From out-of-distribution detection to quality control

Benjamin Lambert, Florence Forbes, Michel Dojat

Chapter 1

From Out-of-distribution detection to Quality Control

ABSTRACT
Quality Control (QC) is an important step of any medical image analysis pipeline to impose safeguards against biased interpretation. Visual QC can be tedious and time-consuming when the volume of data is important and a branch of work has thus focused on providing automated QC algorithms. In the context of computerized image analysis, such algorithms can be categorized according to the domain on which they operate, namely input (i.e. image) or output (i.e. prediction). Input QC is akin to out-of-distribution detection, aiming at the detection of images that are unusual due for example to the presence of artifacts. Output QC, in contrast, focuses on detecting automated predictions that do not meet expectations. These two facets of QC are intertwined, as noisy images are likely to produce poor predictions. However, they are generally considered as separate problems in the literature and tackled with different methodologies and evaluation procedures. In this chapter, a taxonomy of QC methods is first proposed, oriented to input or output checking. Then, a general framework to jointly combine these two QC facets is proposed and illustrated on two tasks, namely binary segmentation of polyps in endoscopic images and multi-class tumor segmentation in multi-modal MRIs.

KEYWORDS
AI Safety, Uncertainty, Out-of-Distribution, Segmentation, Interpretation, Medical imaging

1.1 INTRODUCTION

Quality Control (QC) is an essential step for medical image acquisition and analysis. The poor resolution, the presence of noise or artifacts can indeed greatly complicate the reading of the exam, leading to misdiagnosis, delayed treatment, and increased healthcare costs. Thus, many QC protocols have been proposed in the general context of radiology [111], or dedicated to specific imaging modalities, such as MRI [109], CT [101] and PET [43], or Diffusion Tensor Imaging [64]. These protocols consist of a set of rules that aims at ensuring the reliability of acquired images by the verification of the imaging device as well as the acquisition sequence parameters and processing steps.

On the other side, automated algorithms based on Machine Learning (ML) or Deep Learning (DL) are widely introduced as key elements in medical image

analysis pipelines [106]. For instance, they can automatize fastidious annotation tasks, such as the manual delineation of lesions or anatomical regions. An illustrative example is volBrain, an open-access tool for brain volumetry [69] in T1-w MRI, accessible via an online platform[1], which has performed more than 500, 000 analyses from several thousands of different institutions at the time of writing this article. It provides a precise parcellation of the brain into more than one hundred classes, a process that would take days for a human to perform. However, DL algorithms are not foolproof, and they are known to perform poorly when the input data are *out-of-distribution* (OOD), meaning that it is not drawn from the same distribution as the training data distribution [127, 70, 86, 128]. Modifications in the image acquisition protocol or device can thus lead to unpredictable behaviors. Additionally, DL models are known to be particularly sensitive to adversarial attacks, corresponding to subtle engineered perturbations applied to the input image, which drastically modify the prediction of the network [82, 6, 92]. Thus, when relying on predictive models within an image analysis pipeline, it is desirable to perform QC not only on the input image but also on the automated prediction to make sure that errors are not introduced in the decision-making.

Visual QC, however, is a time-consuming process, prone to inter-rater variability [120], which becomes intractable when the quantity of data to control becomes very large. Yet the tendency is towards an increasing volume of medical images to process [123, 108]. This naturally leads to the replacement of manual QC with automated QC tools.

Automatic QC methods can be categorized into two classes: methods that operate in the input image domain, and methods that operate in the output prediction domain. Another possible denomination is pre-analysis (input) and post-analysis QC [31]. Input QC usually does not make any assumption regarding the downstream task (e.g. registration, classification, or segmentation), and focuses on detecting poor-quality images, generally by using rules defined by human experts to control the objective quality of the data. In contrast, output QC is defined with respect to a target task aiming at detecting poor predictions (e.g. erroneous segmentation or registration). These two aspects are intertwined, as a poor-quality image is more likely to produce a poor prediction, compared to a noise-free one. However, this is not always the case. Indeed, DL models can be trained with a data augmentation procedure to enhance their robustness, which makes them able to correctly generalize to noisy data [104]. Thus, some noisy images can be flagged as incorrect by an input QC algorithm, while leading to a correct prediction. As an illustrative example, Figure 1.1 presents the predictions of a brain tumor segmentation model trained on the BraTS 2023 dataset [73, 9]. The model was trained using a data augmentation strategy including Gaussian noise. At test time, the model was able to process an image corrupted with Gaussian noise with a performance similar to the corresponding

