IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 August 2023, accepted 26 August 2023, date of publication 1 September 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311134

RESEARCH ARTICLE

Toward Detecting and Addressing Corner Cases in Deep Learning Based Medical Image Segmentation

SRIVIDYA TIRUNELLAI RAJAMANI[1], KUMAR RAJAMANI[2], ASHWIN VENKATESHVARAN[3], ANDREAS TRIANTAFYLLOPOULOS[1], ALEXANDER KATHAN[1], AND BJÖRN W. SCHULLER[1,4,5], (Fellow, IEEE)

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[2]Marwadi Education Foundation's Group of Institutions, Rajkot, Gujarat 360003, India
[3]Department of Clinical Physiology, Department of Clinical Sciences Lund, Lund University, Skåne University Hospital, 221 85 Lund, Sweden
[4]Centre for Interdisciplinary Health Research, University of Augsburg, 86159 Augsburg, Germany
[5]GLAM—Group on Language, Audio and Music, Imperial College London, SW7 2BX London, U.K.

Corresponding author: Srividya Tirunellai Rajamani (srividya.tirunellai@informatik.uni-augsburg.de)

ABSTRACT Translating machine learning research into clinical practice has several challenges. In this paper, we identify some critical issues in translating research to clinical practice in the context of medical image segmentation and propose strategies to systematically address these challenges. Specifically, we focus on cases where the model yields erroneous segmentation, which we define as corner cases. One of the standard metrics used for reporting the performance of medical image segmentation algorithms is the average Dice score across all patients. We have discovered that this aggregate reporting has the inherent drawback that the corner cases where the algorithm or model has erroneous performance or very low metrics go unnoticed. Due to this reporting, models that report superior performance could end up producing completely erroneous results, or even anatomically impossible results in a few challenging cases, albeit without being noticed. We have demonstrated how corner cases go unnoticed using the Magnetic Resonance (MR) cardiac image segmentation task of the Automated Cardiac Diagnosis Challenge (ACDC) challenge. To counter this drawback, we propose a framework that helps to identify and report corner cases. Further, we propose a novel balanced checkpointing scheme capable of finding a solution that has superior performance even on these corner cases. Our proposed scheme leads to an improvement of 44.6 % for LV, 46.1 % for RV and 38.1 % for the Myocardium on our identified corner case in the ACDC segmentation challenge. Further, we establish the generalisability of our proposed framework by also demonstrating its applicability in the context of chest X-ray lung segmentation. This framework has broader applications across multiple deep learning tasks even beyond medical image segmentation.

INDEX TERMS Corner-case handling, medical image segmentation, research to clinical practice, cardiac MRI, chest X-ray.

I. INTRODUCTION

Medical image segmentation is arguably one of the most influential and widely researched application fields of artificial intelligence (AI) in the healthcare domain [1], [2], [3], [4], [5], [6], [7]. It corresponds to the segmentation of organs, tissues, or pathologies of interest in medical images, obtained through X-ray, ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), mammography and further more. At the heart of medical image segmentation is the correct identification of a *region of interest* (ROI) that needs to be found in medical images. For example, in cardiac magnetic resonance (MR) image segmentation, the ROI corresponds to the different anatomical parts of the heart.

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

A correct, automated segmentation can vastly accelerate the time to diagnose and relieve medical practitioners from an overburdening workload. Naturally, this vast potential comes with its assortment of risks, since missing critical medical findings can have a detrimental impact on patient outcomes. This in turn leads to much more stringent evaluation protocols and higher robustness standards for medical AI solutions than for other AI applications. Accordingly, the community has spent a lot of effort in devising proper evaluation methods. Nevertheless, even though several evaluation metrics have been proposed over the years, there still remain numerous 'blind spots' with each of them.

In particular, existing evaluation protocols suffer from one major downside: they are computed on an aggregate basis over a population of patients. While this serves the purpose of providing a quick gauge of the performance, they run the risk of masking 'corner cases'. This hidden risk has insofar eluded the attention of the community, but may have serious downstream repercussions when medical practitioners come to rely on them for their daily work.

What is more, missing out on 'corner cases' circumvents the requirement to provide a uniform standard of care for all prospective patients. This is a fundamental ethical requirement for fair treatment: the performance of an algorithm should be the same for all individuals, irrespective of their characteristics. Disaggregated evaluations, applied down to the individual level, can help single out flaws and discrepancies of a model across different patients.

In this work, we show how aggregated evaluations, which are the gold standard of evaluating performance of medical image segmentation models, can lead to misleading interpretations of model performance. Specifically, we focus on identifying corner-cases in the evaluation of a state-of-the-art model using a standardised heart image segmentation database. We show how even one of the most widely used evaluation metric, the *Dice score*, could fail to capture corner cases where the model prediction dramatically diverges from the target ROI when averaged over the entire dataset. We then proceed to propose a procedure for monitoring the training process that can mitigate this issue by highlighting those cases. Our work is thus directly connected to the broader literature on machine learning transparency and accountability, and in particular the need to truthfully, and proactively, identify potential shortcomings of production models [8]. This is particularly critical for medical applications, since blind spots on corner-cases where models can fail directly translate to worse or even potentially dangerous clinical outcomes.

II. RELATED WORK

Identifying appropriate evaluation protocols that holistically measure performance in a fair manner is challenging for most research fields. Nevertheless, it constitutes a critical requirement when it comes to real-world applications, especially in the medical domain, where they are crucial in facilitating a transfer to clinical practice. Naturally, this topic has attracted increasing attention from the community, as the advent of deep learning has rapidly accelerated research in medical image segmentation.

For example, [9] explored the lack of reliability in medical image segmentation performance assessments. Typically used metrics are often overoptimistic of model performance and fail to reveal potential weaknesses [10], [11]. As a consequence, clinical teams repeatedly encounter problems when it comes to transferring beyond research environments [11], [12]. To cope with the opaqueness of medical image segmentation evaluation metrics, [9] provided an overview of often-used evaluation scores, such as the Dice similarity coefficient, Jaccard, or Cohen's Kappa. Furthermore, they proposed a set of guidelines for interpretation and a standardised evaluation. To further advance standardisation and reproducibility, [13] proposed MISeval, a metric library for evaluation.

Similarly, [14] explored a set of boundary overlap metrics to capture a wider range of segmentation errors, covering the most frequently used classes of segmentation metrics: size, overlap, and boundary distance approaches. In their work, they also demonstrated that there are large differences between existing evaluation scores as well as high dependencies on the clinical use case. Therefore, there is a gap between high values of well-known metrics, such as the Dice score, and the applicability to real-world data.

