2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433781

DEEP BAYESIAN IMAGE SEGMENTATION FOR A MORE ROBUST EJECTION FRACTION ESTIMATION

Mohammad H. Jafari [1 ⋆], *Nathan Van Woudenberg* [1 ⋆],
Christina Luong [1 2 ⋆], *Purang Abolmaesumi* [1 ⋆⋆], *Teresa Tsang* [1 2 ⋆⋆]

[1] University of British Columbia, Vancouver, Canada.
[2] Vancouver General Hospital, Vancouver, Canada.

ABSTRACT

The left ventricular ejection fraction (LVEF) is one of the most commonly measured cardiac parameters in echocardiography (echo). While the clinical guidelines suggest that apical views are used for LVEF assessment, these views are difficult to obtain for less experienced operators. Specifically, in point-of-care imaging, parasternal views are commonly used for rapid assessment of LVEF, since those views can be easier to obtain. However, robust LVEF estimation is challenging due to high variability of echo quality and cardiovascular structures across different patients. In this paper, we formulate a Bayesian deep learning approach for fully automatic LVEF estimation based on segmentation of the left ventricle (LV) in parasternal short-axis papillary muscles (PSAX-PM) level. The proposed approach exploits the LV segmentation uncertainty to improve the robustness of the reported LVEF. The experiments using a dataset of 2,680 patients show that the proposed approach could increase the LVEF estimation's R2 score with a noticeable relative margin of 17.9%, while automatically detecting and discarding around 6% of cases with the highest predictive uncertainty.

***Index Terms*—** Uncertainty Estimation, Bayesian Deep Learning, Image Segmentation, Left Ventricular Ejection Fraction, Echocardiography.

1. INTRODUCTION

Transthoracic echocardiography (echo) is the main imaging technique used to estimate the left ventricular ejection fraction (LVEF). The LVEF measures the percentage of blood pumped outside the left ventricle (LV) during a cardiac cycle and is one of the primary indicators of the cardiovascular conditions [1]. Accurate automated assessment of LVEF is of high clinical interest. Recently, several deep learning models have been proposed for LVEF estimation in apical echo views, such as methods based on LV segmentation [2] and direct video analysis [3]. In previous works from our group, we demonstrated LVEF estimation models based on visual assessment of apical echo views [4, 5], and a mobile platform for real-time biplane apical EF estimation according to Simpsons method of disks [6]. The acquisition of apical echo views can be challenging for the novice sonographers [7], specifically in point-of-care settings where the echo operators are usually less experienced. Recent works such as [8] show feasibility of LVEF assessment in more easily obtainable echo views such as parasternal short-axis (PSAX). The PSAX views are usually used in point-of-care imaging to provide a rapid assessment of the LV function.

Accurate segmentation of LV is an important step for the LVEF estimation. Normally, the ground truth LV masks are only available at two points in the cardiac cycle [9], namely the frames at the end-diastolic (ED) and end-systolic (ES) phases. The lack of training samples over the span of the entire cardiac cycle can result in high model uncertainty when estimating LVEF. Quantifying the predictive uncertainty of deep learning models is an active area of research [10] and is of critical importance in applications such as computer-assisted diagnosis. Model uncertainty, aka epistemic uncertainty, accounts for uncertainty in the model parameters and can be decreased given enough training samples. A model will have higher epistemic uncertainty for input data far away from its training data. Recent work [11] presents a comparison of accuracy and calibration of conventional approaches for epistemic uncertainty estimation, such as ensemble algorithms, stochastic variational Bayesian neural networks, and Bayesian approximation by Monte Carlo (MC) dropout [10].

In this paper, we propose a deep Bayesian framework for accurate LVEF estimation in PSAX papillary muscles (PM) echo view. A U-Net [12] with MC dropout is used to segment the LV at ED and ES frames. The variations in size and appearance of the MC dropout predictions at the detected ED and ES phases of the heart are used in the proposed pipeline to obtain a measure of uncertainty in LVEF estimation. The methodology is evaluated with a dataset of echo cine series from 2,680 patients with available ground truth LVEF labels, showing that the proposed algorithm can noticeably improve the accuracy of LVEF assessment from the PSAX-PM view.

⋆ Joint first authors; the order is selected randomly.

⋆⋆ Joint senior authors.

