Informatics in Medicine Unlocked 63 (2026) 101754

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Lightweight multiscale feature refinement network for breast cancer histopathology segmentation

Zaka-Ud-Din Muhammad *, Vincenzo Della Mea

Department of Mathematics, Computer Science and Physics, University of Udine, Udine, 33100, Friuli–Venezia Giulia, Italy

ARTICLE INFO

Keywords:
Breast cancer segmentation
WSI segmentation
Computer vision
Tumor segmentation
Medical image processing

ABSTRACT

The likelihood of developing breast cancer (BC) in women is approximately 12.24%, the highest among all cancers, and clinicians are adopting AI-based technologies to improve early diagnosis and reduce the mortality rate. The present study focuses on histopathological image analysis, particularly the segmentation of Gigapixel Whole Slide Images (WSIs). However, designing a clinically deployable model requires careful consideration to achieve high segmentation accuracy, fast inference speed, and a reduced model size. To address these requirements, this study introduces LightMuSeg, a novel WSI segmentation framework. LightMuSeg utilizes a pre-trained MobileNet-V2 encoder for feature extraction, thereby reducing model size and achieving faster inference speeds. Additionally, the network integrates two specialized modules: the Multi-Scale Feature Fusion Module (MSFFM), which captures multi-scale feature representations to mitigate scale variance across tissue structures, and the Feature Refinement and Fusion Module (FRFM), which enhances salient feature representation to handle segmentation issues related to object camouflage properties.LightMuSeg was trained using the BCSS dataset for binary segmentation and the BCSS-WSSS dataset for weakly supervised multiclass segmentation, both of which are publicly available. In our experimental study for binary segmentation, the LightMuSeg-V2 variant achieved the highest mean Dice (76.51%) and IoU (66.43%) outperforming other models. When evaluated on the BCSS-WSSS dataset for multi-class segmentation, LightMuSeg achieved state-of-the-art performance in the Lymphocytic_infiltrate and Necrosis_or_debris classes, with IoU scores of 77.2% and 93.89%, respectively. Furthermore, a faster inference speed (44.51 mTps), better tumor segmentation accuracy, and a small model size (7.05 million parameters) make the proposed model suitable for deployment in real-world clinical scenarios. The code is publicly available at https://github.com/ZAKAUDD/LightMuSeg.

1. Introduction

The introduction of deep learning (DL) models such as U-Net [1] has led to significant advancements in the domain of medical imaging. In recent years, many researchers have made several modifications to the original U-Net architecture to deal with different issues in different imaging domains. In the case of medical imaging applications, currently, these models are being trained using different imaging modalities, including mammography, ultrasound, magnetic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT), and histopathology images [2,3] for diagnosis purposes. Each modality presents its advantages and limitations; however, among all these, histopathological image analysis is regarded as the gold standard for obtaining detailed tissue and cellular-level information, particularly in cancer treatment (i.e., breast cancer). It plays a crucial role in confirming positive pre-screening findings and offers critical insights that can reduce overdiagnosis and overtreatment [4].

Among all cancers, breast cancer is the leading cause of cancer-related death in women. A recent report by the World Health Organization's International Agency for Research on Cancer reported 2.3 million new cases globally in 2020, accounting for 11.7% of all new cancer diagnoses in women [5]. Despite these concerning statistics, early detection and treatment programs have led to a significant reduction in mortality rates, although this trend is primarily observed in countries with a high Human Development Index [6]. In this context, DL methods are playing a crucial role in addressing key challenges in AI-driven diagnosis using whole-slide images (WSI) [7]. Among various DL techniques, semantic segmentation is becoming increasingly important in medical imaging for disease diagnosis, as it performs pixel-level classification by assigning a class label to each pixel or group of pixels [8], making it particularly valuable for histopathological analysis, where precise segmentation is crucial for accurate diagnosis and treatment. Precise delineation at the pixel level is essential for

* Corresponding author.
E-mail address: muhammad.zaka-ud-din@uniud.it (Z.-U.-D. Muhammad).

https://doi.org/10.1016/j.imu.2026.101754
Received 23 October 2025; Received in revised form 9 April 2026; Accepted 12 April 2026
Available online 20 April 2026

disease characterization and supports pathologists in achieving faster and more reliable diagnoses; however, segmenting histological breast tissue poses a number of challenges. Some arise from the inherent morphological variability and complexity of the tissue, and others stem from limitations in model architectures when addressing issues such as variability in cell size and shape, indistinct boundaries, heterogeneous staining, and increasing image resolutions.

When addressing these challenges, many models become computationally complex due to their millions of trainable parameters (e.g., MS-UNet [9]), resulting in slower inference speeds and limiting their applicability in clinical settings with limited computational resources. A high-throughput pathology laboratory processing approximately 1000 slides per day would necessitate the automated analysis of all slides, posing substantial computational demands. Therefore, beyond conventional performance metrics, the computational efficiency and scalability of models are critical factors for their integration into real-world clinical workflows; however, in most cases, high-performing models contain a large number of trainable parameters, making their real-time application challenging. Furthermore, WSI-annotated data are scarce, making it harder to evaluate model performance on heterogeneous data, as in most cases, the same dataset is divided into train, test, and validation sets after extracting tiles from ROIs. This situation necessitates patient-level stratification during dataset partitioning to prevent information leakage between the training, validation, and test sets.

To address these problems, we developed a fast and lightweight network for the image segmentation of breast cancer histopathology (LightMuSeg), shown in Fig. 1, which takes into account the target objectives of previous studies on the importance of global and local, spatial and channel features to handle the segmentation of objects of varying sizes, shapes, and camouflaging properties at different noise levels. Keeping in mind the computational complexity and inference speed, the proposed approach focuses on a multi-scale feature extraction and manipulation strategy, using convolution operations applying different-sized kernels to extract multi-scale feature maps from larger and varying receptive fields in order to deal with objects of varying sizes and shapes. Instead of using computationally expensive transformers architectures [10], we incorporated lightweight attention operations at different levels to reduce grading variability and increase the correlation between local and global feature maps, helping to improve model performance in segmenting different target objects with hazy boundaries. A hierarchical feature extraction strategy is employed in the decoder to maintain optimal performance in segmenting objects of different shapes and sizes and to improve the model's generalization. Furthermore, to deal with the issue of model evaluation on completely unseen datasets, we divided the data on case levels, making the test set completely unseen for the model to evaluate its performance. In summary, our contributions include:

- a Multi-Scale Feature Fusion Module (MSFFM) to extract multi-scale information through multiple receptive fields to deal with scale variance issues in medical imaging, and also to deal with the issue of vanishing and exploding gradients to achieve generalized performance.
- a Feature Refinement and Fusion Module (FRFM) to improve the model performance by reducing the effect of redundant feature maps possibly generated through the use of skip connections and up-sampling operations.
- multiple experiments have been performed on the BCSS and BCSS-WSSS datasets for both binary and multi-class segmentation tasks to evaluate the performance and computational efficiency of our proposed model against several versions of the U-Net architecture and other state-of-the-art approaches.

