
Bayesian Ensembles for Medical Image Segmentation and Uncertainty Estimation

*A thesis submitted in fulfillment of the requirements
for the degree of Master of Technology*

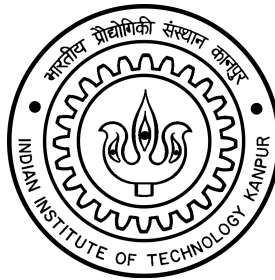
by

Arvapalli Sai Susmitha

18111273

under the guidance of

Sunil Simon and Vinay P. Namboodiri



to the

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

October 2024

Certificate

It is certified that the work contained in the thesis titled “**Bayesian Ensembles for Medical Image Segmentation and Uncertainty Estimation**” has been carried out under my supervision by **Arvapalli Sai Susmitha** and that this work has not been submitted elsewhere for a degree.



Sunil Simon

Associate Professor

Department of CSE

Indian Institute of Technology Kanpur

Kanpur, 208016



Vinay P. Namboodiri

Senior Lecturer

Department of Computer Science

University of Bath

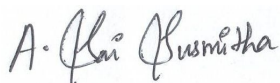
Bath, UK

October 2024

Declaration

This is to certify that the thesis titled “**Bayesian Ensembles for Medical Image Segmentation and Uncertainty Estimation**” has been authored by me. It presents the research conducted by me under the supervision of **Sunil Simon and Vinay P. Namboodiri**.

To the best of my knowledge, it is an original work, both in terms of research content and narrative, and has not been submitted elsewhere, in part or in full, for a degree. Further, due credit has been attributed to the relevant state-of-the-art and collaborations with appropriate citations and acknowledgments, in line with established norms and practices.



Name: Arvapalli Sai Susmitha

Roll No.: 18111273

Program: M.Tech - PhD

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

October 2024

Page intentionally left blank

ABSTRACT

Name of student: **Arvapalli Sai Susmitha** Roll no: **18111273**

Degree for which submitted: **Master of Technology**

Department: **Department of Computer Science and Engineering**

Thesis title: **Bayesian Ensembles for Medical Image Segmentation and Uncertainty Estimation**

Name of Thesis Supervisors: **Sunil Simon and Vinay P. Namboodiri**

Month and year of thesis submission: **October 2024**

Medical image segmentation plays a pivotal role in diagnosing and treating various diseases, enabling precise anatomical structure modeling, image-based diagnosis, surgical planning, and guidance. Despite advancements in deep learning techniques, traditional neural networks often produce overconfident predictions and struggle with ambiguities inherent in medical images. These limitations underscore the need for robust uncertainty estimation methods to enhance the reliability and interpretability of segmentation models.

Bayesian Neural Networks (BNNs) offer a promising solution by incorporating uncertainty into the model through a probabilistic framework. By placing priors over network parameters and performing inference to obtain posterior distributions, BNNs enable the quantification of uncertainty. This estimation is crucial for identifying regions where the model is less confident, thereby improving the trustworthiness of segmentation results.

In this work, we develop nine distinct Bayesian ensemble methods, each employing different encoder-decoder architectures while adhering to the same Bayesian principles. The combination of strengths from non-Bayesian ensembles and Bayesian networks inspires these methods. We introduce Bayesian ensembles for medical image segmentation with uncertainty estimation, where multiple Bayesian networks are trained, and their predictions are aggregated to calculate the segmentation and capture the uncertainty. The proposed methods demonstrate improved uncertainty estimation and offer better or comparable performance across different segmentation tasks.

We conducted extensive experiments on three datasets: ISBI-EM neuronal structure segmentation, MoNuSeg (histopathological images for multi-organ nuclei segmentation), and Lung CT scans. Our enhanced uncertainty estimation strongly correlates with misclassified regions, ensuring that areas of high uncertainty align with potential errors. Consequently, this method provides a more reliable and interpretable solution for medical image segmentation, ultimately contributing to safer and more effective clinical decision-making.

Page intentionally left blank

Acknowledgements

I would like to sincerely thank my advisors, Dr. Vinay P. Namboodiri and Dr. Sunil Simon, for their constant guidance and support throughout my research.

I owe a huge thanks to my family Arvapalli Sravana Kumar, Arvapalli Pushpa Rani, and Arvapalli Sai Srikanth who have always encouraged and supported me in every step of my life, helping me reach this point in my academic journey. They have been a continuous source of guidance, for which I am truly grateful.

Finally, I want to thank my friends Pathlavath Srikanth, Sanket, Navya Chikkam, Tentu Venkatesh, Anusha Motamarri, Yasasvi, for being there for me and motivating me along the way.

Arvapalli Sai Susmitha

Page intentionally left blank

Contents

Acknowledgements	vi
List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 Contributions	2
2 Related Works	3
2.1 Medical Image Segmentation	3
2.2 Uncertainty estimation	5
3 Bayesian Ensembles	6
3.1 Bayesian Uncertainty	6
3.2 Proposed Bayesian Ensemble Architectures	8
3.3 Bayesian Ensemble’s Loss function	16
3.4 Bayesian Ensembles Implementation Details	17
4 Experimentation Details	18
4.1 Datasets Description	18
4.2 Training and Inference	19
4.3 Evaluation Metrics	20
5 Results	22
5.1 Performance analysis on EM Dataset	22
5.2 Performance analysis on Monuseg Dataset	25
5.3 Performance analysis on Lung Dataset	27
5.4 Predictive Uncertainty maps	28
6 Conclusion	33
Bibliography	34

Page intentionally left blank

List of Figures

3.1	Bayesian Ensemble architecture	9
3.2	Bayesian U-Net	10
3.3	Attention based Bayesian U-Net	11
3.4	VGG11-based Bayesian U-Net	12
3.5	ResNet34-based Bayesian U-Net	12
3.6	PSPNet-based Bayesian U-Net	13
3.7	Single encoder-Multi-decoder	14
3.8	Multi-encoder-Single decoder	14
3.9	Deeply supervised Bayesian U-Net	15
3.10	DeepLabV3+ based Bayesian U-Net	16
5.1	Precision, Recall, Miss Detection, False Detection bar diagram for EM Dataset . .	25
5.2	Precision,Recall,Miss Detection, False Detection bar diagram for Monuseg Dataset	26
5.3	Precision, Recall, Miss Detection, False Detection bar diagram for Lung Dataset .	28
5.4	Predictive Uncertainty maps of all models on EM Dataset	29
5.5	Kwon and Gall results on EM Dataset	30
5.6	Predictive Uncertainty maps of all models on Monuseg Dataset	31
5.7	Predictive Uncertainty maps of all models on Lungs Dataset	32

Page intentionally left blank

List of Tables

4.1	MoNuSeg Dataset	18
5.1	Dice and Jaccard values for the EM dataset.	23
5.2	Performance Metrics for EM Dataset	24
5.3	Dice and Jaccard values for the Monuseg dataset.	25
5.4	Performance Metrics for Monuseg Dataset	26
5.5	Dice and Jaccard values for the lung dataset.	27
5.6	Performance Metrics for Lung Dataset	27

Chapter 1

Introduction

Medical image segmentation is fundamental in modern healthcare, aiding in applications such as anatomical structure modelling, diagnosis, and surgical planning. This process involves the identification of boundaries within medical images, which was traditionally done by trained radiologists. However, with the advent of machine learning, particularly deep learning techniques, the field has seen remarkable advancements [25, 28]. Despite these improvements, neural networks have notable limitations, especially in their tendency to make overconfident predictions, even in uncertain regions of the image[14]. This overconfidence can lead to critical errors in diagnosis, particularly when the context of the full image is insufficient to resolve ambiguities an issue that is prevalent in medical imaging [13, 21].

Another key challenge in medical image segmentation is the limited availability of high-quality annotated data. Given the complexity and variability of medical images, neural networks often fail to capture all relevant features, leading to poor segmentation performance[37]. Factors such as poor image quality, variability among patients, and inconsistencies in clinical protocols further exacerbate these issues. This highlights the need for models that not only achieve high accuracy but also provide reliable uncertainty estimates to guide clinical decision-making[20].

Bayesian neural networks (BNNs) offer a principled approach to quantifying uncertainty by introducing probability distributions over the model’s weights. This enables the model to provide more reliable uncertainty estimates by capturing both aleatoric (data-related) and epistemic (model-related) uncertainty. However, Bayesian networks can be computationally intensive, as they require multiple forward passes during inference to approximate the posterior distribution of the model weights, commonly achieved using techniques like Monte Carlo (MC) dropout [11]. Despite the computational cost, Bayesian approaches have shown great promise in improving uncertainty estimation, making them particularly suitable for medical image segmentation tasks where understanding the model’s confidence is essential [20, 23].

While Bayesian networks address the challenge of uncertainty estimation, non-Bayesian ensemble models as proposed by [24], have also proven effective in capturing uncertainty. These ensemble methods combine the predictions of multiple independently trained models, leveraging the diversity in their outputs to enhance robustness and provide more reliable uncertainty estimates.

In our work, we combine the strengths of Bayesian models[20] with the robustness of non-Bayesian ensemble methods [24], creating Bayesian ensembles that generate more reliable uncertainty estimates while maintaining the accuracy benefits typical of ensemble methods similar to [12]. We experimented with nine distinct Bayesian ensemble architectures and provided an analysis demonstrating the flexibility and effectiveness of this approach in addressing uncertainty in complex medical image segmentation tasks. These architectures include a basic Bayesian U-Net ensemble, attention-based Bayesian U-Net ensemble, deeply supervised Bayesian U-Net ensemble, PSPNet-based Bayesian U-Net ensemble, single encoder-multi-decoder Bayesian U-Net ensemble, multi-encoder-single decoder Bayesian U-Net ensemble, VGG11-based Bayesian U-Net ensemble, ResNet34-based Bayesian U-Net ensemble, and DeepLabV3+ based Bayesian U-Net ensemble.

Our experiments were conducted across three datasets: the ISBI EM dataset, the Multi-organ Hematoxylin and Eosin (H&E) stained nuclei images from the MoNuSeg dataset, and a collection of CT images with manually segmented lungs and 2D/3D measurements from the LUNA Data Science Bowl 2017 dataset. The results show slightly improved segmentation performance, with uncertainty estimations strongly correlating with misclassification. Regions with incorrect classifications exhibit higher uncertainty, indicating that Bayesian ensemble models consistently outperform previous methods for uncertainty estimation while maintaining superior segmentation accuracy.

1.1 Contributions

- We developed an ensemble of Bayesian neural networks for capturing predictive uncertainty, which provides more reliable uncertainty estimates than the non-Bayesian and non-ensemble models discussed, leading to better performance in challenging scenarios.
- We experimented with nine different Bayesian ensemble methods, revealing a strong correlation between captured uncertainty and misclassification rates, demonstrating improved reliability of predictions.
- We conducted extensive experiments and analysis across three datasets: the ISBI EM dataset, the MoNuSeg dataset, and the Lung CT dataset. Our method consistently produced better predictive uncertainty maps while maintaining segmentation accuracy.

Chapter 2

Related Works

2.1 Medical Image Segmentation

Medical image segmentation is a fundamental task in medical imaging, aiming to identify and delineate anatomical structures and regions of interest within medical images such as MRI, CT, and ultrasound scans. This process is essential for various clinical applications, including diagnosis, treatment planning, and monitoring of diseases. Accurate segmentation is crucial as it directly impacts the quality of subsequent analyses and interventions.

Before the advent of machine learning, medical image segmentation (MIS) relied heavily on classical image processing techniques. One of the earliest and simplest methods was thresholding, where a threshold value was chosen to divide pixel intensities into foreground (e.g., tissues, organs, or lesions) and background [1, 39]. Thresholding worked well for images with clear intensity differences, but it struggled in cases where intensity transitions were smooth or the image was noisy. Another technique was edge detection, which aimed to identify boundaries between different structures in the image by detecting changes in pixel intensity. Filters such as Sobel, Canny, and Prewitt were commonly used for this purpose. The main disadvantage is that it does not work well when images have many edges, and it cannot easily identify a closed curve or boundary. There was also Region-based segmentation, which includes methods like region growing, identified regions by starting from a seed point and expanding outward based on pixel similarity. The integration of edge-based and region-based techniques helped mitigate their respective limitations, resulting in a more robust segmentation method [41, 30]. However, this approach was computationally expensive and sensitive to the choice of seed point, leading to inconsistent segmentation results. These classical techniques, while foundational in early MIS, were limited by their reliance on manual tuning, sensitivity to noise, and inability to handle complex anatomical structures or variations in image quality.

