A dynamic uncertainty-aware ensemble model: Application to lung cancer segmentation in digital pathology

Applied Soft Computing Journal 165 (2024) 112081

Contents lists available at ScienceDirect

Applied Soft Computing

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

A dynamic uncertainty-aware ensemble model: Application to lung cancer segmentation in digital pathology

Massimo Salvi [a,*], Alessandro Mogetta [a], U. Raghavendra [b], Anjan Gudigar [b], U. Rajendra Acharya [c,d], Filippo Molinari [a]

[a] Biolab, PoliToBIOMed Lab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin 10129, Italy
[b] Department of Instrumentation and Control Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India
[c] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Australia
[d] Centre for Health Research, University of Southern Queensland, Australia

HIGHLIGHTS

- Adaptive uncertainty-based ensemble model (AUE) proposed for tumor segmentation.
- AUE outperformed traditional ensemble models by a significant margin.
- Utilizing uncertainty estimates enhances segmentation performance.

ARTICLE INFO

Keywords:
Deep learning
Ensemble model
Lung cancer
Monte Carlo dropout
Uncertainty

ABSTRACT

Ensemble models have emerged as a powerful technique for improving robustness in medical image segmentation. However, traditional ensembles suffer from limitations such as under-confidence and over-reliance on poor performing models. In this work, we introduce an Adaptive Uncertainty-based Ensemble (AUE) model for tumor segmentation in histopathological slides. Our approach leverages uncertainty estimates from Monte Carlo dropout during testing to dynamically select the optimal pair of models for each whole slide image. The AUE model combines predictions from the two most reliable models (K-Net, ResNeSt, Segformer, Twins), identified through uncertainty quantification, to enhance segmentation performance. We validate the AUE model on the ACDC@LungHP challenge dataset, systematically comparing it against state-of-the-art approaches. Results demonstrate that our uncertainty-guided ensemble achieves a mean Dice score of 0.8653 and outperforms traditional ensemble techniques and top-ranked methods from the challenge by over 3 %. Our adaptive ensemble approach provides accurate and reliable lung tumor delineation in histopathology images by managing model uncertainty.

1. Introduction

Digital pathology and Artificial Intelligence (AI) integration have enabled major advancements in medical image analysis [1,2] particularly for the critical task of tumor delineation in histopathological images [3,4]. Convolutional Neural Networks (CNNs) now form the foundation of many segmentation frameworks [5,6] leveraging their unmatched capabilities in feature learning and pattern recognition [7,8].

Lung cancer is a major global health problem and a leading cause of cancer-related mortality worldwide [9]. In digital pathology, accurate segmentation of lung lesions from Whole Slide Images (WSIs) - high-resolution scans of entire histopathological slides - is important for critical applications such as tumor grading, surgical planning, and treatment response evaluation [10]. However, automatic segmentation of lung cancer in WSIs faces multiple challenges. Lung lesions exhibit high heterogeneity in appearance, with variations in size (from small nodules to large masses), shape (from round to irregular), and histological characteristics (e.g., cell morphology, tissue architecture) [11]. WSIs can also vary substantially in staining and imaging quality due to differences in tissue preparation and scanning protocols. Moreover, delineating the irregular boundaries between tumor and normal tissue

* Correspondence to: Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin 10129, Italy.
E-mail address: massimo.salvi@polito.it (M. Salvi).

https://doi.org/10.1016/j.asoc.2024.112081
Received 21 November 2023; Received in revised form 6 July 2024; Accepted 31 July 2024
Available online 5 August 2024

requires pixel-level precision from segmentation algorithms. Inaccurate segmentation can lead to misclassification of tumor extent and severity, potentially impacting clinical decision-making and patient outcomes [12].

Addressing these complexities is crucial for lung cancer segmentation tools to achieve the high diagnostic performance needed for clinical adoption in pathology workflows. Ensemble methods, which combine predictions from multiple models, have shown promise in improving segmentation accuracy and robustness [13]. Additionally, quantifying the uncertainty of segmentation predictions can provide valuable information for clinical decision-making and help identify areas requiring further expert review [14]. In 2019, the ACDC@LungHP challenge [15] was introduced to evaluate computer-aided diagnosis systems for segmenting lung cancer WSIs. This challenge highlighted two different paradigms: single-model approaches using individual models, and multi-model approaches integrating diverse models.

Single CNN models have demonstrated success in accurately delineating tumors. However, relying on a single architecture may limit the ability to capture the intricacies and variations in tumor morphology across diverse WSIs. To address this limitations, recent research has explored ensemble learning techniques that combine predictions from multiple models. By integrating diverse model architectures such as CNNs, ResNets, and Transformers into a heterogeneous ensemble, we can leverage their complementary strengths and improve overall accuracy in tumor segmentation tasks [16,17]. The multi-model approach has yielded superior lung tumor segmentation results by acknowledging the complexity of this task. Ensemble techniques have demonstrated consistent improvements in lung tumor segmentation by harnessing complementary insights from individual models [18,19].

Traditional ensemble techniques, such as simple averaging [20] and majority voting [21], often treat all models equally during inference, which fails to address inter-model performance variability on different input images. Averaging predictions in this manner can lower overall performance if certain models underperform significantly for a given input. For example, if one model in the ensemble has low accuracy for a specific class of images, its predictions will negatively impact the final output even if the other models perform well on those images [22]. Additionally, traditional ensembles do not selectively leverage individual model strengths, missing opportunities to optimize based on intrinsic capabilities. One method that can be employed in ensemble models is the STAPLE (Simultaneous truth and performance level estimation) algorithm [23]. STAPLE takes as input a collection of maps, which may come from different segmentation algorithms, and estimates both a probabilistic true segmentation as well as a performance level for each input segmentation simultaneously. However, the performance estimates produced for each algorithm do not account for image-specific variability. To better handle inter-model performance differences on a per-image basis, more advanced ensemble techniques that dynamically weight or select models based on their predicted performance for each input would be beneficial. Such adaptive ensembles could potentially improve overall accuracy by leveraging the strengths of individual models and mitigating the impact of their weaknesses in an input-dependent manner.

To address traditional ensemble limitations, recent research has explored integrating uncertainty estimation during testing [24,25]. Uncertainty estimation involves assigning confidence metrics to the predictions made by deep learning models [26,27]. These uncertainty estimates help indicate how reliable a model's outputs are for a given input image. For example, Maadi et al. [28] proposed a two-stage selective bagging model that considers both uncertainty and accuracy for classifier selection in ensemble learning. In another study, Maadi et al. [29] introduced a performance measure that integrates uncertainty and accuracy for feature selection in classification. While uncertainty estimation is largely used for classification tasks, it is not yet widely employed for segmentation networks [30]. In this paper, we introduce an Adaptive Uncertainty-based Ensemble (AUE) model for lung tumor segmentation in histopathological images. Our AUE model leverages estimated uncertainty to select the best-performing individual models during inference for each WSI. To the best of our knowledge, this work represents the first attempt to leverage predictive uncertainty within an ensemble pipeline for the task of image segmentation. The key contributions of this work can be summarized as:

- We propose an AUE model that leverages estimated uncertainty to dynamically select the optimal individual models during inference for each WSI. By identifying the best-performing models for the current image, our approach enhances overall segmentation performance.
- We introduce a novel technique to quantify uncertainty for each model. These model-specific uncertainty estimates are utilized to calibrate the AUE during testing. By assessing each model's uncertainty, we can better predict segmentation accuracy and perform refined tumor delineation for a given WSI.
- We incorporate color normalization and smart patch extraction into our pipeline to focus model attention on salient regions of interest. This improves segmentation precision while reducing computational costs by excluding non-informative areas.
- We validate our methodology on the public ACDC@LungHP dataset and benchmark against state-of-the-art approaches from the challenge. We also systematically compare the ensemble to individual models and analyze the impact of different aggregation techniques like majority voting and median ensemble.

This paper is structured as follows: Section 2 provides a comprehensive overview of the proposed method, while Section 3 details the experimental results. Finally, Sections 4 and 5 offer a thorough discussion of the overall work.

2. Materials and methods

In this paper, we present an AUE method for lung cancer segmentation. The workflow of our approach is illustrated in Fig. 1. Our AI pipeline for lung cancer segmentation consists of three key steps: i) training the individual models, ii) testing and calibrating each model based on uncertainty estimates, and iii) utilizing the AUE model for semantic segmentation of cancerous regions. We provide a detailed description of our methodology in the following sections.