While these issues are present throughout the general medical image segmentation field [15], [16], [17], [18], [19], specific facets of the problem appear for individual applications – in our case, Cardiac MR Image segmentation. Bernard et al. [20] present a comprehensive summary of how state-of-the-art deep learning methods perform in the context of Cardiac MR Image segmentation and diagnosis. They further identify several challenges that still exist in this field, the most prominent of them being:

- Right Ventricle (RV) segmentation and calculation of the RV ejection fraction .
- Myocardium segmentation at the End Systole (ES) phase: The difficulty to precisely delineate LV and RV walls.
- Segmenting slices near the apex and base: Challenges in the apex pertain to small structures while the challenge at the basal slices is about how to differentiate between multiple structures.
- Inter-observer variability among experts in segmenting apex and basal slices.
- Generation of anatomically impossible results: Deep learning based segmentation methods resulted in 82 % of patients having anatomically impossible segmentation in at least one slice.

In light of the understanding that cardiac MR segmentation is technically challenging, it is imperative to precisely identify the boundary conditions and limitations of each method before using them in clinical context.

To that end, a consortium of multiple academia and industry researchers as well as practitioners have teamed up

to analyse the flaws in machine learning algorithm validation. In their seminal work in this area, Maier-Hein et al. [21] have identified various pitfalls in the choice of validation metrics, namely:

- the inappropriate phrasing of the problem
- poor metric selection
- poor metric application

To address these challenges, they propose their ''Metrics Reloaded'' framework comprising of problem fingerprinting as well as a metrics selection methodology.

Furthermore, Maier-Hein et al. [22] emphasise that care has to be exercised while interpreting the outcomes of large-scale international challenges that benchmark different models. They highlight that aspects such as the choice of metrics as well as the criteria used for aggregated ranking across metrics could influence the determination of the winning method. They show that a metric-based vs a case-based ranking scheme is a significant design choice and that winners could change based on the aggregation method chosen. In our current work, we discover that aggregation of results even across patients has to be done with care, especially in the presence of corner-cases.

Specifically, identifying corner-cases, that could potentially remain hidden when only average metrics are considered, still remains an unexplored area. We consider this an extremely important, yet grossly overlooked, aspect of metric application – especially in the context of semantic segmentation. Even though researchers tend to report very high performance metrics , these may still end up performing poorly on a few particularly challenging scenarios. While performance on corner-cases is not of high significance in research where only averages are reported, blind utilisation of such solutions for clinical diagnosis/intervention could have severe consequences. Therefore, an awareness of the pitfalls of deep learning methods on different corner-cases is vital when considering their usage in clinical practice. It is of prime importance for researchers to discover and transparently report such corner cases for any solution – in short, to acknowledge the Achilles' heel of their method.

We note that there is some broader literature on evaluating disagreggated model performance beyond the field of medical image segmentation. Typically, this concerns the evaluation of model fairness with respect to different sub-populations (e. g., age and gender groups), but there is also some existing work which evaluates how models perform across different individuals [23], [24]. This is also related to the notion of 'individual fairness' which contests that ''similar individuals should receive equal treatment'' [25]. Ouyang et al. [26] also explored corner cases for classification tasks in their work. In doing so, they introduced a metric developed on the basis of modified 'surprise' adequacy, which targets the characteristics of corner cases. Furthermore, they also generated artificial corner cases which could be used for improving a model, resulting in a fairer classification performance for all subjects within a dataset. Wu et al. [27] proposed a ''Deep Validation'' framework for classification tasks, which identifies error-inducing inputs and has them flagged for human intervention when the system is perceived working incorrectly. For medical image segmentation, this translates to ensuring that models generalise well to different patients, irrespective of anatomical or pathological differences. To the best of our knowledge, there exists no evaluation procedure that explicitly accounts for the detection of model failures on individual cases. Our work attempts to address this gap in the existing medical image segmentation evaluation practice.

As a significant step towards addressing these challenges and bridging this gap between research and clinical practice, our novel contributions in this paper are the following:

- A methodology for detecting and reporting of corner-cases.
- A strategy for gaining further insight into these corner cases.
- An approach for identifying a balanced checkpoint.

The rest of the paper is organised as follows. Section III describes the dataset used in our experiments and the baseline network architecture. In Section IV, our proposed framework for detecting and addressing corner cases in deep learning based medical image segmentation is presented. This is then followed by Results in Section V, benchmarking with other metrices in Section VI and generalizability of the proposed framework in Section VII. Finally, Section VIII presents a discussion followed by conclusion and directions for future work in Section IX.

III. DATASET AND BASELINE NETWORK ARCHITECTURE

The dataset as well as the baseline network architecture on which we conduct our investigation is detailed next:

A. THE ACDC SEGMENTATION DATASET

We conduct our experiments on the Automated Cardiac Diagnosis Challenge (ACDC)'s segmentation dataset [20]. The objective of the challenge is to evaluate the efficacy of deep learning methods at assessing Cardiac MRI, specially in segmenting the myocardium and the two ventricles, as well as classifying pathologies. The training dataset of this challenge contains 3D cine-Magnetic Resonance (MR) cardiac scans of 100 unique patients from the University Hospital of Dijon. Of these 100, there are 20 patients each belonging to five classes, namely,

1) Normal case
2) Heart failure with Infarction
3) Dilated Cardiomyopathy
4) Hypertrophic Cardiomyopathy
5) Abnormal Right Ventricle

For each patient, the End Systole (ES) and End Diastole (ED) frames are provided, identified based on the motion of the mitral valve from the long axis orientation by a single expert, resulting in a total of 200 volumes. Additionally, the ground truth segmentation masks for the Left Ventricle (LV), Right Ventricle (RV), and Myocardium (MYO) are also made

available for these 100 patients. The test set of the challenge comprises another 50 patients, with 10 patients per class.

B. SAUNET ARCHITECTURE

SAUNet – Shape Attentive U-Net for Interpretable Medical Image Segmentation [28], is one of the recent U-Net based methods that achieves high average Dice scores along with good interpretability in Cardiac MR image segmentation on the ACDC challenge dataset. SAUNet comprises 2 streams, a texture stream and a gated shape stream. The texture stream has the same structure as a U-Net [29], but with the encoder replaced with dense blocks from DenseNet-121 [30], similar to the Tiramisu Network proposed by Jegou et al. [31]. The decoder block is a dual attention decoder block. Furthermore, it incorporates learning of shape features through a secondary stream that processes shape features of the image. Additionally, the interpretability of features is enabled at every resolution of the U-Net using spatial and channel-wise attention paths in the decoder block. We therefore utilize SAUNet as the baseline architecture in our experiments. We use the same training-validation split as well as hyperparameters as in [28].