As medical imaging evolved and datasets became more complex, and with the above-mentioned limitations, classical techniques often struggled to manage the variety and challenges of medical images. This led to the adoption of machine learning (ML) techniques like clustering, decision tree, support vector machine etc, which could learn from data and adapt through experience, providing a more flexible approach to medical image segmentation. K-means clustering is an unsupervised machine learning algorithm that groups pixels into distinct clusters based on their features, such as intensity values or texture, without requiring manually defined thresholds. This can be particularly useful for identifying regions with similar intensities, such as segmenting tumors, organs, or lesions from MRI [36] or CT scans. Another widely adopted approach was support vector machines (SVMs), which were used to classify pixels based on extracted features, effective for distinguishing complex tissue boundaries [15]. The work [2] does a 2-stage brain tumor segmentation using support vector machine (SVM) and genetic algorithm. Methods like Random Forests and Decision Trees can be trained to segment images by classifying each pixel based on input features such as intensity, texture, or gradient. In Random Forests, each tree makes an independent classification decision, and the final decision is determined by majority voting across all trees. In [19], a machine learning approach is presented that leverages binary decision trees and the Random Forest technique to detect and accurately segment brain tumors from multispectral volumetric MRI data. However, while these traditional ML algorithms improved segmentation by learning from data, they still relied heavily on handcrafted features and required extensive preprocessing, which limited their generalizability across different medical imaging modalities.

The introduction of deep learning, particularly convolutional neural networks (CNNs), revolutionized medical image segmentation by eliminating the need for handcrafted features and enabling models to learn hierarchical representations from the data automatically. The Fully Convolutional Neural Network (FCN)[31] was among the pioneering deep learning models used for image segmentation tasks. [28] advanced this architecture with the introduction of U-Net, which achieved impressive segmentation performance by using an encoder-decoder structure. This design allowed the model to capture both fine details and high-level context in images, while also addressing the challenge of requiring large training datasets. U-Net became widely adopted for various medical segmentation tasks due to its ability to handle small datasets with limited annotations while delivering high accuracy. Self-attention techniques have been introduced in [18, 40] to extract region proposals by highlighting salient features for specific tasks. A notable example is Attention U-Net [27], applied for pancreas segmentation, where attention gates are proposed to automatically focus on target structures of different shapes and sizes while suppressing irrelevant regions. These deep learning models provided significant improvements over traditional ML methods, offering more accurate, reliable, and scalable solutions for a wide range of MIS challenges.

2.2 Uncertainty estimation

Incorporating uncertainty into medical image segmentation is crucial for providing confidence estimates and enhancing clinical decision-making. While many methods in medical image segmentation often achieve accuracy levels comparable to medical experts, these approaches typically lack information about incorrectly segmented regions, which can have serious implications given the critical nature of accurate segmentation for diagnosis. To address this gap, few works have introduced techniques to estimate the uncertainty associated with model outputs [24, 20, 23]. Specifically, these methods provide uncertainty estimates for each segmented pixel, reflecting the model’s confidence in its predictions. Various techniques have been proposed to estimate uncertainty in regression, classification, and segmentation, with a focus here on medical image segmentation.

Various methods have been proposed to tackle uncertainty estimation, achieving differing levels of success[34, 9]. Among these, Bayesian Neural Networks[3] the basis of our work are known for their ability to quantify uncertainty by modeling distributions over network parameters which we use in our work. However, they tend to be complex to implement and computationally intensive, making them slower to train compared to non-Bayesian models. To overcome these challenges, Gal and Ghahramani[11] proposed *Monte Carlo dropout* (MC-dropout), a method that estimates predictive uncertainty by applying dropout at test time[35]. This technique leverages dropout as a form of approximate Bayesian inference, providing a practical and efficient way to estimate uncertainty without the full complexity of Bayesian neural networks. In addition, [20] introduced a framework designed to capture both aleatoric and epistemic uncertainty in regression and classification tasks. Aleatoric uncertainty refers to the inherent noise in the data, while epistemic uncertainty represents the model’s uncertainty about the data due to limited knowledge or insufficient training. To address aleatoric uncertainty without adding extra parameters, [23] proposed a Bayesian method that exploits the relationship between the variance and mean of a multinomial random variable, offering a more streamlined approach to uncertainty estimation. Another significant advancement is the deep ensemble approach [24], which provides a non-Bayesian method for capturing predictive uncertainty. This method involves training multiple models and aggregating their predictions to estimate uncertainty. Deep ensembles have shown to outperform previous methods in terms of accuracy and reliability by leveraging the diversity among the models to better capture the variability in predictions. [12] focused on uncertainty estimation for medical images using a basic Bayesian U-Net ensemble and an interactive user-based segmentation tool to leverage the estimated uncertainty. In contrast, our work extends this foundation by proposing nine different Bayesian ensemble architectures and analyzing their performance in medical image segmentation. After comparing and analyzing these various configurations, we observed significant improvements over non-Bayesian ensemble and non-ensemble methods discussed above.

Chapter 3

Bayesian Ensembles

3.1 Bayesian Uncertainty

Bayesian neural networks (BNNs) aim to model uncertainty in predictions by treating the network's weights, ω , as random variables. Instead of assigning fixed values to weights, a BNN places a prior distribution over these parameters, resulting in a posterior distribution that incorporates both the prior belief and the likelihood derived from the data. The fundamental problem in BNNs is that the posterior distribution $p(\omega|D)$ is intractable for most practical applications. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ represents the input and y_i is its corresponding output, the posterior distribution is expressed as:

$$p(\omega|D) = \frac{p(D|\omega)p(\omega)}{p(D)} = \frac{p(D|\omega)p(\omega)}{\int p(D|\omega)p(\omega)d\omega}$$

Here, $p(D|\omega) = \prod_{i=1}^N p(y_i|x_i, \omega)$ is the likelihood, $p(\omega)$ is the prior distribution over weights, and $p(D)$ is the marginal likelihood or evidence.

The challenge in BNNs lies in computing the posterior $p(\omega|D)$, which requires integration over the parameter space. This integral is typically intractable due to the high dimensionality of the weight space and the complexity of neural networks. Several methods have been developed to approximate the posterior. One such method is the Laplace Approximation, which approximates the posterior around the mode with a Gaussian distribution. While computationally efficient, it often suffers from poor approximation quality. Another approach is Hamiltonian Monte Carlo (HMC), which utilizes Markov Chain Monte Carlo (MCMC) sampling to approximate the posterior, yielding accurate samples; our work particularly focuses on this method. Additionally, Variational Inference is another technique that reformulates the problem as an optimization task by approximating the posterior with a simpler distribution.

[11] proposed MC Dropout as an efficient approximation to BNNs. They demonstrated that using dropout during both training and inference is equivalent to approximating the posterior in a variational framework. Dropout provides a scalable way to capture model uncertainty without explicitly calculating the posterior. In this approach, multiple forward passes with dropout activated provide samples from the posterior distribution, allowing the model to estimate epistemic uncertainty.

The predictive distribution for a new input x^* is given by:

$$p(y^*|x^*, D) = \int p(y^*|x^*, \omega) p(\omega|D) d\omega$$

In MC Dropout, this is approximated by averaging predictions from several stochastic forward passes. In Bayesian modeling, as mentioned earlier it has two distinct types of uncertainty are often discussed: *aleatoric* and *epistemic*. where aleatoric uncertainty captures the inherent noise within the observations, while epistemic uncertainty reflects uncertainty in the model parameters. A novel method for estimating both types was proposed by [20]. They utilized a Bayesian neural network that outputs both a mean (μ) and variance (σ^2) for the logits. The variance term represents aleatoric uncertainty, while epistemic uncertainty is captured by performing dropout variational inference. This technique, referred to as *Monte Carlo dropout*, samples from the approximate posterior during test time to estimate epistemic uncertainty.

In this approach, the aleatoric uncertainty is modeled using a Gaussian likelihood. For a classification task, the network predicts a vector of logits μ_i for each pixel i . This vector, after being passed through a softmax function, produces a probability vector p_i . A Gaussian prior is placed on the logits as follows:

$$\hat{y}_i|W \sim \mathcal{N}(\mu_i^W, (\sigma_i^W)^2)$$

$$\hat{p}_i = \text{Softmax}(\hat{y}_i)$$

Here, μ_i^W and $(\sigma_i^W)^2$ are the outputs of the network, parameterized by W . The vector μ_i^W is corrupted with Gaussian noise, and the resulting logits \hat{y}_i are transformed using softmax to obtain the probabilities p_i . The expected log-likelihood is then:

$$\log E_{\mathcal{N}(\hat{y}_i; \mu_i^W, (\sigma_i^W)^2)}[\hat{p}_{i,c}] \quad (3.1)$$

where c is the observed class for pixel i . This objective function is approximated using Monte Carlo integration.

Kwon et al. [23] proposed an alternative approach to capture aleatoric and epistemic uncertainty without the need for learning an additional variance parameter (σ^2). Their method defines the total predictive uncertainty as the sum of aleatoric and epistemic uncertainties, which can be computed as follows:

$$\underbrace{\frac{1}{T} \sum_{t=1}^T \text{diag}(p^{W_t}) - (p^{W_t})^{\otimes 2}}_{\text{Aleatoric}} + \underbrace{\frac{1}{T} \sum_{t=1}^T (p^{W_t} - \bar{p})^{\otimes 2}}_{\text{Epistemic}}$$

Here, $\bar{p} = \frac{1}{T} \sum_{t=1}^T p^{W_t}$ and $p^{W_t} = \text{Softmax}(f^W(x^*))$.

3.2 Proposed Bayesian Ensemble Architectures

Ensemble models have been employed to enhance predictive performance, but their role in uncertainty estimation within the medical field remains largely unexamined. Deep ensembles, a non-Bayesian method used to estimate uncertainty in predictions by [24]. When this concept is extended to Bayesian neural networks, it leads to improved uncertainty estimation compared to standard model ensembles. Here we employ a Bayesian ensemble approach to improve segmentation performance and provide reliable uncertainty estimates. Each model in the ensemble follows an encoder-decoder architecture, and similar to the approach by [20], the final layer of each model outputs two components: the segmentation map and the log variance (transformed by the softplus function), which represents aleatoric uncertainty.

In this method, N Bayesian models are trained independently as shown in 3.1, each with its own prior. Each network learns a posterior based on the training data. While dropout can be viewed as an ensemble of smaller neural networks, there is some dependency between these networks. In contrast, Bayesian model ensembles are made up of independent models that are more likely to converge to different modes, allowing for better diversity in learning compared to MC-dropout [11]. We treat the Bayesian ensemble as a uniformly-weighted mixture model, combining predictions for segmentation as:

$$p(\hat{y}_i|x_i) = \frac{1}{N} \sum_{m=1}^N p_{\theta_m}(y_i|x_i, \theta_m) \quad (3.2)$$

Here, x_i represents pixel i of image x , and y_i represents the class of that pixel. For segmentation, the mean of the segmentation outputs from the ensemble of n models is computed to generate the final prediction. For predictive uncertainty, we calculate the variance of the segmentation outputs across the ensemble:

$$\frac{1}{N} \sum_{m=1}^N (p_{\theta_m}(y_i|x_i, \theta_m) - p(\hat{y}_i|x_i))^2 \quad (3.3)$$

This equation produces an uncertainty value for each pixel, which is used to generate uncertainty maps for the output. This ensemble strategy is consistently applied across different encoder-decoder architectures, ensuring both improved segmentation accuracy and reliable uncertainty quantification.

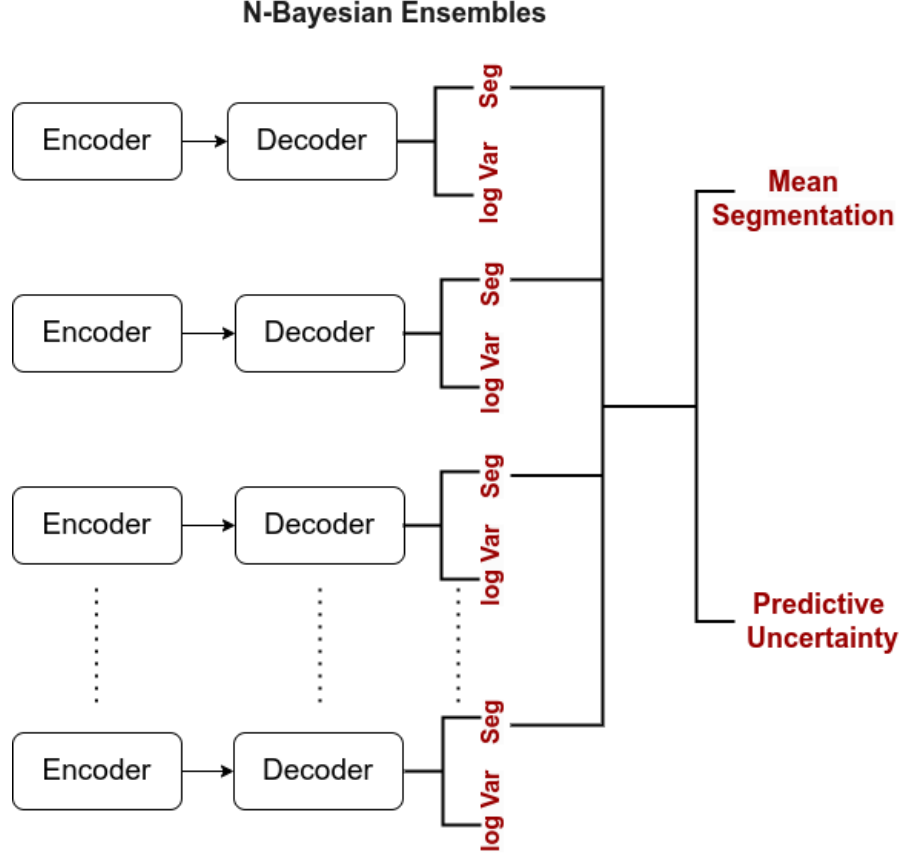


Figure 3.1: Bayesian Ensemble architecture

3.2.1 Bayesian U-Net (BUNet)

The Bayesian U-Net (BUNet) is built upon the classic U-Net architecture[28] but introduces Bayesian principles through dropout layers and uncertainty estimation. The model follows a typical U-shaped structure with an encoder-decoder design 3.2. The encoder downsamples the input image by passing it through convolutional layers followed by max-pooling operations, gradually extracting hierarchical feature representations. The decoder then upsamples the encoded feature maps using transpose convolutions to reconstruct the segmentation mask. Skip connections between the encoder and decoder allow the model to retain fine-grained details at various scales, which is crucial for precise segmentation.