2.1. Dataset

The dataset used in this study is from the ACDC@LungHP challenge [15], comprising 200 WSIs stained with Hematoxylin and Eosin (H&E). This dataset covers the major lung cancer types, including squamous cell carcinoma, small cell carcinoma, and adenocarcinoma, with an approximate distribution ratio of 6:3:1. The organizers of the challenge divided the entire dataset into two groups: 150 WSIs were randomly selected for the training set, while the remaining 50 were used as the test set. As the test set for this challenge is not publicly accessible, we divided the available 150 WSIs into three subsets for the development of our method: 110 WSIs for the training set, 15 WSIs for the validation set, and 25 WSIs for the test set.

2.2. Data preparation and models training

The original WSIs in the ACDC@LungHP dataset have a resolution of 0.5 μm/pixel (20x magnification). Stain normalization was applied to the entire WSI at this original resolution to reduce color variability across WSIs, irrespective of the specific lung cancer subtype. Stain normalization is commonly used as a pre-processing step in deep learning frameworks to reduce stain variability and improve diagnostic algorithms [16,27]. This operation involves normalizing the color profile of an input image to match a template image.

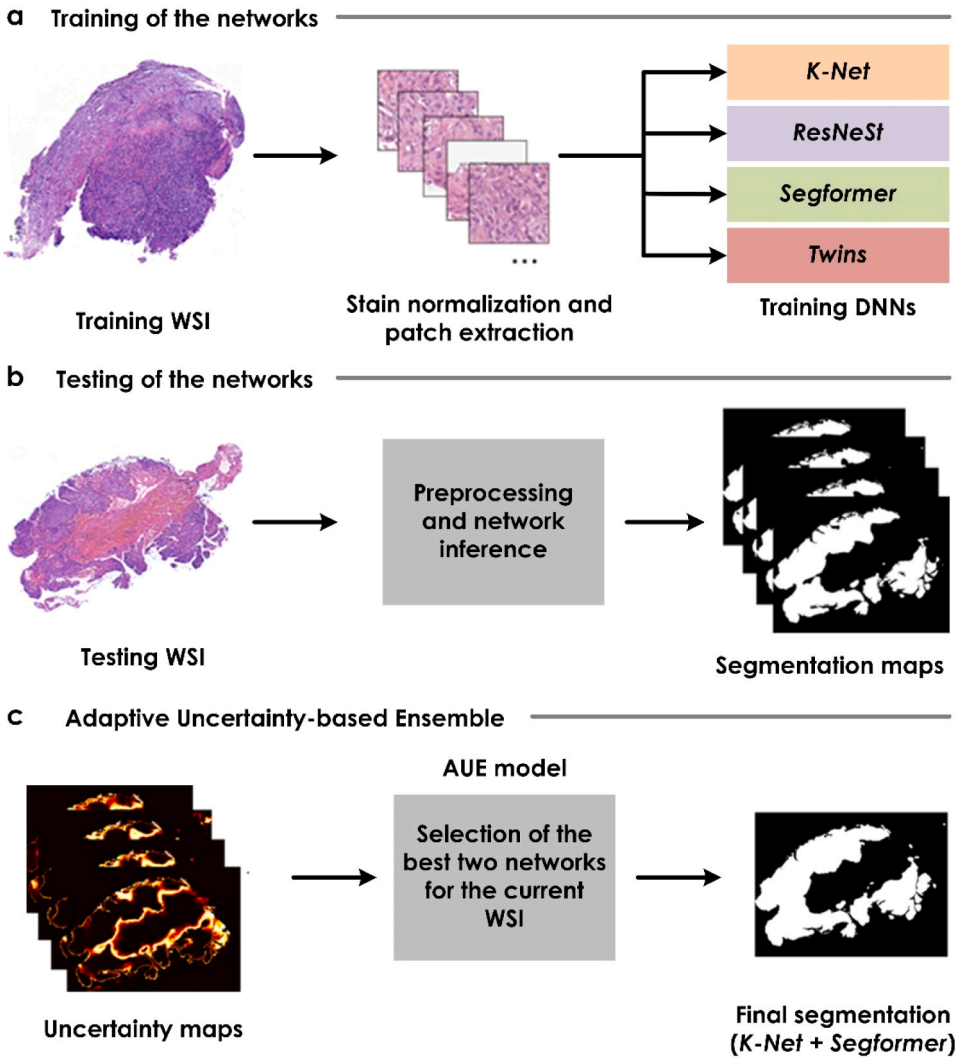

Fig. 1. Schematic overview of the process for constructing the ensemble model. (a) After preprocessing and patch extraction, the individual models are trained for the segmentation task. (b) Each deep neural network (DNN) is tested and calibrated based on the uncertainty estimates. (c) During inference, the AUE selects the best performing networks for the current WSI to perform tumor segmentation.

For stain normalization, we employ a Generative Adversarial Network (GAN) -based approach similar to that proposed by Bentaieb et al. [31], which demonstrated state-of-the-art performance for color normalization tasks in histology images. The goal is to translate the color pattern of images from one domain (domain A) to the color pattern in another domain (domain B), where domain A exhibits a wide range of color patterns and domain B has a relatively uniform color pattern. The model used in this work is a Pix2Pix GAN model with a U-net generator and a PatchGAN discriminator [32]. Fig. 2a shows the stain normalization process on a sample image. The proposed algorithm is designed to extract patches from informative regions, which are most likely to contain cancerous tissue. We define informative regions as areas within the WSI with a high density of cell nuclei, as changes in nuclei morphology and architecture are key indicators of lung cancer. Our patch extraction technique is inspired by strategies introduced by Janowczyk et al. [33], who developed targeted extraction methods focused on nuclear regions to improve training efficiency. Specifically, the algorithm segments the image to detect unstained areas and nuclear regions. Since simple thresholding may be ineffective for segmenting all the white areas, our algorithm employs Gaussian filtering to smooth the original image before thresholding at 95 % of the maximum filtered value. To detect nuclei positions, it uses the segmentation algorithm from our previous work [34].

To reduce the computational cost of subsequent processing steps, we extracted tiles from the normalized WSIs. These tiles had a dimension of 3840×3840 pixels at 20x magnification (0.5 μm/pixel). From these tiles, we extracted patches of size 768 × 768 pixels with a 50 % overlap. The patch size was chosen to provide sufficient context for the segmentation model while keeping computational requirements manageable. The 50 % overlap ensures adequate coverage of all tile regions and helps capture spatial dependencies between adjacent patches during segmentation. Specifically, a smart patch extraction is adopted that exploits the segmentation masks (nuclei and white) previously identified. All patches that show a minimum of 10 % area covered by nuclei and a maximum of 80 % white areas are selected by the algorithm to train the deep network. In this way, patches with a high nuclei density are considered informative and are selected for further processing, while patches containing mostly stromal tissue or large areas of white background are excluded. The pseudo-code and example results obtained with this selective patch extraction strategy are illustrated in Fig. 2b.

Fig. 3 illustrates the four different architectures that compose our ensemble model: K-Net, ResNeSt, Segformer, and Twins. These architectures were selected because they represent state-of-the-art approaches in semantic segmentation and utilize diverse mechanisms for feature extraction and decoding. K-Net and ResNeSt are based on CNNs, while Segformer and Twins are built upon vision transformers. K-Net