2. The proposed method

We propose "LightMuSeg", motivated by three key challenges of existing approaches, including computational complexity, performance degradation caused by objects of varying sizes with ambiguous boundaries, and information gaps created between feature maps at different layers due to the up-sampling operations in skip connections. The proposed model (Fig. 1) employs a pre-trained MobileNet-V2 as the encoder, chosen for its computational efficiency and lightweight architecture, which are advantageous for real-time applications. To address the issue of shape and scale variations in medical imaging objects (e.g., tumors and lymphocytes in WSIs), the model incorporates a Multi-Scale Feature Fusion Module (MSFFM) after each encoder layer. As shown in Fig. 2, the MSFFM employs multiple convolutional kernels to extract multi-scale feature maps. It mitigates scale variability by extracting and integrating these features through channel and spatial attention operations.

Furthermore, a Feature Refinement and Fusion Module (FRFM) is integrated into the decoder (Fig. 1B) to reduce representation gaps between layers, which commonly arise from up-sampling in U-Net-based architectures. FRFM helps the model to deal with the issue of smaller objects segmentation having ambiguous boundaries. FRFM achieves this by enhancing the representation of discriminating features through suppressing redundant feature responses and alleviating camouflage effects, thereby improving the model segmentation performance for fine and small-scale structures.

A detailed description of each component of the proposed LightMuSeg is provided in the following subsections.

2.1. Multi-Scale Feature Fusion Module (MSFFM)

As mentioned earlier, the model is designed by considering multiple factors, including computational complexity related to model size and performance objectives, such as handling variations in object size and the camouflage characteristics of WSIs. To address these challenges, the proposed model incorporates MSFFM to specifically tackle the variation in object size in medical images. MSFFM extracts and integrates multi-scale feature representations, thereby enhancing segmentation performance across objects of varying sizes.

For this purpose, MSFFM performs convolution operations using kernels of varying sizes, such as (1×1, 3×3, and 5×5), to extract local and global contexts from multiple effective receptive fields. We adopt kernels of different sizes instead of dilated convolutions or pooling operations, such as SPP, in order to reduce artifact generation and minimize computational complexity. The resulting multiscale feature maps are then passed through a layer of channel attention (using a Squeeze-and-Excitation block) to emphasize informative feature maps, and through spatial attention to capture inter-spatial relationships between the feature maps. MSFFM performs this set of operations on the feature maps of each layer of the encoder, and this multi-scale feature extraction process can be described as follows.

Suppose F is the input feature map extracted from an encoder layer and passes through multi-scale convolution operations with kernel sizes k_1, k_3, k_5 to extract multi-scale feature maps F_1, F_3, F_5 from different effective receptive fields. These feature maps are concatenated and then passed through a channel attention operation to further enhance the feature map representation.

$$F = Conv_{1\times1}(Concat(Convk_1(F), Convk_3(F), Convk_5(F))) \quad (1)$$

The re-weighting process of feature maps through this attention operation can be represented as follows:

$$A_c = \sigma(MLP(GAP(F))), \quad (2)$$

Here, GAP refers to global average pooling, MLP represents the two consecutive multi-layer perceptron operations performed through a

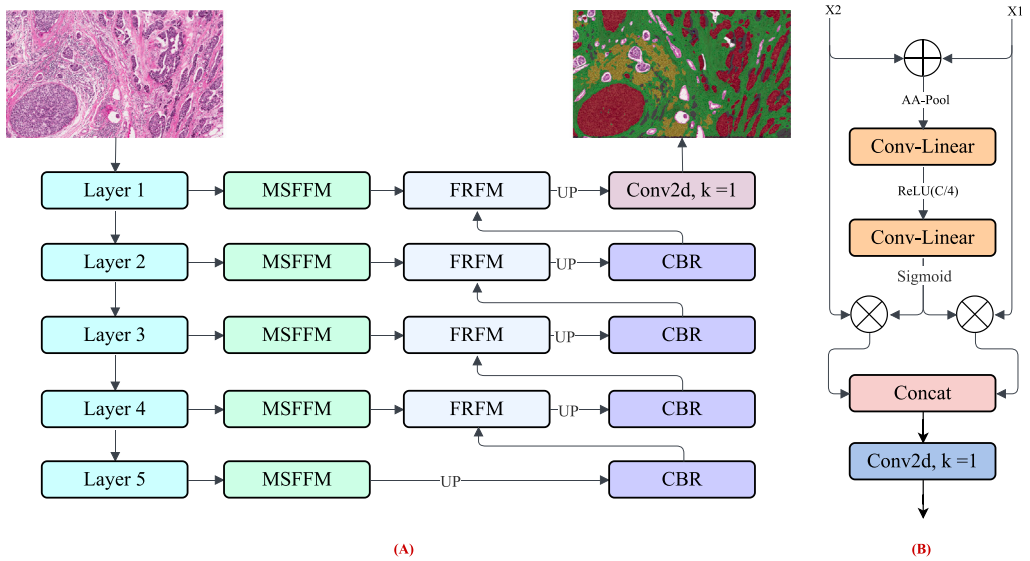

Fig. 1. Graphical representation of the proposed network and its components: (A) represents the overall model, (B) the Feature Refinement and Fusion Module (FRFM), and (CBR) represents the convolution operation followed by Batch Normalization and ReLU activation.

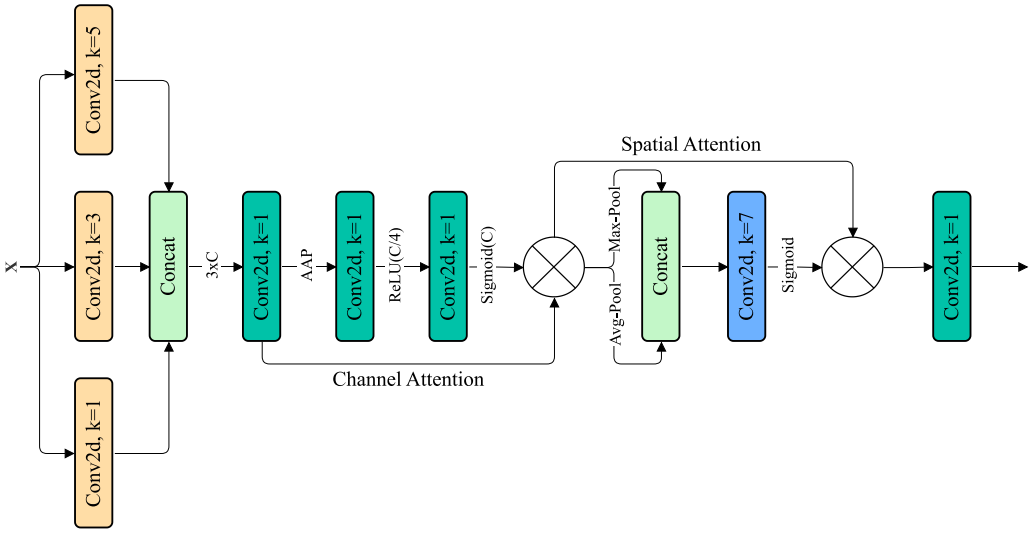

Fig. 2. Graphical representation of the proposed Multi-Scale Feature Fusion Module (MSFFM).

convolution operation, and σ represents the sigmoid activation function. The final output of the channel attention operation can be expressed as follows:

$$F_{ca} = F \times A_c. \tag{3}$$

In order to enhance inter-spatial relationships, these feature map representations are further refined using the spatial attention operation (F_{sa}). This operation is applied to the generated feature map $F_{ca} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width of the feature map, respectively. The overall spatial attention operation is defined as follows:

$$F_{avg} = \sum_{c=1}^{C} F_{ca}(c,:,:), \text{ where } F_{avg} \in \mathbb{R}^{1\times H\times W} \tag{4}$$

$$F_{max} = \max_{c=1}^{C} F_{ca}(c,:,:) \text{ where} F_{max} \in \mathbb{R}^{1\times H\times W} \tag{5}$$

$$F_{\text{concat}} = Concat(F_{\text{avg}}, F_{\text{max}}), where F_{\text{concat}} \in \mathbb{R}^{2\times H\times W}, \tag{6}$$

$$F_{sa} = \sigma(\text{Conv}_{7\times7}(F_{\text{concat}})), where F_{sa} \in \mathbb{R}^{1\times H\times W} \tag{7}$$

$$F_{out} = F_{ca} \cdot F_{sa}, where F_{out} \in \mathbb{R}^{C\times H\times W} \tag{8}$$

$$F_{output} = Conv_{1\times1}(F_{out}), \quad where F_{output} \in \mathbb{R}^{C_{out}\times H\times W} \tag{9}$$

These are the refined form of feature maps generated from each encoder layer and are now ready to be fed into the decoder part of the proposed model to be further refined to obtain the outcome of the model.

