SIAM/ASA J. Uncertainty Quantification
Vol. 13, No. 1, pp. 90–113

Uncertainty Quantification in Machine Learning Based Segmentation: A Post-Hoc Approach for Left Ventricle Volume Estimation in MRI[*]

Felix Terhag[†][‡], Philipp Knechtges[†], Achim Basermann[†], and Raúl Tempone[‡][§]

Abstract. Recent studies have confirmed cardiovascular diseases remain responsible for the highest mortality rate among noncommunicable diseases. The accurate left ventricular (LV) volume estimation is critical for valid diagnosis and management of various cardiovascular conditions, but poses a significant challenge due to inherent uncertainties associated with the segmentation algorithms in magnetic resonance imaging. Recent machine learning advancements, particularly U-Net-like convolutional networks, have facilitated automated segmentation for medical images, but struggles under certain pathologies and/or different scanner vendors and imaging protocols. This study proposes a novel methodology for post-hoc uncertainty estimation in the LV volume prediction using Itô stochastic differential equations to model pathwise behavior for the prediction error. The model describes the area of the left ventricle along the heart's long axis. The method is agnostic to the underlying segmentation algorithm, facilitating its use with various existing and future segmentation technologies. The proposed approach provides a mechanism for quantifying uncertainty, enabling medical professionals to intervene for unreliable predictions. This is of utmost importance in critical applications such as medical diagnosis, where prediction accuracy and reliability can directly impact patient outcomes. The method is also robust to dataset changes, enabling application for medical centers with limited access to labeled data. Our findings highlight the proposed uncertainty estimation methodology's potential to enhance automated segmentation robustness and generalizability, paving the way for more reliable and accurate LV volume estimation in clinical settings as well as opening new avenues for uncertainty quantification in biomedical image segmentation, providing promising directions for future research.

Key words. machine learning, uncertainty quantification, cardiovascular MRI, Itô stochastic differential equations, U-Net, neural networks, left ventricle volume estimation, biomedical image segmentation, convolutional neural networks

MSC codes. 68T07, 62P10, 92C55, 68T05, 65C20, 62M45

DOI. 10.1137/23M161433X

1. Introduction.

Recent studies have confirmed cardiovascular diseases remain responsible for the highest mortality rate among noncommunicable diseases [16]. The left ventricle

[*] Received by the editors November 7, 2023; accepted for publication (in revised form) August 29, 2024; published electronically January 8, 2025.
https://doi.org/10.1137/23M161433X

Funding: This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) and received funding from the Helmholtz Association of German Research Centres. This work was partially supported by the Alexander von Humboldt Foundation and the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under award OSR-2019-CRG8-4033.

[†] Department for High-Performance Computing, Institute of Software Technology, German Aerospace Center (DLR), Cologne, Germany (felix.terhag@dlr.de, philipp.knechtges@dlr.de, achim.basermann@dlr.de).

[‡] Chair of Mathematics for Uncertainty Quantification, RWTH Aachen University, Aachen, Germany (tempone@uq.rwth-aachen.de).

[§] Computer, Electrical and Mathematical Sciences and Engineering Division, KAUST, Thuwal, Saudi Arabia.

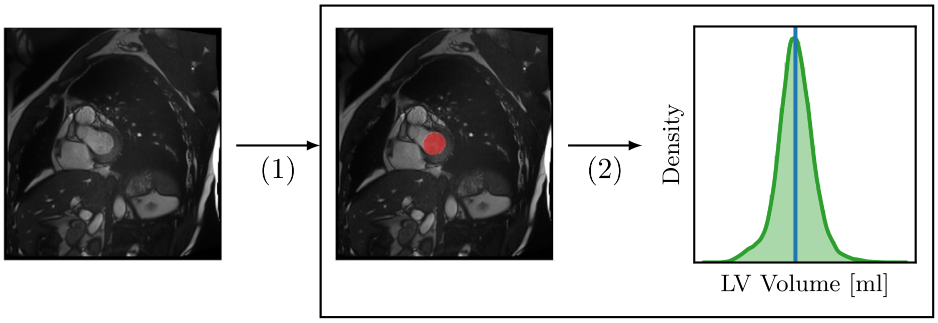

Figure 1. *Proposed approach to model left ventricle (LV) volume from short-axis cardiac MRI (the example image is from the ACDC dataset* [2]*).* (1) *An arbitrary algorithm, segments the left ventricle.* (2) *Our method creates a distributional volume prediction from deterministic segmentations.*

(LV) volume is an essential and well-studied parameter to help measure or assess many cardiovascular related conditions, including hypertension [14], cardiovascular mortality [21], and heart failure [14]. The gold standard for LV volume measurement is manually segmenting the LV by experienced investigators in short-axis view magnetic resonance images (MRIs) [15], which is both a tedious and costly procedure. Short-axis MRIs consist of several slices along the short-axis of the heart.

Recent machine learning developments, notably U-Net-like convolutional networks [18], have enabled better automatic segmentation for medical images, achieving almost expert-level segmentation for short-axis cardio MRIs in several public challenges [2, 5, 9]. However, the prediction quality tends to fall off for some pathologies [2] and has relatively poor generalizability across the various MRI scanner vendors and/or imaging protocols [5]. Accurate and automated machine learning segmentation provides beneficial results; however, particularly for critical applications, i.e., medical diagnosis, accurate uncertainty descriptors are essential in assisting medical experts to infer reasonable outcomes from potentially unreliable predictions.

Therefore, this study proposes a post-hoc uncertainty estimation method for LV volume prediction that is agnostic to the underlying segmentation algorithm (see Figure 1). Our method takes inspiration from [4], where a similar approach with a different base model was chosen to quantify the uncertainty in wind power forecasting. The proposed approach can be used with existing and future segmentation algorithms, and requires only eight parameters to be fit on datasets. Thus, the proposed method can be applied for smaller datasets rather than retraining a neural network with several million parameters.

Automatic LV segmentation is a widely studied field, not least due to the segmentation challenges [2] and [5]. Several segmentation algorithms have been developed, most employing U-Net neural network architecture [9, 23], e.g., the ten best performing methods from the 2021 M&Ms Challenge all relied on this architecture [5]. Although several attempts have been proposed to improve the neural network prediction uncertainty, they usually involve changes to the architecture and/or training routine [11, 7]. Thus, one cannot generally apply these methods onto current state-of-the-art methods and they would usually require at least as many training examples as retraining a neural network from scratch [11, 7, 3]. Retraining neural networks is problematic for real-word applications since medical centers have different MRI machines and/or postprocessing techniques, making it difficult to generalize over research centers [5]. Thus, these techniques are not practical for smaller medical centers, which cannot obtain a sufficient number of labeled MRIs. The proposed approach to provide uncertainty quantification on a derived parameter of medical importance from segmented MRIs has, to our knowledge, not been previously reported.

The remainder of this paper is organized as follows: Section 2 introduces the LV volume prediction problem, and we subsequently describe the proposed modeling approach in section 3. Sections 3.1, 3.2, and 3.3 differentiate between modeling inner and outer slices and how to combine these, respectively. Section 4 describes fitting model parameters using maximum likelihood, with experimental results presented in section 5. Finally, section 6 summarizes and concludes the paper.

2. Volume prediction from cardio MRIs.

Short-axis cardiac MRIs comprise several parallel slices recorded perpendicular to the long axis, i.e., the plane intersecting the base of the heart and the apex. In particular, these slices show the LV cross-sectional area. Typical MRI spatial resolution is much higher than resolution along the slices, e.g., spatial resolution $\approx 0.85-1.32$ mm, whereas slice thickness can be as much as ten times higher. We are interested in LV volumes for end-diastolic and end-systolic phases to calculate essential parameters for cardiac diagnosis [5].

This study employs the ACDC [2] and M&Ms datasets [5], containing short-axis view cardiac MRIs, with manually annotated end-systolic and end-diastolic MRIs for 100 and 150 patients, respectively. Both datasets also include patients with numerous pathologies. Patients in the ACDC datasets are evenly distributed over five characteristics and pathologies classes, e.g., patients with normal cardiac anatomy and function, patients with systolic heart failure and infarction, and patients with abnormal right ventricle. In contrast, the M&Ms dataset includes seven classes for the most frequent pathologies, with a separate class for healthy patients and one for other pathologies which are not represented in the most frequent pathologies. The M&Ms dataset contains MRIs from different medical centers with different MRI machine vendors and imaging protocols, whereas the ACDC dataset includes MRIs from a single medical center with the same imaging protocol. Thus the M&Ms dataset provides not only higher pathologic diversity but also MRI vendor and imaging protocol diversity.[1]

Current state-of-the-art automatic segmentation methods for MRIs are convolutional neural networks with U-Net architecture [2, 5]. The U-Net architecture was introduced by Shelhamer, Long, and Darrell in [18]. Following the characteristic U-Net approach, the first half condenses information using convolutional networks and subsequently by pooling layers, reducing spatial information while typically increasing the number of filters. The second half comprises upsampling deconvolutions, such that the last level produces images with the same resolution as the input image. Downsampling and upsampling paths are connected with skip connections, where feature maps from the downsampling path are passed to the corresponding step in the upsampling path. Isensee et al. [9] proposed the nnU-Net framework for automatic architecture search specifically for medical segmentation tasks, and this method won both ACDC and M&Ms segmentation challenges. Therefore, we employ a pretrained nnU-Net version, trained on the M&Ms dataset since this is the richer dataset, and use the ACDC dataset

[1] More detailed descriptions for the ACDC and M&Ms datasets can be found in [2, 5].

for evaluation; hence we test the method on many unseen examples. This also enables realistic applications, where practitioners were typically limited to data from their research center and use a pretrained net that was trained on a very diverse dataset. The nnU-Net architecture [9] commonly employs an ensemble of five neural networks for inference. For simplicity, we take a single net since this only requires approximately one-fifth of the compute time. The pretrained net can predict four classes: *left ventricle, right ventricle, myocard,* and *background,* i.e., the prediction for every voxel i, j contains four values. We focused on the LV volume since this is the most important medical parameter [14], but the method could also be applied to the other classes.