In the Bayesian U-Net, dropout is applied at both training and inference stages, enabling the model to output a distribution over possible segmentation maps rather than a deterministic one. Additionally, the final layer outputs two branches: one for the segmentation prediction and another for the aleatoric uncertainty, which is obtained by applying the Softplus function over the log variance same as in [12]. The ensemble of such Bayesian U-Nets results in the final segmentation as the mean of individual predictions, and predictive uncertainty is derived from the variance of these predictions across the ensemble.

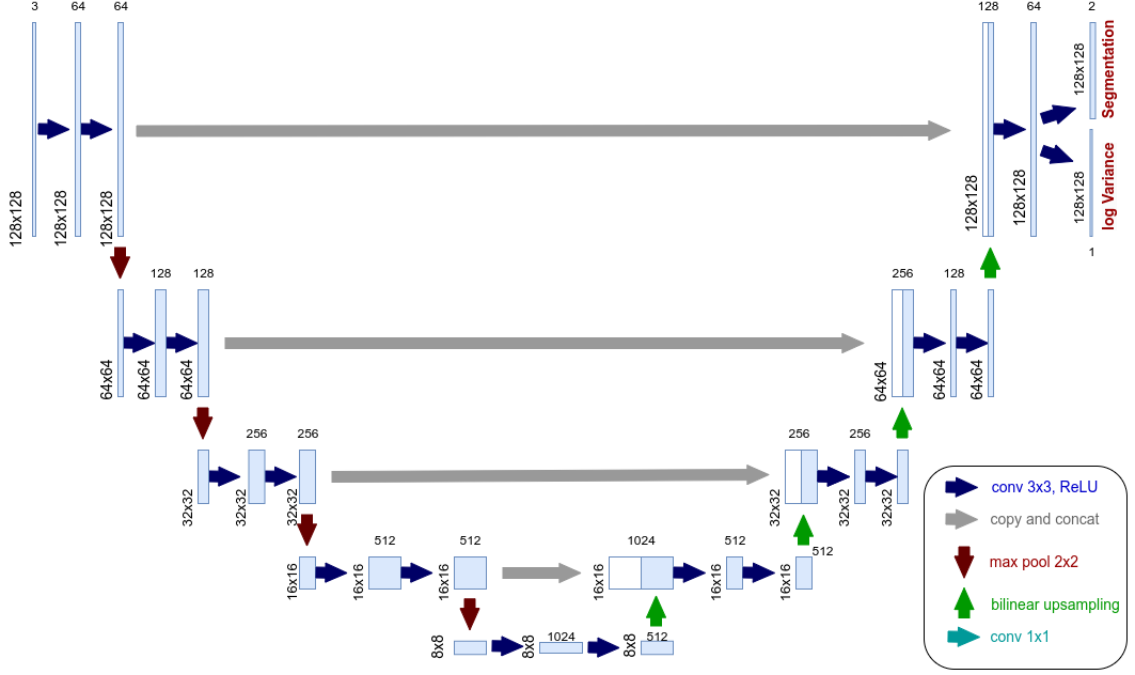


Figure 3.2: Bayesian U-Net

3.2.2 Attention-based Bayesian U-Net

The Attention U-Net introduces an attention mechanism[18] to the Bayesian U-Net architecture to improve the model’s focus on relevant areas of the image. The attention mechanism is embedded between the encoder and decoder blocks, refining the concatenation of feature maps from these stages as shown in 3.3.

In the attention module, attention gates are applied to each skip connection. These gates learn to focus on the most relevant parts of the image by weighting features based on their importance for the segmentation task. This mechanism helps suppress irrelevant features while enhancing crucial regions, especially in complex medical images. The final segmentation and aleatoric uncertainty are generated in the same manner as the base Bayesian U-Net, with Softplus applied to the log variance. The ensemble predictions are averaged to yield the final segmentation map, while the variance across ensemble outputs represents the predictive uncertainty.

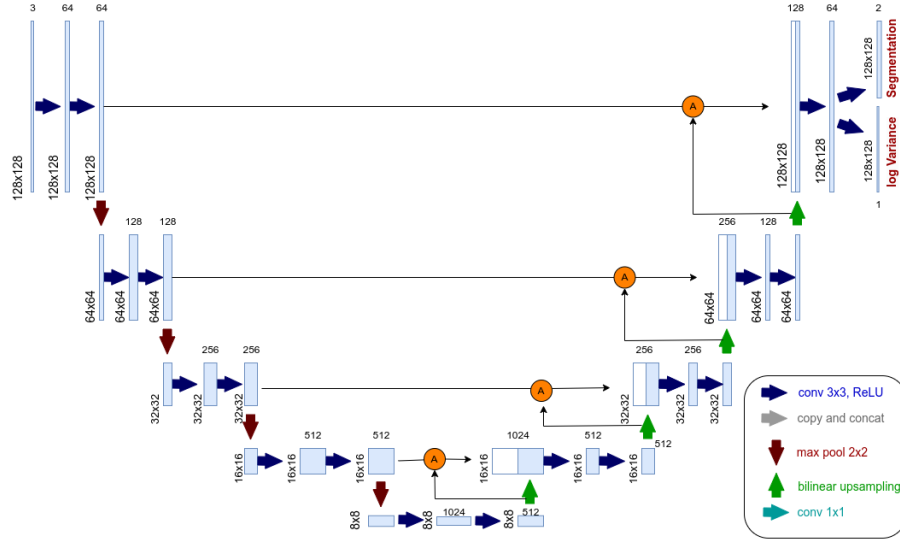


Figure 3.3: Attention based Bayesian U-Net

3.2.3 VGG11-based Bayesian U-Net

VGG11[33] is a well-established convolutional neural network known for its simplicity and efficiency. In this Bayesian U-Net variant, the VGG11 network serves as the encoder, replacing the traditional U-Net encoder [17]. The VGG11 architecture consists of 11 layers, mainly composed of convolutional layers followed by max-pooling layers. These layers capture multi-level hierarchical features by progressively reducing spatial dimensions while increasing feature depth.

The decoder mirrors the upsampling and feature reconstruction process from the base U-Net, using transpose convolutions to produce the final segmentation map 3.4. The model also outputs log variance for aleatoric uncertainty, using the same two-branch output approach. Ensembled VGG11-based Bayesian U-Nets results in the final segmentation as the mean of individual predictions, and predictive uncertainty is derived from the variance of these predictions across the ensemble.

3.2.4 ResNet34-based Bayesian U-Net

ResNet34 introduces residual connections [16], a key innovation in deep learning, which allows the network to efficiently train deeper models without suffering from vanishing gradients. In the ResNet34-based Bayesian U-Net, the encoder is replaced with ResNet34. This architecture consists of residual blocks that allow gradients to flow more easily during training, improving convergence and feature extraction, especially for deep hierarchical representations.

The residual blocks consist of skip connections, which enable the network to retain the identity mapping while transforming the input through convolutional layers. The decoder follows a similar design as the traditional U-Net but works on feature maps extracted from the residual blocks[32]. The final segmentation is generated from the decoder output, while aleatoric uncertainty is modeled

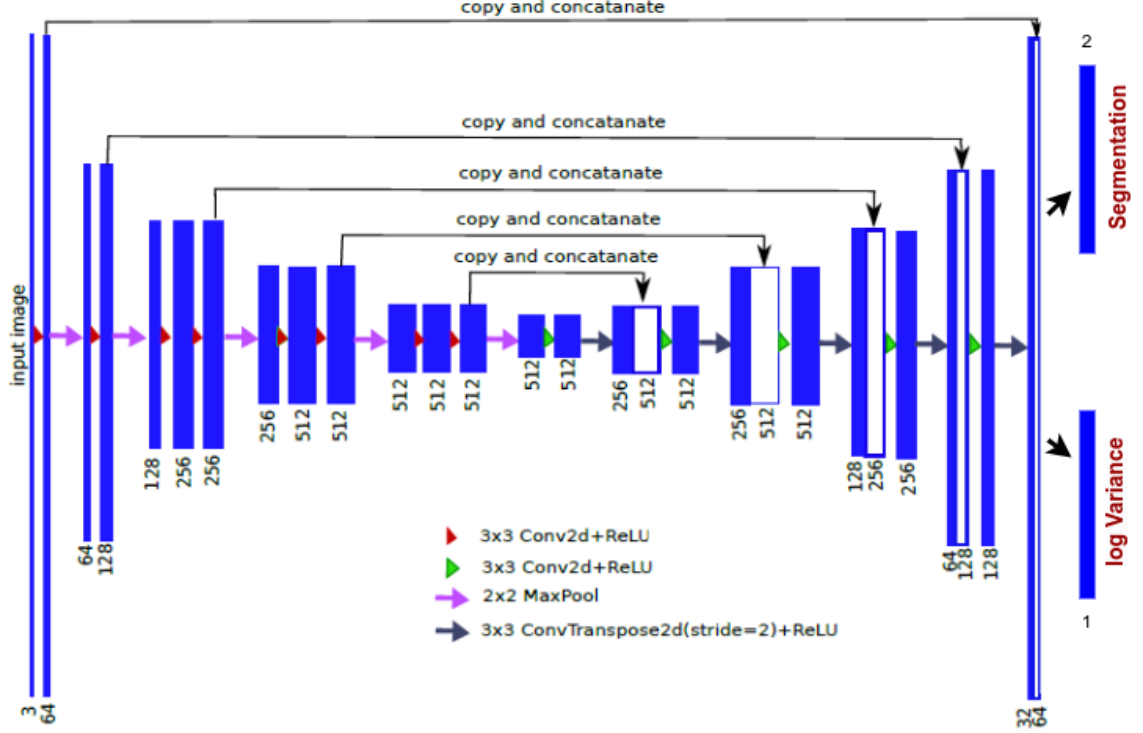


Figure 3.4: VGG11-based Bayesian U-Net

using the log variance output as shown in 3.5. Ensembles of ResNet34-based Bayesian U-Nets result in the final segmentation as the mean of individual predictions, and predictive uncertainty is derived from the variance of these predictions across the ensemble.

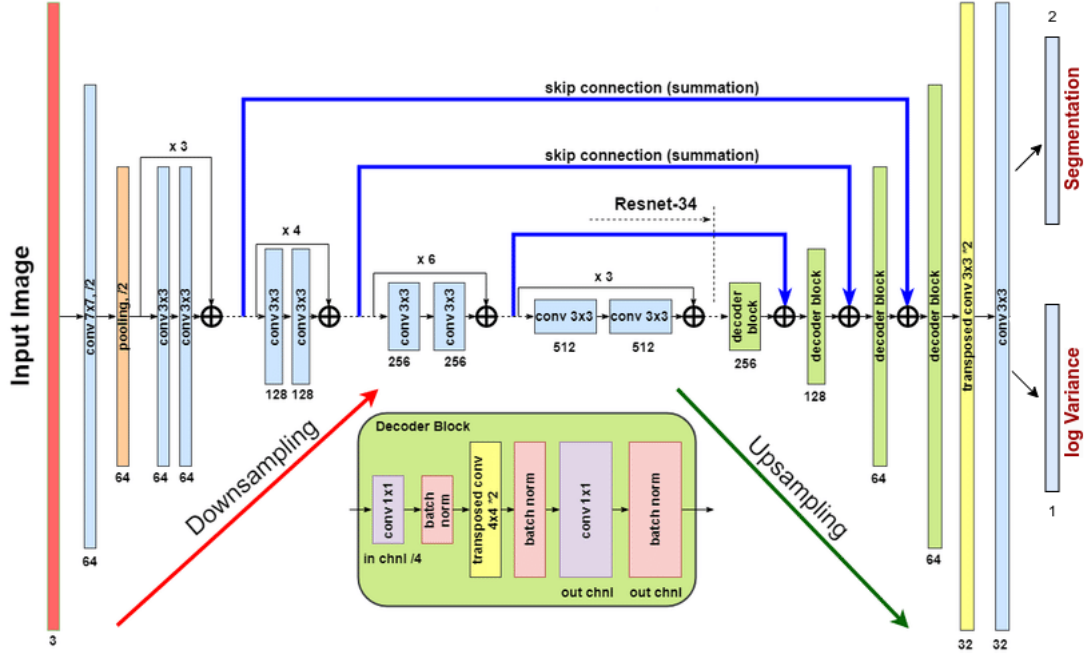


Figure 3.5: ResNet34-based Bayesian U-Net

3.2.5 PSPNet-based Bayesian U-Net

PSPNet (Pyramid Scene Parsing Network) [42] enhances segmentation performance by introducing a Pyramid Pooling Module (PPM), which captures multi-scale contextual information. In this model, after the initial encoding stages of the Bayesian U-Net, the feature maps are passed through the PPM 3.6. The PPM performs adaptive pooling at multiple scales (e.g., 1×1 , 2×2 , 4×4 , and 6×6 pooling), enabling the model to capture context from various receptive fields.