2.2. Feature Refinement and Fusion Module (FRFM)

In medical imaging, varying object sizes and camouflage properties (i.e.,low-contrast boundaries) contribute to existing models' poor performance, while feature upsampling and skip connections make segmenting objects with unclear boundaries even more difficult [11]. Through the skip-connection operation, the up-sampling process results in pixel-level distortion of the feature maps. In this way, feature maps lose their fine-grained details, reducing the accuracy of models in identifying small objects with camouflage properties.

To address this issue, many strategies are now being employed, such as relying solely on deep-layer feature maps to eliminate the need

for multiple up-sampling and skip-connection operations. Furthermore, attention operations are now also being adopted to improve feature map representations.

Motivated by the performance improvements observed in attention-based architectures, we incorporated a similar mechanism into our proposed model. For this purpose, instead of using conventional attention modules, we introduce the Feature Refinement and Fusion Module (FRFM), specifically designed to enhance the integration and refinement of feature maps.

By addressing the artifacts that result from up-sampling and skip connections, FRFM improves the quality of the features, thereby enhancing LightMuSeg's segmentation accuracy. The module combines up-sampled features from the decoder's previous layer with skip-connection feature maps from MSFFM, which is performed using element-wise addition of the feature maps. These feature maps are then passed through a channel attention operation to improve the inter-channel relationship and to reduce the information representation gaps between input feature maps. These refined feature maps are then multiplied with input feature maps to enhance their representation and to reduce the risk of overfitting. The refined feature maps are subsequently concatenated and passed through a convolution operation to adjust the number of channels in the generated output feature maps. The proposed model incorporates four FRFM (as shown in Fig. 1 (B)), helping the model to achieve accurate segmentation. Mathematically, this overall process can be represented as follows:

The process begins by adding the input feature maps x_1 and x_2 ($x_1, x_2 \in \mathbb{R}^{C\times H\times W}$) through element-wise addition operation ($x_{\text{sum}} = x_1 + x_2, x_{\text{sum}} \in \mathbb{R}^{C\times H\times W}$). The resulting feature maps then go through a Global Average Pooling layer to improve the local and global feature relationships in feature maps.

$$x_{\text{gap}}[c] = \sum_{i=1}^{H}\sum_{j=1}^{W} x_{\text{sum}}[c,i,j], \forall c \in \{1,2,\ldots,C\}, x_{\text{gap}} \in \mathbb{R}^{C}, \tag{10}$$

This pooling operation is followed by a channel attention operation $A_c = \sigma(W_2(\text{ReLU}(W_1(x_{gap}))))$, to enhance the feature representations, where $W_1 \in \mathbb{R}^{C\times C/r}, W_2 \in \mathbb{R}^{C/r\times C}$, r is the reduction ratio to lower computing cost, and σ is the sigmoid activation function. These attention maps (A_c) are multiplied by their input resources separately ($x_1' = x_1 \cdot A_c, x_2' = x_2 \cdot A_c,$) and then concatenated $x_{concat} = Concat(x_1', x_2') \in \mathbb{R}^{2C\times H\times W}$. The outcome with refined representation from these feature maps is then generated by a 1×1 convolution operation such as $x_{\text{fused}} = \text{Conv}_{1\times1}(x_{\text{concat}}) \in \mathbb{R}^{C\times H\times W}$.

3. Experimental setup

The proposed model is trained on the Breast Cancer Semantic Segmentation (BCSS) dataset [12] for binary segmentation using a supervised learning approach. In addition, for multi-class segmentation, the model is trained from scratch on the BCSS-WSSS dataset [13] using a weakly supervised learning strategy.

The BCSS-WSSS dataset consists of extracted tiles derived from the original BCSS dataset and is specifically designed for weakly supervised segmentation experiments. As previously noted, many existing methods suffer from data leakage due to the limited availability of publicly accessible WSI datasets. This issue often arises when tiles are extracted from all slides prior to splitting the data into training, validation, and testing sets.

To mitigate this problem, we adopt an ROI-level partitioning strategy on the BCSS dataset, followed by tile extraction performed separately within each split (training, validation, and testing).

This approach is chosen to more rigorously evaluate the model's generalization to unseen cases, thereby reflecting its true learning capability. However, it introduces challenges when comparing results with prior studies using the same dataset, as many do not employ slide-level or patient-level stratification, potentially inflating performance estimates.

In this case, to test our proposed model against existing state-of-the-art models, we trained our model from scratch using the weakly supervised approach on the BCSS-WSSS dataset for the multi-class segmentation task.

Furthermore, we examined our model's performance for multiclass segmentation on fully unseen cases-ROI rather than extracted tiles. To do so, we trained the model with the same BCSS data distribution used for binary segmentation to assess its performance on unseen test cases. It is important to highlight that in all of these studies, we trained our model individually for each target objective.

To train the proposed model on the mentioned dataset, the used loss function, evaluation metrics, the details of the mentioned datasets, and the implementation details are further described in the subsections below. Section 3.1 presents the loss functions that define the objective criteria for training and guide the learning process. Section 3.2 discusses the evaluation metrics used to assess model performance and ensure reliable comparisons. Section 3.3 outlines the datasets used for training, validation, and testing, along with their properties and pre-processing methods. Finally, Section 3.4 elaborates on the implementation details, such as the used framework, training configuration, and optimization strategies.

3.1. Loss function

As the proposed model was designed for both binary and multiclass segmentation tasks, we adopted different loss functions to train and evaluate the model's performance. To train the model for the binary segmentation task, we employed the structure loss $\mathcal{L}_{Str}$ [14], which combines the weighted intersection over union loss $\mathcal{L}_{IoU}^{w}$ and weighted binary cross-entropy $\mathcal{L}_{BCE}^{w}$. In $\mathcal{L}_{Str}$, the $\mathcal{L}_{IoU}^{w}$ increases the weights of hard pixels to emphasize their significance, while $\mathcal{L}_{BCE}^{w}$ prioritizes hard pixels by assigning them higher weights instead of treating all pixels equally. This combination enhances the model's robustness in learning complex situations and improves performance. The effectiveness of this approach has been validated in various segmentation and salient object detection tasks.

$$\mathcal{L}_{BCE}^{w} = -\frac{1}{N}\sum_{i\in I}\left[g_m[i]\log\left(p_m[i]\right) + \left(1-g_m[i]\right)\log\left(1-p_m[i]\right)\right] \tag{11}$$

$$\mathcal{L}_{IoU}^{w} = 1 - \frac{\sum_{i\in I} g_m[i]p_m[i]}{\sum_{i\in I} g_m[i] + p_m[i] - g_m[i]p_m[i]} \tag{12}$$

$$\mathcal{L}_{Str} = \mathcal{L}_{IoU}^{w}(p_m, g_m) + \mathcal{L}_{BCE}^{w}(p_m, g_m) \tag{13}$$

where N is the total number of image pixels, $i \in I$ refers to a pixel in the output (prediction map) and ground truth, (p_m) represents the prediction, and (g_m) denotes the ground truth maps.