Hence, the only information we get from the outputs of the neural net is one value $o_{i,j}$ per voxel. To build a model directly from the output of the neural network, one could consider each output $o_{i,j}$ as the predicted probability that voxel (i, j) belongs to the LV. To mimic this, the volume can be modeled as a sum of independent Bernoulli random variables $O_{i,j}$ with probability $o_{i,j}$ describing whether or not voxel (i, j) belongs to the LV. The predicted volume for one MRI slice Vol_{output} can be described as

$$Vol_{output} = \nu \cdot \sum_i \sum_j O_{i,j}, \tag{2.1}$$

where ν is the volume of one voxel. The expected value is the sum of the probabilities multiplied by the the volume of one voxel: $\mathbb{E}[Vol_{output}] = \nu \sum_i \sum_j o_{i,j}$. This approach reveals the difficulties of the neural network to accurately predict the uncertainties. To see this, we compare the standard deviation from (2.1) with observed error from ground-truth labels, as shown in Figure 2. To obtain the ground-truth labels, experts manually assigned each voxel

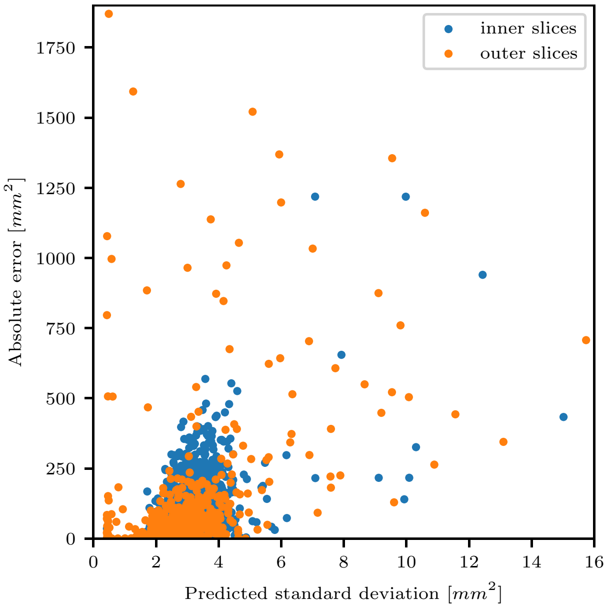

Figure 2. *Left ventricle area absolute error and predicted standard deviation (from (2.1)); outer and inner slices are shown as orange and blue, respectively.*

to the corresponding class. Not only is the observed error several magnitudes larger than the modeled variation, but there is no apparent correlation between higher modeled variation and higher absolute error. Of particular concern is that some examples exhibit very low predicted standard deviation but high observed error. For example, the highest absolute error slice has one of the smallest modeled standard deviations. Thus, the outputs of the neural network cannot accurately capture the underlying uncertainties, consistent with previous studies showing that deep neural networks for classification [8] and segmentation problems [17] are overly confident. Therefore, a different modeling approach is required.

3. Modeling the volume under uncertainty. Section 2 establishes that treating neural network outputs as predicted probabilities results in a poor model. Therefore, we only use the expected value of (2.1) as a point estimate for the volume prediction for each slice. Thus, if MRI $j \in \{1, \dots, M\}$ contains $N_j + 1$ slices, we obtain $N_j + 1$ deterministic point predictions $p_0^{(j)}, \dots, p_{N_j}^{(j)}$, with M being the number of hearts in the dataset.

Even human experts find slices through the apical or basal heart sections, which we call outer slices, to be the hardest to segment. In those slices, it is often unclear whether a voxel belongs to the LV or adjacent tissue, and this is echoed in neural network predictions. Although outer slices only comprise 25% of all slices, they account for 27 of 31 slices with absolute error > 600 mm^2. We define the *outer slices* as two MRI slices where neighboring slices comprise one zero and one nonzero prediction. Since none of the predictions is effectively zero, we define zero predictions where the area is smaller than a threshold ϵ. Figure 3 shows the deterministic predictions for each slice. We see that $\epsilon = 10$ mm^2 is a reasonable threshold to separate very low predictions from the majority of predictions. We then preprocess the data such that all

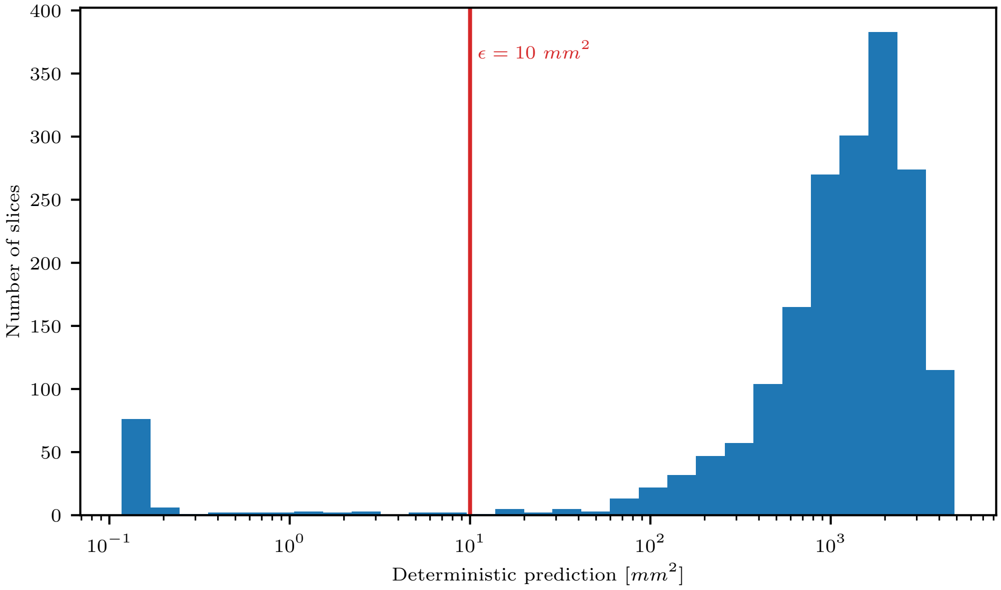

Figure 3. *Deterministic predictions of the left ventricle area for all slices in the dataset. The chosen threshold for zero predictions is shown in red. This is necessary as the neural network never predicts exactly zero.*

predictions start and end with a zero prediction and the next prediction is nonzero, which requires appending zeros or truncating consecutive zero predictions. Truncating consecutive zero predictions does not lose information, since the corresponding labeled data also contains no LV volume, hence $p_0^{(j)} = p_{N_j}^{(j)} = 0$, where $p_1^{(j)}, \ldots, p_{N_j-1}^{(j)} > 0$ for all MRIs $j \in \{1, \ldots, M\}$.

3.1. Modeling the inner slices.

Section 2 establishes that the error behaves fundamentally differently for outer and inner heart slices. Although inner slices generally exhibit smaller deviation from the prediction, outer slices exhibit more categorical error, such that the full predicted area can be assigned to the wrong class. Therefore, we model the inner and outer sections independently. Excluding the first and last slices, we model the inner slices as follows: Neighboring slices for any given heart have an inherent dependency due to locality, which we want to reflect by constructing a stochastic differential equation (SDE) for the slice area X_t over a fake time t, with t ranging over the depth of the heart and X_t being measured at integer values of t. More specifically, we model the area X_t over the depth of the heart t by the stochastic process $X = \{X_t, t \in [0,T]\}$,

$$\begin{cases} dX_t = a(X_t, p_t, \dot{p}_t, \boldsymbol{\theta})dt + b(X_t, p_t, \dot{p}_t, \boldsymbol{\theta})dW_t, \\ t \in [0,T], X_0 > 0, \end{cases} \tag{3.1}$$

where $a(\cdot, p_t, \dot{p}_t, \boldsymbol{\theta}) : \mathbb{R}^+ \to \mathbb{R}$ is the drift function, $b(\cdot, p_t, \dot{p}_t, \boldsymbol{\theta}) : \mathbb{R}^+ \to \mathbb{R}^+$ is the diffusion function, $\boldsymbol{\theta}$ is a parameter vector, $(p_t)_{t\in[0,T]}$ is a depth-dependent $\mathbb{R}^+$-valued deterministic function with time derivative $(\dot{p}_t)_{t\in[0,T]}$, and $\{W_t, t \in [0,T]\}$ is a standard real-valued Wiener process. Thus, $(p_t)_{t\in[0,T]}$ is the linear interpolation of the deterministic point predictions omitting superscript (j). Following [4], we choose the drift function,

$$a(X_t, p_t, \dot{p}_t, \boldsymbol{\theta}) = \dot{p}_t - \theta_t(X_t - p_t), \tag{3.2}$$

to make the process X_t bias-free with respect to the predictions p_i. More precisely, the process will satisfy $\mathbb{E}[X_t] = p_t$ if $\mathbb{E}[X_0] = p_0$ holds initially. We also want the process to be mean-reverting at least with rate $\theta_0 > 0$, hence $\theta_t > \theta_0$.

The LV area over the depth of the heart should not become negative, i.e., the state space for X_t is $\mathbb{R}^+$. Let $\boldsymbol{\theta} = (\theta_0, \alpha)$ and $\alpha > 0$ be parameters controlling the path variability. Then the state-dependent diffusion

$$b(X_t, \boldsymbol{\theta}) = \sqrt{2\alpha\theta_0 X_t}, \tag{3.3}$$

in combination with

$$\theta_t \geq \frac{\alpha\theta_0 - \dot{p}_t}{p_t},$$

ensures that $X_t \in \mathbb{R}^+$ almost surely. Hence we choose

$$\theta_t = \max\left(\theta_0, \frac{\alpha\theta_0 - \dot{p}_t}{p_t}\right). \tag{3.4}$$

To give proof of the last claim, we will prove the following theorem.

Theorem 3.1. *Assume a probability space rich enough to accommodate Brownian motion W_t with $t \in [0,T]$ and a thereof independent random variable X_0 with $X_0 > 0$ almost surely; function $p : [0,T] \to \mathbb{R}^{>0}$ with piecewise continuous derivative; and $\alpha, \theta_0 > 0$ and a piecewise continuous function $\theta : [0,T] \to \mathbb{R}^{>0}$, with*

$$\theta_t \geq \max\left(0, \frac{\alpha\theta_0 - \dot{p}_t}{p_t}\right). \tag{3.5}$$

Then the SDE

$$dX_t = (\dot{p}_t - \theta_t(X_t - p_t))dt + \sqrt{2\alpha\theta_0 X_t}\, dW_t \tag{3.6}$$

has a unique strong solution, and $0 < X_t < \infty$ holds for all $t \in [0,T]$ almost surely.