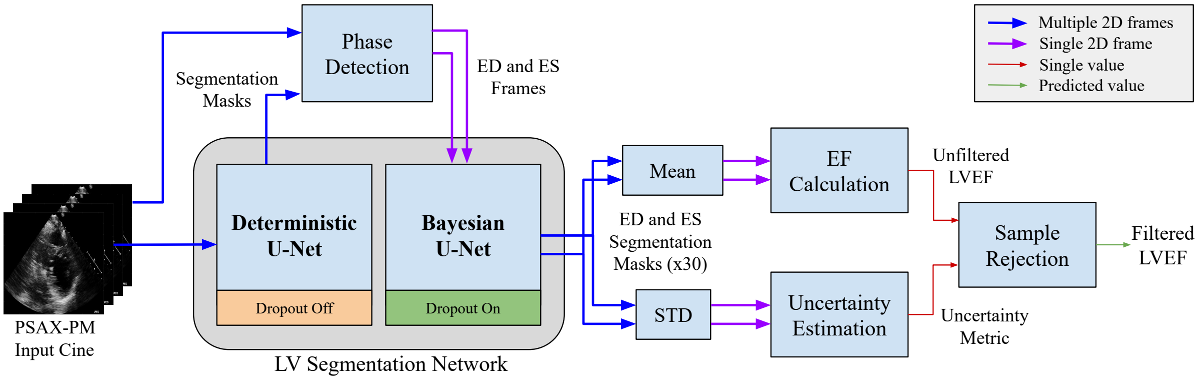


Fig. 1: Block diagram of the proposed framework for LVEF estimation. A deterministic U-Net (dropout off) is used to detect ED and ES frames based on the LV area. The ED and ES frames are fed into a Bayesian U-Net (dropout on) to estimate both the LVEF and the model uncertainty. Statistical sample rejection is used in the pipeline to filter out the predictions with high uncertainty.

2. MATERIALS AND METHOD

2.1. Dataset

The echo data used in this research is collected from Picture Archiving and Communication System at Vancouver General Hospital (VGH), Canada, following approval from Medical Research Ethics Board in coordination with the privacy office. Data from a total of 2,680 patients with PSAX-PM echo cines are used for the experiments. The echo cines are acquired by a variety of ultrasound devices, namely iE33, Vivid i/7/9/95, Sonosite, and Sequoia. For the number of 554 echo cines, the ground truth LV mask is delineated by an experienced level III echocardiographer. For the remaining dataset (2,126 PSAX-PM echo cines), while the LV segmentation is not available, the ground truth LVEF values are acquired from the patients' archived information. These LVEF values were originally measured using biplane Simpsons method by expert cardiologists using apical echo views.

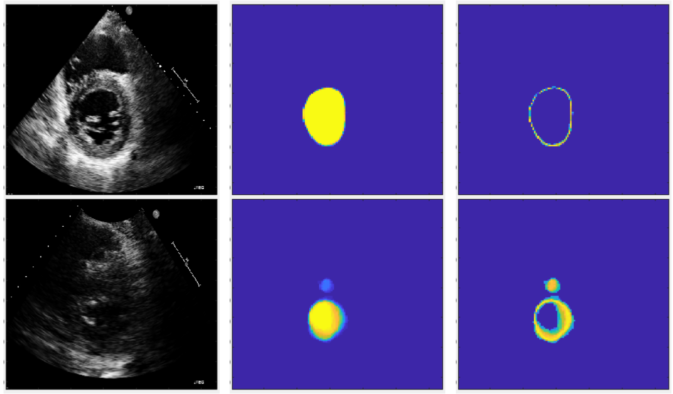
2.2. Methodology

An overview of the proposed framework is shown in Fig. 1. The components and the steps of the proposed approach are explained in details in the rest of this section.

Bayesian U-Net. We train a U-Net [12] model for LV segmentation in the PSAX-PM view. The network architecture and training details are explained in Section 3. Here, we discuss the methodology used to measure the segmentation uncertainty. In Bayesian neural networks (BNN), the goal is to learn a distribution over the network's parameters, as opposed to the deterministic set of parameter values learned in conventional neural networks. Subsequently, BNNs produce a distribution over outputs for a given input, which provides an estimation for the model uncertainty. In practice, Bayesian inference for BNNs is not computationally tractable. Thus, various approximations of BNNs are developed that are feasible to train [11, 10]. In this work, we adapt MC dropout [10] to U-Net (called Bayesian U-Net). Dropout is conventionally used in deep neural networks only at training time as a regularization technique; each node is ignored (dropped out) with probability of p. Gal and Ghahramani [13] show that dropout layers give a Bernoulli-approximate variational inference of BNNs. In dropout variational inference, referred to as MC dropout, the dropout is also performed in the test time to sample from the approximate posterior. Dropout in test time results in stochastic forward passes; each test run is obtained through a random subset of the network. The deviations of the predicted LV segmentation masks across different MC dropout test runs is interpreted as a measure of model uncertainty.