Similarly, to train the model for the multiclass segmentation task, we used the Dice loss and Cross-Entropy loss together. Dice loss ($\mathcal{L}_{Dice}$) is used to address class imbalance issues in multiclass segmentation tasks and is commonly employed in medical image segmentation applications. Cross-Entropy loss ($\mathcal{L}_{CE}$) is used to minimize the divergence between the predicted and target probability distributions across all classes, thereby reducing the likelihood of overfitting the model.

3.2. Evaluation metrics

Based on the intended objectives of the proposed model, we evaluated its performance using a range of standard metrics to ensure a comprehensive assessment. For all experiments, the reported results reflect the mean values of the respective evaluation metrics. In the binary segmentation task, seven widely used metrics were employed: Dice Score, Intersection over Union (IoU), Mean Absolute Error (MAE), structural similarity measure (S_α) [15], weighted F-measure (F_β^i) [16], enhanced alignment measure (E_α) [17], and Sensitivity.

For multiclass segmentation, we further computed class-wise metrics, such as IoU, to capture both per-class performance and global segmentation quality.

We also included the model size (expressed in millions of parameters) and inference speed, measured in tiles processed per second (Tps), together with the other evaluation results detailed in the quantitative performance evaluation.

3.3. Benchmark dataset

For our experiments, we adopted a slide-level data partitioning strategy to promote variability and prevent data leakage, thereby improving model generalization. Before splitting, the data were anonymized by assigning numerical identifiers, and subsequently divided into training, validation, and test sets using a fixed random seed of 42. The original annotation masks were converted into multiclass labels for multiclass segmentation and binarized for binary segmentation. Tiles of size (512×512) pixels were then extracted from the slides after the data split, ensuring proper separation between sets.

For the binary segmentation task, the dataset consisted of 6737 training tiles, 816 validation tiles, and 609 test tiles. For multiclass segmentation, the training, validation, and test sets contained 6345, 526, and 820 tiles, respectively.

To further enhance dataset diversity and model robustness, data augmentation techniques were applied during training. For binary segmentation, augmentations included center cropping, random cropping, horizontal and vertical flips, scaling, arbitrary rotations, and brightness adjustments. For multiclass segmentation, the augmentations consisted of random rotations, color jittering, and Gaussian blur.

In addition, to further evaluate our model's learning capabilities against other state-of-the-art approaches, we conducted additional experiments using the BCSS-WSSS dataset [13], which is derived from the BCSS dataset and provides weak supervision through pseudolabels. The model was trained from scratch for 100 epochs on this pseudolabeled data, and all images in the dataset are of size 224 × 224 pixels. The dataset includes 23,230 patches for training, 3472 for validation, and 4713 for the test set, where the validation and test sets include ground-truth masks, while the training set relies on pseudolabels generated using the authors' WSI-trained model, enabling the use of the published dataset for supervised training under weak supervision.

3.4. Implementation details

The proposed model is implemented using the PyTorch framework and trained on an NVIDIA RTX 3090. For training, we used a fixed tile size of 512 × 512 pixels for binary and multiclass segmentation tasks. For both tasks, augmentation techniques are applied on the fly to increase the diversity of the dataset; however, the input images are randomly scaled within a range of 0.75 to 1.25 to mitigate overfitting in the binary segmentation task. We used the Adam optimizer with $lr = 0.001$ and $weight_decay = 0.0001$ to ensure stable and effective training dynamics. Taking into account the available resources and the enhanced training images, the proposed model was trained for 150 epochs using a batch size of 3 for the binary segmentation task. In contrast, the model for the multiclass segmentation task was trained for 100 epochs using a batch size of 6 to avoid overfitting. Instead of using an early-stopping strategy, we employed the idea of saving the model based on the best dice score to avoid overfitting while letting the model learn, if possible, until the final epochs.

4. Experiments and discussion

In addition to the objective of achieving a smaller model size with lower computational complexity and higher inference speed, this study addresses two segmentation tasks for breast histopathology images: binary segmentation and multi-class segmentation, using the BCSS and BCSS-WSSS datasets, respectively. Section 4.1 presents the experimental results for binary segmentation on the BCSS dataset. Section 4.2 details the experiments and performance analysis for multi-class segmentation on the BCSS-WSSS dataset and also reports the qualitative performance of the proposed model on full ROIs, where the model is trained on the BCSS dataset and evaluated on an unseen BCSS test set. Finally, Section 4.3 presents an ablation study to demonstrate the contribution of each component to model performance.

4.1. Performance evaluation for binary segmentation

Due to the lack of existing experimental work on slide-level data distribution using the BCSS dataset, this study aims to provide an experimental baseline for future research on this dataset, particularly for binary segmentation of the tumor class. To evaluate the performance of our proposed model, we conducted extensive experiments using the U-Net architecture with various pre-trained encoders. The selected encoders included ResNet34 [18], ResNeXt50 [19], ResNeSt50 [20], Res2Net50 [21], VGG19 [22], EfficientNet-B5 [23], MobileNet-V3-L [24], MobileOne-S3 [25], and MobileNet-V2 [26]. All experiments were conducted using the encoders' pre-trained weights without backbone-specific pre-processing to ensure consistency and facilitate fair comparisons across models.

Quantitative Evaluation: In this comparison, the proposed model performance is assessed using a comprehensive set of evaluation metrics, as described above. In this context, ↑ indicates that higher values correspond to better performance, whereas ↓ denotes that lower values are preferable. The performance comparison of the proposed model against selected U-Net variants (based on the chosen backbone) is presented in Table 1. Among the U-Net baselines, the variant employing the MobileNet-V3-L encoder achieved the highest mDice and mIoU scores (68.8% and 55.77%, respectively). However, all versions of the proposed LightMuSeg architecture outperform the baselines across most evaluation metrics. In particular, LightMuSeg-V2 achieves the best overall performance, with an mDice of 76.51% and an mIoU of 66.43%, corresponding to improvements of 7.71% and 10.66%, respectively, over the best-performing U-Net variant. In addition, it records the lowest MAE of 19.18%. This comparison further demonstrates that all versions of the proposed model attain higher mean Dice scores while requiring fewer trainable parameters, although they exhibit average inference speed.

Qualitative Evaluation: To complement the quantitative results, qualitative analyses are also conducted. For binary evaluation, five representative tiles are selected from different slides extracted from unseen test-set ROIs, each representing particularly challenging cases. A qualitative comparison between the proposed LightMuSeg model and multiple U-Net variants, each employing different encoder backbones, is illustrated in Fig. 3. In this figure, the first two columns present the original image and its corresponding ground-truth mask, followed by the outputs of the three LightMuSeg variants. In the proposed model, LightMuSeg-V3 incorporates deep supervision and allows the MobileNet-V2 encoder weights to be fine-tuned during training. LightMuSeg-V2 follows the same architectural design but excludes deep supervision, while LightMuSeg-V1 serves as the baseline model with frozen MobileNet-V2 encoder weights. Among the proposed variants, the prediction maps produced by LightMuSeg-V2 closely resemble the ground truth, a trend that is consistently observed in both quantitative and qualitative evaluations, as reported in Table 1 and Fig. 3.