Proof. Without loss of generality, assume $\sqrt{2\alpha\theta_0 X_t}$ to be padded by zeros for negative values of X_t. Then by potentially limiting us to the segments of $[0,T]$ where $\dot{p}_t$ and θ_t are continuous, applying [19, p. 59] successively proves the existence of a strong solution with $X_t < \infty$ for each of these segments, and the Yamada–Watanabe theorem (see [10, p. 291], [22]) confirms the uniqueness of this solution. Thus, we only need to show that $X_t > 0$ almost surely. We follow the McKean argument as given in [1, p. 23] and [13, Lemma 4.2]. At first, we define a stopping time $\tau_0(\omega) = \inf\{t | X_t(\omega) = 0\}$ with $\tau_0(\omega) = \infty$ for $\{t | X_t(\omega) = 0\} = \emptyset$. Then, using Itô's formula for $f(x) = \log(x)$, it follows that

$$X_t = X_0 \exp\left(\int_0^t \frac{\dot{p}_s + \theta_s p_s - \alpha\theta_0}{X_s}\, ds - \int_0^t \theta_s\, ds + \int_0^t \sqrt{\frac{2\alpha\theta_0}{X_s}}\, dW_s\right)$$

for $t \in [0, \tau_0)$, and using (3.5) yields

$$X_t \geq X_0 \exp\left(-\int_0^t \theta_s\, ds + \int_0^t \sqrt{\frac{2\alpha\theta_0}{X_s}}\, dW_s\right).$$

Thus, the set $\{\omega \mid \tau_0(\omega) \neq \infty\}$ is a subset of

$$A := \left\{\omega \;\middle|\; \lim_{t \to T \wedge \tau_0(\omega)} \int_0^t \sqrt{\frac{2\alpha\theta_0}{X_s(\omega)}}\, dW_s(\omega) = -\infty\right\}.$$

However, $M_t := \int_0^t \sqrt{\frac{2\alpha\theta_0}{X_s}}\, dW_s$ is a continuous local martingale on $[0, \tau_0)$, for which [13, Lemma 4.2] yields that A has probability zero. Hence, $\tau_0 = \infty$ almost surely. ■

With a simple change of variables, we obtain a model for the prediction error $V_t = X_t - p_t$ given by

$$\begin{cases} dV_t = -\theta_t V_t dt + \sqrt{2\alpha\theta_0(V_t + p_t)} dW_t, \\ t \in [0,T], V_0 = v_0 \geq -p_0. \end{cases} \tag{3.7}$$

This formulation provides us a compact description for the prediction error, and Theorem 3.1 confirms that by choosing appropriate θ_t we obtain a unique strong solution for the SDE which does not leave $\mathbb{R}^+$ almost surely. Hence the modeled LV area over the inner slices does not become zero almost surely.

3.2. Modeling the outer slices. As discussed above, outer slices are especially hard to predict (see Figure 2). Particularly large errors occur due to systematic error in the outer slices, and hence some outer slices are falsely segmented as LV, or conversely, some outer slices show LV where the nnU-Net assumes only LV adjacent tissue is visible. The SDE is not designed to incorporate large jumps, particularly where the prediction is near zero. The mean-reverting parameter θ_t is large when p_t is close to zero, as shown in (3.4). This causes the trajectories to stay close to prediction p_t. Therefore, outer MRI slices must be modeled separately due to this systematic error and the SDE's inability to handle those errors well. Critical slices are $p_0 = 0$ and $p_1 > 0$ and $p_N = 0$ and $p_{N-1} > 0$. We treat the first and last two slices equivalently with the same parameters, and hence it suffices to only consider the first two slices. We want the model to be able to cover three mutually exclusive cases as follows:

1. $x_0 = 0,\ x_1 > 0$ (no jump).
2. $x_0 > 0,\ x_1 > 0$ (jump up).
3. $x_0 = 0,\ x_1 = 0$ (jump down).

Case 1. The first case exhibits relatively small error. Slice zero is modeled with no volume and slice one contains LV volume. In this case the model agrees with the neural network.

The subsequent cases describe systematic errors that could occur in the outer slices.

Case 2. We model the possibility that, contrary to the neural network prediction, there is already LV volume in slice zero, which often occurs when the neural network falsely classifies the LV in slice zero as surrounding tissue.

Case 3. We model the possibility that there is no LV volume in slice one.

Case 2 and 3 errors can have a large effect on the final prediction. To be exhaustive in the logical combinations, the fourth possible case ($x_0 > 0$ and $x_1 = 0$) is not modeled since we assume the LV to be a connected volume, which negates the possibility for zero predictions between nonzero predictions.

Let X_0 and X_1 be random variables describing the volume at slices 0 and 1, respectively. We choose a hierarchical approach based on a categorical distribution to model their behavior. This distribution models which of the three depicted cases occurs: jump-up, jump-down, or no-jump with probabilities λ_u, λ_d, and $1 - \lambda_u - \lambda_d$, respectively; where $\lambda_d, \lambda_u \in [0,1]$ and $\lambda_u + \lambda_d \leq 1$. After determining the case in the first layer of the hierarchical model, we draw from a specific distribution for this case, which then composes the second layer of the hierarchical model. We cannot model X_0 and X_1 independently because we have to prevent x_0 from being nonzero while $x_1 = 0$, which would lead to the model having a zero volume in slice one between two nonzero slices. The distribution also must depend on the predicted volume $p_0 + p_1 = p_1$ in the first two slices since we want the model to inherit model properties for the inner slices—more specifically, that the expected volume matches the segmentation algorithm prediction. These considerations lead to the probability density function

$$f_{X_0,X_1}(x_0, x_1 \mid \Theta_J, p_1) = \begin{cases} \lambda_d & \text{for } x_0 = x_1 = 0, \\ \lambda_u \cdot f_{u0}(x_0 \mid \beta_{u0}, p_1) \cdot f_{u1}(x_1 \mid \beta_{u1}, p_1) & \text{for } x_0, x_1 > 0, \\ (1 - \lambda_d - \lambda_u) \cdot f_n(x_1 \mid \beta_n, p_1) & \text{for } x_0 = 0, x_1 > 0, \end{cases} \tag{3.8}$$

where f_{u0}, f_{u1}, and f_n are density functions for the jump-up and no-jump case distributions, respectively, defined by predicted volume p_1 and some parameter β. There is no distribution for the jump-down case since the result $x_0 = x_1 = 0$ is deterministic. For simplicity, we abbreviate the notation $\Theta_J = (\beta_n, \beta_{u0}, \beta_{u1}, \lambda_u, \lambda_d)$ as all parameters necessary for the jump distribution. In order to keep the cases in (3.8) disjunct, we require $f_{u0}(0 \mid \beta_{u0}) = f_{u1}(0 \mid \beta_{u1}) = f_n(0 \mid \beta_n) = 0$.

The probability density function (3.8) can also be expressed as

$$f_{X_0,X_1}(x_0, x_1 \mid \Theta_J) = \lambda_d\,\delta(x_0)\delta(x_1) + \lambda_u \cdot f_{u0}(x_0 \mid \beta_{u0}, p_1) \cdot f_{u1}(x_1 \mid \beta_{u1}, p_1) \\ + (1 - \lambda_d - \lambda_u) \cdot f_n(x_1 \mid \beta_n, p_1)\delta(x_0) \tag{3.9}$$

with the Dirac delta distribution δ. Hence, we can derive the marginal distributions

$$\begin{aligned} f_{X_0}(x_0 \mid \Theta_J) &= \int_0^\infty f_{X_0,X_1}(x_0, x_1 \mid \Theta_J) dx_1 \\ &= \lambda_d \cdot \delta(x_0) + (1 - \lambda_u - \lambda_d) \cdot \delta(x_0) + \lambda_u \cdot f_{u0}(x_0 \mid \beta_{u0}, p_1) \\ &= (1 - \lambda_u) \cdot \delta(x_0) + \lambda_u \cdot f_{u0}(x_0 \mid \beta_{u0}, p_1) \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} f_{X_1}(x_1 \mid \Theta_J) = \int_0^\infty f_{X_0,X_1}(x_0, x_1 \mid \Theta_J) dx_0 \\ = \lambda_d \cdot \delta(x_1) + (1 - \lambda_u - \lambda_d) \cdot f_n(x_1 \mid \beta_n, p_1) + \lambda_u \cdot f_{u1}(x_1 \mid \beta_{u1}, p_1); \end{aligned} \tag{3.11}$$

and their expected values

$$\mathbb{E}[X_0] = \lambda_u \cdot \int_0^\infty x_0 \cdot f_{u0}(x_0 \mid \beta_{u0}, p_1) dx_0 \tag{3.12}$$

and

$$\begin{aligned} \mathbb{E}[X_1] = \lambda_u \cdot \int_0^\infty x_1 \cdot f_{u1}(x_1 \mid \beta_{u1}, p_1) dx_1 \\ + (1 - \lambda_u - \lambda_d) \cdot \int_0^\infty x_1 \cdot f_n(x_1 \mid \beta_n, p_1) dx_1. \end{aligned} \tag{3.13}$$

To keep the model bias-free with respect to the prediction, expected values for the first two slices should be equal to the prediction in those slices, i.e., we require $\mathbb{E}[X_0 + X_1] = p_0 + p_1 = p_1$. The model should have the same mean as the prediction p_1 in slice one if the simulation is nonzero, hence we require $\mathbb{E}_{X_1 \sim f_{u1}}[X_1] = \mathbb{E}_{X_1 \sim f_n}[X_1] = p_1$. With (3.12), (3.13), and $\mathbb{E}[X_0] + \mathbb{E}[X_1] = \lambda_u \mathbb{E}_{X_0 \sim f_{u0}}[X_0] + (1 - \lambda_d) \cdot p_1$, we require

$$\lambda_u \mathbb{E}_{X_0 \sim f_{u0}}[X_0] + (1 - \lambda_d) \cdot p_1 = p_1 \tag{3.14}$$

and

$$\mathbb{E}_{X_0 \sim f_{u0}}[X_0] = \frac{\lambda_d}{\lambda_u} p_1. \tag{3.15}$$

Therefore, the desired means $\mu_{u0}(p_1)$, $\mu_{u1}(p_1)$, and $\mu_n(p_1)$ for the distributions only depend on the prediction p_1 at slice one.

We choose the gamma distribution for f_{u0}, f_{u1}, and f_n in (3.8), which has support of $(0,\infty)$. For the choice of other distributions the corresponding density functions has to be replaced in (3.12) and (3.13), and the following likelihood formulation (4.4) would need to be adapted. The gamma distribution has inverse scale and shape parameters β and k, respectively. We let k be determined globally and adjust β per sample such that the distribution expected value satisfies (3.14) and (3.15). Define $k = \mu(p_1) \cdot \beta$; then $\mu(p_1)$ is the desired mean for the corresponding distributions.

To summarize: This procedure lets us model the critical outer slices, which are difficult even for human experts, and hence can induce large prediction errors when significant large heart areas are misclassified. The introduced distribution can model those particular errors, while remaining bias-free with respect to neural network predictions. The model also remains nonnegative and prevents zero volume between nonzero slices, which would lead to disconnected volumes throughout the heart.

3.3. Combining SDE and jump distribution.

Combining the SDE for intermediate slices and the jump distribution for apical and basal sections requires these parts to be independent, since errors in the respective areas are presumably caused by independent underlying error sources.