These pooled features are then upsampled to match the size of the original feature map and concatenated with the original features to provide a rich, multi-scale representation. This ensures that the model can accurately segment both large objects and fine details within an image. The decoder reconstructs the segmentation mask from these multi-scale features, while the aleatoric uncertainty is calculated using log variance, the predictive uncertainty is calculated as done in other Bayesian U-Net variants.

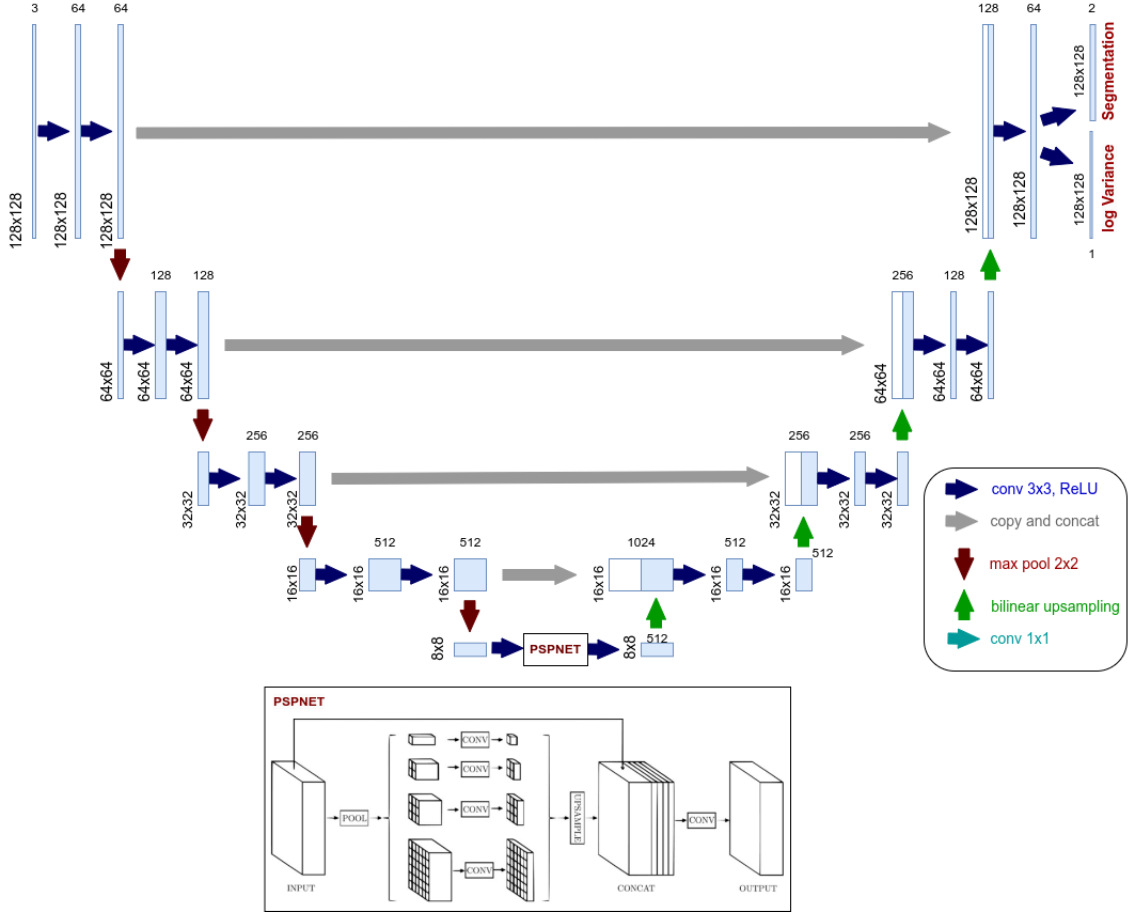


Figure 3.6: PSPNet-based Bayesian U-Net

3.2.6 Single Encoder-Multi Decoder & Multi Encoder-Single Decoder Bayesian U-Net

In these architectures, we extend the traditional U-Net by introducing either multiple decoders or multiple encoders to capture richer variations in feature representation.

Single Encoder-Multi Decoder

The single encoder is shared across multiple decoders 3.7. Each decoder processes the encoded features independently, producing its own segmentation map. These multiple segmentation maps are then averaged to produce the final segmentation. The log variance output from each decoder contributes to the aleatoric uncertainty, and the ensemble variance provides predictive uncertainty.

Multi Encoder-Single Decoder

Multiple encoders independently process the input image, each capturing distinct feature representations. These encoded features are each individually passed through a single decoder to generate multiple segmentation maps as shown in 3.8 and are then averaged to produce the final segmentation, while the predictive uncertainty is the variance of the generated segmentation maps.

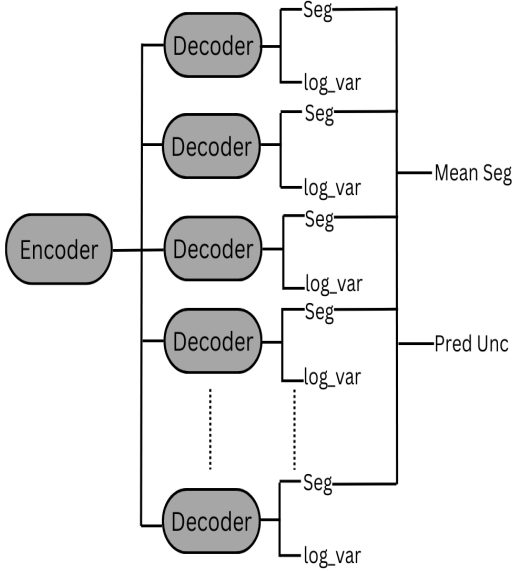


Figure 3.7: Single encoder-Multi-decoder

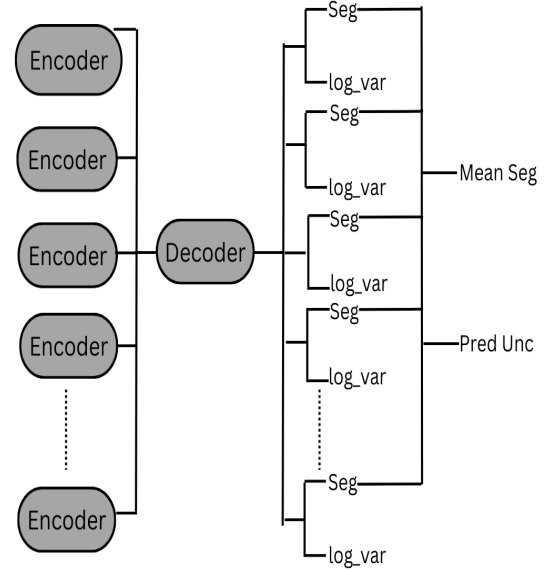


Figure 3.8: Multi-encoder-Single decoder

3.2.7 Deeply Supervised Bayesian U-Net

In the Deeply Supervised Bayesian U-Net, we introduce deep supervision by applying auxiliary loss functions to the intermediate outputs, specifically at the centre layer after the encoder. As seen in the diagram 3.9, this approach encourages the model to learn meaningful feature representations at different scales, enhancing the accuracy of the final segmentation map. An additional cross-entropy loss is computed between the deep supervision network output and the ground truth, promoting better alignment during training. The centre layer’s output is passed through a smaller decoder that produces a segmentation map, which is used to compute auxiliary losses during training. The final output consists of both the main segmentation map and aleatoric uncertainty, derived from the Softplus-applied log variance. Ensemble predictions from multiple deeply supervised Bayesian U-Nets are averaged to yield the final segmentation, and predictive uncertainty is calculated from the variance across ensemble outputs.

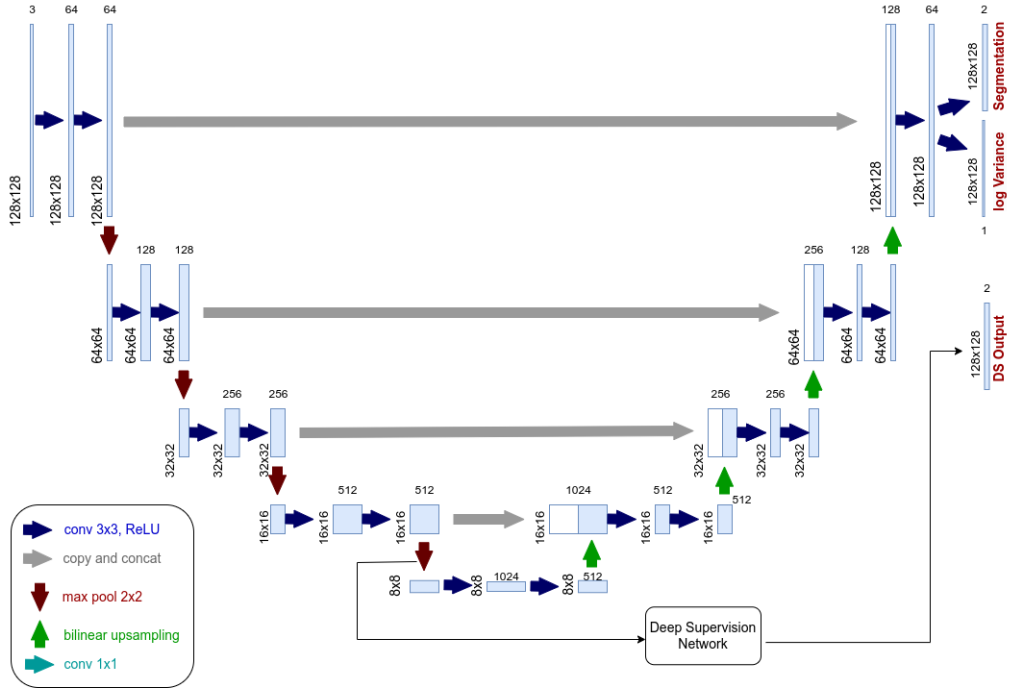


Figure 3.9: Deeply supervised Bayesian U-Net

3.2.8 DeepLabV3+ based Bayesian U-Net

DeepLabV3+ [6, 29] is a state-of-the-art architecture for semantic segmentation, incorporating an encoder-decoder structure along with Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context without losing spatial resolution. The encoder is based on the Xception network, which uses depthwise separable convolutions to efficiently learn complex features.

The ASPP module applies atrous (dilated) convolutions with different dilation rates to capture multi-scale context from the feature maps. The output of the ASPP is combined with low-level

features from the Xception [7] encoder via skip connections, preserving both global context and local details. The decoder refines this information into a detailed segmentation map, while a second branch outputs log variance for aleatoric uncertainty.

DeepLabV3+ excels at handling large variations in object scale due to the ASPP module. The ensemble of DeepLabV3+ Bayesian U-Nets 3.10 averages predictions for final segmentation, while predictive uncertainty is derived from the variance of the ensemble outputs.