Across all samples, all versions of LightMuSeg consistently outperform the U-Net baselines, underscoring the effectiveness of the proposed decoder modifications. The final column in Fig. 3 emphasizes this contrast, highlighting the value of the custom decoder modules introduced in LightMuSeg, an enhanced variant of U-Net utilizing the

Table 1
Performance comparison of the models in our experiments for binary segmentation. ↑ represents higher the better, while ↓ denotes lower, the better, and the best performance is highlighted using boldface letters.

Metric	ResNet34	ResNext50	ResNest50	Res2Net50	VGG19	EffNet-B5	MNV3-L	MobileOne	MobileNet-V2	LightMuSeg.V1	LightMuSeg.V2	LightMuSeg.V3
mDice% ↑	57.44	61.31	60.75	62.92	52.01	64.51	68.80	57.72	59.75	76.15	**76.51**	75.25
mIoU% ↑	43.47	48.48	47.94	50.11	38.70	49.90	55.77	43.79	46.73	65.45	**66.43**	64.81
F_β^w% ↑	50.98	49.78	49.33	51.74	43.70	57.82	62.33	49.61	49.02	70.62	**71.29**	70.83
S_α% ↑	38.55	26.11	26.20	31.40	21.77	46.16	52.18	32.20	27.03	61.41	**62.40**	62.10
E_ϕ^{mean}% ↑	49.88	29.58	29.72	33.46	32.75	57.32	63.11	44.61	32.47	69.85	70.64	**71.30**
mSen% ↑	68.33	97.40	95.78	96.14	73.70	74.63	78.03	74.59	90.19	83.86	83.95	80.19
MAE% ↓	38.76	51.10	51.53	46.59	55.31	33.79	28.09	45.19	50.09	20.07	**19.18**	19.54
Inf-speed% ↑	**175.45**	99.87	94.06	108.49	76.92	76.82	150.94	123.16	161.77	111.91	106.37	94.52
Model-size(millions)% ↓	24.44	31.99	34.45	32.66	37.37	29.06	6.69	16.25	**6.63**	4.83	6.64	6.664

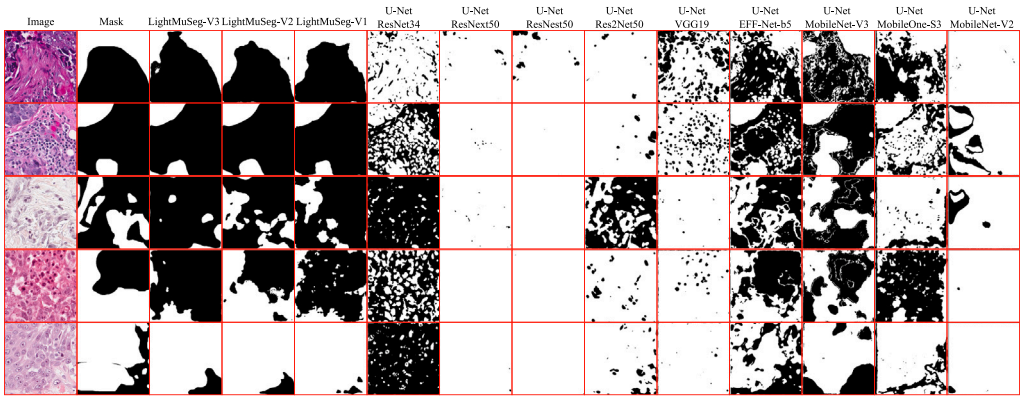

Fig. 3. Qualitative comparison of the proposed model LightMuSeg against the U-Net using different backbones as an encoder. All of the images are from different slides from the test set. LightMuSeg-V3 represents the model with deep supervision with unfrozen encoder weights, while LightMuSeg-V2 is the same model without deep supervision, and LightMuSeg-V1 is the model without deep supervision and frozen weights.

MobileNet-V2 encoder. Additionally, the figure illustrates that models employing heavier encoders with residual connections (e.g., U-Net with ResNet50, ResNeXt50, and Res2Net50) tend to produce redundant feature maps, reducing segmentation precision. Conversely, MobileNet and EfficientNet-based U-Nets demonstrate superior performance, suggesting their suitability for binary segmentation tasks in WSI analysis.

4.2. Performance evaluation for multi-class segmentation

Quantitative Comparison: Following the performance evaluation of the proposed model on the binary segmentation task, we further assessed the proposed model's performance for the multi-class segmentation task using the BCSS-WSSS dataset to evaluate its effectiveness under weak supervision. We performed this performance comparison of our model against several published weakly supervised segmentation methods trained on the BCSS-WSSSS dataset, including HistoSegNet [27], TransWS [28], OEEM [29], MLPS [13], HAMIL [30], TPRO [31], SEAM [32], SC-CAM [33], C–CAM [34], WSSS-Tissue [13], MSRMMP [35], CGNet [36], PCSFormer [37], PDSeg [38], and UAM-WSTS [39].

Training on the BCSS-WSSS dataset constitutes a weakly supervised approach, as the training set incorporates pseudo-labels as ground truth. Notably, our model (V2) is trained directly on raw pseudo-labels, whereas many of the compared approaches applied pre-processing steps to refine the pseudo-labels before training their models. Similarly, most of the studies reported their results after post-processing; however, our results are obtained directly from raw prediction maps, ensuring a fair and transparent evaluation.

The quantitative comparison on the BCSS-WSSS test set [?] only reports the mean-IoU scores for each class and the overall mean-IoU. The performance of existing approaches is taken either directly from their original publications or from reliable comparative studies, and the complete results are summarized in Table 2.

From the quantitative comparison, it is clear that OEEM achieved higher segmentation performance for the Tumor and Stroma classes, while our model achieves the highest overall mean IoU (mIoU) and the best performance specifically for Lymphocytes and Necrosis. This is a good sign for our proposed method, as it presents a generalized performance in terms of dealing with different scales and sizes of objects in a more balanced way than the reported models in this comparison. Thus, our performance is well balanced, and achieving the best scores in the segmentation of lymphocytes and necrosis regions is particularly noteworthy, as these regions are relatively rare in breast cancer and can sometimes lead to false-positive predictions.

Due to the effective combination of multiple strategies within the proposed architecture, the model achieves a mean IoU of 74.87%, surpassing the previous state-of-the-art model, MSR-Net, by 3%. In addition, we achieved a performance gain of over 13% for lymphocytes and over 23% for necrosis segmentation. The proposed approach obtained an overall mean IoU score of 74.87%, compared to the stated best score of 71.32% among current models.

In this comparison, it is also important to note that several models uses transformers as encoders, and almost all of them have more than 20 million trainable parameters. From this, we can conclude that our proposed model is the lightweight option (less than 7 million trainable parameters) among all of them, and these improvements not only reduce computational cost but also make the model an ideal candidate for real-time applications.

