D-LEMA: Deep Learning Ensembles from Multiple Annotations - Application to Skin Lesion Segmentation

Zahra Mirikharaji Kumar Abhishek Saeed Izadi Ghassan Hamarneh
School of Computing Science, Simon Fraser University, Canada
{zmirikha, kabhishe, saeedi, hamarneh}@sfu.ca

Abstract

Medical image segmentation annotations suffer from inter- and intra-observer variations even among experts due to intrinsic differences in human annotators and ambiguous boundaries. Leveraging a collection of annotators' opinions for an image is an interesting way of estimating a gold standard. Although training deep models in a supervised setting with a single annotation per image has been extensively studied, generalizing their training to work with datasets containing multiple annotations per image remains a fairly unexplored problem. In this paper, we propose an approach to handle annotators' disagreements when training a deep model. To this end, we propose an ensemble of Bayesian fully convolutional networks (FCNs) for the segmentation task by considering two major factors in the aggregation of multiple ground truth annotations: (1) handling contradictory annotations in the training data originating from inter-annotator disagreements and (2) improving confidence calibration through the fusion of base models' predictions. We demonstrate the superior performance of our approach on the ISIC Archive and explore the generalization performance of our proposed method by cross-dataset evaluation on the PH^2 and DermoFit datasets.

1. Introduction

The semantic segmentation task in computer vision involves partitioning an image into a set of multiple non-overlapping and semantically interpretable regions [10], and this entails assigning pixel-wise class labels to the entire image, making it a dense prediction task. Segmentation is a crucial task in the visual computing pipeline and is often used to improve several downstream tasks such as classification and depth estimation [39]. Following the seminal work of Long et al. [24], deep learning-based semantic image segmentation models have gained prominence because of their superior performance over traditional approaches. The majority of deep learning-based semantic segmentation models, however, rely on supervised learning of dense pixel annotations for the labels in images. State of the art supervised learning algorithms rely upon training using large volumes of data to yield acceptable results, and previous work has shown the importance of sufficient annotated data for visual tasks [28, 12, 34]. Particularly, Sun et al. [34] showed that the performance of segmentation models in terms of overlap based measures exhibits a logarithmic relationship with the amount of training data used for representation learning for semantic segmentation.

Collecting ground truth annotations for semantic segmentation is considerably more expensive than doing so for other visual tasks such as classification and object detection because of the dense annotations involved. While this can partly be ameliorated by crowd-sourcing the annotation process to non-experts, the presence of multiple object classes in a scene, coupled with factors such as illumination, shading, and occlusion, makes delineating the exact object boundaries an ambiguous and tedious task, leading to inter-annotator disagreements. The presence of multiple annotations (Figure 1) further leads to the challenge of deciding upon an ideal ground truth against which the model's performance is assessed. Moreover, there exists a tradeoff between the precision and the generalizability of an 'ideal' segmentation ground truth, since aoverly precise delineation may not be reflective of the typical uncertainty encountered in practice when localizing the boundary [38]. A similar trade-off exists between the quality and the efficiency of these annotations: High quality dense annotations, although useful, take up more time to collect than relatively less informative approximate annotations (e.g., bounding boxes or simplified polygons). These problems are exacerbated further for medical images since medical imaging datasets with accurate pixel-level annotations are much smaller than their natural image counterparts [35], which can be attributed to the high cost associated with expert annotations, the difficulty in quantifying a true reference standard, the laborious nature of making dense annotations, which is even more difficult for 3D medical image volumes, and patient data privacy concerns. To add to

this, the manual annotation of anatomical regions of interest can be very subjective and presents considerable inter- and intra-annotator disagreements even amongst experts across multiple medical imaging modalities [37, 7, 36, 29, 9], making it difficult to converge on a single gold standard annotation for model training and evaluation.

One of the seminal works on comparing a segmentation model's performance by comparing against a collection of (human-annotated) segmentations is that proposed by Warfield et al. [38], where they proposed an expectation maximization algorithm for the simultaneous truth and performance level estimation (STAPLE). Given a collection of segmentation masks, STAPLE generates a probabilistic estimate of the true segmentation mask as well as the segmentation performance of each of the segmentations in the collection. This was followed by several other extensions of STAPLE which addressed its limitations such as susceptibilities to large variations in inter-annotator uncertainty and annotator performance [3, 14, 21, 23].

More recently, Mirikharaji et al. [27] showed that leveraging different levels of annotation reliability, using spatially-adaptive reweighting while learning deep learning based segmentation model parameters, helps improve performance, and demonstrated superior segmentation accuracy using a large number of low quality, 'noisy' annotations along with only a small fraction of precise annotations. Hu et al. [11] used a modified probabilistic U-Net [17] model to generate quantifiable aleatoric and epistemic uncertainty estimates for segmentation using a supervised learning framework which modeled inter-annotator variability as aleatoric uncertainty ground truth. Ribeiro et al. [29] proposed an approach to improve inter-annotator agreement by conditioning the segmentation masks using morphological image processing operations (opening and closing), convex hulls and bounding boxes to remove details specific to any single particular annotator. They argue that the conditioning could be deemed as denoising operations, removing the annotator specific details from the segmentation masks. The same authors then proposed to train their segmentation model on a subset of the images, derived by filtering out all samples whose mean pairwise Cohen's kappa score was less than 0.5, thus using only those segmentations which largely agree between annotators [30].