IV. METHODOLOGY

The schematic of our proposed methodology for identifying and addressing corner-cases is presented in Figure 1 and explained in the following sections.

A. METHODOLOGY FOR DETECTING AND REPORTING OF CORNER-CASES

Deep learning based medical image segmentation methods currently report average metrics. We propose to analyse the characteristics of patient-wise metrics to determine potential outliers. One of the recent unsupervised approaches for outlier detection in large, high-dimensional datasets is Empirical-Cumulative-distribution-based Outlier Detection (ECOD) [32].

ECOD is a multivariate statistical anomaly detection method. It derives inspiration from the fact that outliers are often the ''rare events'' that appear in the tails of a distribution (right-tail and left-tail). In this method, an empirical cumulative distribution is first computed along each data dimension. In the next step, this empirical distribution is utilized to estimate the left and right tail probabilities ($\hat{\mathbf{F}}^{(j)}_{left}$ and $\hat{\mathbf{F}}^{(j)}_{right}$). Finally, by aggregating the estimated tail probabilities across all dimensions, the outlier score is computed in a non-parametric way.

Given input data $\mathbf{X} = \{X_i\}_{i=1}^{n} \in \mathbb{R}^{n\times d}$ with n samples and d features where $X_i^{(j)}$ refers to the value of j-th feature of the i-th sample,

$$\hat{\mathbf{F}}^{(\mathbf{j})}_{\mathbf{left}}(\mathbf{z}) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbb{1}\{\mathbf{X}_\mathbf{i}^{(\mathbf{j})} \leq \mathbf{z}\} \quad \text{for } z \in \mathbb{R} \tag{1}$$

$$\hat{\mathbf{F}}^{(\mathbf{j})}_{\mathbf{right}} = \frac{1}{\mathbf{n}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbb{1}\{\mathbf{X}_\mathbf{i}^{(\mathbf{j})} \geq \mathbf{z}\} \quad \text{for } z \in \mathbb{R} \tag{2}$$

where $\mathbb{1}\{.\}$ is the indicator function that is 1 when its argument is true and is 0 otherwise [32].

For cardiac image segmentation, we propose to jointly analyse the Dice scores of LV, RV and MYO by representing them as a 3-dimensional (3D) vector. This 3D vector is computed for every patient and analysed using the ECOD algorithm to determine the corner cases. Outliers detected by this approach are flagged for detailed analysis. Furthermore, the segmentation outcomes should be reported for these flagged cases to enable clinicians to gain insights into understanding where the model fails to segment correctly.

B. STRATEGY FOR GETTING FURTHER INSIGHTS INTO THE CORNER CASES

Generally, the average Dice scores across the different training epochs are plotted to monitor the training process. However, this does not give any insights on how the model performs on corner-cases. To address this gap, we propose that further insights should be obtained by analysing the characteristics of the Dice score curves of the corner-cases, across different training epochs. For this analysis, we utilize the ECOD algorithm [32] to detect the presence of any outliers across the different training epochs. While in the previous step, the analysis is across patients, in this step, the analysis is done using the 3-dimensional (LV, RV, MYO) Dice scores across different training epochs of the corner-cases.

C. APPROACH FOR IDENTIFYING A BALANCED CHECKPOINT

In scenarios where the corner cases are observed to have large Dice score variations across different epochs, the traditional approach of model checkpointing based on least-loss or highest average-IoU (Intersection Over Union) could end up compromising the performance on corner-cases. Also, utilisation of such solutions could result in anatomically impossible outcomes in clinical practice which could lead to disastrous consequences. Hence, an active quest for a more balanced checkpointing solution is crucial for enabling deep learning based medical image segmentation approaches to be used in clinical context.

Our proposal to identify a more balanced checkpoint is to first exclude all epochs that are identified as outlier epochs for the corner-case in the previous step. Then, from the remaining epochs, we propose to utilize the final epoch as the balanced checkpoint.

V. RESULTS

A. CORNER CASE DETECTION AND REPORTING

In Table 1, we report the average Dice scores obtained using our model trained with a SAUNet network architecture [28] on the ACDC segmentation challenge dataset (column 2). In addition, we compute patient-wise Dice scores for LV, RV and MYO and identify outliers by providing these 3-dimensional scores to the ECOD algorithm [32]. We utilise the default contamination rate of 0.1 of ECOD algorithm from the PyOD