The SDE requires prediction $(p_t)_{t=0,\ldots,N}$ to be continuous and differentiable: As noted earlier, we define $p(t)$ as the linear interpolation between points, and we need to keep the integral of $p(t)$ over t the same as the sum over all points. Thus, we need to ensure we start and end with 0, which is achieved by appending $p_{-1} = p_{N+1} = 0$. This leads to

$$\begin{aligned}\int_{-1}^{N+1} p(t)dt &= \sum_{i=-1}^{N} \int_{i}^{i+1} p(t)dt \\ &= \sum_{i=-1}^{N} \left(p_i + \frac{p_{i+1} - p_i}{2} \right) \\ &= \frac{p_0}{2} + \frac{p_{N+1}}{2} + \sum_{i=0}^{N} p_i \\ &= \sum_{i=0}^{N} p_i.\end{aligned} \tag{3.16}$$

For each simulation $x(t)$ for $t \in [-1, N+1]$ we first draw values x_0, x_1, x_{N-1}, x_N from the jump distribution as described in subsection 3.2. We require a starting value to initialize the SDE, where p_1 is not optimal, since the SDE assumes consistent error transitions, but the error at p_1 is caused by a different mechanism. Starting at p_2 would mean the model always has $x_2 = p_2$, which is also undesirable because we also want to model deviations from prediction at slice two. Therefore, we introduce an additional parameter δ, following [4, section 4.5]. To ensure the model has similar error behavior for the second slice as subsequent slices, we start the process at an assumed point $\tilde{p}_{2-\delta} = p_2$ at $t = 2 - \delta$, and evaluate the SDE

$$\begin{cases} dX_t = (\dot{p}(t) - \theta_t(X_t - p_t))dt + \sqrt{2\alpha\theta_0 X_t} dW_t, \\ t \in [2-\delta, N-2], X_{2-\delta} = p_2 > 0, \end{cases} \tag{3.17}$$

where p_t is the linear interpolation of $\tilde{p}_{2-\delta}, p_2, p_3, \ldots, p_{N-2}$ for $t \in [2-\delta, N-2]$.

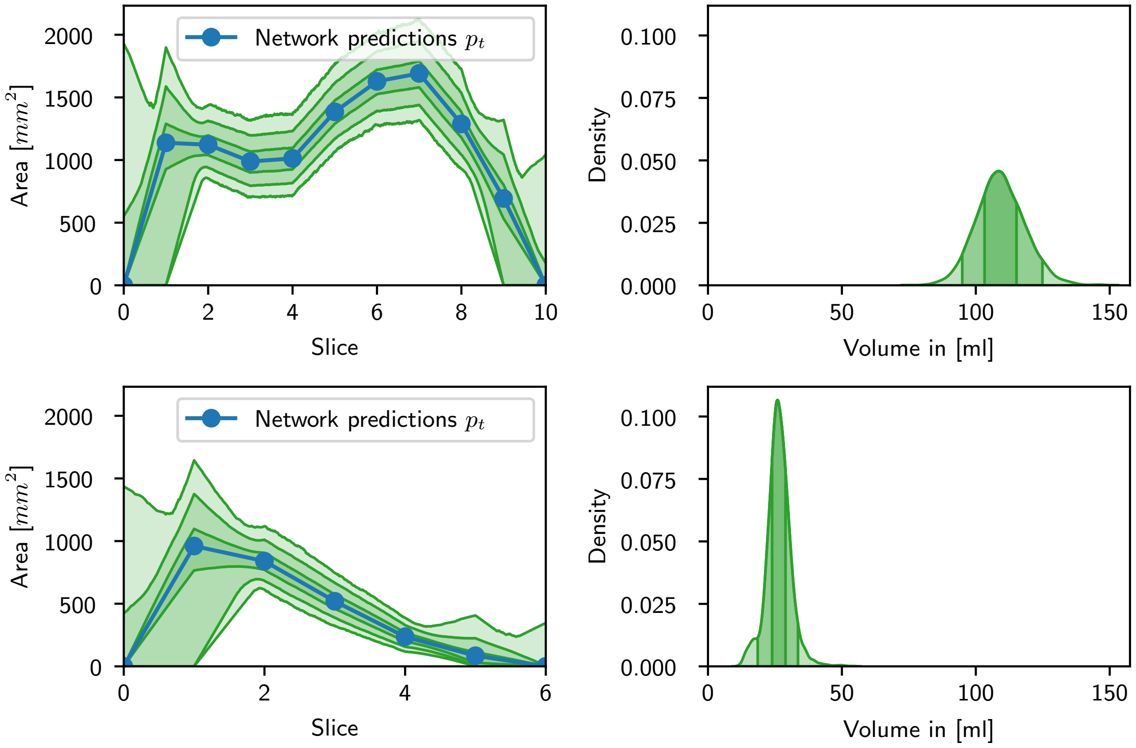

Figure 4. *The results of the proposed method on two different sized hearts. On the left: The neural network prediction* p_t *in blue. The shaded areas depict the quantiles of* 10,000 *paths, containing* 50%, 90%, *and* 99% *of the paths. On the right: A kernel density estimate of the resulting volume prediction of these* 10,000 *paths. The green shades correspond to the same quantiles as on the left.*

We integrate $x(t)$ over t to obtain the volume. For $t \in [-1,1] \cup [N-2, N+1]$, $x(t)$ contains the linear interpolation of $x_{-1} = 0, x_0, x_1, x_2$ and $x_{N-2}, x_{N-1}, x_N, x_{N+1} = 0$. Between these piecewise linear functions $x(t)$, $t \in (2, N-2)$, is given by one simulation of the SDE (3.17). In Figure 4 we can see the bands created by 10,000 trajectories of the combined approach with SDE and jump distribution along with the final volume prediction.

4. Likelihood formulation.

We adopt a maximum likelihood approach to find our model parameters, which requires parameter likelihoods depending on the segmentation algorithm predictions $P^{(j)} = \{p_0^{(j)}, \dots, p_{N_j}^{(j)}\}$ and labeled ground-truth data $G^{(j)} = \{g_0^{(j)}, \dots, g_{N_j}^{(j)}\}$, with $p_i^{(j)}, g_i^{(j)} \geq 0$ and $j \in \{1, \dots, M\}$. Algorithm 4.1 shows the proposed procedure.

The complete likelihood $\mathcal{L}_C$ factors nicely into three parts: $\mathcal{L}_{Jump}$ for the first and last two slices, $\mathcal{L}_{SDE}$ for inner slices, and $\mathcal{L}_\delta$ for the connection between the first jump and the inner slices:

$$\mathcal{L}_C = \mathcal{L}_{Jump}(\Theta; G, P)\, \mathcal{L}_{SDE}(\Theta; G, P)\, \mathcal{L}_\delta(\Theta; G, P)\,, \tag{4.1}$$

with Θ being the set of all parameters.

Algorithm 3.1. Simulation.

```
Require: Preprocessed predictions $P = p_0, \dots, p_N$
Require: Parameters $\lambda_u$, $\lambda_d$, $\beta_{u0}$, $\beta_{u1}$, $\beta_n$, $\alpha$, $\theta_0$, $\delta$
  Calculate $\theta_t$ for $t \in [2, N-2]$
  Set $x_{-1} = 0, x_{N+1} = 0$
  for desired number of simulations do
    Draw uniformly from:
      Jump down with probability $\lambda_d$
        $x_0 = 0$, $x_1 = 0$
      Jump up with probability $\lambda_u$
        $x_0 \sim \text{Gamma}\left(\frac{\lambda_d \cdot \beta_{u0}}{\lambda_u} p_1, \beta_{u0}\right)$, $x_1 \sim \text{Gamma}\left(p_1 \cdot \beta_{u0}, \beta_{u0}\right)$
      No jump with probability $1 - \lambda_u - \lambda_d$
        $x_0 = 0$, $x_1 \sim \text{Gamma}\left(p_1 \cdot \beta_n, \beta_n\right)$
    Draw uniformly from:
      Jump down with probability $\lambda_d$
        $x_N = 0$, $x_{N-1} = 0$
      Jump up with probability $\lambda_u$
        $x_N \sim \text{Gamma}\left(\frac{\lambda_d \cdot \beta_{u0}}{\lambda_u} p_{N-1}, \beta_{u0}\right)$, $x_{N-1} \sim \text{Gamma}\left(p_{N-1} \cdot \beta_{u0}, \beta_{u0}\right)$
      No jump with probability $1 - \lambda_u - \lambda_d$
        $x_N = 0$, $x_{N-1} \sim \text{Gamma}\left(p_{N-1} \cdot \beta_n, \beta_n\right)$
    Simulate starting point $\tilde{x}_2$:
      Simulate $dX_t = \theta_t(X_t - p_2)dt + \sqrt{2\alpha\theta_0 X_t} dW_t$ for $t \in [2-\delta, 2]$ with $x_{2-\delta} = p_2$
      Set $\tilde{x}_2 = x_2$
    Simulate intermediate slices
      Define $p_t$ in between the points $i$ and $i+1$ as the linear interpolation of $p_i$ and $p_{i+1}$
      Simulate $dX_t = (\dot{p} - \theta_t(X_t - p_t))dt + \sqrt{2\alpha\theta_0 X_t} dW_t$ for $t \in [2, N-2]$ with $x_2 = \tilde{x}_2$
    $Volume = \int_{-1}^{N+1} x_t dt$
  end for
```

4.1. Likelihood of the jump parameters. First, consider the jump parameters $\Theta_J = (\lambda_d, \lambda_u, \beta_{u0}, \beta_{u1}, \beta_n)$. Jump parameter likelihoods on the observed dataset G, P with M observations can be expressed as

$$\mathcal{L}_{Jump}\left(\Theta_J; G, P\right) = \prod_{i=1}^{M} \left(f_{X_0, X_1}\left(g_0^{(i)}, g_1^{(i)} \,\middle|\, p_1^{(i)}, \Theta_J\right) \cdot f_{X_0, X_1}\left(g_N^{(i)}, g_{N-1}^{(i)} \,\middle|\, p_{N-1}^{(i)}, \Theta_J\right)\right), \tag{4.2}$$

where f_{X_0, X_1} is the joint jump distribution (3.9). Initial and final jumps share the same parameters. Therefore, reindexing the ground-truth labels for the latter edge as $g_0^{(i+M)} := g_N^{(i)}$, $g_1^{(i+M)} := g_{N-1}^{(i)}$, and $p_1^{(i+M)} := p_{N-1}^{(i)}$ can simplify the likelihood notation to

$$\mathcal{L}_{Jump}\left(\Theta_J; G^{(i)}, P^{(i)}\right) = \prod_{i=1}^{2M} f_{X_0, X_1}\left(g_0^{(i)}, g_1^{(i)} \,\middle|\, p_1^{(i)}, \Theta_J\right). \tag{4.3}$$