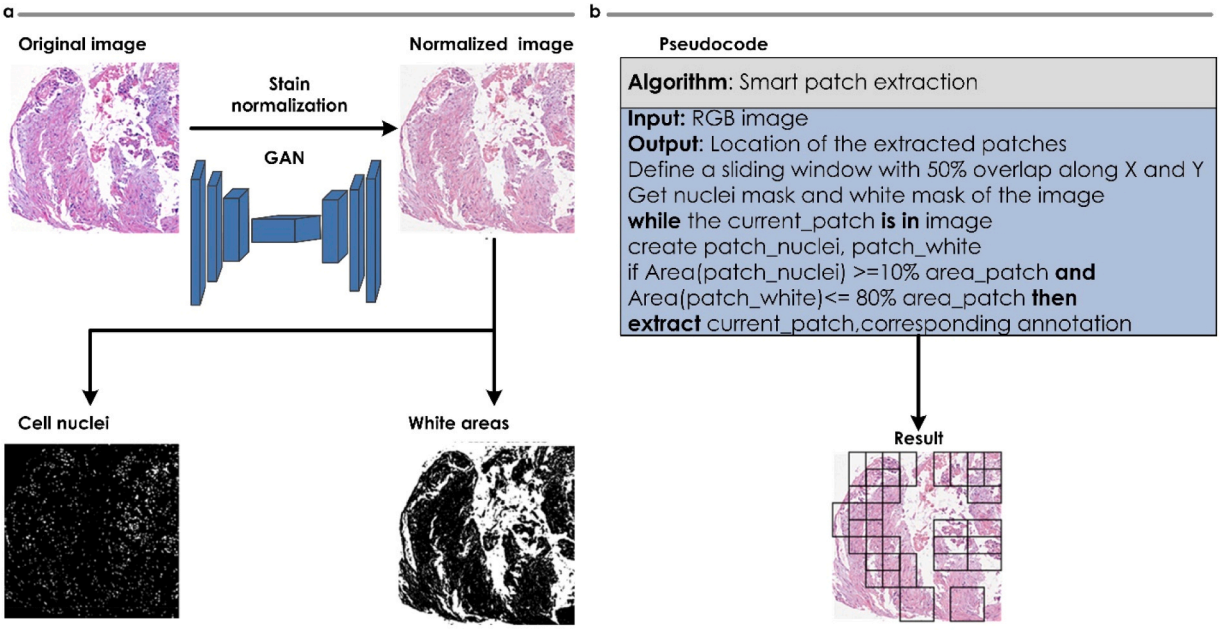

Fig. 2. Patch extraction employed in this work. (a) First, stain normalization is employed to standardize the color appearance of the histological images. Then, all the relevant structures (cell nuclei and white areas) are extracted to guide the patch extraction. (b) Pseudocode and results obtained with our smart patch selection approach.

[35] employs a dual-branch structure to extract multi-scale features while ResNeSt model [36] leverages a split-attention mechanism to capture both local and global dependencies. Segformer [37] utilizes a hierarchical transformer encoder to extract features at different scales, and Twins [38] employs a two-branch architecture with self-attention and locally-grouped self-attention to capture both global and local context. By incorporating these diverse architectures into our ensemble, we aim to exploit their complementary strengths and improve the overall segmentation performance. The combination of CNN-based and transformer-based models allows our ensemble to benefit from both the local feature extraction capabilities of CNNs and the global context modeling abilities of transformers. All four networks were trained using the hyperparameters listed in Table 1.

2.3. Adaptive uncertainty-based ensemble (AUE)

The construction of the AUE model consists of 2 phases: ensemble calibration and model implementation.

2.3.1. AUE calibration

In this work, we quantify the uncertainty of each individual model's predictions using Monte Carlo (MC) dropout [39]. MC dropout involves applying dropout regularization during inference to generate multiple predictions for a given input, allowing us to estimate the model's uncertainty by analyzing the variability in these predictions.

We strategically place dropout layers at different depths within each network architecture to capture uncertainty at various stages of the feature extraction process [40]. In K-Net, dropout is applied after layers 3, 6, and 9; in ResNeSt, after the split attention blocks; in Segformer, following the multi-scale feature fusion stages; and in Twins, after the transformer encoders. This approach allows us to assess the confidence of the learned features at different levels of abstraction, from low-level to high-level representations.

For each model, we generate five random MC samples per WSI of the validation set by applying dropout with a probability of 0.5 [41]. The probability p_r of the tumor region r is then computed by averaging the T samples using the following equation [41]:

$$p_r = \frac{1}{T}\sum_{t=1}^{T} p_{r,t} \tag{1}$$

From these probabilities, we compute the normalized entropy [41], defined as:

$$H = -[p_r \log p_r + (1-p_r)\log(1-p_r)]\frac{1}{\log(2)} \mathrm{H} \in [0,1] \tag{2}$$

From the MC dropout, we generate an uncertainty map for each WSI in the validation set (Fig. 4a). The uncertainty estimates obtained from MC dropout reflect the reliability of each model's predictions for a given input image. In our AUE model, we leverage these uncertainty estimates to dynamically select the two most confident models for each input image during inference. By choosing the models with the lowest uncertainty, we prioritize the predictions that are most likely to be accurate for the specific input. It is important to note that the uncertainty in our AUE model is derived from the variability in each individual model's predictions when MC dropout is applied, rather than from a difference map between two models.

To quantify the uncertainty for each WSI, we calculate the median of the non-zero values in the uncertainty map. We then compare the Dice similarity coefficient from the segmentation map to the uncertainty value to identify potential correlations between uncertainty and segmentation accuracy. The motivation behind using the correlation between the Dice score and the median entropy stems from the fundamental principles of uncertainty quantification in deep learning models. In the context of semantic segmentation, we expect that a model with low uncertainty would produce more accurate and consistent segmentations (Dice score) across multiple forward passes. On the other hand, the choice of using the median entropy as a summary statistic for the uncertainty map is based on its robustness to outliers and its ability to capture the central tendency of the uncertainty distribution.

Since each of the four architectures has different dropout placements, we calibrate the uncertainty individually for each network. The goal is to relate the uncertainty from MC dropout to the Dice score, which indicates segmentation quality. Fig. 4b shows the calibration curves learned to relate estimated uncertainty to the Dice score for each

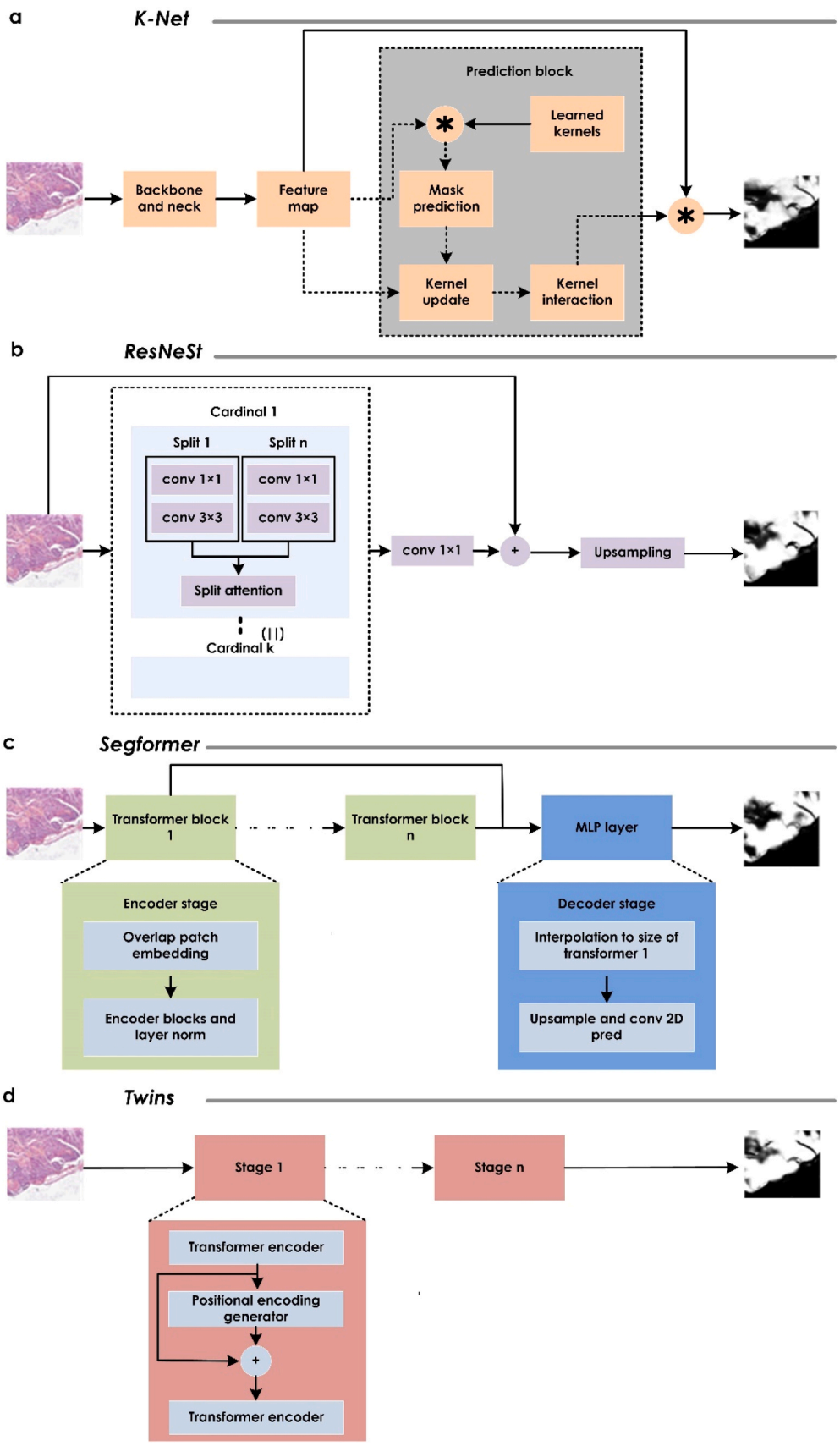

Fig. 3. Segmentation networks that compose our ensemble model. (a) K-net, (b) ResNeSt, (c) Segformer, (d) Twins. In K-Net architecture, a Swin Transformer backbone is employed with a UperNet neck module.