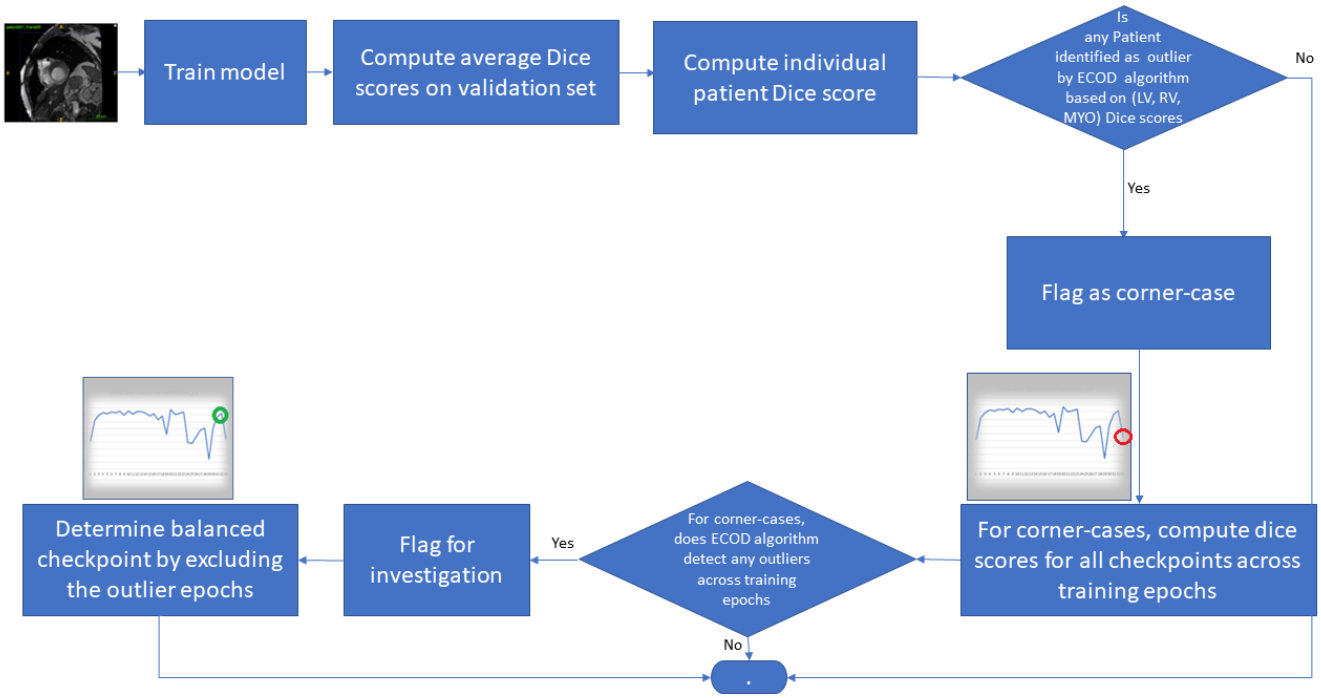

FIGURE 1. Schematic of our proposed framework for detecting and addressing corner cases in deep learning based medical image segmentation.

TABLE 1. Average Dice score and Dice scores for the corner-case identified for LV, RV, and MYO on ACDC validation set.

Organ	(a). Avg Dice score	(b). Dice score for corner case Patient057_ES	Dice score difference (a-b)
LV	0.912	0.351	0.561
RV	0.833	0.201	0.632
MYO	0.848	0.441	0.407

toolbox [33]. Patient057_ES is the only one to be detected as an outlier using our approach. In column 3 of the table, we report the Dice-scores of this corner case patient. We also report the difference between the average Dice scores and the Dice scores of Patient057_ES which is 56.1 % for LV, 63.2 % for RV and 40.7 % for MYO in column 4.

In Figure 2, the segmentation results for the corner-case Patient057_ES for all the 8 slices at End Systole are presented. We observe that for the first 4 slices the predicted segmentation is completely incorrect and also anatomically impossible. In these 4 slices, the left ventricle region is identified as the myocardium, whereas the myocardium region is identified as the right ventricle.

B. INSIGHTS INTO THE CORNER CASES

Using our proposed approach of analysing the 3-dimensional (LV, RV, MYO) Dice scores across the training epochs with the ECOD algorithm [32], outliers are also observed across the training epochs for Patient057, unlike the other patients. Hence, our approach flags Patient057 for careful investigation by clinicians and researchers.

We also compute and plot the Dice scores for the entire validation set as well as for Patient057. The results are visualised in Figure 3. The top row depicts the average Dice score plotted for the entire validation set. The bottom row depicts the individualised Dice score plot for the corner-case, Patient057_ES. The columns contain the plots for LV, RV, MYO, and a consolidated view of the 3 anatomies. The Dice scores are captured for the epochs where the model was checkpointed. We use least average-loss as the criteria to create these checkpoints.

In this figure, we observe that all the curves in the first row seem to indicate that the model is training effectively. Typically, this is how model performance and metrics are reported. However, in the bottom row, we observe that for the corner-case, Patient057_ES, the Dice scores varies considerably across the training epochs for LV, RV and MYO. For instance, the Dice score between the 24th and 25th checkpoint has a very large variation of 71.89 % for