Despite the obvious benefits of improving segmentation performance, it is also crucial to analyze the predictive uncertainty of deep networks in medical image segmentation. In machine learning, the uncertainty has been classified into aleatoric and epistemic types. The aleatoric, which reflects the inherent noise in the data, has been estimated using a second auxiliary output in the network [16]. Bayesian neural networks (BNNs) have adopted Monte Carlo (MC) dropout [8] to reflect the epistemic uncertainty associated with the network parameters. Thanks to their simplicity, MC dropout uncertainty estimation has been studied in the context of general semantic segmentation [15] as well as medical image segmentation [18, 33]. However, the uncertainty estimates obtained using MC dropout tend to be miscalibrated, i.e., they do not correspond well with the model error [22]. Recently, there have been efforts to improve the uncertainty calibration using ensemble learning. Particularly, Lakshminarayanan et al. [19] demonstrated the advantage of ensemble learning, i.e., averaging a collection of models trained from different initializations, in yielding more accurate predictive uncertainty estimates for classification and regression tasks. Mehrtash et al. [25] studied the performance of ensemble learning for predictive uncertainty in medical image segmentation. Particular to skin lesion segmentation, Jungo et al. [13] thoroughly studied the reliability of existing uncertainty estimation methods and showed their benefits and limitations [13].

Deep neural networks have been shown to potentially overfit to noisy labels [40] and our motivation for this work is to avoid single annotator bias [20]. Therefore, we seek training deep segmentation models to learn from multiple annotations as available instead of discarding some annotations. Rather than selecting a subset of images to learn from Ribeiro et al. [30], we instead propose a generalized approach of annotation weighting by leveraging different groups of consistent annotations in an ensemble method towards efficiently learning from all available annotations. We also utilize uncertainty estimates [16, 19] in an ensemble learning framework to improve predictive uncertainty and calibration confidence in the final prediction.

Contribution claims: We consider two major factors in the aggregation of multiple ground truth annotations: (1) handling contradictory annotations in the training data originating from inter-annotator disagreements, and (2) improving the model's confidence calibration through deep ensembling. Our hypothesis is that given a new image, leveraging different experts' skills independently and fusing them in an ensemble model, while considering their estimated uncertainty, makes for a more reliable final prediction.

2. Method

2.1. Problem Statement and Method Overview

Let $\mathcal{X} = \{X_n\}_1^N$ and $\mathcal{Y} = \{Y_n\}_1^N$ be a set of N images and segmentation ground truth masks, respectively. In a supervised learning scheme, a network is trained to learn a function $f_\theta : X_n \mapsto \hat{Y}_n$ parameterized by θ, which maps an image X_n to the corresponding estimated segmentation mask $\hat{Y}_n$. Approximating the mapping function f_θ using a single annotation per image has been well studied in the literature. However, training supervised models in the presence of multiple annotations remains largely unexplored.

Let us assume that K annotators have independently an-

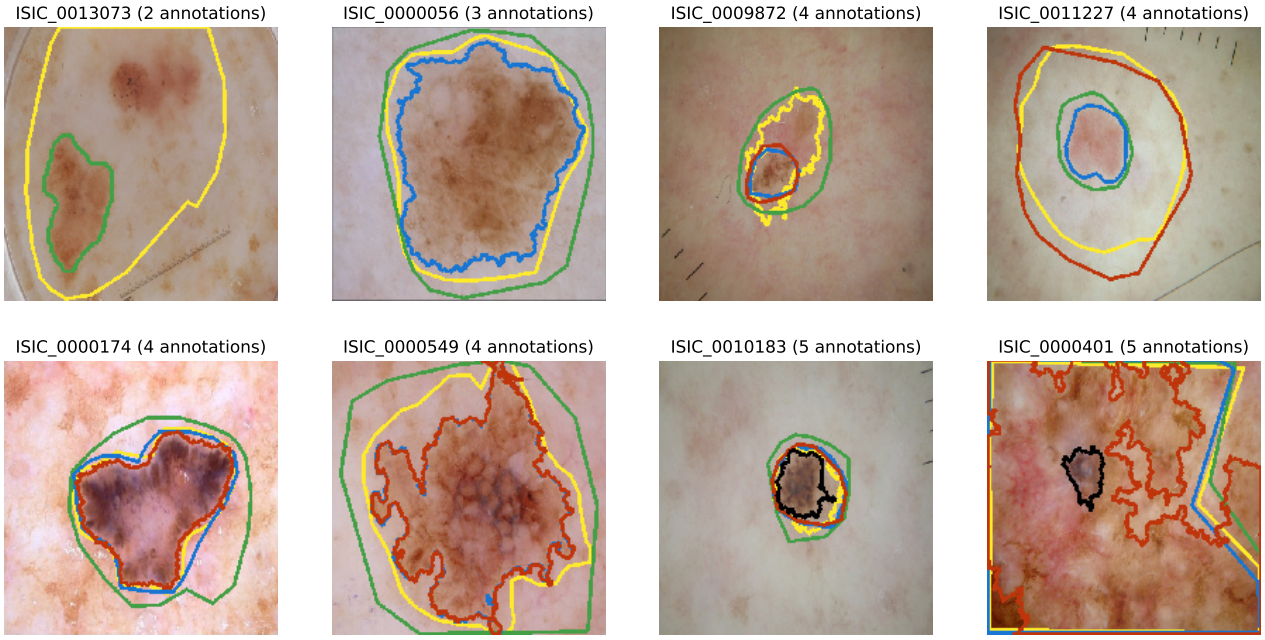

Figure 1: Sample skin lesion images from the ISIC Archive which contain multiple lesion boundary annotations (denoted by different colors).