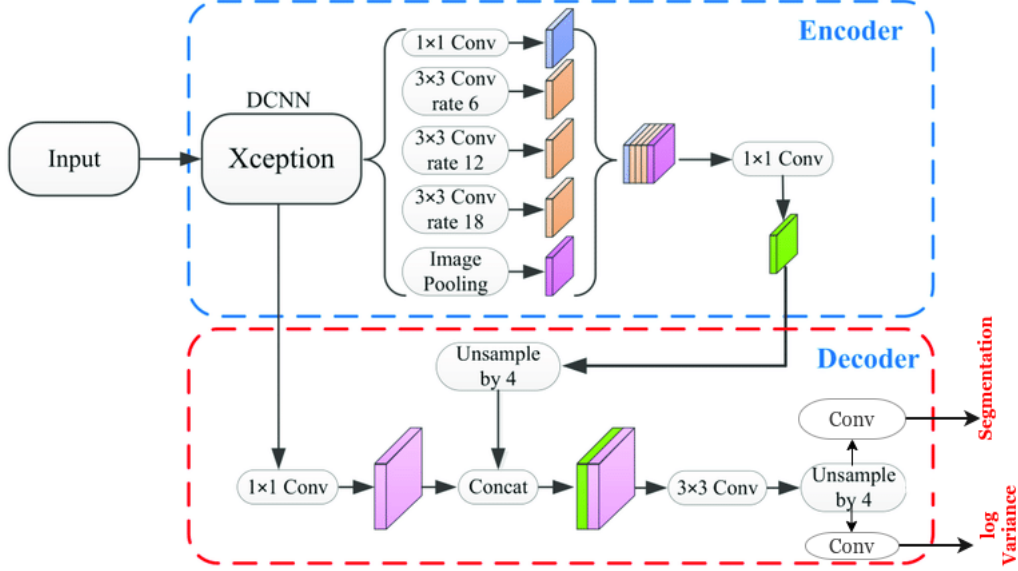


Figure 3.10: DeepLabV3+ based Bayesian U-Net

3.3 Bayesian Ensemble's Loss function

We approximate the objective of equation 3.1 using Monte Carlo integration over T samples, leveraging a weighted cross-entropy loss. To improve numerical stability, we modified equation 3.1. The final loss formulation is as follows:

$$\hat{y}_{i,t} = \mu_i^W + \sigma_i^W \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 1) \quad (3.4)$$

$$\mathcal{L} = \sum_i \frac{1}{T} \sum_t w_c \log \left(\frac{\exp(\hat{y}_{i,t,c})}{\sum_{c'} \exp(\hat{y}_{i,t,c'})} \right) \quad (3.5)$$

In this expression, w_c represents the weight for the observed class c at the i -th pixel of the output y , and $\hat{y}_{i,t,c'}$ refers to the c' -th component of the vector $\hat{y}_{i,t}$.

3.4 Bayesian Ensembles Implementation Details

The Bayesian model ensemble consists of N Bayesian Encoder-Decoder models, where in our case $N = 5$. Figure 3.1 illustrates the Bayesian ensemble architecture, where each component represents one of the nine Bayesian Encoder-Decoder models we have discussed above. These models can vary in their specific architecture, but they all follow the same general principle of encoding and decoding layers to produce segmentation outputs. where each Encoder-Decoder model consists of encoding and decoding layers. The encoding layers extract contextual features at multiple levels, and the decoding layers combine the extracted features to produce the final segmentation outputs. All convolutional filters in the models have a size of 3×3 .

The central layers of each Bayesian Encoder-Decoder incorporate Dropout, with a dropout rate of 0.2, to estimate uncertainty. ReLU activation functions are used throughout the network. For training, the Adam optimizer is employed, which adaptively adjusts the learning rate. Additionally, the model learns a parameter σ^2 , representing aleatoric uncertainty, which is integrated into the loss function as described in 3.3, improving the model’s robustness to noise. During training, 30 samples are drawn from the Gaussian distribution in the loss function. Similarly, during inference, 30 samples are used for Monte Carlo sampling with Dropout.

In each variation of the Bayesian ensemble architectures, the N models are trained independently. The final prediction is obtained by averaging the segmentation outputs from the ensemble, while the variance between the outputs is used to estimate predictive uncertainty, resulting in uncertainty maps.

Chapter 4

Experimentation Details

4.1 Datasets Description

4.1.1 ISBI EM Dataset

We utilized the ISBI 2012 EM neuronal segmentation challenge dataset[5, 4], which includes 30 publicly available images of size 512×512 pixels, each accompanied by pixel-wise binary segmentation labels. The dataset was divided into training and testing sets with an 80:20 ratio, with images randomly assigned to each split. Data augmentation techniques, including elastic distortion, cropping, and rotation, were applied uniformly across all models compared. No additional preprocessing was required for this dataset.

4.1.2 MoNuSeg Dataset

Table 4.1: MoNuSeg Dataset

Data subset	Breast	Liver	Kidney	Prostate	Bladder	Colon	Stomach	Total
Train	4	4	4	4	-	-	-	16
Test	2	2	2	2	2	2	2	14
Total	6	6	6	6	2	2	2	30

We utilized publicly available data of multiorgan Hematoxylin and Eosin (H&E) stained nuclei images [22] obtained from various hospitals. This dataset was generated by downloading H&E stained tissue images captured at 40x magnification from the TCGA archive. H&E staining is a standard procedure that enhances the contrast of tissue sections and is frequently employed for tumor evaluation, including grading and staging. The dataset includes 30 images and approximately 22,000 annotations of nuclear boundaries. We partitioned the data into training and test sets, as detailed in Table 4.1.

Data Preprocessing for MoNuSeg

The histological composition of tissue comprises stroma, adipose tissue, lumen, and epithelium. The characteristics of the epithelium and stroma—such as color, shape, size, and the presence of glands and nuclei—provide valuable insights into the tissue’s health. Hematoxylin and Eosin (H&E) staining is the most commonly used and cost-effective staining technique in medical diagnostics. Typically, hematoxylin stains cell nuclei a deep bluish-purple, while eosin imparts a pink hue to the cytoplasm.

Normalization of H&E images is essential because the staining results can vary significantly due to differences in reagents, staining protocols, scanning equipment, and the expertise of the technician. In our study, we utilized the normalized images from [38], which employed sparse non-negative matrix factorization (NMF) to align the color space of the input images with that of a reference image. By selecting a specific H&E-stained image as the target, the colors of all other images were adjusted to ensure consistent color representation across the dataset.

4.1.3 Lung CT Dataset

This dataset consists of CT images of the lungs, along with manually segmented lung regions and measurements in both 2D and 3D. The dataset is extracted from Kaggle [26] and includes a subset of the LUNA16 dataset, which was featured in the Kaggle Data Science Bowl in 2017, where the goal is to detect lung lesions or abnormalities. We use 80% of the data for training and the remaining 20% for testing. The resolution of each image is 512×512 pixels. Accurate segmentation of the lungs is a critical preprocessing step in identifying lung nodules.

4.2 Training and Inference

The training process for all three datasets was straightforward. We independently trained five variations of Bayesian Ensembles for 100 epochs on each dataset, using median frequency class balancing as outlined by [10]. The entire network was trained in an end-to-end manner.

In the case of the **MoNuSeg** dataset, we employed a patch-based training strategy as described in [22, 8]. Patches of size 128×128 were extracted from the images for training, with each patch processed individually for segmentation prediction. We followed the preprocessing and normalization steps outlined by [38].

For the **EM dataset** and **Lung Dataset**, images of size 512×512 pixels were used with a batch size of 1. The final predictions were obtained by averaging all segmentation outputs from each model, and the predictive uncertainty was computed as the variance among these outputs.

4.3 Evaluation Metrics

To demonstrate the effectiveness of our methods, we evaluate the performance of all models using several image segmentation metrics: Dice score, Jaccard index, precision, recall, miss detection rate, and false detection rate.

4.3.1 Metrics Description

Dice Coefficient The Dice coefficient measures the overlap between the predicted segmentation and the ground truth. It is defined as:

$$D(X, Y) = 2 \frac{|X \cap Y|}{|X| + |Y|} \quad (4.1)$$

A high Dice score indicates that the model’s predictions closely align with the actual segmentation areas, demonstrating effective capture of relevant regions.

Jaccard Index The Jaccard index assesses the similarity between the predicted segmentation and the ground truth:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (4.2)$$

This metric provides a measure of how much the predicted and actual segments overlap relative to their combined size, offering insight into the model’s accuracy.

Precision Precision quantifies the accuracy of positive predictions made by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.3)$$

The precision reflects the proportion of correctly identified pixels belonging to the target class out of all pixels that the model predicted as positive, highlighting the model’s reliability in avoiding false positives.

Recall Recall evaluates the model’s ability to identify all relevant instances:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.4)$$

for our task, recall indicates the proportion of true positive predictions relative to the total actual positive instances, emphasizing the model’s effectiveness in capturing all relevant regions, including those it may have missed.

Miss Detection Rate (MD) Miss Detection Rate indicates the proportion of actual positive cases that were missed by the model:

$$\text{Miss Detection Rate} = \frac{FN}{TP + FN} \quad (4.5)$$

This metric highlights the percentage of relevant instances that the model failed to identify, serving as a measure of how many true positives were overlooked.

False Detection Rate (FD) False Detection Rate quantifies the proportion of negative cases incorrectly identified as positive:

$$\text{False Detection Rate} = \frac{FP}{TN + FP} \quad (4.6)$$

In segmentation, this reflects how often the model incorrectly classifies pixels as belonging to the target class when they do not, indicating the potential for false alarms in the model's predictions.

Chapter 5

Results

5.1 Performance analysis on EM Dataset

The table 5.1 presents the Dice and Jaccard scores of 12 models on the EM dataset, focusing primarily on segmentation performance. So the first 9 models are the Bayesian ensemble models we proposed, followed by a non-Bayesian ensemble model[24], and finally, the non-ensemble Bayesian models proposed by Kwon.et.al[23] and Gal.et.al [20]. The VGG-based Bayesian ensemble outperforms all other models, achieving the highest scores for both metrics (Dice: 0.9372, Jaccard: 0.9017). This implies that VGG11’s strong feature extraction capabilities, coupled with Bayesian ensembling, result in more accurate segmentation and are able to capture the fine details present in the small and complex EM dataset. The ResNet34-based Bayesian ensemble is the next best performer (Dice: 0.9289, Jaccard: 0.8911). Similar to VGG, ResNet34’s skip connections allow it to maintain fine-grained information across layers, benefiting the segmentation task. Next, the 1dec.5enc Bayesian ensemble also performs well (Dice: 0.9261, Jaccard: 0.8872). The use of multiple encoders ensures that diverse feature representations are captured, leading to a strong ensemble performance. Interestingly, the Gall Unet, though a single model (not an ensemble), has competitive Dice and Jaccard scores (Dice: 0.9250, Jaccard: 0.8851). However, despite these good segmentation metrics, Gall Unet suffers from poor uncertainty estimation, indicating a lack of confidence when it makes incorrect predictions. The Deeply Supervised Bayesian Unet ensemble follows closely with solid scores (Dice: 0.9249, Jaccard: 0.8858). Deep supervision helps refine the feature maps at various levels, contributing to the model’s accuracy.

The Bayesian Unet ensemble, the basic architecture, is next (Dice: 0.9236, Jaccard: 0.8839), followed by the PSPNet-based and Deeplab-based Bayesian Unet ensembles. These models (Dice around 0.922) incorporate advanced multi-scale learning techniques but seem to underperform slightly on the EM dataset, possibly due to its small size and the highly detailed nature of the

Model	Dice	Jaccard
1enc_5dec	0.9257	0.8863
attention_unet	0.9210	0.8806
Deeply supervised unet	0.9249	0.8858
Bayesian_unet	0.9236	0.8839
1dec_5enc	0.9261	0.8872
pspnet_unet	0.9220	0.8819
resnet_unet	0.9289	0.8911
vgg_unet	0.9372	0.9017
deeplab_unet	0.9220	0.8814
Non_Bayesian ensemble[24]	0.8981	0.8498
Kwon Unet[23]	0.8979	0.8496
Gall Unet[20]	0.9250	0.8851

Table 5.1: Dice and Jaccard values for the EM dataset.

images. Multi-scale pooling techniques like those in PSPNet and Deeplab might not capture intricate details as effectively on this small dataset.

Finally, the Attention-based Bayesian Unet ensemble ranks the lowest among Bayesian models (Dice: 0.9210, Jaccard: 0.8806). While attention mechanisms often improve performance by focusing on key areas, their contribution may be less impactful for the EM dataset’s complexity.

The non-Bayesian ensemble (Dice: 0.8981, Jaccard: 0.8498) and the Kwon U-Net (Dice: 0.8979, Jaccard: 0.8496) models show the lowest performance in terms of Dice and Jaccard scores, struggling with accurate predictions. This demonstrates the importance of Bayesian ensembling, as the predictive uncertainty captured by the Bayesian ensembles likely leads to better segmentation decisions, underlining that our proposed models are more reliable for this dataset.

While Dice and Jaccard provide a general measure of segmentation accuracy, focusing solely on these metrics can mask other important aspects of model performance. Precision and Recall help distinguish how well the model balances false positives (Precision) and false negatives (Recall), which is crucial for medical image segmentation. Additionally, Miss Detection Rate (MD) and False Detection Rate (FD) offer insights into how often the model misses true positives or predicts false positives, which can further inform its reliability.

The 5.2 table (as also shown in 5.1) introduces these additional metrics. Here, we observe that the non-Bayesian ensemble has the highest Precision (0.9815), but its Recall is significantly lower (0.8282), leading to a high Miss Detection Rate (MD) of 0.1717. This suggests that while the non-Bayesian ensemble is good at avoiding false positives, it misses many true positives, which reduces its overall segmentation quality. The False Detection Rate (FD) is also the lowest (0.0184), indicating the model’s hesitance to make false predictions, but this comes at the cost of missing important details.