Algorithm 4.1. Fitting.

```
Require: Preprocessed predictions $P^{(j)} = p_0^{(j)}, \dots, p_{N_j}^{(j)}$, $j = 1, \dots, M$
Require: Preprocessed ground-truth $G^{(j)} = g_0^{(j)}, \dots, g_{N_j}^{(j)}$, $j = 1, \dots, M$
  Compute jump parameters $\lambda_u$, $\lambda_d$, $\beta_{u0}$, $\beta_{u1}$, $\beta_n$:
    Find index sets $I_n, I_u, I_d$ as in (4.4)
    $\lambda_n = \frac{|I_n|}{2M}$, $\lambda_u = \frac{|I_u|}{2M}$, $\lambda_d = \frac{|I_d|}{2M}$ according to (4.5)
    Use gradient-free maximization scheme to find $\beta_{u0}$, $\beta_{u1}$, $\beta_n$ with (4.4):
        $\underset{\beta_{u0},\beta_{u1},\beta_n}{\operatorname{argmax}} \prod_{i \in I_u} f_{u0}(g_0^{(i)} \mid \beta_{u0}) \cdot f_{u1}(g_1^{(i)} \mid \beta_{u1}) \cdot \prod_{i \in I_n} f_n(g_1^{(i)} \mid \beta_n)$
  Compute SDE parameters $\alpha$, $\theta_0$:
    Use gradient free maximization scheme to maximize approximate likelihood (4.8):
        $\underset{\alpha,\theta_0}{\operatorname{argmax}} \prod_{i=1}^{M} \left( \prod_{j=3}^{N_i-2} f_\gamma(g_j^{(i)} \mid \beta_\gamma(j^-), k(j^-)) \right)$
    $\beta_\gamma(j^-), k(j^-)$ are obtained by solving the initial value problem (4.7) for each slice $j$
  Compute combining parameter $\delta$:
    Use gradient free maximization scheme to maximize the approximate likelihood (4.9)
```

Now consider mutually disjoint index sets $I_n, I_u, I_d \subset \{1, \dots, 2M\}$ with $I_n \cup I_u \cup I_d = \{1, \dots, 2M\}$. Each set contains the indices for the different jump cases: no-jump, jump-up, jump-down. Hence for $i \in I_d \Rightarrow g_0^{(i)} = 0, g_1^{(i)} > 0$, $i \in I_u \Rightarrow g_0^{(i)}, g_1^{(i)} > 0$, and $i \in I_d \Rightarrow g_0^{(i)} = g_1^{(i)} = 0$. There is no observation i in the dataset with $g_0^{(i)} > 0$ and $g_1^{(i)} = 0$, which confirms our assumption that the LV volume cannot be interrupted with slices with no volume. Therefore, the three cases above cover all possible cases and are mutually exclusive. Thus, using those index sets and (3.9), we can express (4.3) as

$$
\begin{aligned}
\mathcal{L}_{Jump}\left(\Theta_J; G^{(i)}, P^{(i)}\right) &= \prod_{i \in I_d} \lambda_d \cdot \prod_{i \in I_u} \lambda_u f_{u0}(g_0^{(i)} \mid \beta_{u0}) \cdot f_{u1}(g_1^{(i)} \mid \beta_{u1}) \\
&\quad \cdot \prod_{i \in I_n} (1 - \lambda_d - \lambda_u) f_n(g_1^{(i)} \mid \beta_n). \\
&= \lambda_d^{|I_d|} \cdot \lambda_u^{|I_u|} \cdot (1 - \lambda_d - \lambda_u)^{2M - |I_d| - |I_u|} \\
&\quad \cdot \prod_{i \in I_u} f_{u0}(g_0^{(i)} \mid \beta_{u0}) \cdot f_{u1}(g_1^{(i)} \mid \beta_{u1}) \cdot \prod_{i \in I_n} f_n(g_1^{(i)} \mid \beta_n).
\end{aligned}
\tag{4.4}
$$

Recall that f_{u0}, f_{u1}, f_n are density functions of the gamma distributions defined for the different jump cases in subsection 3.2. Our goal is to find parameters Θ_J that maximize likelihood $\mathcal{L}_{Jump}\left(\Theta_J; G^{(i)}, P^{(i)}\right)$. Using the formulation in (4.4) and knowing that all factors are greater than zero, we can separate the likelihood into two maximization problems: one for λ_u and λ_d and one for the parameters β_{u0}, β_{u1}, and β_n. First consider the lambdas,

$$\operatorname*{argmax}_{\substack{\lambda_u\geq 0,\lambda_d\geq 0\\ \lambda_u+\lambda_d\leq 1}} \lambda_d^{|I_d|}\cdot\lambda_u^{|I_u|}\cdot(1-\lambda_d-\lambda_u)^{2M-|I_d|-|I_u|}. \tag{4.5}$$

The optimal solution is $\lambda_d = \frac{|I_d|}{2M}$ and $\lambda_u = \frac{|I_u|}{2M}$, as can be seen by maximizing the logarithm of (4.5). It follows that the model's probability for jump-ups and jump-downs must be the same as the observed frequency of the respective jumps in the data.

The likelihood for β_{u0}, β_{u1}, and β_n can be determined by the likelihood of the gamma distributions u_0, u_1, and n. Note that the likelihood can also be separated for each of the distributions.

4.2. Likelihood of the SDE parameters. For the SDE parameters $\Theta_{SDE} = \{\alpha, \theta_0\}$, we can formulate the likelihood as a product of transition densities in the V-space (see (3.7)). Therefore, we transfer the labeled ground-truth data into V-space $V^{(i)} = (v_0^{(i)}, v_1^{(i)}, \dots v_{N_i}^{(i)})$, with $v_j^{(i)} = g_j^{(i)} - p_j^{(i)}$, for MRI $i \in \{1, \dots, M\}$. Similarly to (12) in [4], we define a transitional density $\rho(v \mid v_{j-1}^{(i)}; \Theta_{SDE})$. Then the likelihood function for $V = V^{(0)}, \dots, V^{(M)}$ can be expressed as

$$\mathcal{L}_{SDE}(\Theta_{SDE}; V) = \prod_{i=1}^{M} \left(\prod_{j=3}^{N_i-2} \rho(v_j^{(i)} \mid v_{j-1}^{(i)}; p_{[j-1:j]}^{(i)}, \Theta_{SDE}) \right). \tag{4.6}$$

We would need to solve an initial-boundary value problem to derive an exact solution for the transition densities (see (13) in [4]), which would be computationally expensive. To avoid this, we approximate the likelihood similarly to [4, section 4.2], selecting the gamma distribution as the surrogate transition density, since this is the invariant distribution of (3.7). For $t \in [j, j+1)$ we can describe the first two moments $m_1(t) = \mathbb{E}[V_t]$ and $m_2(t) = \mathbb{E}[V_t^2]$ with the initial value problem as follows:

$$\begin{cases} \frac{dm_1(t)}{dt} = -m_1(t)\theta_t, \\ \frac{dm_2(t)}{dt} = -2\theta_t m_2(t) + 2\alpha\theta_0 m_1(t) + 2\alpha\theta_0 p_t, \end{cases} \tag{4.7}$$

with initial conditions $m_1(j) = v_j$ and $m_2(j) = v_j^2$. System (4.7) can be deduced using Itô's lemma. We obtain shape and scale parameters (k, β_γ) by matching the gamma distribution with the first two moments,

$$\beta_\gamma(j+1) = \frac{\mu_{j+1}}{\sigma_{j+1}^2},$$

$$k(j+1) = \frac{\mu_{j+1}^2}{\sigma_{j+1}^2},$$

where $\mu_{j+1} = m_1(j+1) + p_{j+1}$, $\sigma_{j+1}^2 = m_2(j+1) - m_1(j+1)^2$; we shifted the mean by p_t, to translate back to the X-space. This is possible as the variance σ_{j+1}^2 is invariant to changes in location, hence we can now express (4.6) as an approximate likelihood $\tilde{\mathcal{L}}_{SDE}$,

$$\tilde{\mathcal{L}}_{SDE}(\Theta_{SDE}; G) = \prod_{i=1}^{M} \left(\prod_{j=3}^{N_i-2} f_\gamma(g_j^{(i)} \mid \beta_\gamma(j^-), k(j^-)) \right), \tag{4.8}$$

where f_γ is the probability density function, and shape $k(j^-)$ and scale $\theta(j^-)$ parameters solely depend on the limits $\mu(j^-, \Theta_{SDE})$ and $\sigma(j^-, \Theta_{SDE})$ for $t \uparrow j$, which can computed by numerically solving the initial value problem (4.7).[2] This enables computing an approximate likelihood for SDE parameters Θ_{SDE}, and we find the optimal parameters by minimizing the negative logarithm of (4.8) with the gradient-free downhill simplex algorithm implemented in SciPy.

Likelihood $\mathcal{L}_\delta$ for δ can be approximated similarly to $\mathcal{L}_{SDE}$. Set $\tilde{v}_{2-\delta} = 0$, as described in subsection 3.3, and again assume a gamma distribution as transition density $\rho(v_2 \mid v_{2-\delta}; p_2, \Theta_{SDE})$. We approximate this transition density with the same moment matching method as in (4.7) to obtain the approximate likelihood $\tilde{\mathcal{L}}_\delta(\delta; G, \Theta_{SDE})$,

$$\tilde{\mathcal{L}}_\delta(\delta; G, \Theta_{SDE}) = \prod_{i=1}^{M} f_\gamma(g_2^{(i)} \mid \beta_\gamma(2^-), k(2^-)), \tag{4.9}$$

where $\beta_\gamma(2^-), k(2^-)$ are solutions for the initial value problem (4.7) at $t = 2$ starting with $\tilde{v}_{2-\delta} = 0$ at $t = 2 - \delta$. We calibrate δ after estimating Θ_{SDE}.

In contrast to [4], our tests were insensitive regarding particular choices of the starting values in the optimization process. However, the method to find initial values from [4] could in principle be adapted to match the SDE (3.7).

5. Experimental results.

The proposed method adds an uncertainty estimate to the point prediction for LV volume, where the point prediction is obtained by an arbitrary machine learning algorithm. The method must capture the uncertainty of the data and adjust for different data regimes. Furthermore, it needs to be stable on smaller datasets.