notated different subsets of the images resulting in a set of segmentation ground truths $\mathcal{Y} = \{\{Y_{mn}\}_{m=1}^{M_n}\}_{n=1}^{N}$, where M_n denotes the number of available annotations for X_n. Inconsistent annotations for a given image could mislead the network and substantially deteriorate the performance of the model. Let M indicate the maximum number of annotations per image over the entire dataset. Instead of aggregating multiple annotations to estimate a single ground truth before the training phase, we propose to (1) learn a set of M mapping functions $\mathcal{F} = \{f_{\theta^i}\}$ through ensembling M base deep models trained over the union of available annotations and (2) minimize the confusion induced from observing multiple annotations through a spatial reweighting scheme during training. (3) Lastly, we demonstrate that our proposed ensemble learning framework not only improves the segmentation performance but also provides a well-calibrated predictive uncertainty. Figure 2 illustrates the overview of our ensemble learning framework for skin lesion segmentation with multiple annotations.

2.2. Detailed Method

Non-contradictory Subsets Selection: To handle contradictory annotations arising from having multiple annotations per image during the training, we partition the entire dataset into M disjoint subsets, denoted by $\{\mathcal{C}^i\}_{i=1}^{M}$, such that each $\mathcal{C}^i$ includes at most one unique annotation for every image. In particular, for each image, with $M_n \leq M$ annotations, we randomly assign the M_n annotations to $\{\mathcal{C}^i\}_{i=1}^{M_n}$ subsets.

A naïve approach is to utilize these disjoint subsets to train individual base models independently. Even though this solution prevents exposing each ensemble base model to multiple annotations per image and encourages a diverse set of model performance, however, each disjoint set includes a small number of training samples which can adversely affect the generalization capability of individual base models. To address this issue, we combine all images along with all available annotations into a *union* dataset, denoted as $\mathcal{U}$, and use it to train M base networks. Following Mirikharaji et al. [27], we utilize these non-contradictory subsets to assess the quality of annotations in $\mathcal{U}$. Specifically, spatially-adaptive weight maps associated with varying annotations in $\mathcal{U}$ are learned to adjust the contribution of each annotated pixel in the optimization of deep network based on its consistency with clean annotations in $\{\mathcal{C}^i\}$.

Learning Models: In more details, for each base model i, $i \in 1, ..., M$, we define a cross-entropy loss, denoted as $\mathcal{L} = \{L_{ce}^{\mathcal{C}^i}\}$ over each non-contradictory set $\mathcal{C}^i$. We then, in a meta-learning paradigm, learn a set of spatial weight maps $\mathcal{W}^i = \{\{W_{mn}^i\}_{m=1}^{M_n}\}_{n=1}^{N}$ for all annotations $\mathcal{U}$ based on the gradients of the cross-entropy losses with respect to the weights maps, i.e. $\nabla_{W^i}\mathcal{L}_{ce}^{\mathcal{C}^i}$. This way, $\mathcal{W}^i$ is optimized to cancel out the contributions of annotations inconsistent with $\mathcal{C}^i$ while optimizing the parameters for i^{th} base network, i.e.