Qualitative Evaluation: Following the quantitative evaluation, we also performed a qualitative assessment for the multiclass segmentation task; however, for this purpose, instead of using a weakly supervised trained model, we trained the model on the BCSS dataset with ROI-level distribution as described in the manuscript. We adopted this strategy as the BCSS-WSSS dataset is pre-distributed, and the tile-level distribution is publicly available, which did not satisfy our objective of evaluating the model performance on unseen cases.

To satisfy our target objective and to assess the proposed architecture's learning ability and generalization capacity, we trained a

Table 2
Performance comparison of the models in our experiments for multiclass segmentation using the BCSS-WSSSS dataset. In this quantitative evaluation, the classes include "tumor", "stroma", "lymphocytic_infiltrate", and "necrosis_or_debris". We skipped the "class 0" (background) as it is not as mandatory as the other classes with their important evaluation information, and the best results are in boldface letters.

Method	Tumor	Stroma	Lymphocytic_infiltrate	Necrosis_or_debris	mIoU% ↑
	mIoU% ↑	mIoU% ↑	mIoU% ↑	mIoU% ↑	
HistoSegNet	33.14	46.46	29.05	01.91	27.64
SEAM	74.37	62.16	50.79	48.43	58.94
SC-CAM	76.79	70.61	58.02	60.07	66.37
C-CAM	75.57	67.96	31.00	49.43	55.99
WSSS-Tissue	77.98	72.95	60.98	66.87	69.70
OEEM	**80.21**	**74.74**	62.60	63.78	70.33
MSRMMP	79.80	74.08	62.63	68.75	71.32
CGNet	68.22	61.77	52.24	56.84	59.77
PCSFormer	75.04	69.84	58.44	62.86	66.54
TransWS	44.71	36.49	41.72	38.08	40.25
TPRO	77.95	65.10	54.55	64.96	65.64
MLPS	74.54	64.45	52.54	58.67	62.5
PDSeg	79.33	73.08	60.45	65.71	69.64
HAMIL	71.65	62.37	51.52	54.29	59.96
UAM-WSTS	79.89	74.66	64.71	70.88	70.88
LightMuSeg-V2(Ours)	76.33	57.95	**77.2**	**93.89**	**74.87**

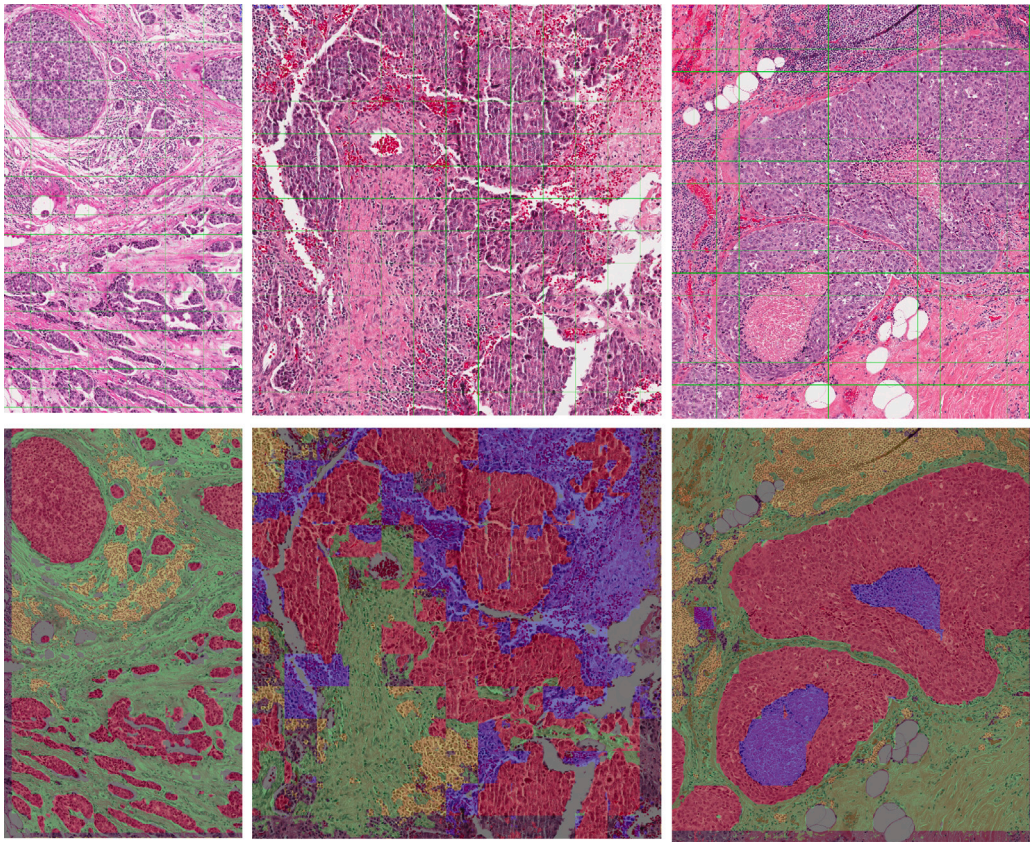

Fig. 4. Qualitative performance of the proposed model for multi-class segmentation on BCSS test set. In this color scheme, red represents tumoral areas, green highlights stromal tissues, yellow is used for lymphocytic infiltrates, and blue denotes necrotic regions.

fresh version of our model (V2) on the BCSS dataset (for multi-class segmentation). We reported the qualitative outcomes on the completely unseen test set to understand the model's generalization capabilities. We conducted this experiment because, to the best of our knowledge, no existing work considers slide-level data distribution and ROI-level evaluation on this dataset. The performance evaluation step automatically extracts the tiles from each case and then rebuilds the tile-level prediction maps into their corresponding ROIs to study the ROI-level performance. The qualitative performance representations of unseen ROIs are presented in Fig. 4.

4.3. Ablation study

Before proceeding to the final selection of the model's architecture, we performed several experimental ablation studies to select an optimal encoder for the proposed model in terms of performance and

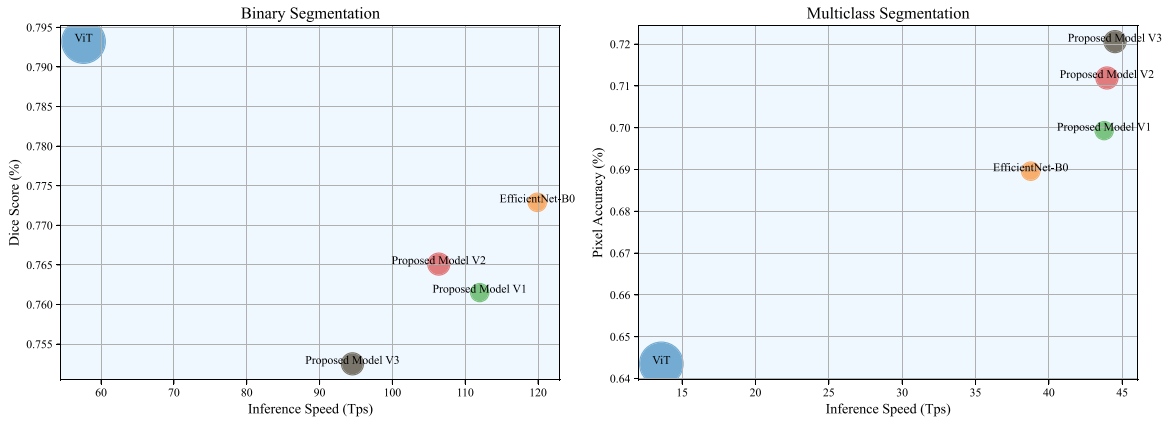

Fig. 5. Ablation Experiments on the proposed model using different encoders and different settings using MobileNet-V2. The bubble size represents the model's trainable parameters. In contrast, their location on the x-axis and y-axis represents the inference speed (tiles per second) and the Dice, Pixel Accuracy scores, respectively.