In contrast, the VGG11-based Bayesian ensemble, despite not having the highest Precision (0.9712), achieves the best Recall (0.9174), leading to the lowest Miss Detection Rate (MD)

Model	Precision	Recall	MD	FD
1enc_5dec	0.9702	0.8854	0.1145	0.0297
Attention_unet	0.9749	0.8733	0.1266	0.0250
Deeply_supervised_unet	0.9737	0.8813	0.1186	0.0262
Bayesian_unet	0.9755	0.8773	0.1226	0.0244
1dec_5enc	0.9732	0.8836	0.1163	0.0267
PSPNet_unet	0.9758	0.8744	0.1255	0.0241
ResNet_unet	0.9715	0.8904	0.1095	0.0284
VGG_unet	0.9712	0.9174	0.0825	0.0387
Deeplab_unet	0.9721	0.8772	0.1227	0.0278
Non_Bayesian_ensemble[24]	0.9815	0.8282	0.1717	0.0184
Kwon_unet[23]	0.9786	0.8308	0.1691	0.0213
Gall_unet[20]	0.9668	0.8873	0.1126	0.0331

Table 5.2: Performance Metrics for EM Dataset

(0.0825). This high Recall means that the VGG ensemble successfully identifies most of the true positives, while its False Detection Rate (FD) is slightly higher (0.0387), indicating a slight tendency to predict false positives. Nonetheless, this balance of high Recall and relatively low FD shows that the VGG ensemble is both accurate and confident in its predictions. The other nine Bayesian ensemble models generally perform better than the non-Bayesian ensemble and single models (Gall and Kwon Unet) across most metrics.

Overall, The VGG11-based Bayesian ensemble consistently outperforms other models across metrics like Dice, Jaccard, and Recall, demonstrating its effectiveness in capturing fine-grained details in the EM dataset, likely due to VGG11’s pre-trained architecture and robust feature extraction capabilities. Almost all Bayesian ensembles outperform non-Bayesian ensemble and non-ensemble models, confirming the benefits of Bayesian ensembling. While the ResNet-based ensemble follows closely behind VGG, models like 1dec_5enc and Deeply Supervised Unet also deliver strong results. Interestingly, the non-Bayesian Unet ensemble excels in Precision but struggles in Recall and other metrics, indicating its conservative prediction style limits feature capture in the EM dataset. Single models like the Kwon Unet and Gall Unet perform reasonably well but fall short of ensemble models, reaffirming ensembling’s advantages in uncertainty estimation(as can be seen in Fig.5.4) and segmentation accuracy.

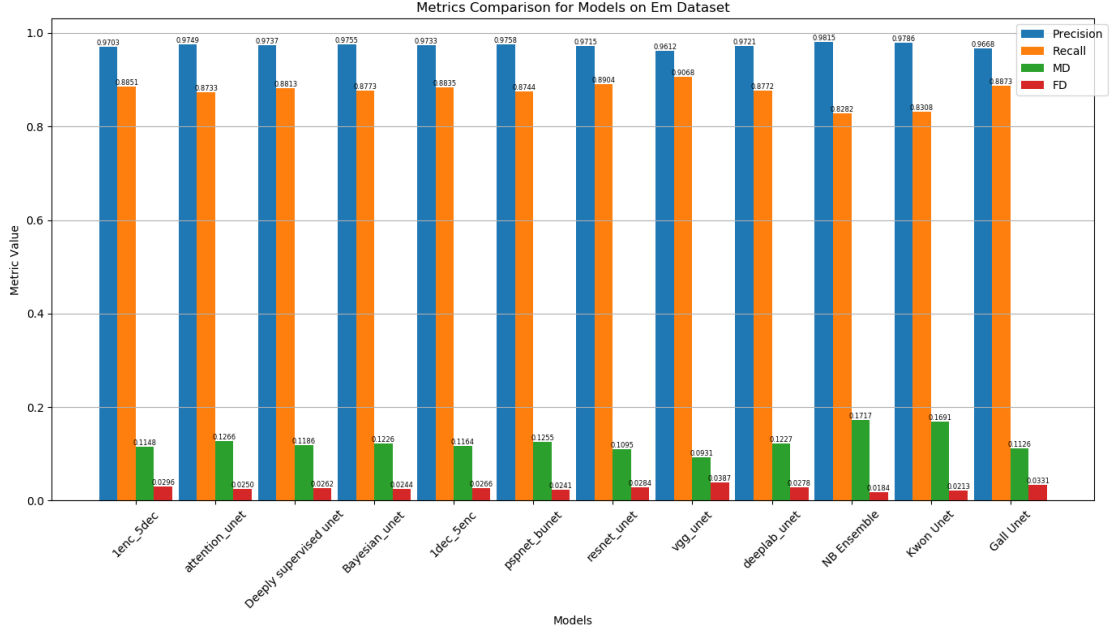


Figure 5.1: Precision, Recall, Miss Detection, False Detection bar diagram for EM Dataset

5.2 Performance analysis on Monuseg Dataset

Model	Dice	Jaccard
1enc_5dec	0.7500	0.8857
attention_unet	0.7675	0.8956
Deeply supervised unet	0.7731	0.8956
Bayesian_unet	0.7569	0.8894
1dec_5enc	0.7763	0.8994
pspnet_unet	0.7470	0.8857
resnet_unet	0.7653	0.8959
vgg_unet	0.7765	0.9003
deeplab_unet	0.7530	0.8890
Non_Bayesian ensemble[24]	0.7515	0.8837
Kwon Unet[23]	0.7396	0.8843
Gall Unet[20]	0.7513	0.8873

Table 5.3: Dice and Jaccard values for the Monuseg dataset.

In the analysis of the 12 models on the Monuseg dataset given in Table 5.3 and 5.4(which is also shown in diagram 5.2), it is clear that the VGG11-based Bayesian ensemble performs the best overall, achieving the highest Dice score (0.7765), Jaccard index (0.9003), precision (0.8410), and lowest false detection (FD) rate (0.1589). This suggests that VGG ensemble is particularly effective at accurately identifying and segmenting regions of interest with minimal false positives. However, its recall (0.7522) and miss detection (MD) rate (0.2460) are not the best, indicating that while it is precise, it may miss some true positive regions. The high precision and FD values indicate strong segmentation accuracy but the slightly lower recall suggests it may not capture all relevant areas, likely due to the complexity of the Monuseg images with their fine-grained edges

Model	Precision	Recall	MD	FD
1enc_5dec	0.7733	0.7642	0.2340	0.2266
Attention_unet	0.8216	0.7550	0.2431	0.1783
Deeply_supervised_unet	0.7998	0.7867	0.2114	0.2001
Bayesian_unet	0.8218	0.7357	0.2624	0.1781
1dec_5enc	0.8213	0.7699	0.2283	0.1786
PSPNet_unet	0.8262	0.7165	0.2817	0.1737
ResNet_unet	0.8321	0.7417	0.2564	0.1678
VGG_unet	0.8410	0.7522	0.2460	0.1589
Deeplab_unet	0.8013	0.7517	0.2465	0.1986
Non_Bayesian_ensemble[24]	0.8125	0.7635	0.2346	0.1874
Kwon_unet[23]	0.8003	0.7287	0.2695	0.1996
Gall_unet[20]	0.7755	0.7709	0.2273	0.2244

Table 5.4: Performance Metrics for Monuseg Dataset

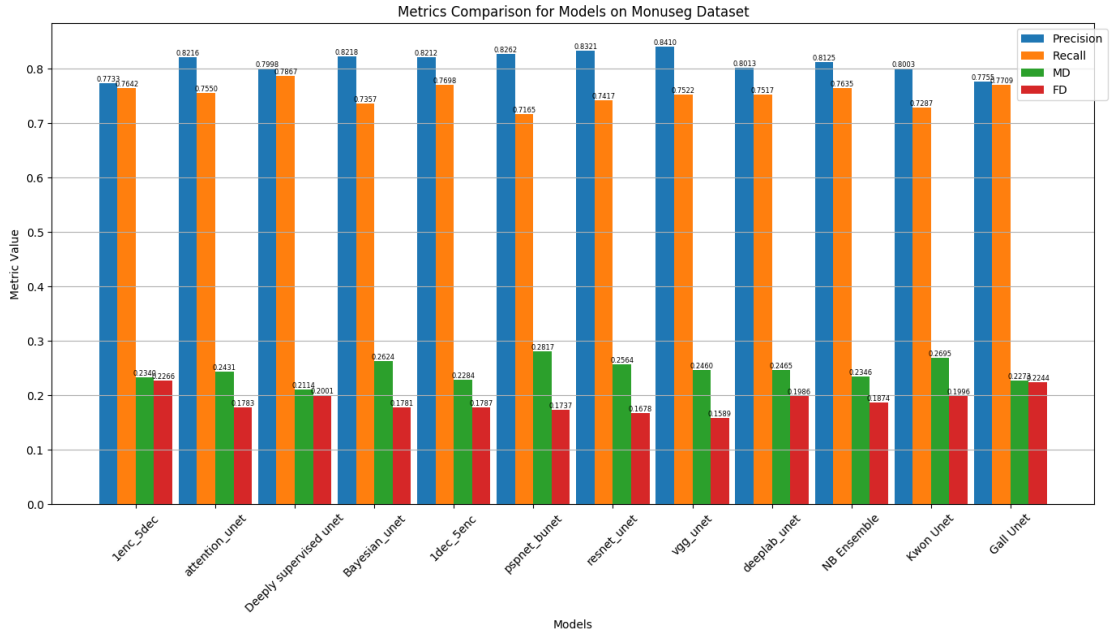


Figure 5.2: Precision, Recall, Miss Detection, False Detection bar diagram for Monuseg Dataset

and overlapping regions.

Close behind, the 1dec_5enc ensemble ranks second, with strong Dice (0.7763), Jaccard (0.8994), and solid recall (0.7699). Its good performance across multiple metrics suggests that utilizing different encoders helps capture diverse features within the images. However, its precision (0.8213) is slightly lower than the VGG ensemble, meaning it might produce slightly more false positives. The Deeply supervised Bayesian unet ensemble also ranks well, especially in recall (0.7867) and miss detection (0.2114), indicating it excels at capturing true positives, likely due to the added deep supervision. On the other hand, models like the Non-bayesian ensemble, Kwon Unet, and Gall Unet underperform in comparison, with lower Dice and Jaccard scores. These models fail to leverage the advantages of Bayesian ensembling, particularly when handling complex images with multiple fine structures in Monuseg.

5.3 Performance analysis on Lung Dataset

Model	Dice	Jaccard
1enc_5dec	0.9765	0.9866
attention_unet	0.9768	0.9867
Deeply supervised unet	0.9767	0.9867
Bayesian_unet	0.9762	0.9865
1dec_5enc	0.9765	0.9866
pspnet_unet	0.9765	0.9866
resnet_unet	0.9735	0.9853
vgg_unet	0.9705	0.9841
deeplab_unet	0.9754	0.9862
Non_Bayesian ensemble[24]	0.9744	0.9857
Kwon Unet[23]	0.9727	0.9850
Gall Unet[20]	0.9726	0.9851

Table 5.5: Dice and Jaccard values for the lung dataset.

In evaluating the lung dataset as seen in Table 5.5 and 5.6(as also shown in 5.3) , the attention-based Bayesian U-Net ensemble emerges as the top performer, achieving the best Dice (0.9768) and Jaccard (0.9867) scores. This model likely excels due to the attention mechanism, which effectively emphasizes important regions in the segmentation task, improving its focus on lung boundaries. Following closely are the 1enc_5dec-based and deeply supervised U-Net Bayesian ensemble, both of which offer comparable performance. The deeply supervised network’s additional loss from intermediate outputs helps guide the model towards better convergence. While the PSPNet-based Bayesian ensemble and other ensemble models also deliver strong results, their slightly lower performance may be due to the relative simplicity of the lung segmentation task, which reduces the impact of advanced multi-scale features or deep supervision.