Phase Detection. For a given echo cine, the ED and ES frames are selected as follows. Firstly, the cine is run through the deterministic U-Net (dropout off), producing LV segmentation masks for each frame. These masks are then sorted by increasing segmentation area. If the segmentation masks were perfectly accurate, we would simply select the mask with the largest area to be the ED frame, and the smallest to be ES. However, as a form of outlier rejection, we instead select the ED frame as the 90^{th} percentile of segmentation areas, and the ES frame as the 10^{th}.

LVEF Calculation. The ED and ES frames are run through the Bayesian U-Net (dropout on) N times, producing N slightly different segmentation masks. We then calculate the mean and standard deviation of each pixel across the N segmentation masks for both key-frames. The standard deviation of predictions is used to produce the uncertainty metric as described in the following subsection, while the average of



Input PSAX-PM Echo Frame

Predicted LV Segmentation Mask

Corresponding Uncertainty Map

Fig. 2: LV segmentation and uncertainty maps for sample low uncertainty (upper row) and high uncertainty (bottom row) predictions.

predictions (middle column of Fig. 2) is used to calculate the LVEF estimation as:

$$LVEF = \frac{V_{ED} - V_{ES}}{V_{ED}} \approx \frac{A_{ED} - A_{ES}}{A_{ED}}, \qquad (1)$$

where V_{ED} and V_{ES} are the ED and ES LV volumes, and A_{ED} and A_{ES} are the areas of the segmentation masks from the ED and ES frames. While LV volume is not well approximated by segmentation area directly, the ratio above approximates LVEF quite well, as shown in Section 3. More complex modelling of the LV were investigated, but no approach showed significant improvement over Equation 1. More accurate models could be developed given additional information from an orthogonal view, such as the length of the LV estimated from a parasternal long-axis (PLAX) or apical (AP2 or AP4) view of the same patient [14]. However in this work, we specifically target single plane LVEF estimation from the PSAX-PM view due to its relative ease of acquisition, particularly for novice echo operators.

LVEF Uncertainty Estimation. The standard deviations of the predicted ED and ES masks are reconstructed into 2D 'uncertainty maps' as seen in the rightmost column of Fig. 2. These maps typically contain 'rings' appearing around the edges of the LV, with thicker rings resulting from more variation across the N segmentation masks output by the Bayesian network, and thus higher uncertainty. We then calculate the 'normalized ring thickness' as our final uncertainty metric. This is done by first finding the center of mass (COM) of the uncertainty map, transforming the surrounding pixels from Cartesian to polar coordinates, and finding the average ring thickness and average ring radius. We estimate these values by considering the radial lines 360° around the COM. The normalized ring thickness is then calculated as:

$$t_{norm} = \frac{\text{mean(ring thickness)}}{\text{mean(ring radius)}} = \frac{\sum_{i=1}^{360} \sum l_i}{\sum_{i=1}^{360} \text{argmax}(l_i)}, \qquad (2)$$

where l_i represents the pixel values of the i^{th} line radiating from the COM. We normalize the pixel sum by the distance to the maximum (argmax(l_i)) in order to account for differing LV sizes. This normalized ring thickness is calculated for both ED and ES frames, and the maximum value is used as the final uncertainty metric. Typically, the ES frames have higher estimated uncertainty, which could be attributed to the increased variability of LV shape during the ES cardiac phase.

Sample Rejection. After calculating both the LVEF estimate and its associated uncertainty metric, we need to decide whether or not to reject the sample, *i.e.,* not to report the predicted LVEF due to high uncertainty. This is done by applying a threshold; all LVEF estimates with normalized ring thicknesses above a certain threshold T are rejected. We set aside a part of unseen training samples as validation set (called threshold-validation set) in order to select the optimal threshold value, T^*. We calculate the correlation between our predicted LVEF and the ground truth labels on the threshold-validation set for a variety of threshold values. The leftmost chart in Fig. 3 shows the R2 scores at different rejection rates between 0-50%. There is a trade-off here as selecting a stricter threshold will result in more accurate LVEF predictions, but will reject more of the cases. We therefore use the elbow method to select a T^* that gives us the largest improvement in R2 score, while also rejecting the least amount of predicted results. This inflection point occurs when we reject approximately 6% of cases, which corresponds to $T^* = 18.14$. Note that this threshold is a tunable parameter; the end-user of this system might, for example, want to select a stricter threshold value in order to produce more certain results, or vice-versa.