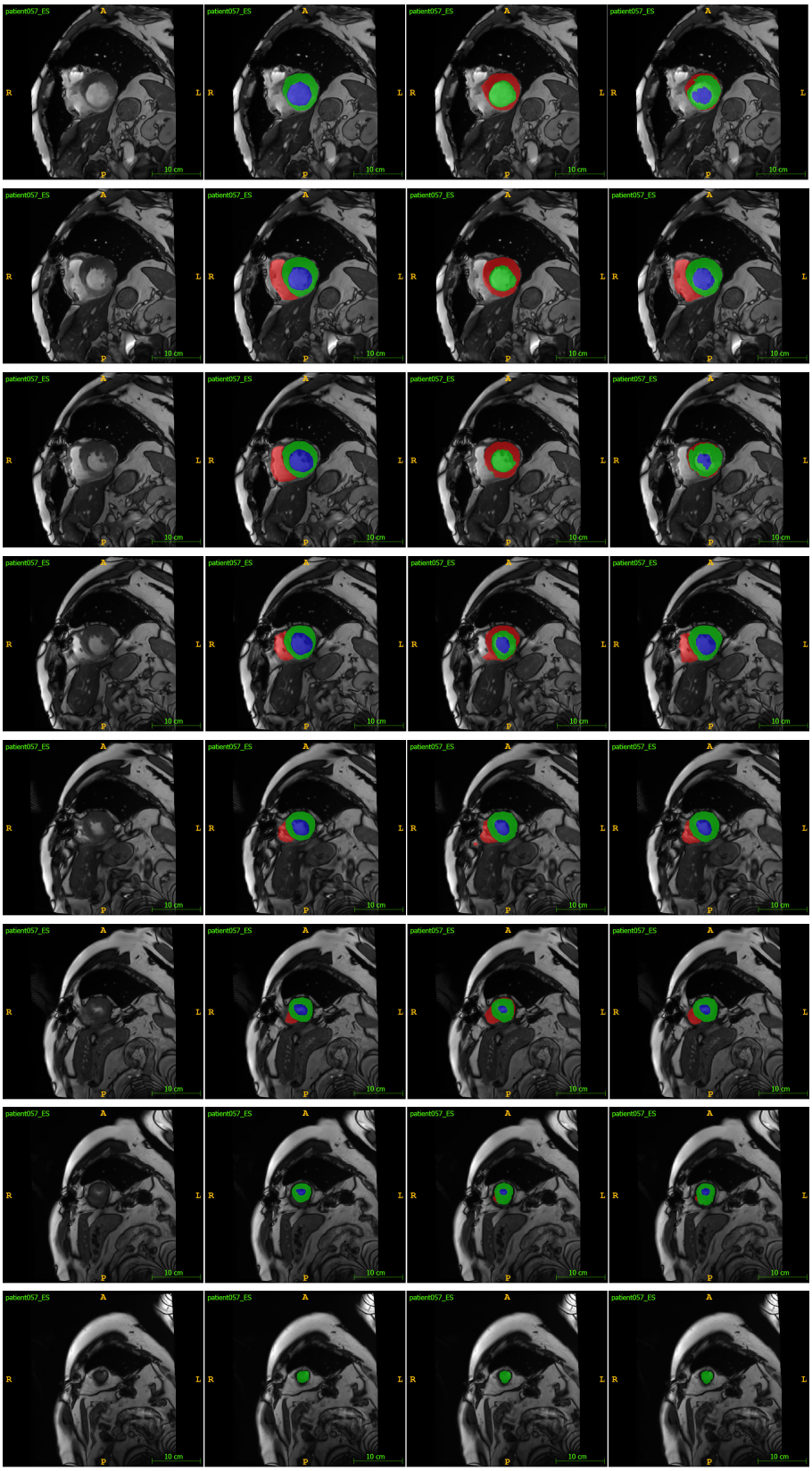

FIGURE 2. The rows contain, from top to bottom, slices 1 to 8 of End Systole frames for Patient057 from ACDC dataset. The columns from left to right are: (a). Original image, (b). Ground truth, (c). Predicted segmentation with the least-loss checkpoint, and (d). Predicted segmentation with the proposed balanced checkpoint, respectively. The colour coding used is blue for LV, green for MYO, and red for RV.

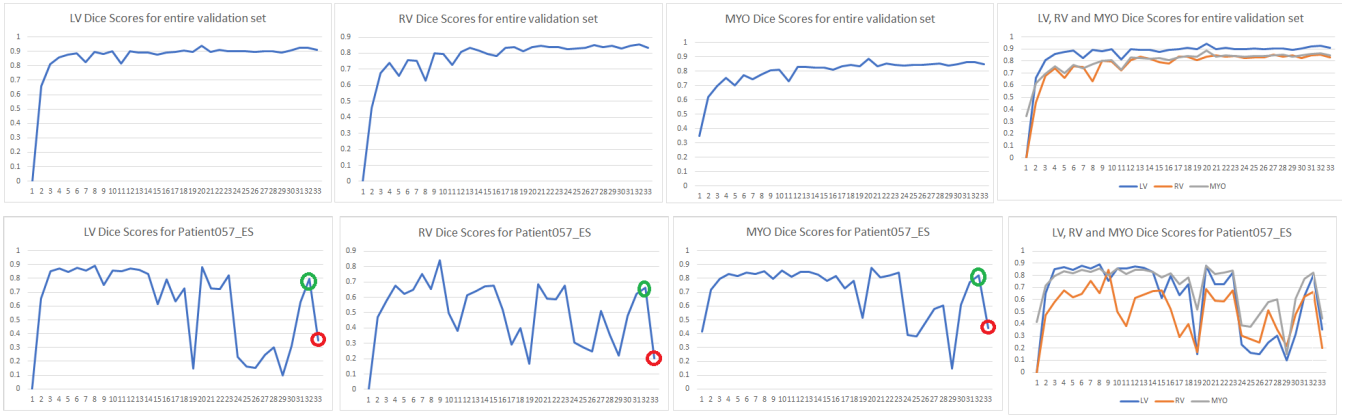

FIGURE 3. Plot of Dice scores at the least-loss based checkpoints over the training epochs. The 4 columns, from left to right, contain Dice scores for (a). LV, (b). RV, (c). MYO, and (d). consolidated-view, respectively. The top row contains the plots of average Dice score for the validation set. The bottom row contains the plots for the corner-case, Patient057_ES (the Dice score at least-loss based checkpoint is marked in red and at the proposed balanced checkpoint marked in green, respectively).

TABLE 2. Table (A). Average Dice scores for entire validation set with least-loss checkpoint and proposed balanced checkpoint. Table (B): Dice scores for the corner case, Patient057. Proposed balanced checkpoint significantly improves performance on corner-case (d-c). Furthermore, average Dice scores also improves (b-a).

(A) Results for entire validation set

Organ	(a). Dice Scores based on Least-loss checkpoint (checkpoint 33)	(b). Dice Scores based on proposed balanced checkpoint (checkpoint 32)	Percentage gain (b - a)
LV	0.912	**0.925**	**1.3**
RV	0.833	**0.856**	**2.3**
MYO	0.848	**0.863**	**1.5**

(B) Results for Patient057

Organ	(c). Dice Scores based on Least-loss checkpoint (checkpoint 33)	(d). Dice Scores based on proposed balanced checkpoint (checkpoint 32)	Percentage gain (d - c)
LV	0.352	**0.798**	**44.6**
RV	0.201	**0.662**	**46.1**
MYO	0.441	**0.822**	**38.1**

LV, 55.04 % for RV, and 53.74 % for the Myocardium. Such atypical variations could indicate erroneous model training or model performance behaviour which gets masked when only looking at the average Dice scores.

C. BALANCED CHECKPOINT DETERMINATION

Based on our proposed approach of balanced checkpoint determination, we excluded the outlier epochs determined by ECOD and chose the final epoch of the remaining ones. With this approach, the checkpoint that gets identified is the penultimate (32nd) checkpoint, as is also visualised in row 2 of Figure 3.

The results of utilising this identified balanced checkpoint is reported in Table 2. As observed in column (a) of the table, the average Dice scores based on least-loss checkpoint for the entire validation dataset are 0.912 for LV, 0.833 for RV, and 0.848 for the Myocardium. At face value, this seems like a reasonably well performing solution. However, as observed in column (c) of the table, for Patient057, this same checkpoint results in extremely low Dice scores of 0.352 for LV, 0.201 for RV, and 0.441 for the Myocardium. Hence, such a classical approach of saving the model based on least-loss compromises the performance on the corner case considerably. As observed in column (d) of the table, at our proposed checkpoint, the corner case Patient057 has a Dice score of 0.798 for LV, 0.662 for RV, and 0.822 for the Myocardium which is an improvement of 44.6 % for LV, 46.1 % for RV, and 38.1 % for the Myocardium as compared to the previously identified checkpoint. Furthermore, at this new identified checkpoint, as observed in column (b), the

TABLE 3. Average Jaccard Coefficient and Jaccard Coefficient for the corner-case identified for LV, RV, and MYO on ACDC validation set.

Organ	(a). Avg Jaccard Coefficient	(b). Jaccard Coefficient for corner case Patient057_ES	Jaccard Coefficient difference (a-b)
LV	0.849	0.213	0.636
RV	0.731	0.112	0.619
MYO	0.745	0.283	0.462

average Dice scores on the entire validation set also increase by about 1 to 2 % for each of LV, RV, and MYO.