Model	Precision	Recall	MD	FD
1enc_5dec	0.9866	0.9730	0.0269	0.0133
Attention_unet	0.9848	0.9753	0.0246	0.0151
Deeply_supervised_unet	0.9865	0.9735	0.0264	0.0134
Bayesian_unet	0.9834	0.9756	0.0243	0.0165
1dec_5enc	0.9863	0.9733	0.0266	0.0136
PSPNet_unet	0.9870	0.9727	0.0272	0.0129
ResNet_unet	0.9894	0.9641	0.0358	0.0105
VGG_unet	0.9817	0.9661	0.0338	0.0182
Deeplab_unet	0.9824	0.9752	0.0247	0.0175
Non_Bayesian ensemble[24]	0.9930	0.9627	0.0372	0.0069
Kwon_unet[23]	0.9922	0.9602	0.0397	0.0077
Gall_unet[20]	0.9811	0.9711	0.0288	0.0188

Table 5.6: Performance Metrics for Lung Dataset

Interestingly, the non-Bayesian ensemble ranks highest in precision (0.9930), indicating fewer false positives, but its overall performance lags slightly in recall, Dice, and Jaccard metrics compared to Bayesian ensembles. The small size and simplicity of the lung dataset, where only two large regions (the lungs) need segmentation, likely result in marginal differences between these

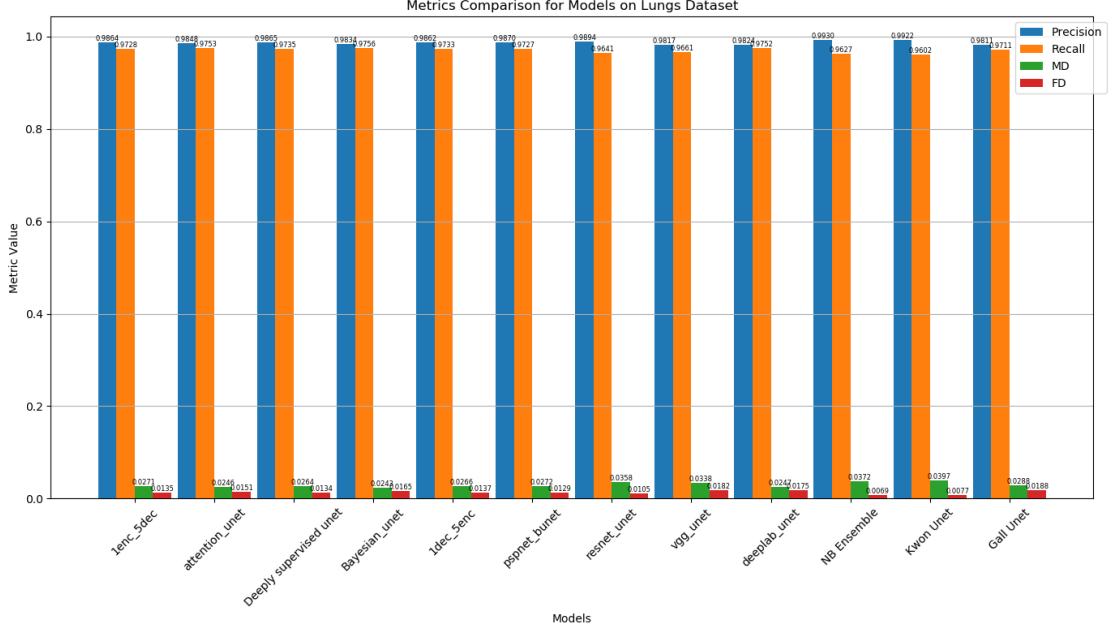


Figure 5.3: Precision, Recall, Miss Detection, False Detection bar diagram for Lung Dataset

models, as even simpler architectures can achieve high accuracy. The VGG11 and ResNet34-based Bayesian U-Nets, which performed well on more complex datasets like EM and Monuseg, did not excel here, suggesting that their feature extraction power is underutilized in such a straightforward segmentation task. Overall, the Bayesian models show slightly improved metrics over the non-Bayesian models, but the simplicity of the lung dataset leads to more uniform performance across all models, including in the uncertainty maps as in 5.7.

5.4 Predictive Uncertainty maps

Here are the predictive uncertainty heatmaps for all models, as shown in Fig.5.4 we observed that the nine Bayesian ensemble models consistently provide better uncertainty estimates than those by Gall Unet, Kwon Unet, and the non-Bayesian ensemble. While all methods struggle to classify certain pixels correctly, the Bayesian ensemble models generally assign higher uncertainty to these misclassified regions.

In contrast, Gall Unet shows rapidly diminishing aleatoric and epistemic uncertainties, with low uncertainty magnitudes in misclassified areas 5.5. This is likely due to the shared parameters across network subsets during dropout, which limits their ability to learn diverse modes. Although Kwon Unet improved aleatoric uncertainty estimation compared to Gall Unet, the issue with epistemic uncertainty remains, and aleatoric uncertainty still tends to converge toward zero 5.5. As a result, the uncertainty maps from Gall Unet and Kwon Unet are less reliable, showing low uncertainties in regions of misclassification. In the non-Bayesian ensemble, the uncertainty maps resemble those of the Bayesian ensembles, but the latter is more precise, with sharper boundaries and stronger

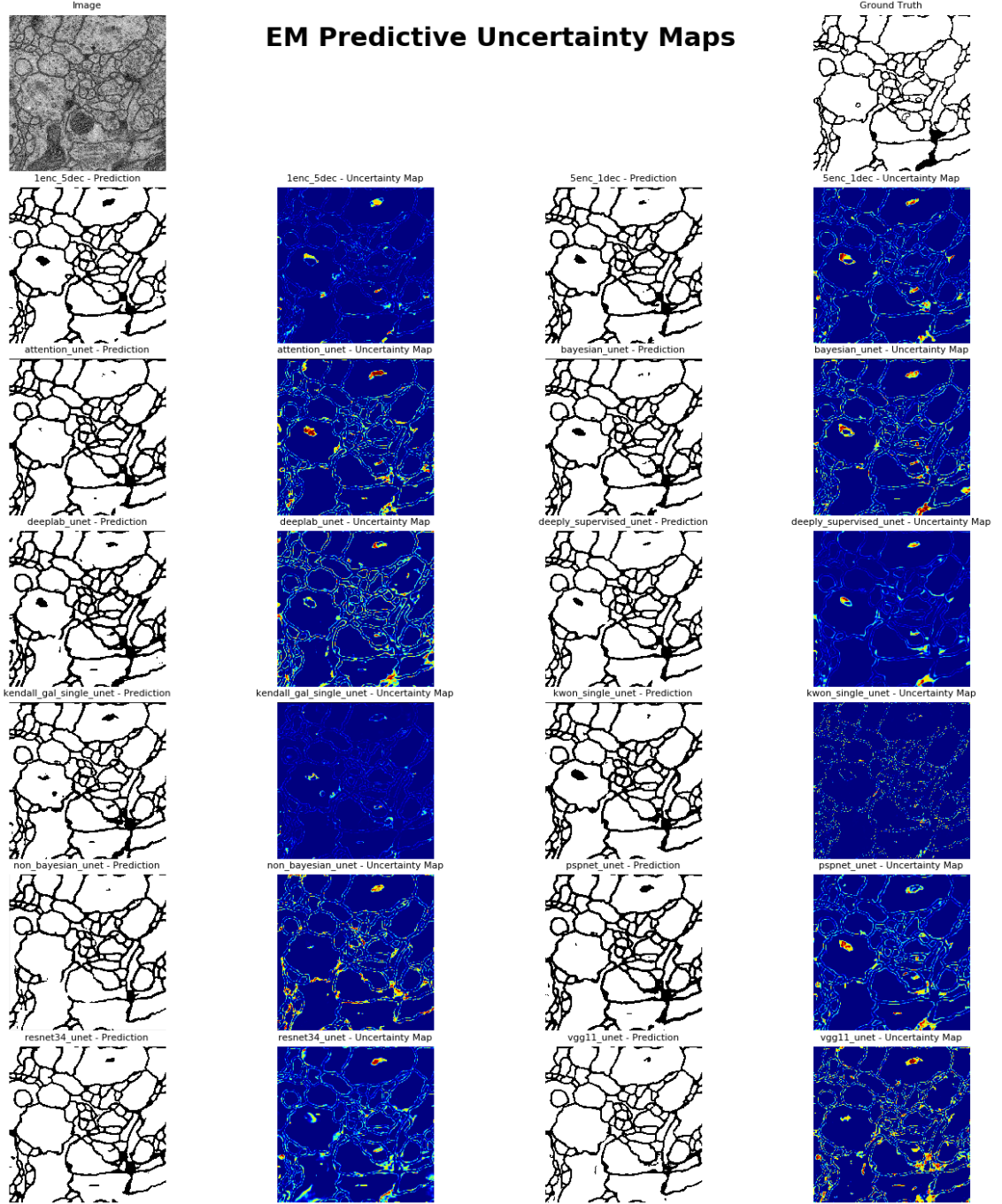


Figure 5.4: Predictive Uncertainty maps of all models on EM Dataset

uncertainties around misclassified pixels, providing smoother segmentation outputs. Although the Bayesian ensembles outperform non-Bayesian ensemble [24] and non-ensemble methods [23, 20], the scores and uncertainty maps are somewhat similar across the different Bayesian ensemble models. This can be attributed to the small dataset size and which may limit the ability of certain models, particularly those that rely on multi-scale learning (like PSPNet and DeepLab), to fully utilize their potential on such fine-grained data. However, the Bayesian ensemble effectively addresses parameter dependence, resulting in better uncertainty estimates and more diverse learning modes.

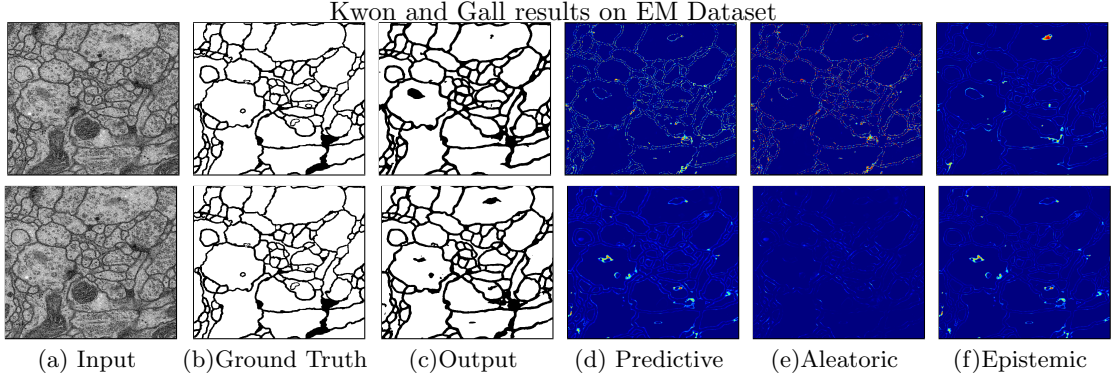


Figure 5.5: The 1st row shows the image segmentation results on the EM dataset from Kwon’s method [23], while the second row presents results from Gal’s method [20]

In Fig. 5.6 the Non-Bayesian ensemble, Kwon Unet, and Gal Unet models struggle with segmentation accuracy. The Non-Bayesian ensemble shows scattered uncertainty across both correct and incorrect areas, reflecting a lack of confidence differentiation, while the Kwon Unet performs poorly with frequent segmentation errors and widespread uncertainty, suggesting low reliability. The Gal Unet similarly misses many cell boundaries and displays high uncertainty within cell interiors, indicating poor predictive performance. The Bayesian ensemble models consistently demonstrate accurate segmentation with well-localized uncertainty focused on cell boundaries. These models confidently handle uncertain regions by capturing predictive uncertainty, making them more reliable for segmentation tasks. Their ability to model uncertainty leads to better predictions, especially in ambiguous or difficult areas.

Given that the lung dataset is relatively simple (segmentation of two large lung areas), the uncertainty maps in 5.7 across all models are somewhat reasonable, with most models focusing their uncertainty at the borders. This is likely due to the nature of lung segmentation tasks where boundary regions tend to pose the greatest challenge, especially in the central areas near anatomical structures like the trachea and bronchi. However, since the task is relatively easy, models might default to showing uncertainty around borders as a safeguard, even if the predictions are largely correct.