3. EXPERIMENTS

Implementation Details. The details of the echo dataset used in this research are explained in Section 2.1. The U-Net model is trained using 554 PSAX-PM echo cines with available ground truth LV masks; 10% of the training data is initially set aside as a validation set (model-validation set) to optimize the hyper parameters of the segmentation model. Also, another 425 unseen PSAX-PM cines with available LVEF ground truth are used as the threshold-validation set to obtain the sample rejection rate T^* as explained in Section 2.2. Finally, 1,701 unseen echo cines with available ground truth LVEF are used as the test set to evaluate the model performance. There is no overlap between the patients among train/threshold-validation/test sets. The trained U-Net for LV segmentation has three down-sampling and

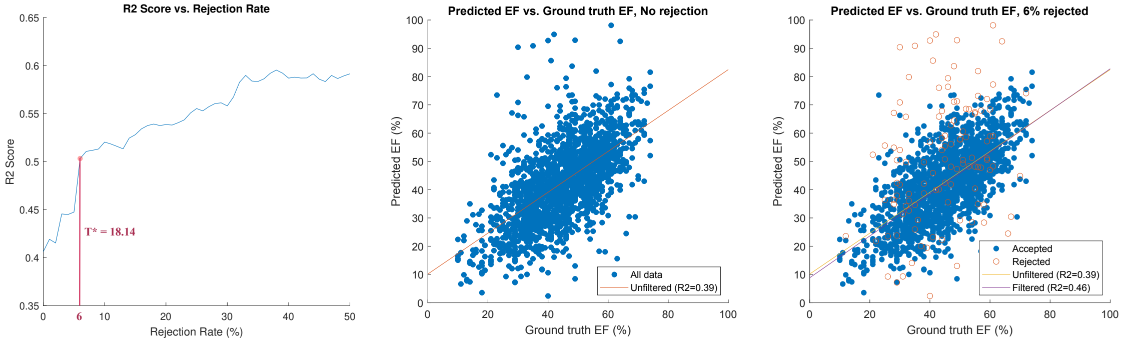


Fig. 3: Left: R2 scores of predicted vs. ground truth LVEF at rejection rates between 0-50% (on threshold-validation set). Using the elbow method, T^* is selected such that $\approx 6\%$ of data is rejected. Middle: Unfiltered predicted vs. ground truth LVEF (test set). Right: Filtered predicted vs. ground truth LVEF (test set), highlighting the rejected predictions whose uncertainty-metric was above threshold T^*. While there are some cases that are being falsely rejected, almost all of the outlying cases (especially those with over-predicted LVEF) are being correctly filtered out using the proposed uncertainty based filtering method.

three reverse up-sampling blocks. Following [12], each block includes two convolutional layers with kernel size 3×3 and a max pooling layer with pooling size 2×2. Each convolutional layer is followed by batch normalization (momentum $= 0.8$), ReLU, and dropout ($p = 0.2$), except the output layer that has a kernel size 1×1 with sigmoid activation. The number of MC samples in test time is $N = 30$. The first convolutional has 32 filters doubled after each down-sampling step. The input images are resized to 128×128. The soft Dice loss is used for training with an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is initially set to $1e - 3$.

Evaluations. In order to evaluate the performance of the proposed method, we compare the calculated LVEF values to the ground truth, for a variety of rejection threshold values. Table 1 shows the R2 score, mean absolute error, and standard deviation of absolute error on the test data. We see clear improvements across all performance metrics as we reject more data, suggesting that the uncertainty value we calculate in Section 2.2 is indeed a good predictor of low-accuracy cases. Table 1 also contains the results at T^*, the optimal threshold calculated in 2.2, and shows that the largest jump in performance occurs when rejecting a small percentage of cases with the highest uncertainty. We visualize which samples are being rejected at T^* by plotting our predicted LVEF values against the ground truth labels, as shown in Fig. 3.