VI. BENCHMARKING WITH OTHER METRICES

So far, we have focused our analysis on the average Dice score as the evaluation metric since it is a commonly used and well-established metric for evaluating segmentation models. It is defined as twice the area of overlap between the predicted segmentation and the actual labels, divided by the sum of the areas of the predicted segmentation and the ground truth labels, leading to a range between 0 (worst) and 1 (best) [34].

In this section, we evaluate other metrics for benchmarking segmentation results to analyse if the failure to detect the low performance in corner cases arises because of averaging across all patients or is a characteristic of Dice score.

One metric that is closely related to the Dice score is the Jaccard Coefficient, also known as the intersection over union, which is often used to determine the performance of image segmentation algorithms [1]. It also calculates the ratio of the overlapping regions, but in contrast to the average Dice score which focuses on balancing precision and recall, the Jaccard Coefficient is more sensitive to false positives.

The balanced Average Hausdorff Distance (bAHD) is another recently introduced, but yet popular metric [35]. It is derived from the Hausdorff distance, which calculates the closeness of each point in a segmentation set to the nearest point in the ground truth label set and vice-versa. The balanced Average Hausdorff Distance (bAHD), however, averages these distances, resulting in a more robust way to account for outlier points in segmentation tasks. Lower bAHD scores indicate higher segmentation quality.

While in Table 1, we present the average Dice score and Dice scores for the corner-cases, in Table 3 and 4, we present the results evaluated using the Jaccard Coefficient and the balanced Average Hausdorff Distance (bAHD), respectively. These metrics were computed using the EvaluateSegmentation tool [36]. When utilising the ECOD algorithm on the patient-wise metrices, Patient057_ES is detected as a corner-case. These results validate that averaging across patients is indeed the major factor for failure in detecting the corner cases, even with other well established and state-of-the-art metrices.

VII. GENERALISABILITY OF THE PROPOSED FRAMEWORK

In this section, we validate the generalisability of our proposed framework on the task of chest X-ray lung segmentation. The NIH chest X-ray dataset [37] contains both posterior-anterior and anterior-posterior views. Tang et al. [38] used 100 abnormal chest X-ray images from this dataset with various severity of lung diseases and manually annotated the lung masks.[1] We perform our experiments on this abnormal chest X-ray dataset.

We utilise the U-Net architecture of Oktay et al. [39], [40] which has four blocks each in the down-sampling and up-sampling path. Each block is composed of $2\times$(Batch Norm - 2D Conv (kernel size 3×3, stride 1, padding 1) - ReLU). A 2D convolution with kernel size 1×1 forms the last block. Max-pooling is used in the down-sampling path to halve the spatial dimension of the feature maps after each block. In the up-sampling path, 2D transposed convolution is utilised to double the size of the spatial dimension of the concatenated feature maps. In the down-sampling path, feature channels are increased as $(1 - 64 - 128 - 256 - 512)$. In the up-sampling path, they are decreased again accordingly. The last layer of the U-Net has feature channels that matches the number of label classes for semantic segmentation.

A criss-cross attention module (CCA) [41] is inserted in the bottleneck of this U-Net architecture. The input for this module is the feature maps from the U-Net's last block within the down-sampling path. The contextual information in the criss-cross path of each pixel is gathered by the criss-cross attention module leading to feature maps $\mathbf{H}'$. The resulting feature maps after 2 iterations of criss-cross attention are then passed through the U-Net's up-sampling path.

The average Dice score obtained using this model on the validation set of 40 patients on the NIH dataset is 0.955. We further compute the patient-wise Dice scores whose scatter plot is visualised in Figure 4. Utilising the ECOD algorithm with the default contamination factor of 0.1, patient NIH_0072 is detected as an outlier and hence flagged for detailed analysis (marked in red in the scatter plot). The segmentation result for an exemplar patient, NIH_0090 and for the detected outlier patient NIH_0072, is presented in Figure 5. From this figure, it is evident that the outlier detected by our framework does have sub-optimal segmentation outcomes.

This demonstrates that our proposed framework for detecting corner cases is generalisable across other modalities, anatomies and network architectures.

[1] Data: https://nihcc.app.box.com/s/r8kf5xcthjvvvf6r7l1an99e1nj4080m

TABLE 4. Average bAHD and bAHD for the corner-case identified for LV, RV, and MYO on ACDC validation set.

Organ	(a). Avg bAHD	(b). bAHD for corner case Patient057_ES	bAHD difference (b-a)
LV	0.152	1.938	1.786
RV	0.851	24.177	23.326
MYO	0.218	1.689	1.471

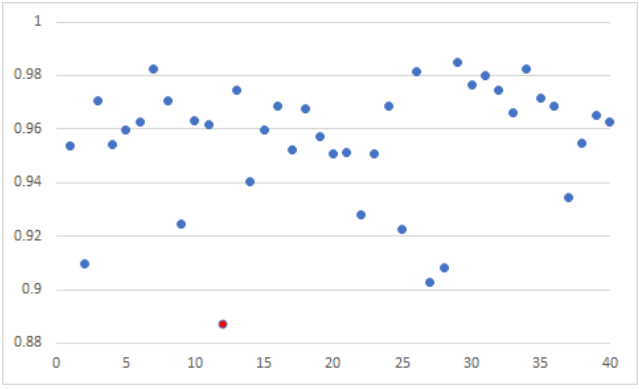

FIGURE 4. Scatter plot of patient-wise Dice scores for the NIH validation set. The outlier Dice score detected with ECOD (which corresponds to patient NIH_0072) is highlighted in red.

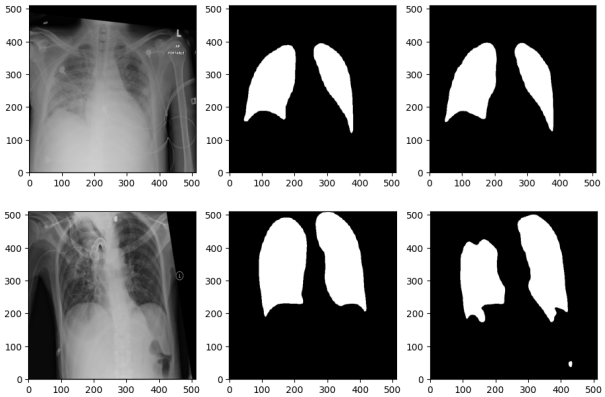

FIGURE 5. Lung segmentation results for couple of images from the NIH dataset. The first row contains results for patient NIH_0090 which is an exemplar patient. The second row contains results for the identified outlier patient NIH_0072. The columns from left to right are (a). Original image, (b). Ground truth and (c). Predicted segmentation.