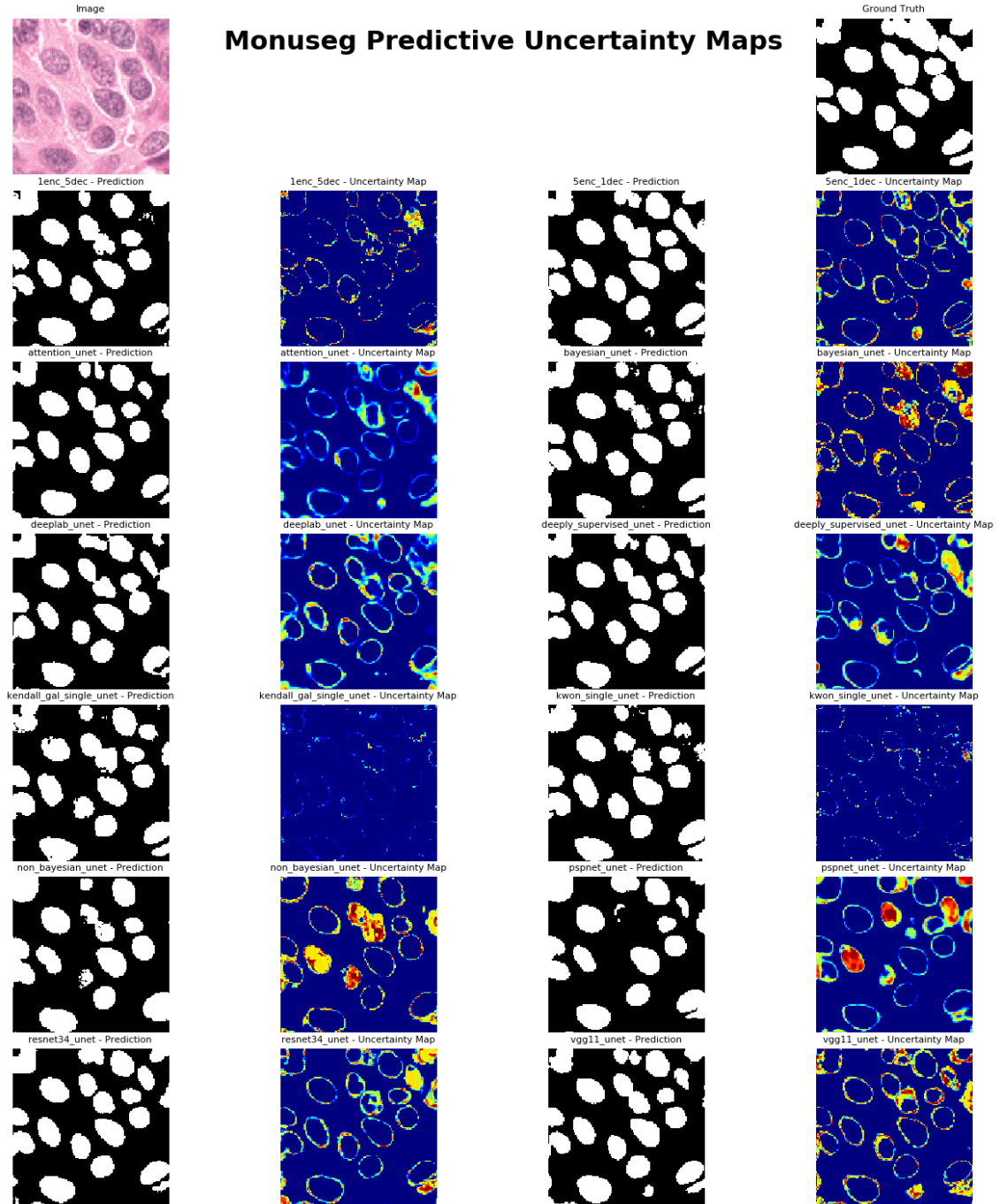


Figure 5.6: Predictive Uncertainty maps of all models on Monuseg Dataset

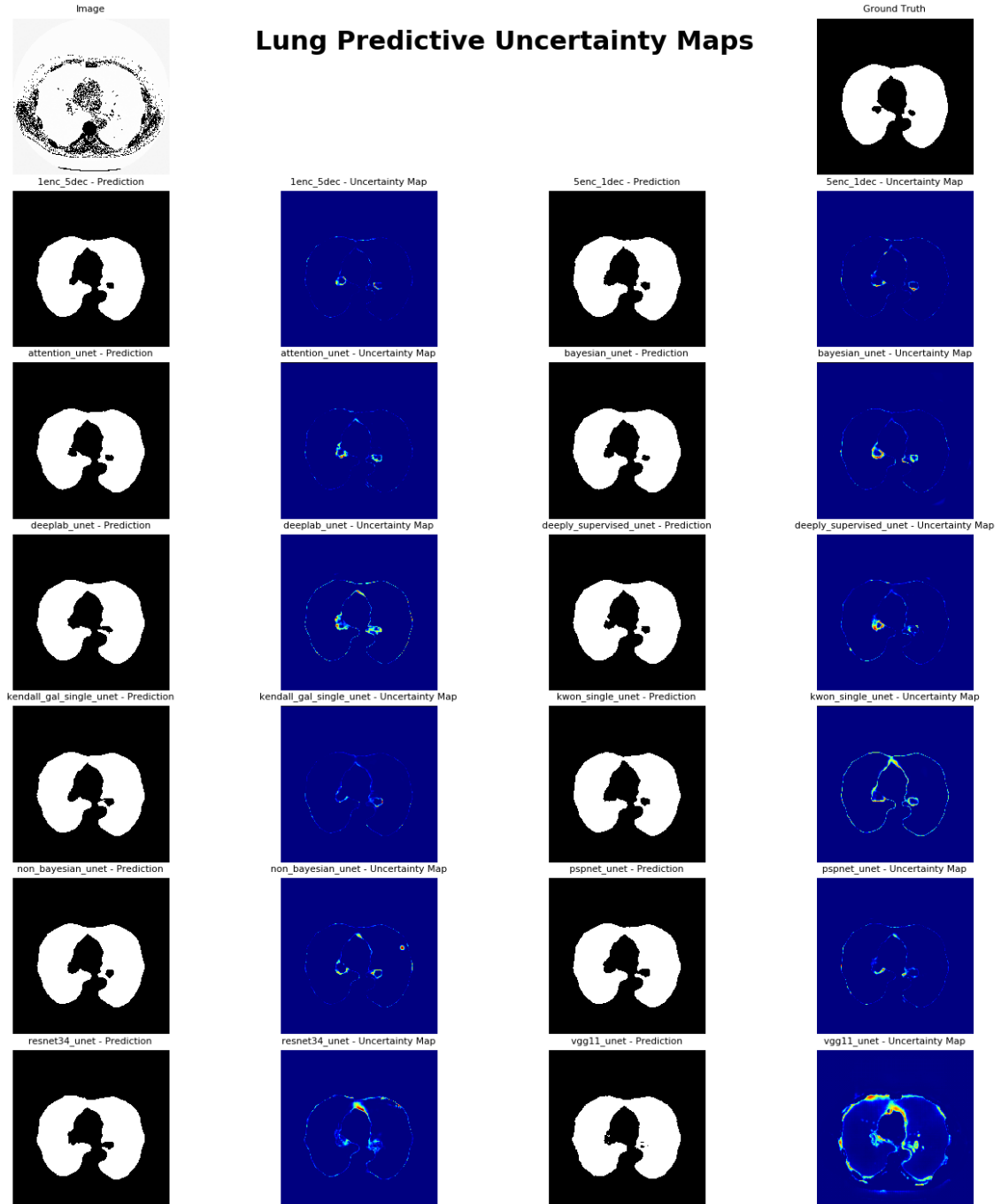


Figure 5.7: Predictive Uncertainty maps of all models on Lungs Dataset

Chapter 6

Conclusion

In this thesis, we explored the challenge of uncertainty estimation in medical image segmentation using various Bayesian ensemble approaches. Our analysis revealed that many existing methods, such as Kwon [23] and Gal[20], lead to an undesirable convergence of uncertainty toward zero, producing misleading confidence maps. In contrast, our proposed Bayesian ensemble models demonstrated improvements in uncertainty estimation while maintaining similar or better segmentation accuracies compared to non-ensemble Bayesian works [23, 20] and non-Bayesian ensemble approach [24]. Through extensive experimentation, we showed that Bayesian ensemble models provide more reliable uncertainty estimates, particularly at misclassified regions, reinforcing the effectiveness of ensembling in uncertainty quantification. On the EM and Monuseg datasets, Bayesian ensemble models consistently provided better uncertainty maps compared to non-Bayesian ensembles, Kwon, and Gal. While non-Bayesian ensembling improved over Kwon and Gal, it still failed to capture uncertainty at key misclassified locations where Bayesian ensembles succeeded, strengthening our claim of the superiority of Bayesian approaches. In simpler tasks like lung segmentation, where the dataset is relatively straightforward, most models delivered reasonable uncertainty maps, concentrating uncertainty at the borders.

Our findings highlight the importance of uncertainty estimation and establish Bayesian ensembles as a valuable tool for improving the reliability of medical image segmentation and an in-depth investigation into the relationship between implicit encoder-decoder architecture variations and uncertainty estimation across these datasets. We recognize that further research can enhance the performance of Bayesian networks in generating more accurate uncertainty estimates. This could involve curating the datasets more carefully, such as incorporating a separate boundary segmentation task. Moreover, exploring alternative posterior approximation methods beyond MC Dropout may lead to more reliable and precise uncertainty estimates in medical image segmentation.

Bibliography

- [1] Santiago Aja-Fernández, Ariel Hernán Curiale, and Gonzalo Vegas-Sánchez-Ferrero. A local fuzzy thresholding methodology for multiregion image segmentation. *Knowl. Based Syst.*, 83:1–12, 2015.
- [2] Devadhas G. Glan Aswathy Sukumaran and S. S. Kumar. An improved tumor segmentation algorithm from t2 and flair multimodality mri brain images by support vector machine and genetic algorithm. *Cogent Engineering*, 5(1):1470915, 2018.
- [3] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, 2009.
- [4] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulkas, Pavel Tomancak, and Volker Hartenstein. An integrated micro- and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLOS Biology*, 8:1–17, 10 2010.
- [5] Ignacio Arganda Carreras, Srinivas C. Turaga, Daniel R. Berger, Dan Cire San, Alessandro Giusti, Luca M. Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M. Buhmann, Ting Liu, Mojtaba Seyedhosseini, Tolga Tasdizen, Lee Kamentsky, Radim Burget, Vaclav Uher, Xiao Tan, Changming Sun, Tuan D. Pham, Erhan Bas, Mustafa G. Uzunbas, Albert Cardona, Johannes Schindelin, and Hyunjune Sebastian Seung. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 9(November):1–13, 2015.
- [6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 833–851, 2018.
- [7] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2016.
- [8] Yuxin Cui, Guiying Zhang, Zhonghao Liu, Zheng Xiong, and Jianjun Hu. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. *Medical & biological engineering & computing*, 57:2027–2043, 2019.
- [9] Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill DF Campbell, and Ivor Simpson. Structured uncertainty prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5477–5485, 2018.
- [10] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.
- [11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- [12] Sanket Gandhi. Uncertainty estimation for medical image segmentation. Accessed: 2019-06-03.

- [13] Florin C. Ghesu, Bogdan Georgescu, Yefeng Zheng, Sasa Grbic, Andreas K. Maier, Joachim Hornegger, and Dorin Comaniciu. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:176–189, 2019.
- [14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- [15] Xu Haixiang, Cao Wanhua, Chen Wei, and Guo Liyuan. Performance evaluation of svm in image segmentation. *2008 9th International Conference on Signal Processing*, pages 1207–1210, 2008.
- [16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.
- [17] Vladimir Iglovikov and Alexey Shvets. Ternaunet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.
- [18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.
- [19] Zoltán Kapás, László Lefkovits, and László Szilágyi. Automatic detection and segmentation of brain tumor using random forest approach. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 301–312. Springer, 2016.
- [20] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, pages 5574–5584, 2017.
- [21] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Led- sam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, pages 6965–6975, 2018.
- [22] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.
- [23] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. *Medical Imaging with Deep Learning*, 4(2), 2018.
- [24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6402–6413, 2017.
- [25] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [26] K. Mader. Finding and measuring lungs in ct data. <https://www.kaggle.com/datasets/kmader/finding-lungs-in-ct-data>. Accessed: 2019-06-27.
- [27] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018.
- [28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

- [29] Ahana Roy Choudhury, Rami Vanguri, Sachin R Jambawalikar, and Piyush Kumar. Segmentation of brain tumors using deeplabv3+. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 154–167. Springer, 2019.
- [30] Gaurav Sethi, Barjinder Singh Saini, and Dilbag Singh. Segmentation of cancerous regions in liver using an edge-based and phase congruent region enhancement method. *Comput. Electr. Eng.*, 53:244–262, 2016.
- [31] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2014.
- [32] Alexey A Shvets, Vladimir I Iglovikov, Alexander Rakhlin, and Alexandr A Kalinin. Angiodysplasia detection and localization using deep convolutional neural networks. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 612–617. IEEE, 2018.
- [33] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *Computational and Biological Learning Society*, pages 1–14, 2015.
- [34] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 560–569. AUAI Press, 2018.
- [35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.
- [36] S.Thylashri, Udutha Mahesh Yadav, and T.Danush Chowdary. Image segmentation using k-means clustering method for brain tumour detection. *International journal of engineering and technology*, 7:97, 2018.
- [37] Carole Helene Sudre, Wenqi Li, Tom Kamiel Magda Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep learning in medical image analysis and multimodal learning for clinical decision support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City, QC*, 2017:240–248, 2017.
- [38] Abhishek Vahadane, Tingying Peng, Shadi Albarqouni, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Amit Sethi, Irene Esposito, and Nassir Navab. Structure-preserved color normalization for histological images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 1012–1015. IEEE, 2015.
- [39] Patil Priyanka Vijay and Nikita Patil. Gray scale image segmentation using otsu thresholding optimal approach. *International Journal for Scientific Research and Development*, 2:20–24, 2016.
- [40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [41] Kebin Wu and David Dian Zhang. Robust tongue segmentation by fusing region-based and edge-based approaches. *Expert Syst. Appl.*, 42:8027–8038, 2015.
- [42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.