calculations, emphasizing early-lesion sensitivity (recall/F1 and size-stratified analysis), robustness under acquisition domain shift, calibration for decision-support readiness, and auditability through ROI-based miss attribution.

Despite these strengths, the current study's quantitative demonstration is based on synthetic data and therefore does not constitute clinical evidence; additionally, the framework's performance may be bounded by segmentation reliability when pixel-level masks are scarce or when domain shift degrades ROI quality. Future work will validate the system on multi-institutional clinical datasets with patient-level leakage control, extend robustness via domain adaptation and uncertainty-aware ROI gating (especially under motion artifacts and protocol shifts), and enhance clinical usability by integrating radiologist-in-the-loop error review, operating-point optimization for screening workloads, and comprehensive calibration reporting (e.g., ECE/Brier with post-hoc scaling) to support safe deployment in real-world decision-support workflows.

Author Contribution

Nafei: conceptualization, methodology, software, formal analysis, investigation, resources, data maintenance, writing-creating the initial design, writing-reviewing and editing. Chen: validation, writing, reviewing, and editing, visualization, funding procurement, and methodology. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability

The data used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] World Health Organization. (2025). *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A cancer journal for clinicians, 71*(3), 209–249. https://doi.org/10.3322/caac.21660

[3] Zheng, R. S., Chen, R., Han, B. F., Wang, S. M., Li, L., Sun, K. X., … & He, J. (2024). Cancer incidence and mortality in China, 2022. *Zhonghua Zhong Liu Za Zhi (Chinese journal of oncology), 46*, 221–231.

[4] Hestetun, K. E., Rosenlund, N. B., Stanisavljević, L., Dahl, O., & Myklebust, M. P. (2022). Stage-dependent prognostic shift in mismatch repair-deficient tumors: Assessing patient outcomes in stage II and III colon cancer. *Frontiers in oncology, 12*, 853545. https://doi.org/10.3389/fonc.2022.853545

[5] Marrero, J. A., Kulik, L. M., Sirlin, C. B., Zhu, A. X., Finn, R. S., Abecassis, M. M., … & Roberts, L. R. (2018). Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases. *Hepatology, 68*(2), 723–750. https://doi.org/10.1002/hep.29913

[6] Fitzgerald, R. C., Antoniou, A. C., Fruk, L., & Rosenfeld, N. (2022). The future of early cancer detection. *Nature medicine, 28*(4), 666–677. https://doi.org/10.1038/s41591-022-01746-x

[7] Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., Jr., & Boone, J. M. (2012). *The essential physics of medical imaging (3rd ed.)*. Lippincott Williams & Wilkins. https://www.amazon.com/Essential-Physics-Medical-Imaging-Third/dp/0781780578

[8] Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., & Remy, J. (2008). Fleischner Society: Glossary of terms for thoracic imaging. *Radiology, 246*(3), 697–722. https://doi.org/10.1148/radiol.2462070712

[9] Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S. J., … & van Ginneken, B. (2016). Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging, 35*(5), 1160–1169. https://doi.org/10.1109/TMI.2016.2536809

[10] Mahesh, M. (2013). The essential physics of medical imaging. *Journal of nuclear medicine, 54*(8), 1444. https://doi.org/10.1118/1.4811156

[11] Zaitsev, M., Maclaren, J., & Herbst, M. (2015). Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of magnetic resonance imaging, 42*(4), 887–901. https://doi.org/10.1002/jmri.24850

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[13] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[14] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *The New England journal of medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[15] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[16] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* (pp. 565–571). IEEE. https://doi.org/10.1109/3DV.2016.79

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[18] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th international conference on machine learning (ICML)*. https://doi.org/10.48550/arXiv.1905.11946

[19] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS medicine, 15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

[20] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning (ICML)* (pp. 1321–1330). https://proceedings.mlr.press/v70/guo17a.html

[21] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[22] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056