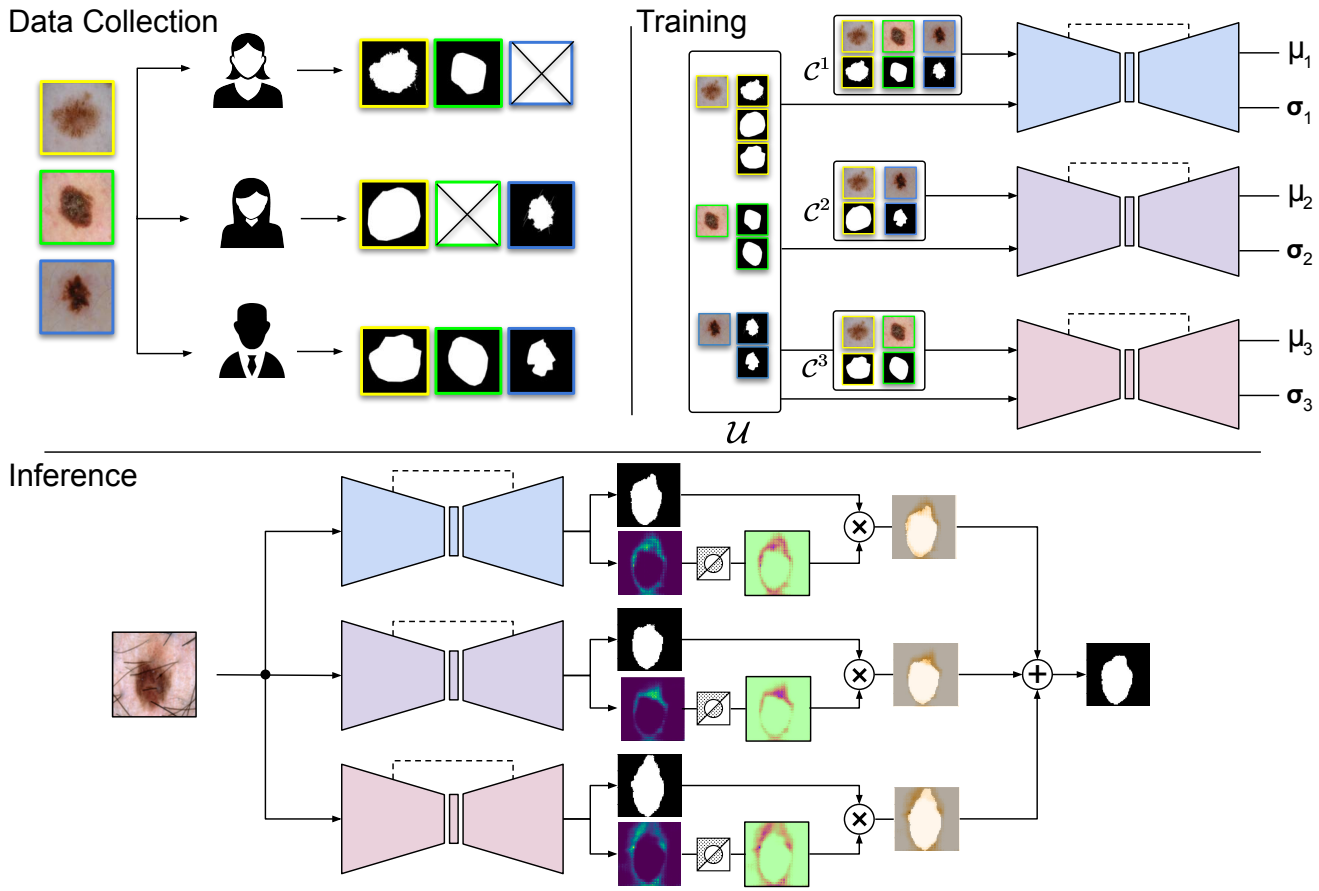

Figure 2: An overview of our proposed framework for skin lesion segmentation with multiple annotations. (top left) Multiple users annotating different, potentially overlapping, subsets of the original data. (top right) Each set of non-contradictory labels is considered as ground truth and, along with the remaining annotations that are deemed potentially noisy, are used to train a different base model. (bottom) At inference, each base model's prediction, along with its estimated aleatoric uncertainty maps are fused to obtain the final prediction.

θ^i. Mathematically:

$$\mathcal{W}^{i*} = \underset{\mathcal{W}^i,\, \mathcal{W}^i \geqslant 0}{\arg\min} \sum_{n \in \mathcal{C}^i} L_{ce}^n(\hat{Y}_n^i, Y_n; \theta^i(\mathcal{W}^i)). \tag{1}$$

Note that every image in $\mathcal{C}^i$ has only one ground truth. $\mathcal{W}^i$ are encoded in $\mathcal{L}$ and they are optimized along with the network parameters θ^i for each individual base model. By integrating the information in the optimized $\mathcal{W}^i$, we can determine the degree by which a pixel-level annotation from any of annotators is considered noisy for model i, depending on how similar this annotation is to the annotations in $\mathcal{C}^i$. Therefore:

$$\mathcal{L}(\hat{Y}_n^i, Y_{mn}; \theta^i, W_{mn}^i) = -\sum_{q \in X_n} W_{mnq}^i Y_{mnq} \log \hat{Y}_{nq}^i, \tag{2}$$

$$\hat{Y}_{nq}^i = \text{softmax}(U_{nq}^i). \tag{3}$$

Fusion of Predictions: Once the individual base models are trained, the final prediction of the entire ensemble for the X_n is obtained by using a weighted fusion [31], that is:

$$\hat{Y}_n = \sum_{i=1}^{M} \alpha_n^i \hat{Y}_n^i, \tag{4}$$

where α_n^i is the combination coefficient for prediction by model i. The simplest way to determine α_n^i is to consider equally weighted averaging and set them to $1/M$. Another

popular technique is to set α_n^i coefficients according to the confidence of the model [32]. In this work, we explore both aggregation techniques in our experimental evaluations.

Uncertainty-driven Aggregation: For the uncertainty-driven aggregation of base models, we leverage aleatoric uncertainty, which models irreducible observation noise, to estimate how confident a base model is about its prediction, and utilize the confidence when combining the base models' prediction maps. Following Kendall et al. [16], we approximate the aleatoric uncertainty for each pixel $q \in X_n$ by placing a Gaussian distribution over the logit space before applying a sigmoid function in the last layer and reformulate the network output as:

$$U_{nq}^i \sim \mathcal{N}\left(f_{nq}^i, (\sigma_{nq}^i)^2\right), \tag{5}$$

where f_i and σ_i are the network i outputs.

We use the aleatoric uncertainty in two forms: (1) considering the pixel-wise uncertainty values as spatially-adaptive coefficients and (2) averaging the pixel-wise uncertainty into a single scalar image-level coefficient.