Table 1
Hyperparameters and settings used during model training.

Hyperparameter	K-Net [35]	ResNeSt [36]	Segformer [37]	Twins [38]
Backbone	PCPVT	ResNetV1c	MixVisionTransformer	PCPVT
Depths	[3, 4, 6, 3]	50	[2, 2, 2, 2]	[3, 8, 27, 3]
Embed Dims	[64, 128, 320, 512]	-	64	[64, 128, 320, 512]
Num Heads	[1, 2, 5, 8]	-	[1, 2, 5, 8]	[1, 2, 5, 8]
Patch Sizes	[4, 2, 2, 2]	-	[7, 3, 3, 3]	[4, 2, 2, 2]
Strides	[4, 2, 2, 2]	[1, 2, 1, 1]	-	[4, 2, 2, 2]
MLP Ratios	[8, 8, 4, 4]	-	4	[8, 8, 4, 4]
SR Ratios	[8, 4, 2, 1]	-	[8, 4, 2, 1]	[8, 4, 2, 1]
Decoder Channels	128	512	256	256
Decoder Dropout Ratio	0.1	0.1	0.1	0.1
Learning Rate	0.0001	0.0001	0.0001	0.0001
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight Decay	0.0005	0.0005	0.0005	0.0005
Loss function	Dice	Dice	Dice	Dice
Batch Size	4	4	4	4
N° epochs	80	80	80	80

model. As can be seen from the same figure, a strong negative correlation is observed between these two variables, showing lower segmentation uncertainty corresponds to higher tumor segmentation accuracy. These curves also demonstrate that our approach is able to align uncertainty with segmentation accuracy through linear regression [42]. Specifically, we establish a linear association between the median non-zero uncertainty values and Dice scores for the full validation set. For each model, we obtain the slope (m) and intercept (q) using:

$$y = mx + q \tag{3}$$

Where 'y' is the Dice score and 'x' is the median non-zero uncertainty value. This calibration aligns the uncertainty estimates with the Dice scores, improving the model's segmentation performance.

2.3.2. AUE implementation

The calibration functions obtained for each of the four models on the validation set are utilized during inference on the test set (Fig. 5). To make predictions on a test whole slide image (WSI), we follow these steps:

1. MC dropout is applied to the test WSI for all four trained models to capture prediction uncertainties. Five MC samples are generated per model, producing uncertainty maps using Eq. 1 and 2.
2. The median of the positive uncertainty values is calculated from each model's uncertainty map, providing a per-model uncertainty measure. The expected Dice scores are estimated using the linear regression functions from the validation set, quantitatively evaluating segmentation performance.

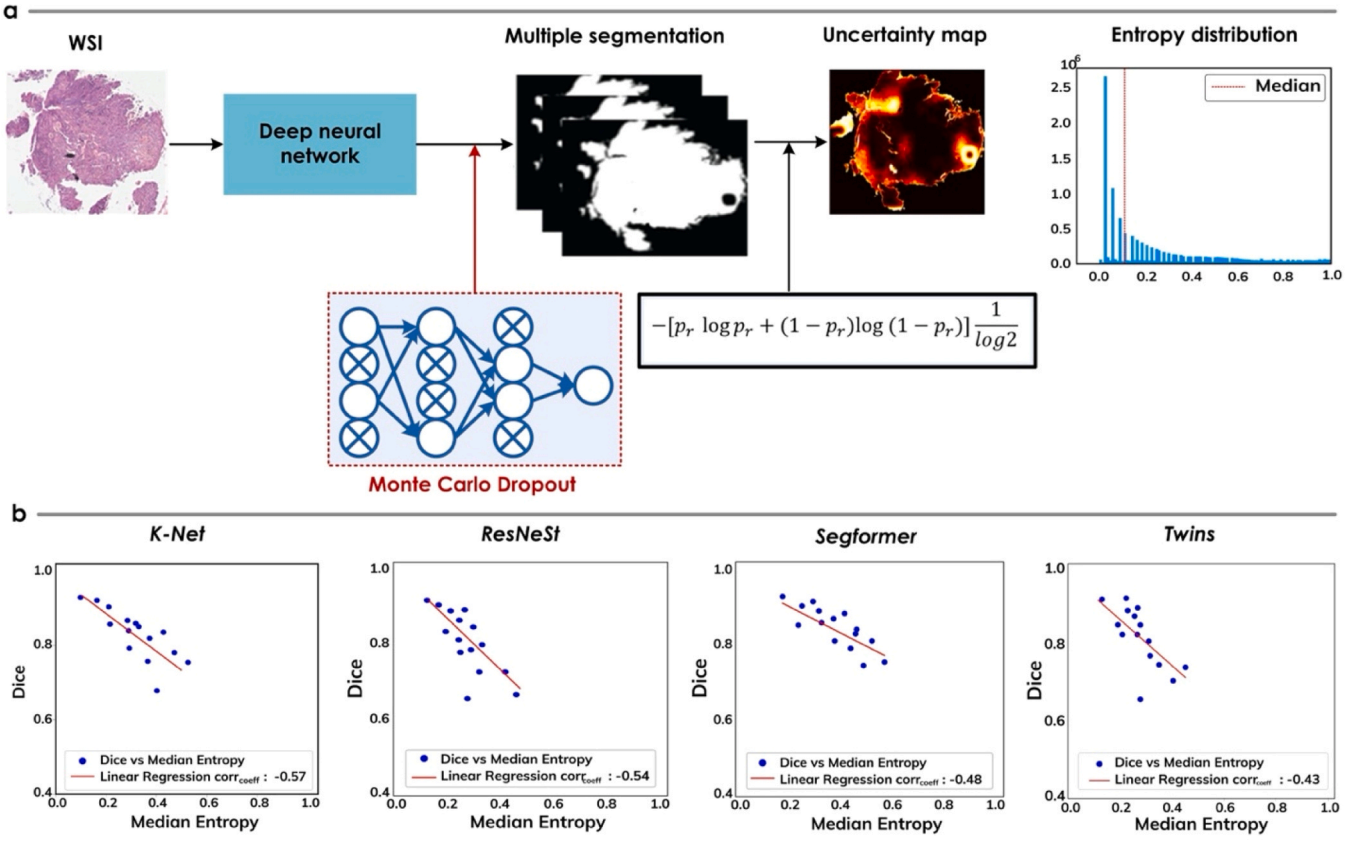

Fig. 4. AUE calibration process. (a) Multiple inferences are performed using Monte Carlo dropout to estimate a normalized entropy uncertainty map. The median uncertainty value is extracted from each WSI as its overall measure of uncertainty. (b) For each of the four neural networks used in the AUE, the plots approximate the relationship between the estimated uncertainty values and the corresponding Dice scores for the WSIs in the validation set. This calibration helps assess the accuracy of the uncertainty estimates for each network.