VIII. DISCUSSION

In this section, we report the clinical insights gained from the corner case that our proposed approach identified on the ACDC cardiac image segmentation dataset. In addition, we also outline other potential solutions for addressing corner-cases. We also elaborate on a few alternatives for optimal checkpoint determination.

A. CLINICAL INSIGHTS INTO THE IDENTIFIED CORNER-CASE

To understand the observed aberration in the predicted segmentation of Patient057, we obtained clinical insights from an experienced cardiac imaging specialist. Careful inspection of the short axis images from the apex to the base of the LV in addition to the corresponding long axis images revealed prominent anterolateral and posteromedial papillary muscles that are generally underrepresented in the dataset. Further, the segmentation prediction based on least-loss checkpoint inaccurately identified this region of pronounced musculature as myocardium. Current international recommendations advise that papillary muscles are included in the LV cavity, as seen in the ground truth analysis where experts carefully cut through this region during cavity delineation. A plausible explanation for this aberration is the under-representation of such variants in the current dataset. This hypothesis, however, requires further investigation in larger databases.

B. CHECKPOINT DETERMINATION USING LEAST-LOSS VS HIGHEST AVERAGE-IOU

The standard approach to checkpoint the model during training is either based on least-loss or highest average-IoU.

We have computed the Dice scores based on both of these approaches on the validation set, the result of which is reported in Table 5. As seen in the 2nd and 3rd column of this table, either of these checkpointing approaches yields comparable performance and hence, we have utilised the least-loss based checkpoint in this current work.

C. OTHER POTENTIAL APPROACHES FOR HANDLING CORNER-CASES

There could be several factors that could lead to subjects/patients ending up as being corner-cases. Identification of these reasons and potential mitigation approach need an active collaboration between researchers and clinical experts. Our current insights are that this could either be due to data characteristics or due to flaws in annotation, or model/network's deficiencies.

Similarly, the resolution to address such corner cases could also be done through various regimes. For instance, if the corner case is due to data being a unique case not well represented in the training dataset, there are the following ways to address it.

TABLE 5. Checkpoint based on least-loss vs highest average-IoU on validation set.

Organ	(a). Dice Scores based on highest average-IoU checkpoint	(b). Dice Scores based on least-loss checkpoint	Percentage difference (b-a)
LV	0.899	0.912	**1.3**
RV	0.848	0.833	**-1.5**
MYO	0.844	0.848	**0.4**

- Using a data approach: In our proposed approach, we have addressed this by separately handling the corner-case. Other approaches for addressing this could be adding more data with similar characteristics to the dataset (real or synthetic). One could also potentially exclude such corner cases from the training and validation data and include a disclaimer that the solution cannot be utilised in such outlier scenarios. This would complement standardised model reporting [8] and provide clinicians a better understanding of model capabilities and potential pitfalls.
- Through the model: Further attributes of the data could be provided as context during the model training. For instance, in the ACDC challenge dataset, there are 5 different classes. This class information could be provided as additional input to the model while training.
- Through ground-truth refining: Regions which confuse the model could be marked as a separate class. For instance, the papillary muscles, when prominently visible, could be labelled as a separate class.
- Through anomaly classification as a precursor to segmentation: A standalone classifier could be built to distinguish between corner and regular cases. This is a challenging research problem since the number of corner-cases could be very few.

D. OTHER POTENTIAL APPROACHES FOR OPTIMAL CHECKPOINT DETERMINATION

In our proposed balanced checkpointing approach, we have suggested to exclude the outlier epochs and chose the final epoch from the remaining epochs to determine a balanced checkpoint so that corner-cases also obtain reasonable results. However, this approach could result in a local-optimum rather than the global optimum. Finding the global optima depends on several factors such as

- the number of corner-cases
- the behaviour of the solution in the corner-cases over the different training epochs
- the behaviour of the solution on the non-corner cases over the different training epochs.

Hence, this is a complex multi-factor optimisation problem which is an area of active research [10], [34], [42].

IX. CONCLUSION AND FUTURE DIRECTIONS

In this research work, we have uncovered a fundamental aspect of deep-learning based segmentation models which has been so far overlooked. Average metrics are indicative of model performances for the majority of the cases. Such approaches tend to overlook the method's performance on the corner-cases. Spotting these corner-cases – or the Achilles' heel of the solution – is crucial when deploying such solutions in a clinical setup.

The strategies we have proposed help to systematically address these challenges. Our framework first helps to easily spot any corner-cases. Additionally, we have elucidated approaches to delve deeper into the specific corner cases and garner further insights. Finally, we have outlined an approach to get a balanced model which yields promising results on the corner-case we identified while also improving average Dice scores.

One possible future direction is to leverage our proposed framework in tasks of biomedical image analysis other than medical image segmentation, such as medical image classification and object detection. The automatic determination of a balanced checkpoint based on global optima is yet another exciting research direction to explore.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA: IEEE Computer Society, Jun. 2015, pp. 3431–3440.

[2] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Honolulu, HI, USA: IEEE Computer Society, Jul. 2017, pp. 5168–5177.

[3] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Washington, DC, USA: IEEE Computer Society, Jun. 2018, pp. 1857–1866.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[5] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "Abstract: nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," in *Bildverarbeitung für die Medizin 2019*. Wiesbaden, Germany: Springer, 2019, p. 22.

[6] K. Rajamani, S. D. Gowda, V. N. Tej, and S. T. Rajamani, "Deformable attention (DANet) for semantic image segmentation," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 3781–3784.

[7] K. T. Rajamani, P. Rani, H. Siebert, R. ElagiriRamalingam, and M. P. Heinrich, "Attention-augmented U-Net (AA-U-Net) for semantic segmentation," *Signal, Image Video Process.*, vol. 17, no. 4, pp. 981–989, Jun. 2023.

[8] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 220–229.