3. Experiments

3.1. Data

For training, we used the International Skin Imaging Collaboration (ISIC) Archive data [1, 6, 5], the largest dermoscopic public dataset with over 13,000 images, captured by diverse devices in international clinical centers. All images are 8-bit RGB color dermoscopy images. Similar to Ribeiro et al. [30], we utilized 2,223 images with more than one segmentation ground truth mask (2,094 with two, 100 with three and 36 with four and 3 with five) to train our models. We split all 2,223 images to 80% for training and 20% for validation. For model selection, we randomly selected which annotation to use in validation set. To create our non-contradictory annotation sets, all training data are randomly and uniformly partitioned into five groups of overlapping images but unique ground truth annotations. ISIC ground truth masks were generated using three different pipelines with different levels of border irregularities all involving a dermatologist with expertise in dermoscopy: (1) an automatic algorithm followed by an expert review; (2) a semi-automatic algorithm controlled by an expert; and (3) manually drawing a polygon by an expert. A large variation of disagreement based on Cohen's kappa scores with the mean 0.67 is reported in Ribeiro et al. [29]. Figure 1 shows some examples of skin lesion images with multiple lesion boundary annotations from this dataset.

To thoroughly assess the segmentation performance of our proposed ensemble framework, we leveraged three publicly available datasets in our evaluations. All the images in the used datasets are resized into 96×96 pixels and normalized using the per-channel mean and standard deviation across the entire dataset. A brief description of these test datasets are provided as follows:

- **ISIC**: Ribeiro et al. [30] randomly selected a test set of 2,000 images with just one segmentation ground truth from ISIC Archive. We used the exact set in our experimental evaluations for fair comparisons.
- **PH^2**: The PH^2 (Pedro Hispano Hospital) dataset contains 200 8-bit RGB color dermoscopic images [26]. All images are acquired under the same condition using Tuebinger Mole Analyzer system at 20$\times$ magnification.
- **DermoFit**: This dataset has 1300 8-bit RGB color clinical images [2]. The images are captured with a Canon EOS 350D SLR camera at the same distance from the lesion under controlled lighting conditions.

3.2. Base Models and Implementation Details

Our architecture is an encoder-decoder architecture with residual and skip connections transferring the information in the encoder modules to the corresponding decoder modules [4]. Since the images in our training dataset are paired with at most five annotations ($M = 5$), our ensemble framework consists of five base deep neural networks. Each network outputs two spatial maps in the last layer: the dense segmentation prediction and the predicted aleatoric uncertainty map. In training the aleatoric loss, 10 Monte Carlo samples from logits are taken. Stochastic gradient descent with an initial learning rate of 10^{-4} is used to optimize the network parameters. The batch size for optimizing the spatial weight maps and network parameters is 64 and 2. The momentum and weight decay are set to 0.99 and 5 $\times 10^5$, respectively.

3.3. Results

Table 1 compares the segmentation performance of our baseline models as well as the individual base models, across different prediction fusion schemes, using the Jaccard index. To train the baseline model, for every image in the training batch, we randomly select which ground truth to use when optimizing the loss function (row A). While it is interesting to consider each annotator separately and evaluate their performance, the assignments between annotators and ground truth are not stated in the ISIC Archive dataset. Instead, we evaluate the performance of each base model trained on non-contradictory annotations simulating an expert knowledge (rows B to F). In addition, we compare the performance of our proposed method against the work of Ribeiro et al. [30] where a subset of samples with small annotator disagreements is taken into account during the train-

Table 1: Comparing the segmentation performance based on Jaccard index reported in percent ($\% \pm$ standard error) on three datasets.

	Method	ISIC Archive [1]	PH^2 [26]	DermoFit [2]
A	baseline	68.00 ± 0.56	81.30 ± 0.77	70.30 ± 0.54
B	model 0	69.22 ± 0.53	82.82 ± 0.75	72.57 ± 0.50
C	model 1	69.75 ± 0.55	82.40 ± 0.75	71.05 ± 0.55
D	model 2	70.33 ± 0.52	83.46 ± 0.74	72.80 ± 0.51
E	model 3	70.37 ± 0.51	83.31 ± 0.70	73.04 ± 0.53
F	model 4	69.73 ± 0.52	82.29 ± 0.72	70.87 ± 0.48
G	equally weighted fusion (ours)	**72.11± 0.51**	84.96± 0.73	74.22± 0.51
H	pixel-level confidence (ours)	71.46± 0.49	84.52± 0.74	73.91± 0.53
I	image-level confidence (ours)	**72.08± 0.49**	**85.20 ± 0.70**	**74.33± 0.50**
J	less is more [30]	69.20	81.25	72.55

Table 2: Comparing predictive uncertainty based on negative log-likelihood (NLL) and Brier score (Br) on three datasets. Lower NLL and Br values correspond to a better predictive uncertainty estimate.

Dataset		ISIC Archive		PH^2		DermoFit	
Method		NLL	Br	NLL	Br	NLL	Br
A	MC dropout model 0	0.073	0.019	0.166	0.048	0.272	0.082
B	MC dropout model 1	0.075	0.020	0.151	0.044	0.310	0.099
C	MC dropout model 2	0.075	0.019	0.149	0.044	0.283	0.087
D	MC dropout model 3	0.078	0.020	0.152	0.042	0.291	0.091
E	MC dropout model 4	0.075	0.019	0.155	0.045	0.312	0.100
F	deep ensemble (ours)	**0.070**	**0.018**	**0.144**	**0.041**	**0.254**	**0.078**

ing.For the fusion stage, we examine three approaches as listed below:

- **Uniformly weighted fusion**: The predictions from the base models are combined by averaging the output probabilities.

- **Pixel level confidence-based fusion**: The predictions from the models are fused using normalized confidence spatial maps computed by inverting the predicted aleatoric outputs.