Table 3
Quantitative performance analysis of the conducted ablation experiments. The first two rows represent the proposed model with ViT and EfficientNet-b0 encoder, and the other three are with MobileNet-V2 with different settings. The final 2 rows are to study the contribution of proposed modules on both binary and multiclass segmentation. ↑ represents higher the better while ↓ denotes lower, the better.

Models	Binary segmentation			Multi class segmentation											
				Tumor			Stroma			Lymphocytic_infiltrate			Necrosis_or_debris		
	mDice% ↑	mIoU% ↑	MAE%↓	mDice% ↑	mIoU% ↑	F1% ↑	mDice% ↑	mIoU% ↑	F1% ↑	mDice% ↑	mIoU% ↑	F1% ↑	mDice% ↑	mIoU% ↑	F1% ↑
LightMuSeg+ViT Encoder	79.32	69.66	18.01	56.15	65.70	78.47	47.50	47.76	63.17	37.10	39.12	49.83	03.14	03.18	04.14
LightMuSeg+EfficientNet-B0 Encoder	77.29	66.7	19.49	56.89	67.46	79.84	51.18	55.11	69.63	35.67	37.88	47.96	28.77	31.94	39.22
LightMuSeg V1	76.15	65.45	20.07	57.49	69.11	81.02	50.75	54.19	68.86	39.97	42.53	53.92	29.84	33.96	41.06
LightMuSeg V2	76.51	66.43	19.18	**58.40**	**71.52**	**82.77**	51.22	54.92	69.61	40.54	43.52	54.82	31.70	37.41	44.14
LightMuSeg V3	75.25	64.81	19.54	58.12	70.85	82.25	**51.59**	**55.69**	**70.26**	**41.08**	**44.02**	**55.49**	**32.81**	**38.73**	**45.69**
LightMuSeg V1*	74.74	63.16	21.67	–	–	–	–	–	–	–	–	–	–	–	–
LightMuSeg V3†	–	–	–	58.26	71.05	82.49	50.85	54.16	68.98	37.38	38.48	49.81	33.79	39.27	46.82

the number of trainable parameters. The primary goal of the encoder selection was to keep the model size smaller and increase inference speed while maintaining segmentation accuracy. Following that, we further conducted some preliminary experiments to understand the roles of the proposed modules in the proposed model.

In terms of encoder selection, we started the experimental study by selecting a ViT encoder, as transformers are well-known for their superior performance. In this study, we found that the model with a pre-trained ViT encoder has a huge overfitting issue in both binary and multiclass segmentation tasks. In addition, it has a lower inference speed compared to other models (U-Net versions with different backbones). From Table 3 and Fig. 5, it is also clear that the ViT encoder has a degraded performance in multiclass segmentation and holds more than 27 million parameters. Based on these results, we decided to use a simple CNN architecture as an encoder and conducted two experiments to select an encoder among EfficientNet-b0 and MobileNet-V2.

After all these experiments, we chose Mobile-Net-V2 as our proposed model encoder after analyzing model size, inference speed, and multiclass segmentation performance, as indicated in Fig. 5 and Table 3. Following the encoder selection, we further conducted two different experimental investigations to examine the impact on model performance of utilizing the MobileNet-V2 encoder with its pre-trained (unfrozen) weights and the use of deep supervision to train the model for the multiclass segmentation task. From Table 3, it is clear that the proposed model with unfrozen weights but without deep supervision is performing best among other versions for binary segmentation tasks. In contrast, the same version with deep supervision performs best for multiclass segmentation.

To further investigate the contributions of the proposed modules, we conducted two separate studies focusing on the model's performance in both binary and multiclass segmentation. For this purpose, we excluded the Multi-Scale Feature Fusion Module (MSFFM) from the model architecture, and then, for binary segmentation, trained the model version with frozen Mobile-Net-V2 encoder weights, and for multiclass segmentation, with unfrozen weights and deep supervision, as this version of the model presented the best performance among all versions. We compared the performances of both experiments with their corresponding full versions of the proposed model. The quantitative results clearly demonstrate the contribution of each experiment to the model's performance, as shown in Table 3.

5. Conclusion

Breast cancer remains the most prevalent malignancy among women, and accurate early diagnosis is essential for effective treatment planning. AI-based methods, particularly those leveraging deep learning, have demonstrated considerable potential in assisting clinicians; however, the morphological complexity and heterogeneity of breast tissue continue to present significant challenges for reliable segmentation. Moreover, practical deployment of AI models in clinical settings must account for the high volume of cases, which imposes substantial demands on computational infrastructure.

This study presented a comparative evaluation of the U-Net architecture with a range of pre-trained encoders, adopting a slide-level data distribution strategy to better approximate real-world clinical scenarios. In addition to segmentation accuracy, model size and inference speed were considered key factors for clinical applicability. The proposed model integrates multi-scale feature extraction from diverse receptive fields and a dedicated fusion module, enhancing robustness in delineating cancerous regions in whole-slide images. In binary segmentation tasks, the proposed model consistently outperformed all baseline models while maintaining a lightweight architecture suitable for near real-time inference.

For multiclass segmentation compared to other state-of-the-art BCSS-WSSS-trained approaches, it achieved slightly lower IoU scores for the Tumor and Stroma classes; however, it demonstrated higher

segmentation accuracy for Lymphocytic Infiltrate and Necrosis, as well as a higher overall mean IoU compared to other models.

The findings also highlight the impact of incomplete annotations in the dataset, which may hinder the full potential of supervised learning. Addressing this limitation through the generation of high-quality pseudolabels offers a viable path forward. Given its balance of performance and efficiency, the proposed model is well-suited to serve as a pseudolabel generator in future efforts aimed at improving annotation quality and training data availability.

CRediT authorship contribution statement

Zaka-Ud-Din Muhammad: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Vincenzo Della Mea:** Writing – review & editing, Supervision, Funding acquisition.

Ethical statement

This study was conducted using the publicly available *Breast Cancer Semantic Segmentation (BCSS) dataset and the Breast Cancer Semantic Segmentation—Weakly Supervised Semantic Segmentation (BCSS-WSSS) dataset*. These datasets contain fully anonymized histopathological images sourced from The Cancer Genome Atlas (TCGA) and are made openly available for research purposes. In this study, no new human or animal data were collected; therefore, ethical approval and informed consent were not required. All experiments were performed according to the relevant guidelines and regulations.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT 4.0 and Grammarly in order to improve the readability and language of the paper. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially funded by HORIZON EUROPE Marie Sklodowska-Curie Actions, BosomShield project, grant number 101073222.

References

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention - MICCAI 2015: 18th international conference. Springer; 2015, p. 234–41. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[2] Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, Davidson B, Mongtomery RC, Crowley MJ, McCrory DC, Kendrick A, Sanders GD. Benefits and harms of breast cancer screening: A systematic review. JAMA 2015;314(15):1615–34. http://dx.doi.org/10.1001/jama.2015.13183.

[3] Shah TA, Guraya SS. Breast cancer screening programs: Review of merits, demerits, and recent recommendations practiced across the world. J Microsc Ultrastruct 2017;5(2):59–69. http://dx.doi.org/10.1016/j.jmau.2016.10.002.