[9] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Res. Notes*, vol. 15, no. 1, pp. 1–8, Dec. 2022.
[10] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Sci. Rep.*, vol. 10, no. 1, Aug. 2020, Art. no. 13724.
[11] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing bias in artificial intelligence in health care," *J. Amer. Med. Assoc.*, vol. 322, no. 24, pp. 2377–2378, 2019.
[12] I. M. E. Naqa, Q. Hu, W. Chen, H. Li, J. D. Fuhrman, N. Gorre, and M. L. Giger, "Lessons learned in transitioning to AI in the medical imaging of COVID-19," *J. Med. Imag.*, vol. 8, no. S1, Oct. 2021, Art. no. 010902.
[13] D. Müller, D. Hartmann, P. Meyer, F. Auer, I. Soto-Rey, and F. Kramer, "MISeval: A metric library for medical image segmentation evaluation," *Stud. Health Technol. Informat.*, vol. 294, pp. 33–37, May 2022, doi: 10.3233/SHTI220391.
[14] V. Yeghiazaryan and I. Voiculescu, "Family of boundary overlap metrics for the evaluation of medical image segmentation," *Proc. SPIE*, vol. 5, no. 1, 2018, Art. no. 015006.
[15] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019.
[16] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 48–56.
[17] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Phys. Medica*, vol. 85, pp. 107–122, May 2021.
[18] A. Reinke et al., "Common limitations of performance metrics in biomedical image analysis," in *Proc. Med. Imag. Deep Learn.*, 2021, pp. 1–3.
[19] A. Reinke et al., "Common limitations of image processing metrics: A picture story," 2021, *arXiv:2104.05642*.
[20] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.
[21] L. Maier-Hein et al., "Metrics reloaded: Recommendations for image analysis validation," 2022, *arXiv:2206.01653*.
[22] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Commun.*, vol. 9, no. 1, pp. 1–13, Dec. 2018.
[23] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, Nov. 1998, pp. 1–4.
[24] A. Kathan, M. Harrer, L. Küster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, I. Grossmann, D. D. Ebert, and B. W. Schuller, "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," *Frontiers Digit. Health*, vol. 4, Nov. 2022, Art. no. 964582.
[25] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth, "Average individual fairness: Algorithms, generalization and experiments," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1–10.
[26] T. Ouyang, V. S. Marco, Y. Isobe, H. Asoh, Y. Oiwa, and Y. Seo, "Corner case data description and detection," in *Proc. IEEE/ACM 1st Workshop AI Eng., Softw. Eng. AI (WAIN)*, May 2021, pp. 19–26.
[27] W. Wu, H. Xu, S. Zhong, M. R. Lyu, and I. King, "Deep validation: Toward detecting real-world corner cases for deep neural networks," in *Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2019, pp. 125–137.
[28] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "SAUNet: Shape attentive U-Net for interpretable medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*. Cham, Switzerland: Springer, 2020, pp. 797–806.
[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015.
[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Washington, DC, USA: IEEE Computer Society, Jul. 2017, pp. 4700–4708.
[31] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*. Washington, DC, USA: IEEE Computer Society, Jul. 2017, pp. 11–19.
[32] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. Chen, "ECOD: Unsupervised outlier detection using empirical cumulative distribution functions," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 16, 2022, doi: 10.1109/TKDE.2022.3159580.
[33] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python toolbox for scalable outlier detection," *J. Mach. Learn. Res.*, vol. 20, no. 96, pp. 1–7, Jan. 2019.
[34] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and Jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Shenzhen, China: Springer, Oct. 2019, pp. 92–100.
[35] O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, and V. I. Madai, "On the usage of average Hausdorff distance for segmentation performance assessment: Hidden error when used for ranking," *Eur. Radiol. Exp.*, vol. 5, no. 1, pp. 1–7, Jan. 2021.
[36] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, Aug. 2015.
[37] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.
[38] Y. Tang, Y. Tang, J. Xiao, and R. M. Summers, "XLSor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2019, pp. 457–467.
[39] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2018, pp. 1–10.
[40] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.
[41] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[42] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: Theory and practice when evaluating with dice score or Jaccard index," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3679–3690, Nov. 2020.

SRIVIDYA TIRUNELLAI RAJAMANI received the master's degree in computer applications from the Amrita School of Engineering, Coimbatore, Tamil Nadu, India, in 2007. She has extensive industry experience as a Solution Architect in the med-tech domain. She is currently a Researcher with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. She has three patents to her credit. Her research interests include machine learning and deep learning methods, medical image segmentation, and affective computing.

KUMAR RAJAMANI received the Ph.D. degree in biomedical engineering from the University of Bern, Switzerland. He was a Postdoctoral Researcher with the Institute of Medical Informatics, University of Lübeck, Germany. He has extensive industrial research experience and was with multinationals, such as Robert Bosch, GE Global Research (GRC), and Philips Research. He has nine patents to his credit. His research interests include medical deep learning, medical image analysis, and health care technologies for emerging markets.

ASHWIN VENKATESHVARAN received the Ph.D. degree in applied medical engineering (specializing in novel clinical applications of cardiac ultrasound) from Kungliga Tekniska Högskolan (KTH), in 2016.

During his postdoctoral research with Karolinska Institutet, from 2017 to 2022, he validated and established novel imaging surrogates that potentially obviate invasive catheterization in a large, multicenter database of patients with heart failure. He is currently a Biomedical Scientist and a Senior Researcher with Lund University, Sweden.

Dr. Venkateshvaran is a fellow of the American Society of Echocardiography and the European Society of Cardiology. For his research that integrate cardiac MR with a multimodal approach to solve real-world clinical issues, he received numerous awards, including the Young Investigator Award, the Clinical Research Award, and the Article of the Year Award for best submission to the *European Heart Journal*, in 2022. He is recognized as an international expert in cardiac imaging.

ANDREAS TRIANTAFYLLOPOULOS received the Diploma degree in ECE from the University of Patras, Greece, in 2017. He is currently pursuing the Ph.D. degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg. His current research interests include deep learning methods for auditory intelligence and affective computing.

ALEXANDER KATHAN received the M.Sc. degree in business analytics from the University of Ulm, Germany, in 2021. He is currently pursuing the Ph.D. degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. His research interests include deep learning and machine learning methods for audio and multimodal signal processing in healthcare applications and personalized machine learning approaches.

BJÖRN W. SCHULLER (Fellow, IEEE) received the Diploma, Ph.D., and Habilitation degrees in electrical engineering and information technology from Technische Universität München, Munich, Germany, in 1999, 2006, and 2012, respectively. He received an Adjunct Teaching Professorship in signal processing and machine intelligence from Technische Universität München, in 2012. He is currently a Professor of AI and the Head of the GLAM, Imperial College London, London, U.K., and a Full Professor and the Head of the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. He co-authored five books and more than 1200 publications in peer-reviewed books, journals, and conference proceedings leading to more than 50000 citations (H-index = 106). He is a fellow of the AAAC, BCS, ELLIS, and ISCA.