For this evaluation, we used two cardiac MRI datasets: The Multi-Centre, Multi-Vendor, and Multi-Disease Cardiac Segmentation (M&Ms) dataset [5] and the Automatic Cardiac Diagnosis Challenge (ACDC) dataset [2]. The ACDC dataset comprises 300 MRIs from 150 patients, where each patient has one MRI in the end-diastolic and one in the end-systolic phase. Patients are evenly distributed between five classes: four pathologies and one extra for healthy patients. All images come from the same research center. In contrast to the ACDC dataset, the M&Ms dataset focuses on generalizing across different medical centers with different scanner vendors. The 375 MRIs comprise images from 6 medical centers with five distinct MRI devices. Pathologies are considerably less evenly distributed in this dataset. Although healthy patients comprise one-third of the data, the second largest pathology class (hypertrophic cardiomyopathy) also contains 103 MRIs. Thus, more than 60% of the data are in the two largest groups. We only used the training sets from both datasets because labels are only available for the training sets.[3]

We selected the nnU-Net [9] as the pretrained predictor.[4] The nnU-Net method automatically finds suitable U-Net structures for a wide range of medical datasets and won both the ACDC segmentation and M&Ms challenges. We used the version trained on the more diverse M&Ms dataset and evaluated the method on the ACDC dataset, providing a whole MRI

[2] The notation $f(a^-)$ describes the one-sided limit $\lim_{x \uparrow a} f(x)$ from below.

[3] The final ACDC dataset contains 200 MRIs from 100 patients, and the M&Ms dataset contains 300 MRIs from 150 patients.

[4] See https://github.com/MIC-DKFZ/nnUNet for implementation details.

image dataset unseen during training. The proposed procedure is applicable for real-world applications, where one would select a suitable pretrained net on the most diverse dataset possible and subsequently evaluate on data from one medical center. Although nnU-Net uses a five neural network ensemble for final prediction, we used only the first of those neural networks for simplicity. Real-world use cases would improve prediction quality by using the full ensemble and the largest dataset possible, e.g., combining the publicly available M&Ms and ACDC datasets.

5.1. Model assessment. As described earlier the nnU-Net employs an ensemble of five neural networks trained on different folds of the training data. Leveraging the predictions of different networks for uncertainty estimation, could be an alternative approach for estimating the uncertainty. As a benchmark model, we will use the five predictions of the full nnU-Net. We will view the prediction of each neural network as a point prediction, as we saw in Figure 2 that the variance within each prediction is negligible. We obtain five point predictions for the volume of each slice. We then assume that the volume of each slice is normally distributed with the empirical mean and standard deviation of those five predictions. Let $N+1$ be the number of slices for one heart and μ_i, σ_i be the mean and standard deviation of the five predictions at slice $i \in \{0, \dots, N\}$. The model for the volume of the full LV is a sum of normal distributions. Hence, it is normally distributed with mean $\sum_{i=0}^{N} \mu_i$ and variance $\sum_{i=0}^{N} \sigma_i^2$.

For our model the parameters were fitted on the whole training dataset, and section 5.2 shows that fitted parameters do not substantially differ on data subsets, confirming that models fitted on data subsets exhibit similar characteristics. We sample 10,000 volume predictions based on our uncertainty model and the benchmark model for each heart and compare them to labeled ground-truth data. We compare the deviation of the models from the predictions of the neural network with the deviation of the expert labels from the predictions of the neural network. To compare this we corrected the bias of the neural network, as we do not want to capture this uncertainty in our model. The bias in the model would need to be addressed separately and is beyond the scope of this work. We use the empirical distribution functions for the comparison to be independent of the choice of bins. To account for the small sample size of the labeled data we also plotted the Kolmogorov–Smirnov (KS) confidence band for the level $\lambda = 0.05$ around the empirical cdf. The KS confidence band utilizes the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality to estimate the interval around the empirical cdf, which contains the true cdf with probability $1-\lambda$ [12, 6].[5] In Figure 5 we see that not only is our model a lot better at describing the uncertainty seen in the data, but it is also able to capture the data quite well, including the width of the deviation from the neural network prediction. Figure 5 shows that our method captures the error of the full heart well. It lies well within the KS confidence band. To investigate further, we examined the volume of the difficult edge slices. Because of the different volumes, it is hard to compare the data over several hearts. To circumvent this problem we divided the expert labeled volume and the predicted volume of our model by the predicted volume of the neural network. This way for each heart the expected volume is one and we can see the multimodality of the jumps in the data. Figure 6 shows that

[5] Let $F(x)$ be the true cdf and $\hat{F}_n(x)$ the empirical cdf for n samples. From the DKW inequality follows with probability $1-\lambda$ that for every x, the interval that contains the true cdf $F(x)$ is a subset of $[\hat{F}_n(x) - c_{n,\lambda}, \hat{F}_n(x) + c_{n,\lambda}]$, where $c_{n,\lambda} = \sqrt{ln(2/\lambda)/(2n)}$.

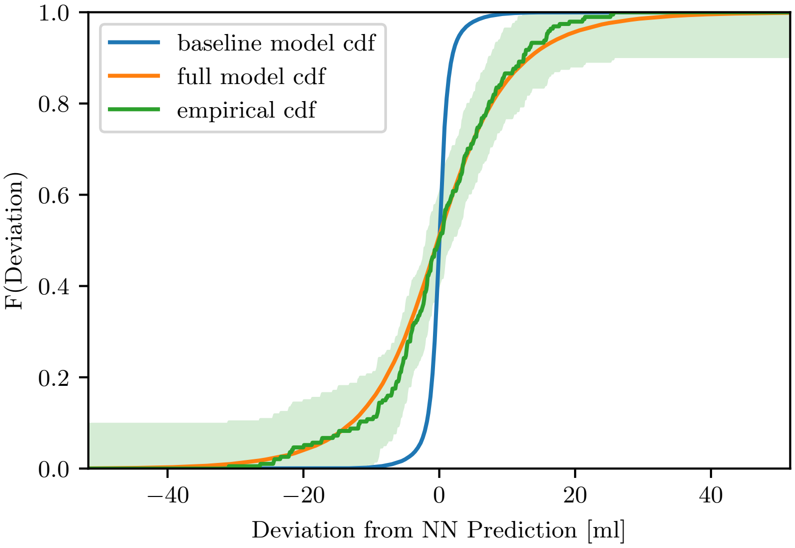

Figure 5. *Comparison of the different model-based and empirical cumulative density functions (cdfs) of the deviation from the neural network prediction. Green shows the deviation of the expert labels from the neural network predictions. Light green shows the KS confidence band, based on the DKW inequality, which includes the true cdf with a probability of* 95%*. Orange shows the cdf of our model, and blue the cdf of the benchmark model, utilizing the five prediction of the neural network ensemble. To obtain the cdfs, the LV volume of each heart is predicted* 10,000 *times. The benchmark model utilizes the predictions of all five neural networks from the nnU-Net.*

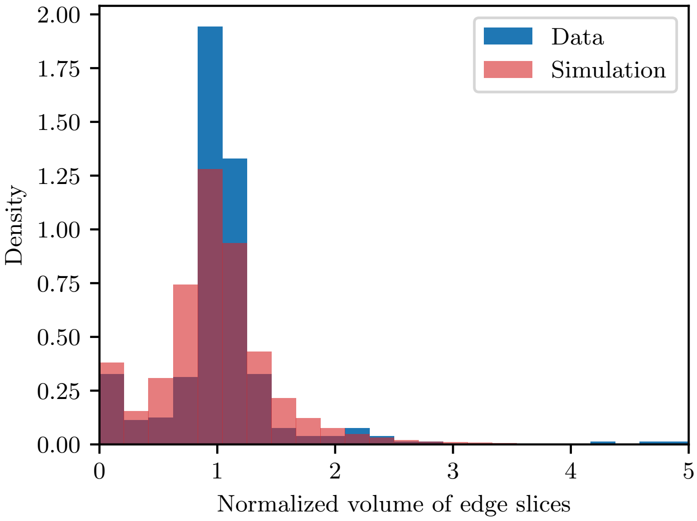

Figure 6. *Comparison of the volumes of the edge slice. In order to make the volumes comparable it is normalized by the predicted volume. Blue shows the ground-truth volume and red shows the simulation volumes. For the simulation each heart is evaluated* 1000 *times. In both the data and the simulation one can see a multimodal distribution in the particularly difficult edge slices.*

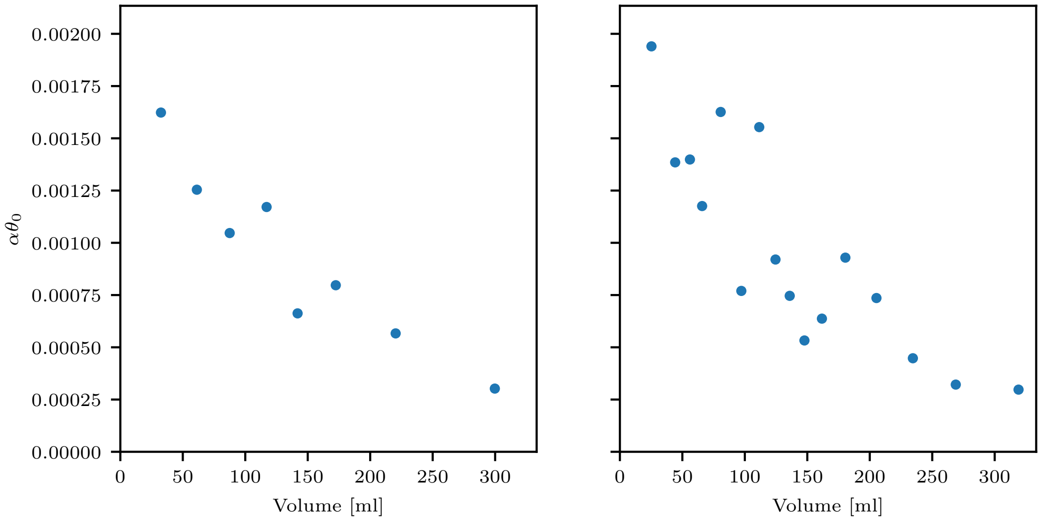

Figure 7. *Heart size effects on $\alpha\theta_0$ obtained by maximizing the likelihood on a subset of the data. The subsets are binned by volume, with each bin receiving approximately the same number of observations: (left)* 8 *bins, (right)* 16 *bins. The scatter points are located at the midpoint for the respective bin. This shows that for larger volumes the optimal $\alpha\theta_0$ get smaller. This trend is consistent, even when very small subsets of data are used.*

our approach of modeling the jumps captures this multimodality well. This multimodality is mostly evident for total volume predictions of one heart where the edge slices have significant influence on overall volume.

Additional to the difficult edge slices, the ACDC dataset contains significant heart variance due to different pathologies and end-systolic and end-diastolic phases included in the data. The LV volumes range from ∼25 to ∼325 ml. Therefore, we investigated model effects due to different heart sizes. The hearts were binned by predicted volume, so we could also use this binning for the inference step, where no ground-truth label is available. We then calculated parameters α_i and $\theta_{0,i}$ for each bin i. Figure 7 confirms $\alpha_i\theta_{0,i}$ decrease consistently with LV volume. The trend remains consistent up to 16 bins, providing only 12 or 13 observations per bin. Larger values for $\alpha_i\theta_{0,i}$ in small hearts lead to a higher path variability and uncertainty estimation for those hearts. These outcomes agree well with the larger relative errors in this data regime (see Figure 8). For smaller hearts the model fitted on the whole dataset underestimates the errors while still being in the KS confidence band for the level $\lambda = 0.05$. The model fitted on the subsets of the small hearts is closer to the empirical cdf in this regime. One could potentially compensate for pretrained net shortcomings, i.e., higher uncertainty for smaller hearts in this case, by adjusting the predicted uncertainty accordingly.