[4] Calhoun BC, Livasy CA. Mitigating overdiagnosis and overtreatment in breast cancer: what is the role of the pathologist? Arch Pathol Lab Med 2014;138(11):1428–31. http://dx.doi.org/10.5858/arpa.2013-0763-ED.

[5] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast 2022;66:15–23. http://dx.doi.org/10.1016/j.breast.2022.08.010.

[6] Kim J, Harper A, McCormack V, Sung H, Houssami N, Morgan E, Mutebi M, Garvey G, Soerjomataram I, Fidler-Benaoudia MM. Global patterns and trends in breast cancer incidence and mortality across 185 countries. Nature Med 2025;31(4):1154–62. http://dx.doi.org/10.1038/s41591-025-03502-3.

[7] Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal 2016;33:170–5. http://dx.doi.org/10.1016/j.media.2016.06.037.

[8] Krithika Alias AnbuDevi M, Suganthi K. Review of semantic segmentation of medical images using modified architectures of UNET. Diagnostics 2022;12(12):3064. http://dx.doi.org/10.3390/diagnostics12123064.

[9] Su R, Zhang D, Liu J, Cheng C. Msu-net: Multi-scale u-net for 2d medical image segmentation. Front Genet 2021;12:639930. http://dx.doi.org/10.3389/fgene.2021.639930.

[10] Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. In: Medical image computing and computer assisted intervention - MICCAI 2021: 24th international conference. Springer; 2021, p. 36–46. http://dx.doi.org/10.1007/978-3-030-87193-2_4.

[11] Zhao Y, Zhang J, Hu D, Qu H, Tian Y, Cui X. Application of deep learning in histopathology images of breast cancer: a review. Micromachines 2022;13(12):2197. http://dx.doi.org/10.3390/mi13122197.

[12] Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MA, Abo Elnasr LS, Sakr RA, Salem HS, Ismail AF, Saad AM, et al. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 2019;35(18):3461–7. http://dx.doi.org/10.1093/bioinformatics/btz083.

[13] Han C, Lin J, Mai J, Wang Y, Zhang Q, Zhao B, Chen X, Pan X, Shi Z, Xu Z, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Med Image Anal 2022;80:102487. http://dx.doi.org/10.1016/j.media.2022.102487.

[14] Fan D-P, Ji G-P, Zhou T, Chen G, Fu H, Shen J, Shao L. Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2020, p. 263–73. http://dx.doi.org/10.1007/978-3-030-59725-2_26.

[15] Cheng M-M, Fan D-P. Structure-measure: A new way to evaluate foreground maps. Int J Comput Vis 2021;129(9):2622–38. http://dx.doi.org/10.1007/s11263-021-01490-8.

[16] Margolin R, Zelnik-Manor L, Tal A. How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, p. 248–55. http://dx.doi.org/10.1109/CVPR.2014.39.

[17] Fan D-P, Ji G-P, Qin X, Cheng M-M. Cognitive vision inspired object segmentation metric and loss function. Sci Sin Inf 2021;51(6):1036–48. http://dx.doi.org/10.1360/ssi-2020-0370.

[18] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8. http://dx.doi.org/10.1109/CVPR.2016.90.

[19] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1492–500. http://dx.doi.org/10.1109/CVPR.2017.634.

[20] Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, He T, Mueller J, Manmatha R, et al. Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, p. 2736–46. http://dx.doi.org/10.1109/CVPRW56347.2022.00309.

[21] Gao S-H, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr P. Res2net: A new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 2019;43(2):652–62. http://dx.doi.org/10.1109/TPAMI.2019.2938758.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv preprint arXiv:1409.1556. Retrieved on March 10, 2026.

[23] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, vol. 97, PMLR; 2019, p. 6105–14, URL https://proceedings.mlr.press/v97/tan19a.html. Retrieved on March 10, 2026.

[24] Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, et al. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 1314–24. http://dx.doi.org/10.1109/ICCV.2019.00140.

[25] Vasu PKA, Gabriel J, Zhu J, Tuzel O, Ranjan A. Mobileone: An improved one millisecond mobile backbone. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 7907–17. http://dx.doi.org/10.1109/CVPR52729.2023.00764.

[26] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 4510–20. http://dx.doi.org/10.1109/CVPR.2018.00474.

[27] Chan L, Hosseini MS, Rowsell C, Plataniotis KN, Damaskinos S. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 10662–71. http://dx.doi.org/10.1109/ICCV.2019.01076.

[28] Zhang S, Zhang J, Xia Y. Transws: Transformer-based weakly supervised histology image segmentation. In: International workshop on machine learning in medical imaging. Springer; 2022, p. 367–76. http://dx.doi.org/10.1007/978-3-031-21014-3_38.

[29] Li Y, Yu Y, Zou Y, Xiang T, Li X. Online easy example mining for weakly-supervised gland segmentation from histology images. In: International conference on medical image computing and computer-assisted intervention. Springer; 2022, p. 578–87. http://dx.doi.org/10.1007/978-3-031-16440-8_55.

[30] Zhong L, Wang G, Liao X, Zhang S. HAMIL: High-resolution activation maps and interleaved learning for weakly supervised segmentation of histopathological images. IEEE Trans Med Imaging 2023;42(10):2912–23. http://dx.doi.org/10.1109/TMI.2023.3269798.

[31] Zhang S, Zhang J, Xie Y, Xia Y. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 109–18. http://dx.doi.org/10.1007/978-3-031-43907-0_11.

[32] Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 12275–84. http://dx.doi.org/10.1109/CVPR42600.2020.01229.

[33] Chang Y-T, Wang Q, Hung W-C, Piramuthu R, Tsai Y-H, Yang M-H. Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 8988–97. http://dx.doi.org/10.1109/CVPR42600.2020.00901.

[34] Chen Z, Tian Z, Zhu J, Li C, Du S. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11676–85. http://dx.doi.org/10.1109/CVPR52688.2022.01138.

[35] Xue Y, Hu Y, Yao Y, Huang J, Wang H, He J. MSRMMP: Multi-scale residual module and multi-layer pseudo-supervision for weakly supervised segmentation of histopathological images. Med Eng Phys 2025;136:104284. http://dx.doi.org/10.1016/j.medengphy.2025.104284.

[36] Kweon H, Yoon S-H, Kim H, Park D, Yoon K-J. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 6994–7003. http://dx.doi.org/10.1109/ICCV48922.2021.00691.

[37] Liu C, Shen Y, Xiao Q, Li G. PCSformer: Pair-wise cross-scale sub-prototypes mining with CNN-transformers for weakly supervised semantic segmentation. Neurocomputing 2024;593:127834. http://dx.doi.org/10.1016/j.neucom.2024.127834.

[38] Li W-H, Hsieh Y-H, Yang H-F, Chen C-S. PDSeg: Patch-wise distillation and controllable image generation for weakly-supervised histopathology tissue segmentation. In: ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing. IEEE; 2025, p. 1–5. http://dx.doi.org/10.1109/ICASSP49660.2025.10888097.

[39] Kang Y, Li H, Shi X, Zhang X, Xing Y, Wen Y, Wang Y, Cui L, Feng J, Yang L. Exploring unbiased activation maps for weakly supervised tissue segmentation of histopathological images. IEEE Trans Med Imaging 2025. http://dx.doi.org/10.1109/TMI.2025.3541115.