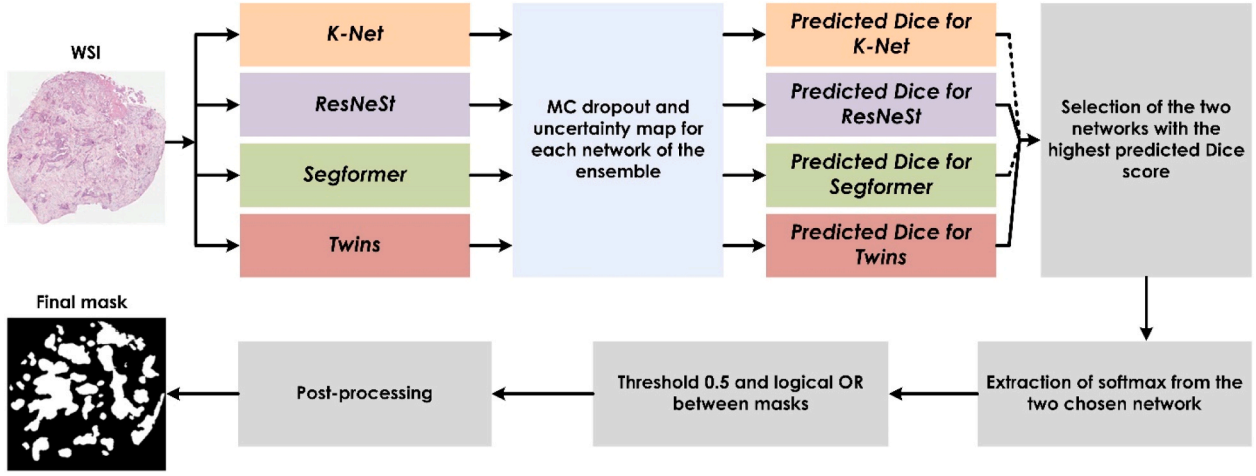

Fig. 5. AUE implementation during test time. Monte Carlo dropout is applied to each of the four trained models to generate uncertainty maps using normalized entropy. Then, the median uncertainty value is extracted from each map. Expected Dice scores are estimated using regression on the validation set and the two models with the highest Dice scores are selected for final segmentation. Softmax maps from these models are extracted and thresholded at 0.5 to create binary masks. A logical OR combines the masks to generate the segmented tumor mask and post-processing is applied to further refine the result.

3. The two models with the highest expected Dice scores are selected to perform the final tumor segmentation, ensuring accurate and reliable model contributions.
4. Without dropout, the probability maps (softmax values) are extracted from the chosen models, generating two binary tumor masks by thresholding at 0.5. A logical OR operation between the masks produces the final tumor segmentation mask.

This uncertainty-guided selection process enhances the segmentation performance by leveraging the most reliable models, while effectively managing model uncertainty. To obtain the final mask, morphological operations are applied to the ensemble output. These morphological operations are necessary to address potential issues, such as disconnected objects or small holes that may arise during the fusion of two predictions. The post-processing step serves to regularize the shape of the final prediction by applying dilation with a 20-pixel disk, filling holes over 10,000 pixels, and erosion with a 10-pixel disk to refine boundaries. We also refined the mask by removing segmented areas outside the tissue regions, following the same approach used by several authors who participated in the challenge [15]. This step helps maintain the consistency of the segmentation mask.

2.4. Evaluation metrics and performance comparison

To assess the performance of the AUE model, we utilize a comprehensive set of segmentation metrics. The Dice score, which serves as our primary evaluation criterion, is the reference metric of the ACDC@LungHP challenge [15]. In addition to the Dice score, our evaluation framework includes secondary metrics. These metrics encompass sensitivity (true positive rate), specificity (the model's ability to identify true negative instances), and accuracy (the overall correctness of predictions). While the Dice score remains our primary metric for ranking and assessing segmentation quality, the inclusion of sensitivity, specificity, and accuracy ensures a comprehensive evaluation, enhancing our understanding of the model's performance in the context of lung cancer segmentation.

In this study, we assess the robustness of our proposed model by incorporating the DICE curve as an additional metric. The DICE curve is generated by examining the relationship between the mean DICE value of the AUE model and different segmentation thresholds applied to the ensemble's output [43]. Furthermore, we compare our proposed AUE model to five traditional ensemble techniques for merging individual network predictions:

1. Mean Ensemble: The probability maps from the four networks are averaged, and the resulting map is thresholded at 0.5 to generate the final tumor mask. This approach treats all networks equally and combines their outputs through a simple averaging operation.
2. Logical AND Ensemble: The softmax outputs from each network are thresholded at 0.5 to create binary masks. These masks are then combined using a logical AND operation, where a pixel is considered part of the tumor only if all networks agree. This strict approach favors precision over recall.
3. Logical OR Ensemble: Similar to the Logical AND Ensemble, the softmax outputs are thresholded at 0.5 to create binary masks. However, in this case, a logical OR operation is applied, where a pixel is considered part of the tumor if any of the networks predict it as such. This approach prioritizes recall over precision.
4. Majority Voting Ensemble: The softmax outputs from each network are thresholded at 0.5 to create binary masks. The final tumor mask is obtained through majority voting, where a pixel is considered part of the tumor if at least half of the networks agree. This approach aims to balance precision and recall.
5. STAPLE: STAPLE algorithm [23] is applied to the thresholded softmax outputs from each network. This algorithm simultaneously estimates the true segmentation and the performance level of each network, producing a final segmentation map based on a threshold of 0.5. STAPLE takes into account the agreement between networks and their estimated performances to generate a weighted combination of the individual predictions.

To assess the statistical significance of our results, we employed a two-tailed paired t-tests to compare the performance of our AUE model against other ensemble methods on the test set (N=25). Prior to applying the t-tests, we verified the underlying assumptions. The normality of the differences between paired observations was confirmed using Shapiro-Wilk tests for each comparison. Our performance metrics (Dice score, accuracy, sensitivity, and specificity) are continuous variables ranging from 0 to 1, satisfying the continuity assumption. All statistical tests were conducted with a significance level of 0.05.

3. Results

3.1. Segmentation performance

Table 2 compares the test set segmentation results of the four individual models described in the Methods with the top-ranked model from

Table 2
Single model performance on test set. Comparison between the top-ranked single model of the challenge and individual models of our ensemble.

Network	Dice score	Accuracy	Sensitivity	Specificity
Single model ranked 1st on ACDC-LungHP [15]	0.7700	0.9375	0.8003	0.9567
Knet	0.8362	0.9526	0.8190	0.9639
Twins	0.8291	0.9500	0.8189	0.9618
Segformer	0.8322	0.9525	0.8079	0.9683
Resnest	0.8239	0.9499	0.8150	0.9743

the ACDC-LungHP challenge. Among the proposed single models, K-Net achieves the highest mean Dice score on the test set at 0.8362 while the ResNeSt model has the lowest at 0.8239. All four individual models outperformed the top ACDC-LungHP model, which had a mean Dice of 0.77.

Table 3 compares the proposed ensemble model (AUE) to the top ACDC-LungHP challenge participant's multi-model approach. The AUE model achieved a mean Dice of 0.8653 on the test set, while the top participant had 0.8373, an improvement of around 3 % for AUE. This demonstrates the superior performance of the AUE ensemble versus the current multi-model approach.

3.2. AUE vs. single models

Fig. 6 and Fig. 7 show two examples of how the AUE model greatly improves tumor segmentation accuracy compared to the individual networks. This enhancement is achieved by intelligently combining the predictions of the two models and leveraging the information provided by their respective prediction uncertainties.

Fig. 8 presents a comparison of the Dice score distributions for the AUE model and the individual networks (K-Net, ResNeSt, Segformer, Twins) on the test set. The AUE model achieves a mean Dice coefficient of 0.8653, which is 2–4 % higher than the performance of each individual network alone (Table 2). This positions AUE as a highly competitive solution, especially in terms of Dice and sensitivity. To confirm the statistical significance of our ensemble's improved accuracy, paired t-tests were conducted between AUE and individual models. All tests reported p-values < 0.05, confirming the significant performance improvement achieved by AUE. Additionally, AUE outperforms all individual networks in DICE curves, with a significantly higher AUC-DICE of 0.84 (Fig. 8b). The Dice coefficient of the AUE model remains stable as the segmentation threshold changes, unlike the substantial decreases observed with the individual networks. This suggests that the proposed AUE model has superior coherence and stability in tumor segmentation compared to the individual networks. This suggests that the proposed AUE model has higher consistency and stability in tumor segmentation compared to the individual networks.

To further analyze the performance variability, we calculated the coefficient of variation (CV) of the Dice score for each model. The CV values are as follows: AUE (7.22 %), K-Net (9.37 %), ResNeSt (10.75 %), Segformer (9.53 %), and Twins (10.27 %). The lower CV value for the AUE model indicates that it exhibits less relative variability in performance compared to the individual networks. This suggests that the AUE model not only improves the overall segmentation accuracy but also provides more consistent results across different test images.

Table 3
Multi-model performance on test set. Comparison between the top-ranked multi-model of the challenge and the proposed AUE model.