5.2. Stability analysis.

We evaluated the proposed method's stability by assessing parameter stability to data changes. Figure 9 shows the negative log-likelihood for SDE parameters α and θ_0 with respect to the entire dataset. The minimum for the relevant parameter range is quite wide, and there are no other local minima. To further study the proposed procedure stability on smaller datasets, we divided the dataset into three disjoint subsets with 64 MRIs in each, then repeated this procedure three times to obtain nine subsets, with always three

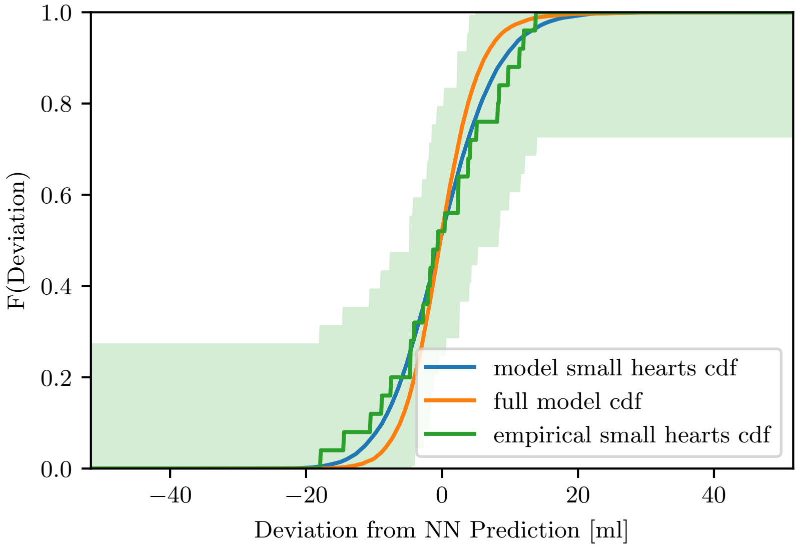

Figure 8. *Comparison of the different model-based and empirical cumulative density functions of the deviation from the neural network prediction on the* 25 *smallest hearts. Green shows the deviation of the expert labels from the neural network predictions. Light green shows the KS confidence band, based on the DKW inequality, which includes the true cdf with a probability of* 95%*. Orange shows the cdf of the model fitted on the full dataset and blue the model fitted on only the small models. While both models lie within the KS confidence band, which is quite wide, as the dataset only consists of* 25 *samples, the model fitted on the small hearts is better at capturing the relative larger deviation for small hearts.*

being pairwise disjoint. Furthermore, our method allows us to choose the direction we go through the MRI slices, since we do not have a preferred or optimal direction, i.e., from basal to apical sections, or the reverse. Therefore, we also generated nine subsets for the reversed dataset following the same procedure. We see that the parameters for all subsets lie in the same region with similar likelihood. This implies that the procedure is stable on the subsets of data. The consistent trend when using small bins of 12–13 observations in Figure 7 implies stability for even smaller datasets.

The main challenge for assessing optimal jump parameters is that jumps are scarce in the data. A jump-down occurred for approximately 5.4% of the outer slices, with the jump-up case slightly more frequent (approximately 7%). Therefore, we modeled jumps in basal and apical sections with the same parameters, λ_u and λ_d, i.e., observed jump frequency in the dataset. Assuming true jump frequency of 6%, we calculated the observed frequency for a given subset with the binomial distribution as shown in Figure 10. To investigate parameter stability for describing gamma distributions, we again split the full dataset into three parts (64 MRIs) and found maximum likelihood parameters for each subset. Figure 11 shows outcomes after repeating this process ten times. Although jumps remain quite rare, parameters for the jump distributions are stable across the various data subsets, with only one potential outlier from 30 subsets.

To test for the combined effect of the jump parameters and the SDE parameters on the stability, we again divided the dataset into 18 subsets, with always three being pairwise disjunct and nine of them being reversed. Figure 12 shows that the models fitted on each of

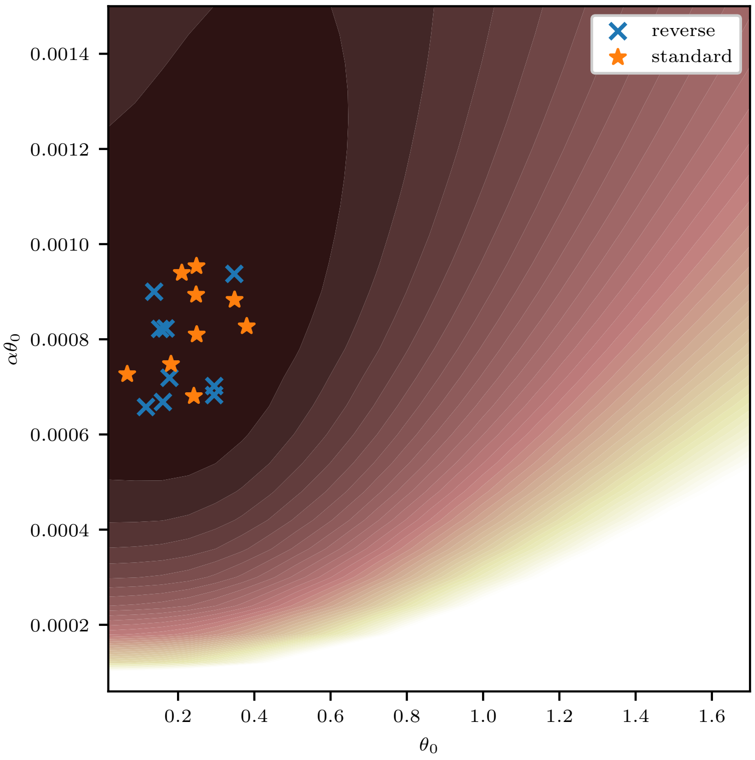

Figure 9. *Fitted parameters for different data subsets. The full dataset was divided three times into three parts, obtaining nine datasets, with always three being disjoint. The same procedure was performed on the dataset with reversed order for the slices for each heart. The contour plot shows negative log-likelihood with respect to the full standard dataset.*

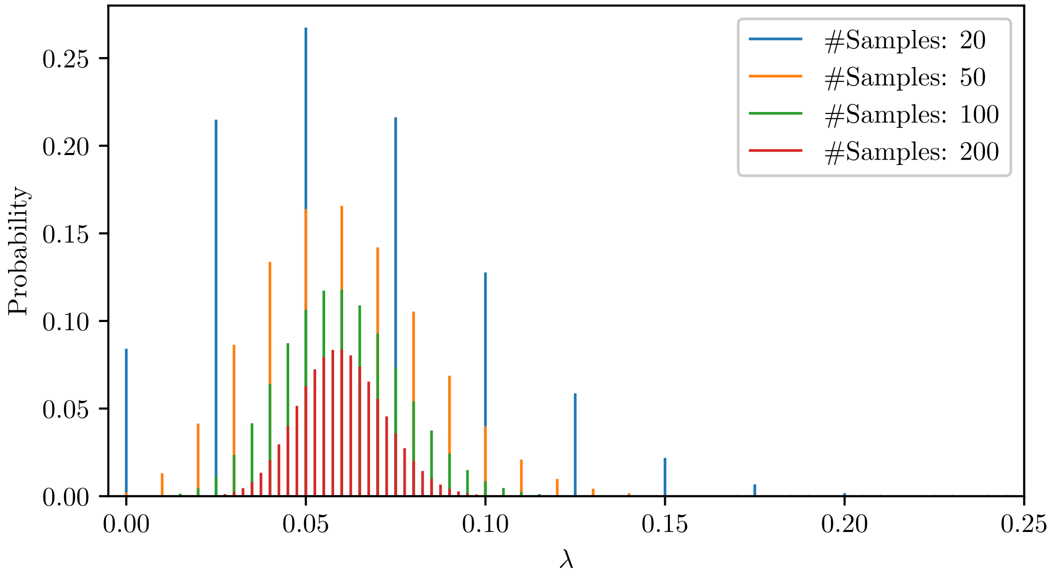

Figure 10. *Probability of estimating different* λ *for different sample sizes, assuming true* $\lambda_{\text{true}} = 0.06$.

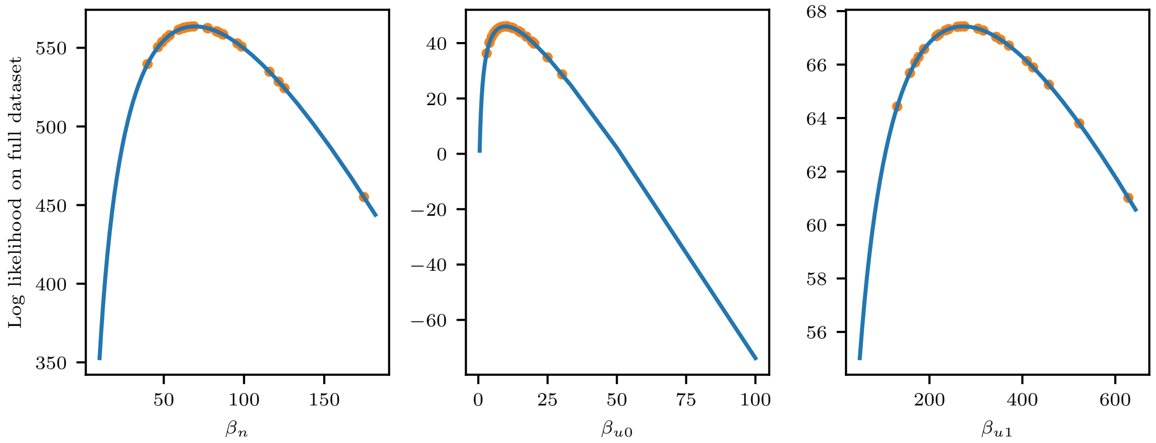

Figure 11. *Fitted parameters for different data subsets. The full dataset is divided ten times into three parts, providing* 30 *datasets, with always three being disjoint. Values are shown with the likelihood calculated from the full dataset.*