4. DISCUSSION AND CONCLUSIONS

In this work, we proposed a deep Bayesian framework for LVEF estimation in PSAX-PM echo view, investigating the utilization of model uncertainty in a pipeline to improve robustness of the predictions. We proposed a formulation to use prediction uncertainty to automatically not report a prediction for cases with high uncertainty, improving the R2 score with noticeable relative margin of 17.9% (t-test on LVEF error; p-value < 0.05) by discarding only 6.6% low confident cases (see Table 1). Specifically, the proposed method can significantly reject over-predicted LVEF values (see Fig. 3), which are critical in clinical diagnosis.

Table 1: Evaluation of LVEF prediction on the test set, at various thresholds T. The threshold values are obtained based on the rejection rates applied to the threshold-validation set (two left-most columns). T^* is selected using the elbow rule (trade-off between R2 score and rejection percentage) on the threshold-validation set.

Rejection rate-val (%)	T	Rejection rate-test (%)	R2 score	Δ(R2) (%)	mean error (%)	STD of error (%)
0 (raw)	n/a	0 (raw)	0.387	n/a	8.91	7.70
6 (T*)	**18.14**	**6.58**	**0.457**	**+17.94**	**8.34**	**6.69**
10	13.64	9.35	0.468	+20.86	8.26	6.58
20	9.49	16.81	0.491	+26.85	8.04	6.35
30	6.88	20.45	0.521	+34.40	7.86	6.17
40	6.03	39.27	0.540	+39.43	7.79	6.01
50	5.32	49.79	0.578	+49.17	7.72	5.92

The LVEF values predicted by the proposed system, even when filtered at low rejection rates, show surprisingly high correlation with the ground truth, especially when considering that they are calculated based on single-plane PSAX-PM cines, and are being compared against the echocardiographic gold standard for LVEF assessment: biplane Simpsons method using LV segmentations from apical view cines. While apical views of the heart may offer increased accuracy for LVEF assessment, they are usually more difficult to obtain than those that can be acquired from the parasternal window, especially for less-experienced sonographers and operators practicing in point-of-care environments. Future work includes extension of the proposed framework to other echo views such as parasternal long axis (PLAX), analysis of the effect of US machines and pathology cases in clinical settings, as well as investigation of other sources of predictive uncertainty, such as heteroscedastic aleatoric uncertainty.

5. REFERENCES

[1] Roberto M Lang, Luigi P Badano, et al., "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] Jeffrey Zhang, Sravani Gajjala, et al., "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[3] David Ouyang, Bryan He, et al., "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[4] Delaram Behnami, Zhibin Liao, et al., "Dual-view joint estimation of left ventricular ejection fraction with uncertainty modelling in echocardiograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2019, pp. 696–704.

[5] Mahdi Kazemi, Christina Luong, et al., "A deep Bayesian video analysis framework: Towards a more robust estimation of ejection fraction," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2020, pp. 582–590.

[6] Mohammad H Jafari, Hany Girgis, et al., "Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training," *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, vol. 14, no. 6, pp. 1027–1037, 2019.

[7] Carol Mitchell, Peter S Rahko, et al., "Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography," *Journal of the American Society of Echocardiography*, vol. 32, no. 1, pp. 1–64, 2019.

[8] Ilaria Russo, Edoardo Micotti, et al., "A novel echocardiographic method closely agrees with cardiac magnetic resonance in the assessment of left ventricular function in infarcted mice," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[9] Mohammad H. Jafari, Hany Girgis, et al., "Semi-supervised learning for cardiac left ventricle segmentation using conditional deep generative models as prior," in *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2019, pp. 649–652.

[10] Alex Kendall and Yarin Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.

[11] Yaniv Ovadia, Emily Fertig, et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 13991–14002.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[13] Yarin Gal and Zoubin Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[14] Edward Folland, AF Parisi, et al., "Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. a comparison of cineangiographic and radionuclide techniques," *Circulation*, vol. 60, no. 4, pp. 760–766, 1979.

6. COMPLIANCE WITH ETHICAL STANDARDS

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The echo data used in this research is collected from Picture Archiving and Communication System at Vancouver General Hospital (VGH), with ethics approval from the Clinical Medical Research Ethics Board, in consultation with the Information Privacy Office.

7. ACKNOWLEDGEMENTS

The authors declare that they have no conflict of interest. This work is supported in part by the Natural Sciences and Engineering Research and Council of Canada (NSERC) and in part by the Canadian Institutes of Health Research (CIHR).