Network	Dice score	Accuracy	Sensitivity	Specificity
Multi-model ranked 1st on ACDC-LungHP [15]	0.8373	0.9505	0.9052	0.9531
AUE (proposed)	0.8653	0.9689	0.9244	0.9673

3.3. AUE vs. traditional ensemble models

We evaluate our novel AUE model in two different configurations: AUE with logical AND between the best estimated predictions, and AUE with logical OR between the best predictions (as described in Section 2.3). This is done to identify the most effective ensemble approach. As shown in Table 4, AUE with OR operator outperforms all five traditional ensembles, achieving superior mean Dice of 0.8653, accuracy of 0.9689, and sensitivity of 0.9244 on the test set. Paired t-tests confirm that AUE (OR) significantly outperforms all traditional ensemble methods ($p < 0.05$) in almost all evaluation metrics (Dice score, Accuracy, Sensitivity). While AUE excelled overall, its specificity of 0.9673 is slightly lower than the logical AND ensemble. However, AUE's specificity remains competitive and reasonably robust, balancing correct non-tumor identification while minimizing false negatives (high sensitivity).

4. Discussion

Combining the predictions of single deep learning models into an effective ensemble model remains an open challenge. Traditional ensemble methods, such as majority voting and average pooling, often yield suboptimal results by oversimplifying how they treat the constituent networks. These traditional methods treat all networks within the ensemble equally, regardless of their varying performance on specific images. This one-size-fits-all approach can have negative effects on ensemble performance, as the limitations of one model can propagate across the entire ensemble.

In our work, we introduce a novel and effective strategy called AUE model to address this limitation during test time. Through our proposed approach, we gain valuable insights during inference, allowing for better control of the individual networks that compose the ensemble. Specifically, thanks to uncertainty estimation, we are able to select the "best" pair of networks for each input data.

We tested our approach on the ACDC-LungHP challenge [15] for lung cancer segmentation from histopathological slides. Lung cancer segmentation from histopathology slides is challenging due to variations in stain intensity and tissue structure between slides. Accurately segmenting tumor boundaries is essential for treatment planning, assessing tumor progression, and monitoring treatment response. Our pipeline contains a stain normalization procedure as a pre-processing step and a "smart patch extraction" technique that extracts patches only from informative tissue regions. Notably, our ensemble leverages a model pool consisting of K-Net, ResNeSt, Segformer and Twins - all of which individually outperform the top-ranked single model [15] from the ACDC-LungHP challenge in terms of Dice score (Table 2). The models' performance is largely due to the effective use of stain normalization and patch extraction techniques during training.

By employing calibration functions and adopting a model selection process based on uncertainty quantification, our ensemble approach provided reliable and accurate tumor segmentation for each WSI during testing. The correlation values between the uncertainty estimates and the Dice scores range from 0.43 to 0.57, indicating a moderate correlation. While a stronger correlation would be ideal, it is important to consider the purpose and impact of this correlation in the context of our AUE model. The primary objective of establishing this correlation is to enable the AUE model to estimate the Dice scores during the testing phase and combine the most reliable predictions adaptively, rather than relying on simple averaging like traditional ensemble models. Even a moderate correlation can provide valuable information for guiding the ensemble's decision-making process, as it allows the AUE model to prioritize the predictions from models that are expected to perform better on a given input image. This approach significantly improved the overall performance and generalization capability of the segmentation models, achieving around a 3 % improvement compared to the multi-

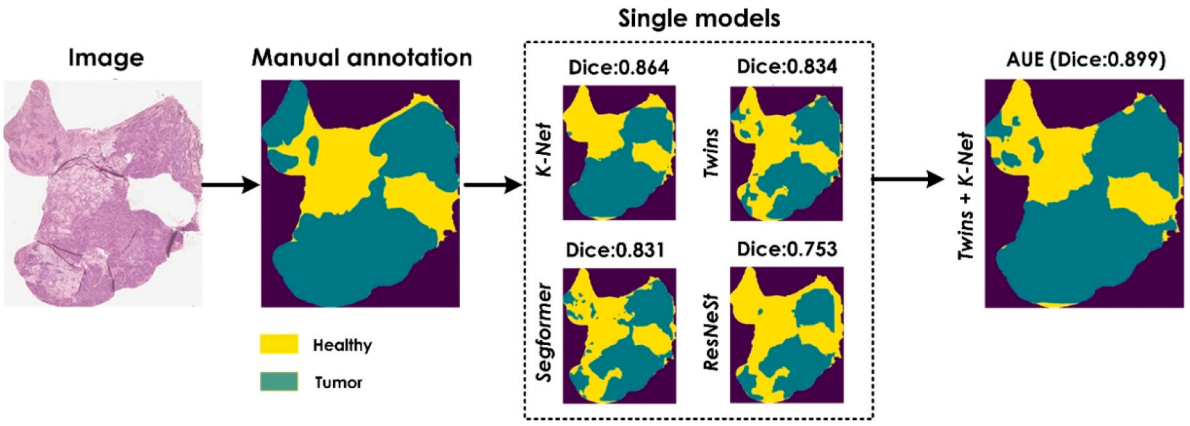

Fig. 6. Sample #1 - Application of AUE on a WSI from the test set. The AUE model is able to select the two best performing networks of the ensemble to produce the final segmentation mask with more than 3 % improvement over the single model.

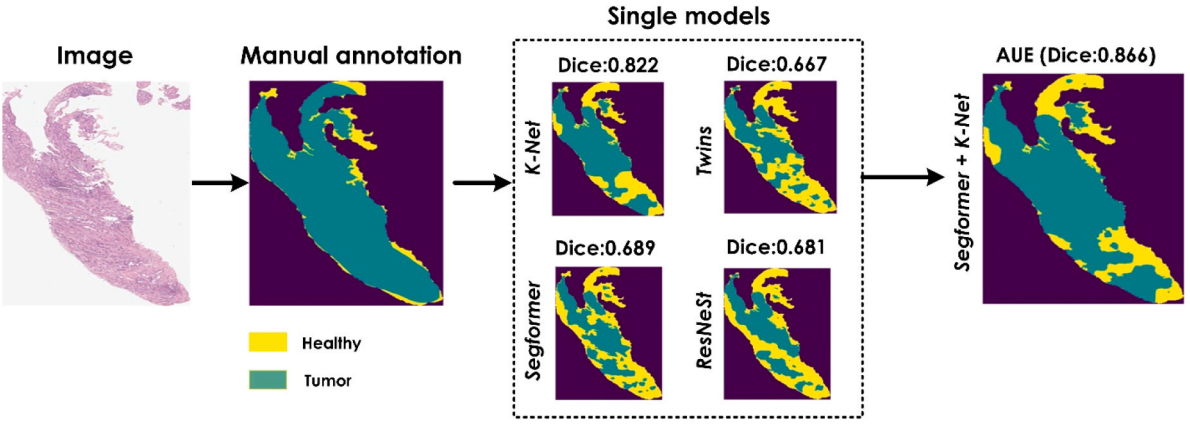

Fig. 7. Sample #2 - Application of AUE on a WSI from the test set. The AUE model is able to select the two best performing networks of the ensemble to produce the final segmentation mask with more than 4 % improvement over the single model.