- **Image level confidence-based fusion**: The aleatoric uncertainty maps are aggregated into an image level aleatoric scalars and the predictions of the base models are combined based on the image-level normalized confidence scalars computed by inverting the uncertainty scalars.

Our results demonstrate that leveraging all available annotations effectively in an ensemble framework consistently improves the performance of the segmentation performance both in a held-out test set and over two other distinct datasets. Looking into different variants of our deep ensemble method, it is evident that aggregating the aleatoric uncertainty into the image-level scalar and leveraging them in the fusion stage (row H) either outperforms or exhibits competitive performance against the uniform averaging scheme (row G).

While modeling predictive uncertainty in clinical applications without a 'real' gold standard is helpful in decision making, miscalibrated uncertainty with overconfident predictions leads to an unreliable outcome. To evaluate the calibration quality of our ensemble annotation aggregation against Bayesian FCNs, we implemented Bayesian epistemic uncertainty using dropout for each base model. Similar to Bayesian SegNet [15], we added five dropout lay-

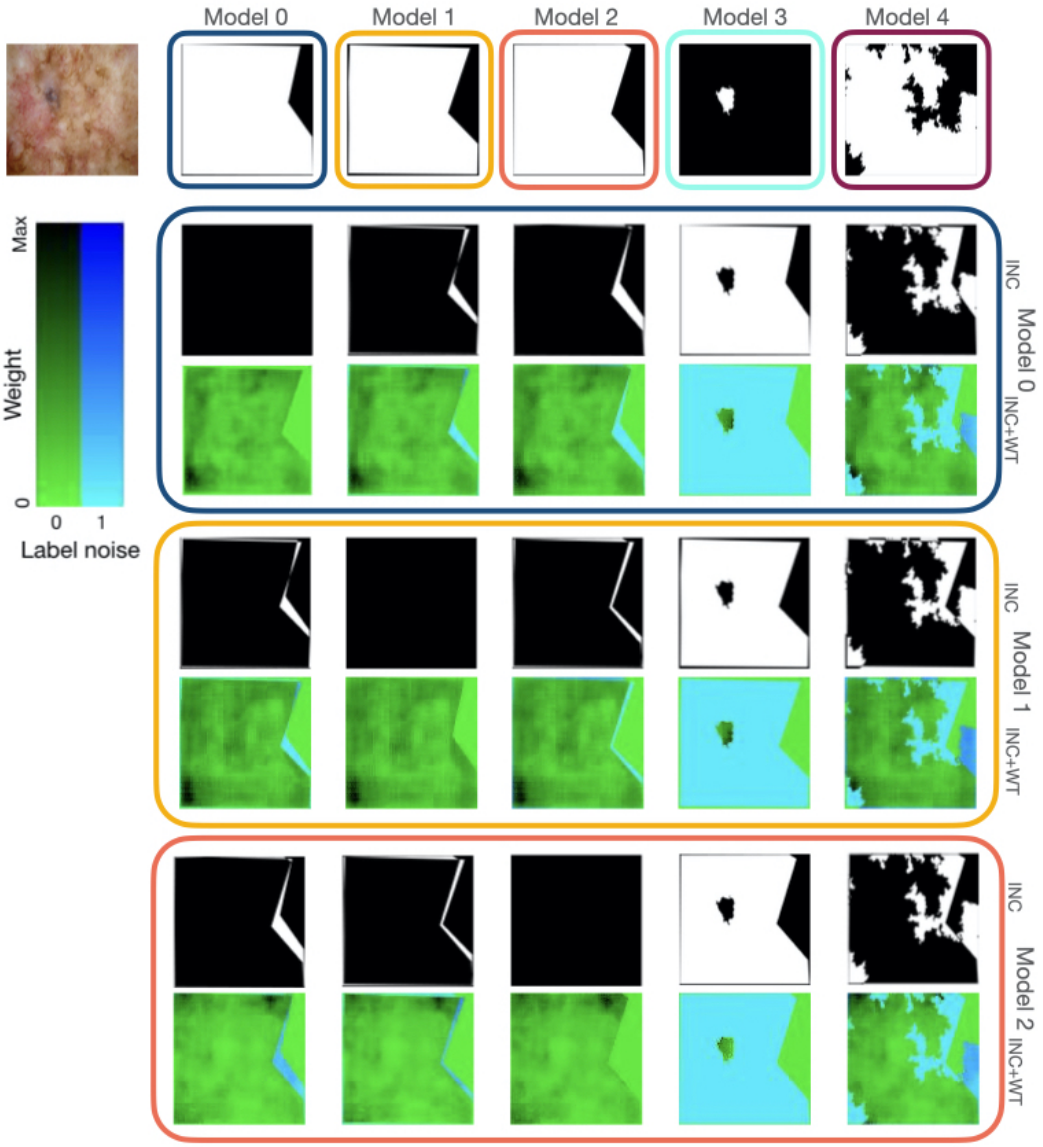

Figure 3: Qualitative evaluation of weighting matrices: (first row) a sample training image and trusted annotations in base models 0 to 4. (second row) inconsistency maps (INC) between the trusted ground truth in Model 0 and other ground truth annotations. (third row) learned weight maps in iteration 100K overlaid over the inconsistency maps (INC+WT). Color-coded boxes indicates the change when the trusted annotations in base models 0, 1 and 2 are different.

ers in the central part of the encoder and the decoder after each convolutional layer. Dropout probability is set to 0.3 and they are kept active at the inference time. Fifteen feed-forwards are executed to perform MC sampling and the output mean is considered as the final segmentation prediction.