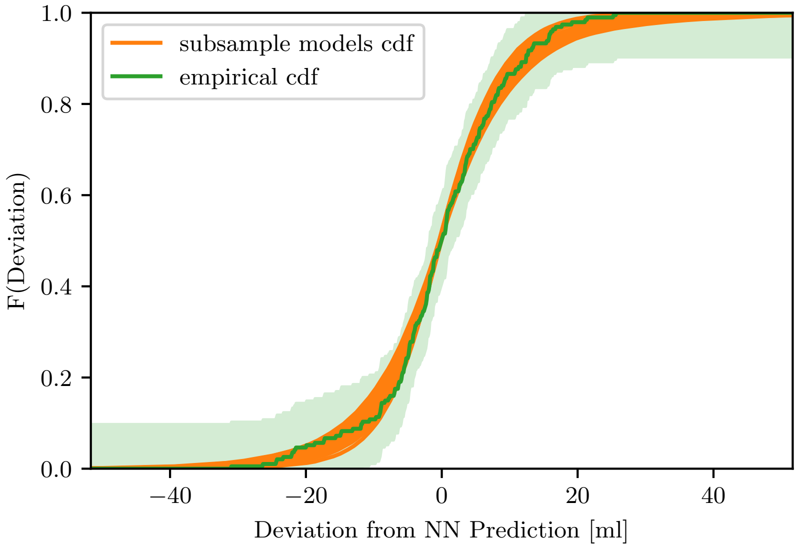

Figure 12. *Comparison of the different model-based and empirical cumulative density functions of the deviation from the neural network prediction on the full dataset. Green shows the deviation of the expert labels from the neural network predictions. Light green shows the KS confidence band, based on the DKW inequality, which includes the true cdf with a probability of* 95*%. Orange shows the cdf of the different models fitted on subsets of the data. In total there are* 18 *subsets of the data, with always three being pairwise disjoint and in nine the direction is flipped; cf. Figure* 9*. The models fitted on all subsets stay within the KS confidence band.*

the 18 subsets lie well within the KS confidence band of the empirical CDF of the full dataset. This indicates that the full model, comprised of the jump distribution and the SDE, is stable on small datasets.

6. Conclusion.

We developed a phenomenological model to describe the uncertainty for left ventricle volume prediction. Accurately measuring LV volume is crucial to properly assess cardiovascular mortality or diagnose heart failure. The proposed method is a post-hoc approach to quantify uncertainty for predictions of any segmentation algorithm.

Our requirements for the model was that it should

1. consist of few parameters, so that it can be fitted on limited data;
2. stay nonnegative and bias-free with respect to the neural network prediction; and
3. most importantly, describe the uncertainty well.

Those requirements should be met with few assumptions. The SDE approach is well suited to model the transitions and naturally model the correlations. We only require two parameters for the SDE to obtain a flexible model, which meets our requirements and is thus very parameter-efficient.

Our model employs the inherent locality, to obtain reliable uncertainty estimations, using heart depth as *fake time* to model deviation from underlying deterministic predictions with an SDE. Outer slice errors arise from different sources than inner slices and must be modeled independently. Therefore, we designed a jump distribution to model systematic error in these slices. Finally, we introduced an additional parameter δ to seamlessly combine the SDE and jump distribution while retaining their independence.

Max likelihood approaches were employed to fit the parameters. We described the likelihood using an analytical expression for the jump distribution and constructed an approximate likelihood formulation for the SDE parameters of the model. Moment matching was employed to describe a gamma distribution for the transition density between slices.

The proposed method is well suited to describe uncertainty for LV volume prediction, with estimated uncertainty matching ground-truth data errors well. The proposed distribution was able to reproduce the observed multimodality in ground-truth data even for particularly difficult outer slices.

There is a wide range of possible phenomenological models. Our approach meets these applications requirements. Different conceivable approaches include Gaussian processes. Similar to our approach, Gaussian processes typically have few parameters but would need to be modified to fulfill the above-mentioned requirements. Especially keeping it nonnegative while also bias-free with respect to the neural network prediction is not trivial. A different possible approach is leveraging the flexibility of diffusion models. Diffusion models share some similarities with the SDE approach and show promising results in a wide range of applications. In [20] the authors integrate Bayesian filtering into learning diffusion models to generate stochastic processes governed by sparse observations. The sparse observations could be the predictions of the slices for our case. The time dynamics are also modeled by an Itô SDE. The drift function is typically modeled by a neural network. This increases the flexibility of the model compared to the drift function defined in (3.2). On the other hand it also drastically increases the number of parameters which need to be fitted, which could be problematic in our use case. Furthermore, the drift and diffusion function would need to be adapted, such that they ensure nonnegativity. The comparison of other phenomenological models, such as Gaussian processes and diffusion models, with our approach would be an interesting direction of future research.

In real-world application, most medical centers only have small labeled datasets and can only use publicly available data with various limitations due to differing vendors and imaging protocols from center to center. The stability of the proposed method with respect to dataset size and its ability to be combined with any segmentation algorithm is crucial and makes the method applicable for real-world scenarios. Our evaluations showed the proposed method is stable even on small datasets.

This proposed method can also compare segmentation algorithm performance on different data subregimes. We found increased relative error for smaller hearts, which suggests one could adapt this method to account for the different behavior in different data regimes using parameters dependent on the predicted volume.

Automatic LV volume prediction would facilitate and speed-up diagnosis, but reliable uncertainty quantification is essential. In contrast to current best practice approaches, the proposed adaptable, post-hoc method provides this critical parameter, and paves the way toward reliable automatic left ventricle volume prediction.

REFERENCES

[1] A. Alfonsi, *Affine Diffusions and Related Processes: Simulation, Theory and Applications*, Springer, 2015, https://doi.org/10.1007/978-3-319-05221-2.

[2] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, *Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?*, IEEE Trans. Med. Imaging, 37 (2018), pp. 2514–2525, https://doi.org/10.1109/TMI.2018.2837502.

[3] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, *Weight uncertainty in neural network*, in International Conference on Machine Learning, PMLR, 2015, pp. 1613–1622.

[4] R. Caballero, A. Kebaier, M. Scavino, and R. Tempone, *Quantifying uncertainty with a derivative tracking SDE model and application to wind power forecast data*, Statist. Comput., 31 (2021), 64, https://doi.org/10.1007/s11222-021-10040-8.

[5] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreño, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Viladés, M. L. Descalzo, A. Guala, L. L. Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Çavuş, S. E. Petersen, S. Escalera, S. Seguí, J. F. Rodríguez-Palomares, and K. Lekadir, *Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge*, IEEE Trans. Med. Imaging, 40 (2021), pp. 3543–3554, https://doi.org/10.1109/TMI.2021.3090082.

[6] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, Ann. Math. Statist., 27 (1956), pp. 642–669, https://doi.org/10.1214/aoms/1177728174.

[7] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*, in ICML 48, PMLR, 2016, pp. 1050–1059.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, *On calibration of modern neural networks*, in International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.

[9] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, *nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation*, Nat. Methods, 18 (2021), pp. 203–211, https://doi.org/10.1038/s41592-020-01008-z.

[10] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Grad. Texts in Math. 113, Springer Science & Business Media, 1998.

[11] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. A. Eslami, D. Jimenez Rezende, and O. Ronneberger, *A probabilistic U-Net for segmentation of ambiguous images*, in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Curran Associates, 2018, pp. 6965–6975, https://proceedings.neurips.cc/paper_files/paper/2018/file/473447ac58e1cd7e96172575f48dca3b-Paper.pdf.

[12] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, Ann. Probab., 18 (1990), pp. 1269–1283, https://doi.org/10.1214/aop/1176990746.

[13] E. Mayerhofer, O. Pfaffel, and R. Stelzer, *On strong solutions for positive definite jump diffusions*, Stochastic Process. Appl., 121 (2011), pp. 2072–2086, https://doi.org/10.1016/j.spa.2011.05.006.

[14] C. G. Missouris, S. M. Forbat, D. R. Singer, N. D. Markandu, R. Underwood, and G. A. MacGregor, *Echocardiography overestimates left ventricular mass: A comparative study with magnetic resonance imaging in patients with hypertension*, J. Hypertens., 14 (1996), pp. 1005–1010, https://doi.org/10.1097/00004872-199608000-00011.

[15] L. Perdrix, N. Mansencal, B. Cocheteux, G. Chatellier, A. Bissery, B. Diebold, E. Mousseaux, and E. Abergel, *How to calculate left ventricular mass in routine practice? An echocardiographic versus cardiac magnetic resonance study*, Arch. Cardiovasc. Dis., 104 (2011), pp. 343–351, https://doi.org/10.1016/j.acvd.2011.04.003.

[16] G. A. Roth, et al., *Global, regional, and national age-sex-specific mortality for* 282 *causes of death in* 195 *countries and territories,* 1980-2017*: A systematic analysis for the global burden of disease study* 2017, The Lancet, 392 (2018), pp. 1736–1788, https://doi.org/10.1016/S0140-6736(18)32203-7.

[17] A.-J. Rousseau, T. Becker, J. Bertels, M. B. Blaschko, and D. Valkenborg, *Post training uncertainty calibration of deep networks for medical image segmentation*, in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 1052–1056, https://doi.org/10.1109/ISBI48211.2021.9434131.

[18] E. Shelhamer, J. Long, and T. Darrell, *Fully convolutional networks for semantic segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 39 (2017), pp. 640–651, https://doi.org/10.1109/TPAMI.2016.2572683.

[19] A. V. Skorokhod, *Studies in the Theory of Random Processes*, Dover, 2017; unabridged and unaltered republication of the original 1965 Addison-Wesley edition.

[20] E. Tamir, M. Trapp, and A. Solin, *Transport with support: Data-conditional diffusion bridges*, Trans. Mach. Learn. Res., 2023, https://doi.org/10.48550/arXiv.2301.13636.

[21] P. Verdecchia, G. Schillaci, C. Borgioni, A. Ciucci, R. Gattobigio, I. Zampi, G. Reboldi, and C. Porcellati, *Prognostic significance of serial changes in left ventricular mass in essential hypertension*, Circulation, 97 (1998), pp. 48–54, https://doi.org/10.1161/01.CIR.97.1.48.

[22] T. Yamada and S. Watanabe, *On the uniqueness of solutions of stochastic differential equations*, J. Math. Kyoto Univ., 11 (1971), pp. 155–167, https://doi.org/10.1215/kjm/1250523691.

[23] Y. Zhang, J. Yang, F. Hou, Y. Liu, Y. Wang, J. Tian, C. Zhong, Y. Zhang, and Z. He, *Semi-supervised cardiac image segmentation via label propagation and style transfer*, in Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges, E. Puyol Anton, M. Pop, M. Sermesant, V. Campello, A. Lalande, K. Lekadir, A. Suinesiaputra, O. Camara, and A. Young, eds., Springer International, Cham, 2021, pp. 219–227, https://doi.org/10.1007/978-3-030-68107-4_22.