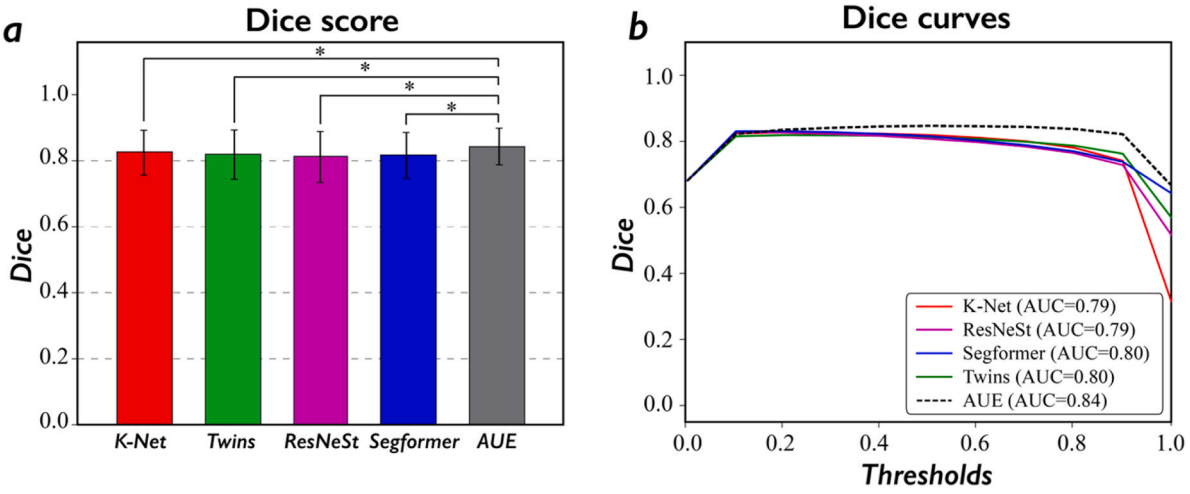

Fig. 8. (a) Comparative analysis of Dice scores: AUE vs. individual models, with statistically significant differences marked by *. (b) Dice Score Curves: AUE vs. Individual Models.

model ranked first in the challenge [15]. This notable improvement underscores the effectiveness of leveraging uncertainty information to optimize ensemble predictions. The obtained results highlight the efficacy of the AUE model in the challenging task of lung tumor segmentation across entire WSIs. Notably, our approach employs precise model selection during inference, effectively exploiting the most optimal models for each specific WSI. Rigorous t-tests confirm the statistical significance of this methodology in combining the two highest-performing models during testing.

Our ensemble model not only enhances segmentation accuracy but

Table 4
Comparison between AUE and traditional ensemble models. Best performance for each metric is highlighted in bold. The asterisk (*) indicates a statistically significant difference between the AUE method and all compared methods based on a paired t-test.

Ensemble method	Dice score	Accuracy	Sensitivity	Specificity
Mean of the four networks	0.8394	0.9383	0.8203	0.9675
Logical AND of the four networks	0.7821	0.9066	0.6927	**0.9853**
Logical OR of the four networks	0.8532	0.9361	0.8972	0.9333
Majority voting between the four networks	0.8254	0.9349	0.7907	0.9826
STAPLE [23]	0.8424	0.9451	0.8235	0.9615
AUE with logical AND (proposed)	0.8355	0.9247	0.7899	0.9745
AUE with logical OR (proposed)	**0.8653 (*)**	**0.9689 (*)**	**0.9244(*)**	0.9673

also outperforms traditional ensemble methods, which often overlook the varied performance of individual models across different images. AUE achieves a remarkable Dice score of 0.8653, surpassing traditional ensembles. Importantly, this superior segmentation performance maintains a critical maximal sensitivity of 0.9244 for tumor segmentation tasks. By correctly managing uncertainty during testing, our model can effectively merge only the reliable predictions for each input data.

It is important to acknowledge that our AUE model relies on a semi-empirical approach, as it leverages the correlation between the uncertainty estimates and the Dice scores to guide the ensemble's decision-making process. While this correlation provides valuable information for combining the most reliable predictions, it is not a perfect measure of segmentation accuracy. The effectiveness of the AUE model is contingent upon the following hypotheses and conditions:

1. The uncertainty estimates obtained through MC dropout are informative and capture the model's confidence in its predictions. We assume that lower uncertainty corresponds to higher segmentation accuracy, which is supported by the observed moderate correlation between the median entropy and the Dice scores.
2. The calibration curve, which relates the uncertainty estimates to the Dice scores, is representative of the relationship between model confidence and segmentation accuracy. We hypothesize that this relationship is consistent across different images within the dataset and can be approximated using linear regression.
3. The AUE model's performance is expected to be optimal when the individual models in the ensemble exhibit diversity in their predictions and uncertainties. If all models consistently produce similar uncertainties, the AUE model may not provide significant improvements over traditional ensemble methods.

Our AUE model is not without limitations. Currently, it relies on a fixed pool of models for ensemble selection. Future work could explore expanding the model pool to further enhance performance. Additionally, while our current calibration method employs the median of entropy value as the uncertainty measure, refinements may involve using additional features from the uncertainty map. Optimizing the calibration process with more features and with non-linear calibration could enable even better performance. Another limitation is the computational complexity of the ensemble approach. The time complexity of our AUE approach depends on several factors, including the number of individual models in the ensemble, the complexity of each model's architecture, and the size of the input images. In the inference phase, our AUE model requires running each individual model on the input image to obtain their segmentation predictions and uncertainty estimates. The time complexity of this step is linear with respect to the number of models in the ensemble. Compared to single-model approaches, our AUE model has a higher computational cost during inference due to the need to run multiple models. However, this increased cost is offset by the improved segmentation accuracy and robustness achieved through the ensemble approach. In comparison to traditional ensemble methods, such as majority voting, our AUE model has a higher time complexity, as it also performs the additional step of uncertainty estimation. However, the number of samples used by MC dropout ($n=5$) is lower than that used in current literature (typically $n \geq 20$) [44,45]. Exploring techniques to optimize the ensemble process or designing efficient architectures can address these computational limitations.

As we continue to refine our AUE model, we aim to address the current limitations and explore potential improvements. One avenue for enhancement is to investigate alternative uncertainty quantification methods beyond MC dropout, such as Bayesian neural networks, which may provide more robust and reliable uncertainty estimates. Another potential area for improvement is to explore more advanced calibration methods for relating the uncertainty estimates to the segmentation accuracy. While our current approach uses a linear regression model to approximate this relationship, non-linear or probabilistic calibration methods may better capture the complex interactions between uncertainty and performance.

Our upcoming research aims to broaden the comparative analysis of our AUE model with a more comprehensive set of state-of-the-art methods for lung cancer segmentation in digital pathology. Additionally, we will explore and compare different ensemble techniques, such as weighted averaging, and stacking [46], to further assess the effectiveness of our uncertainty-based ensemble strategy. Furthermore, we will investigate the performance of our AUE model in comparison to other uncertainty-aware approaches, such as Bayesian neural networks [47], and ensemble methods that incorporate uncertainty estimates [48].

Future applications of our approach may extend beyond lung cancer segmentation to encompass other challenging tasks in digital pathology, such as gland [49] or vessel [2] segmentation. Leveraging uncertainty maps more broadly opens the possibility of application to other imaging modalities as well. Investigating and integrating additional uncertainty map features could refine the calibration process and potentially achieve even more accurate and reliable ensemble predictions. This ongoing research holds great promise for the advancement of medical image analysis and, ultimately, for the improvement of diagnosis and treatment of various medical conditions.

5. Conclusion

In this work, we have introduced a novel adaptive uncertainty-based ensemble model for robust lung tumor segmentation in whole slide images. To the best of our knowledge, this work represents the first attempt to leverage predictive uncertainty within an ensemble pipeline for the task of image segmentation. Our approach dynamically selects the best pair of models during testing by leveraging uncertainty estimates from Monte Carlo dropout. Experiments on the ACDC-LungHP challenge dataset demonstrate superior performance over single models, traditional ensembles, and top-ranked methods from the challenge. The proposed ensemble model provides an effective pipeline to manage uncertainty across models and optimally combine predictions, enhancing lung cancer delineation. Our methodology could also be extended to other histopathology segmentation tasks, providing a valuable tool for computer-aided diagnosis. Further work should investigate expanding the model pool and refining the uncertainty calibration process. Overall, this research demonstrates that leveraging predictive uncertainty can optimize ensemble-based medical image analysis by integrating heterogeneous models. Our findings thus advance the development of reliable computer-assisted diagnosis techniques toward improved personalized healthcare.

CRediT authorship contribution statement

Massimo Salvi: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **U Raghavendra:** Writing – review & editing, Visualization, Validation, Conceptualization. **Alessandro Mogetta:** Writing – original draft, Validation, Formal analysis, Data curation. **U Rajendra Acharya:** Writing – review & editing, Supervision. **Anjan Gudigar:** Writing – review & editing, Visualization, Validation. **Filippo Molinari:** Writing – review & editing, Supervision, Investigation.

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.

Data availability

Data will be made available on request.

Acknowledgements

This research was supported by Fondazione Compagnia di San Paolo (Italy). Grant ID: E13C23001660007.