To evaluate the quality of the predictive uncertainty, we use two widely used metric in the literature [19, 8]; negative log-likelihood (NLL) and Brier score (Br). Given a segmentation network with sigmoid non-linearity in the output layer, NLL and Br for X_n are calculated as follows:

$$NLL = \frac{-1}{|X_n|} \sum_{q \in X_n} Y_{nq} \log \hat{Y}_{nq} + (1 - Y_{nq}) \log(1 - \hat{Y}_{nq}) \tag{6}$$

$$Br = \frac{1}{|X_n|} \sum_{q \in X_n} [Y_{nq} - \hat{Y}_{nq}]^2 \tag{7}$$

Consistent with prior studies on deep ensembling [19, 25], Table 2 indicates that our annotation aggregation en-

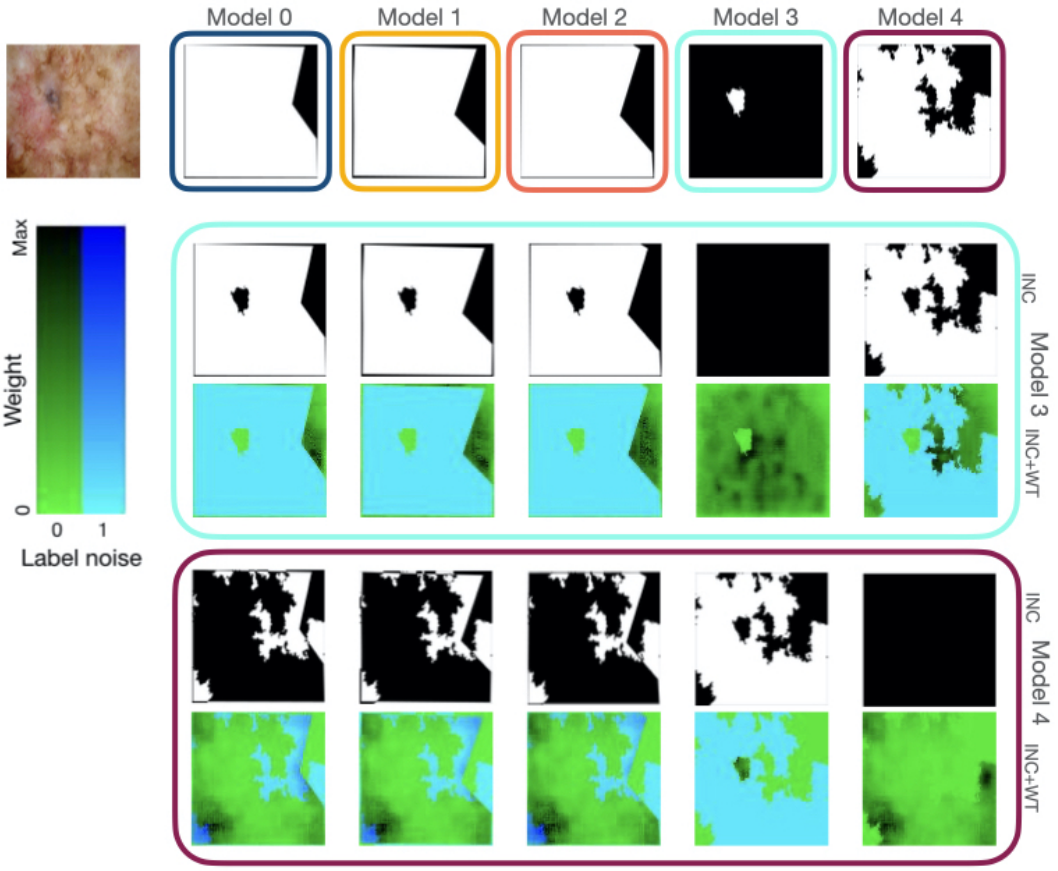

Figure 4: Qualitative evaluation of weighting matrices: (first row) a sample training image and trusted annotations in base models 0 to 4. (second row) inconsistency maps (INC) between the trusted ground truth in Model 3 and other ground truth annotations. (third row) learned weight maps in iteration 100K overlaid over the inconsistency maps (INC+WT). Color-coded boxes indicate the changes when the trusted annotations in base models 3 and 4 are different.

semble with five base models consistently improves the confidence calibration and predictive uncertainty for three datasets in comparison to modeling epistemic uncertainty by MC dropout.

The spatially adaptive weight maps for model i, $\mathcal{W}^i$, are learned to prevent penalizing the pixels whose feature maps are similar to the feature maps of data in $\mathcal{C}^i$ while their gradient direction is not similar to the direction of loss gradient on annotations in $\mathcal{C}^i$. To qualitatively evaluate matrices $\mathcal{W}^i$, in Figures 3 and 4, we overlay the learned weight maps, in training iteration 100K, over the inconsistency maps (absolute differences of ground truth masks). Looking into the color-coded boxes shows how the location of the cyan pixels matches the inconsistency maps (zero or very close to zero weights are assigned to inconsistent annotated pixels), which results in exclusively leveraging the experts knowledge in $\mathcal{C}^i$ when learning θ^i.

4. Conclusion

Approaches to train deep segmentation models do not trivially generalize to datasets with multiple image annotations. We propose an ensemble paradigm to deal with discrepancies in segmentation annotations. A robust-to-annotation-noise learning scheme is utilized to efficiently leverage the multiple experts' opinions toward learning from all available annotations and improve the generalization performance of deep segmentation models. The quality of predictive uncertainty in clinical applications without true gold standards is critical. Our model captures two types of uncertainty, aleatoric uncertainty modeled in the training loss function and epistemic uncertainty modeled in the ensemble framework to improve confidence calibration.

Acknowledgments. We gratefully thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding and the NVIDIA Corporation for the donation of the Titan X GPU used for this research.

References

[1] International Skin Imaging Collaboration: Melanoma Project. https://www.isic-archive.com/. [Online. Accessed December 11, 2020]. 5, 6

[2] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013. 5, 6

[3] Alberto M Biancardi and Anthony P Reeves. TESD: a novel ground truth estimation method. In *Medical Imaging 2009: Computer-Aided Diagnosis*, volume 7260, page 72603V. International Society for Optics and Photonics, Feb. 2009. 2

[4] Abhishek Chaurasia and Eugenio Culurciello. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *IEEE VCIP*, pages 1–4. IEEE, 2017. 5

[5] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 5

[6] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 5

[7] Michael C. Fu, Rafael A. Buerba, William D. Long, Daniel J. Blizzard, Andrew W. Lischuk, Andrew H. Haims, and Jonathan N. Grauer. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions. *The Spine Journal*, 14(10):2442–2448, Oct. 2014. 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 2, 7

[9] Manu Goyal, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. *IEEE Access*, 8:4171–4181, 2020. 2

[10] Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edition, 1991. 1

[11] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *MICCAI*, pages 137–145. Springer, 2019. 2

[12] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016. 1

[13] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *MICCAI*, pages 48–56. Springer, 2019. 2

[14] Joni-Kristian Kamarainen, Lasse Lensu, and Tomi Kauppi. Combining multiple image segmentations by maximizing expert agreement. In *MLMI-MICCAI*, pages 193–200. Springer, 2012. 2

[15] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv:1511.02680*, 2015. 2, 6

[16] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584, 2017. 2, 5

[17] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *NeurIPS*, 31:6965–6975, 2018. 2

[18] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020. 2

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30:6402–6413, 2017. 2, 7

[20] Thomas A. Lampert, Andre Stumpf, and Pierre Gancarski. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *IEEE Transactions on Image Processing*, 25(6):2557–2572, June 2016. 2

[21] Thomas Robin Langerak, U A van der Heide, A N T J Kotte, M A Viergever, M van Vulpen, and J P W Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, Dec. 2010. 2

[22] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019. 2

[23] Xiang Li, Ben Aldridge, Robert Fisher, and Jonathan Rees. Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In *2011 IEEE ISBI: From Nano to Macro*, pages 1438–1441. IEEE, Mar. 2011. 2

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, pages 3431–3440, 2015. 1

[25] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020. 2, 7

[26] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. PH2 - a dermoscopic image database for research and benchmarking. In *IEEE EMBC*, pages 5437–5440. IEEE, 2013. 5, 6

[27] Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. Learning to segment skin lesions from noisy annotations. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 207–215. Springer, 2019. 2, 3

[28] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE CVPR*, pages 1717–1724, 2014. 1

[29] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Handling inter-annotator agreement for automated skin lesion segmentation. *arXiv:1906.02415*, 2019. 2, 5

[30] Vinicius Ribeiro, Sandra Avila, and Eduardo Valle. Less is more: Sample selection and label conditioning improve skin lesion segmentation. In *IEEE CVPR Workshops*, pages 738–739, 2020. 2, 5, 6

[31] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. 4

[32] Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999. 5

[33] Suman Sedai, Bhavna Antony, Dwarikanath Mahapatra, and Rahil Garnavi. Joint segmentation and uncertainty visualization of retinal layers in optical coherence tomography images using Bayesian deep learning. In *Computational Pathology and Ophthalmic Medical Image Analysis*, pages 219–227. Springer, 2018. 2

[34] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE ICCV*, pages 843–852, 2017. 1

[35] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, June 2020. 1

[36] Saeid Asgari Taghanaki, Noirin Duggan, Hillgan Ma, Xinchi Hou, Anna Celler, Francois Benard, and Ghassan Hamarneh. Segmentation-free direct tumor volume and metabolic activity estimation from PET scans. *Computerized Medical Imaging and Graphics*, 63:52–66, Jan. 2018. 2

[37] Hilke Vorwerk, Gabriele Beckmann, Michael Bremer, Maria Degen, Barbara Dietl, Rainer Fietkau, Tammo Gsänger, Robert Michael Hermann, Markus Karl Alfred Herrmann, Ulrike Höller, Michael van Kampen, Wolfgang Körber, Burkhard Maier, Thomas Martin, Michael Metz, Ronald Richter, Birgit Siekmeyer, Martin Steder, Daniela Wagner, Clemens Friedrich Hess, Elisabeth Weiss, and Hans Christiansen. The delineation of target volumes for radiotherapy of lung cancer patients. *Radiotherapy and Oncology*, 91(3):455–460, June 2009. 2

[38] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004. 1, 2

[39] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE CVPR*. IEEE, 2018. 1

[40] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016. 2