References

[1] G. Litjens, C.I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen - van de Kaa, P. Bult, B. van Ginneken, J. van der Laak, Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis, Sci. Rep. 6 (2016) 26286, https://doi.org/10.1038/srep26286.
[2] M. Salvi, A. Mogetta, K.M. Meiburger, A. Gambella, L. Molinaro, A. Barreca, M. Papotti, F. Molinari, Karpinski score under digital investigation: a fully automated segmentation algorithm to identify vascular and stromal injury of donors' kidneys, Electronics 9 (2020) 1644.
[3] M.K.K. Niazi, A.V. Parwani, M.N. Gurcan, Digital pathology and artificial intelligence, Lancet Oncol. 20 (2019) e253–e261.
[4] J. Van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic, Nat. Med. 27 (2021) 775–784.
[5] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[6] P.K. Gupta, M.K. Siddiqui, X. Huang, R. Morales-Menendez, H. Panwar, H. Terashima-Marin, M.S. Wajid, COVID-WideNet—a capsule network for COVID-19 detection, Appl. Soft Comput. 122 (2022) 108780.
[7] B. Kayalibay, G. Jensen, P. van der Smagt, CNN-based segmentation of medical imaging data, ArXiv Preprint ArXiv:1701.03056 (2017).
[8] Y. Gao, X. Fu, Y. Chen, C. Guo, J. Wu, Post-pandemic healthcare for COVID-19 vaccine: tissue-aware diagnosis of cervical lymphadenopathy via multi-modal ultrasound semantic segmentation, Appl. Soft Comput. 133 (2023) 109947.
[9] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (2021) 209–249.
[10] P. Giraud, M. Antoine, A. Larrouy, B. Milleron, P. Callard, Y. De Rycke, M.-F. Carette, J.-C. Rosenwald, J.-M. Cosset, M. Housset, Evaluation of microscopic tumor extension in non–small-cell lung cancer for three-dimensional conformal radiotherapy planning, Int. J. Radiat. Oncol. Biol. Phys. 48 (2000) 1015–1024.
[11] W.D. Travis, E. Brambilla, A.P. Burke, A. Marx, A.G. Nicholson, Introduction to the 2015 world health organization classification of tumors of the lung, pleura, thymus, and heart, J. Thorac. Oncol. 10 (2015) 1240–1242.
[12] B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, L.H. Schwartz, Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (2016) 23428.
[13] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, Ensembles of multiple models and architectures for robust brain tumour segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3, Springer, 2018: pp. 450–462.
[14] T. Nair, D. Precup, D.L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Med. Image Anal. 59 (2020) 101557.
[15] Z. Li, J. Zhang, T. Tan, X. Teng, X. Sun, H. Zhao, L. Liu, Y. Xiao, B. Lee, Y. Li, Deep learning methods for lung cancer segmentation in whole-slide histopathology images—the acdc@ lunghp challenge 2019, IEEE J. Biomed. Health Inf. 25 (2020) 429–440.
[16] M. Salvi, F. Molinari, S. Iussich, L.V. Muscatello, L. Pazzini, S. Benali, B. Banco, F. Abramo, R. De Maria, L. Aresu, Histopathological classification of canine cutaneous round cell tumors using deep learning: a multi-center study, Front Vet. Sci. 8 (2021) 640944.
[17] Y. Xiao, J. Wu, Z. Lin, X. Zhao, A deep learning-based multi-model ensemble method for cancer prediction, Comput. Methods Prog. Biomed. 153 (2018) 1–9.
[18] M. Bilal, R. Jewsbury, R. Wang, H.M. AlGhamdi, A. Asif, M. Eastwood, N. Rajpoot, An aggregation of aggregation methods in computational pathology, Med Image Anal. (2023) 102885.
[19] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (2010) 1–39.
[20] C. Ju, A. Bibaut, M. van der Laan, The relative performance of ensemble methods with deep convolutional neural networks for image classification, J. Appl. Stat. 45 (2018) 2800–2818.
[21] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Anal. Appl. 6 (2003) 22–31.
[22] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: opportunities and challenges, J. King Saud. Univ. Comput. Inf. Sci. 35 (2023) 757–774.
[23] S.K. Warfield, K.H. Zou, W.M. Wells, Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation, IEEE Trans. Med. Imaging 23 (2004) 903–921.
[24] A. Jungo, R. Meier, E. Ermis, M. Blatti-Moreno, E. Herrmann, R. Wiest, M. Reyes, On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation, September 16-20, 2018, Proceedings, Part I, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Springer, Granada, Spain, 2018, pp. 682–690. September 16-20, 2018, Proceedings, Part I.
[25] T. DeVries, G.W. Taylor, Leveraging uncertainty estimates for predicting segmentation quality, ArXiv Preprint ArXiv:1807.00502 (2018).
[26] A. Loquercio, M. Segu, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, IEEE Robot Autom. Lett. 5 (2020) 3153–3160.
[27] F.G. Zanjani, S. Zinger, B.E. Bejnordi, J.A.W.M. van der Laak, P.H.N. de With, Stain normalization of histopathology images using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 573–577.
[28] M. Maadi, H.A. Khorshidi, U. AickelinUncertainty in Selective Bagging: A Dynamic Bi-objective Optimization Model, in: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), SIAM, 2023: pp. 235–243.
[29] M. Maadi, H.A. Khorshidi, U. AickelinAn Uncertainty-Accuracy-Based Score Function for Wrapper Methods in Feature Selection, in: 2023 IEEE International Conference on Fuzzy Systems (FUZZ), IEEE, 2023: pp. 1–6.
[30] S. Seoni, V. Jahmunah, M. Salvi, P.D. Barua, F. Molinari, U.R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023), Comput. Biol. Med (2023) 107441.
[31] A. BenTaieb, G. Hamarneh, Adversarial stain transfer for histopathology image analysis, IEEE Trans. Med. Imaging 37 (2017) 792–802.
[32] M. Salvi, F. Branciforti, F. Molinari, K.M. Meiburger, Generative models for color normalization in digital pathology and dermatology: advancing the learning paradigm, Expert Syst. Appl. 245 (2024) 123105.
[33] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases, J. Pathol. Inf. 7 (2016) 29.
[34] M. Salvi, F. Molinari, N. Dogliani, M. Bosco, Automatic discrimination of neoplastic epithelium and stromal response in breast carcinoma, Comput. Biol. Med. 110 (2019) 8–14.
[35] W. Zhang, J. Pang, K. Chen, C.C. Loy, K-net: Towards unified image segmentation, Adv. Neural Inf. Process Syst. 34 (2021) 10326–10338.
[36] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: pp. 2736–2746.
[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process Syst. 34 (2021) 12077–12090.
[38] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: revisiting the design of spatial attention in vision transformers, Adv. Neural Inf. Process Syst. 34 (2021) 9355–9366.
[39] Y. Gal, Z. Ghahramani, Bayesian convolutional neural networks with Bernoulli approximate variational inference, ArXiv Preprint ArXiv:1506.02158 (2015).
[40] H. Wu, X. Gu, Towards dropout training for convolutional neural networks, Neural Netw. 71 (2015) 1–10.
[41] A. Jungo, F. Balsiger, M. Reyes, Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation, Front Neurosci. 14 (2020) 282.
[42] E. Bisong, E. Bisong, Linear Regression, Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners (2019) 231–241.
[43] S.R. González, I. Zemmoura, C. Tauber, 3D brain tumor segmentation and survival prediction using ensembles of convolutional neural networks, , October 4, 2020, Revised Selected Papers, Part II 6, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Springer, Lima, Peru, 2021, pp. 241–254. , October 4, 2020, Revised Selected Papers, Part II 6.
[44] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process Syst. 30 (2017).

[45] F.C. Maruccio, W. Eppinga, M.-H. Laves, R.F. Navarro, M. Salvi, F. Molinari, P. Papaconstadopoulos, Clinical assessment of deep learning-based uncertainty maps in lung cancer segmentation, Phys. Med. Biol. 69 (2024) 035007.
[46] M. Mohammed, H. Mwambi, I.B. Mboya, M.K. Elbashir, B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on TCGA data, Sci. Rep. 11 (2021) 15626.
[47] D.J.C. MacKay, Bayesian neural networks and density networks, Nucl. Instrum. Methods Phys. Res A 354 (1995) 73–80.
[48] J.A. Vrugt, B.A. Robinson, Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging, Water Resour. Res. 43 (2007).
[49] M.A. Inamdar, U. Raghavendra, A. Gudigar, S. Bhandary, M. Salvi, R.C. Deo, P. D. Barua, E.J. Ciaccio, F. Molinari, U.R. Acharya, A novel attention based model for semantic segmentation of prostate glands using histopathological images, IEEE Access (2023).