1. https://www.volbrain.net

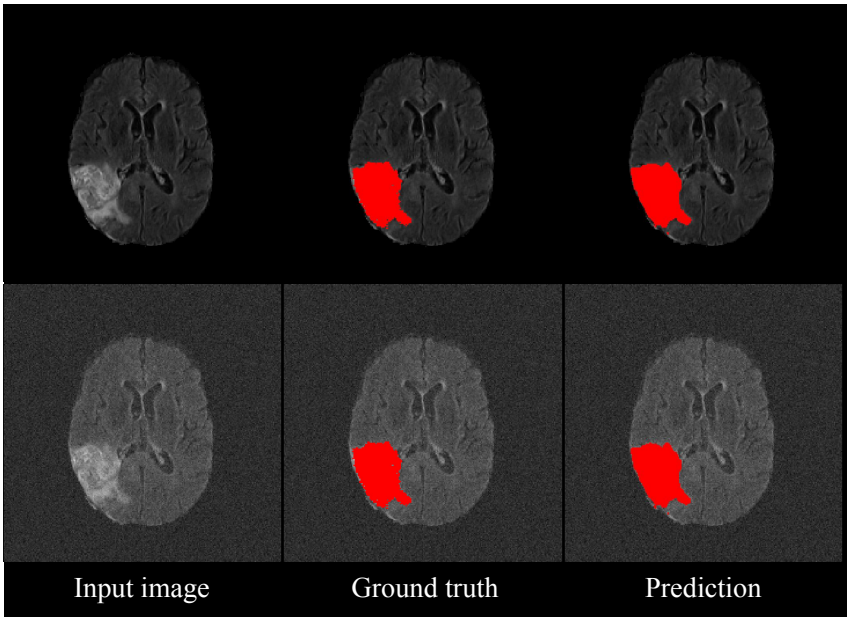

FIGURE 1.1 Brain tumor segmentation for two variants of a T2w FLAIR MRI: the noise-free image (top row) and a noisy image obtained by applying Gaussian Noise (bottom row). The top row segmentation achieves a Dice score of 0.907. In the presence of noise, the segmentation remains valid (bottom row, Dice score of 0.904). The model has been trained using a data augmentation strategy including Gaussian noise, making it robust to this type of noise.

noise-free image. In this setting, raising the alarm for the noisy input image may be considered a false alarm as the model is able to process it seamlessly.

Based on this simple example, it appears that QC can lead to two different conclusions depending on the domain considered (input or output). It can be considered that input and output QC are complementary tasks that can be combined to boost the informativeness of the global QC procedure. For instance, knowing that the image is poorly segmented (failed output QC) **and** out-of-distribution (failed input QC) gives insights regarding the source of errors. Alternatively, when the output QC indicates failures **and not** the input QC, the user may be confronted with images causing in-domain errors that could be included in the training set. However, these two QC facets are generally tackled in separate literature and few works have been proposed to merge the corresponding QC strategies.

In this chapter, an overview of existing methods for both input and output QC is first proposed. In the second part, a practical demonstration of the introduced paradigms is provided in the context of image segmentation.

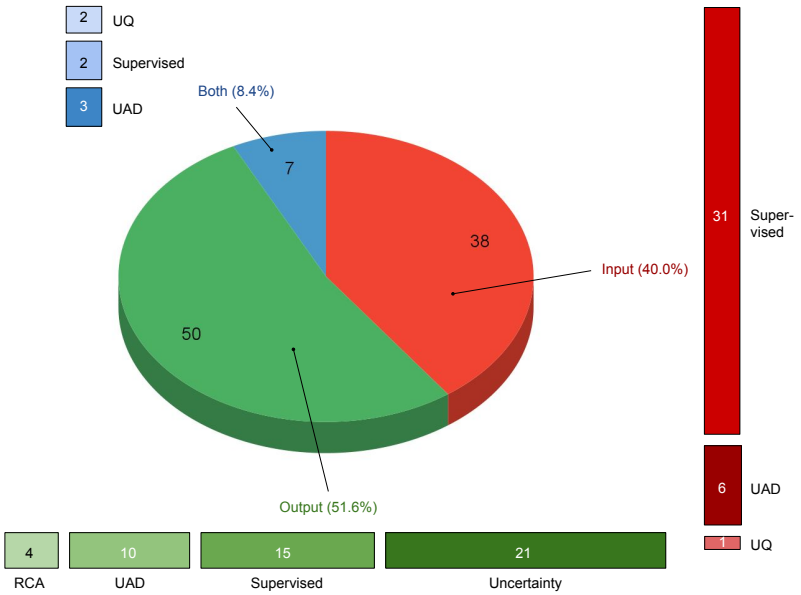

FIGURE 1.2 Pie chart of the studies according to the QC level: input (red), output (green), or both (blue). The corresponding number of papers is mentioned in parentheses. UAD: Unsupervised Anomaly Detection; UQ: Uncertainty Quantification; RCA: Reverse Classification Accuracy

1.2 A TAXONOMY OF QC METHODS FOR ML-BASED MEDICAL IMAGE ANALYSIS

In this section, a review of existing methods for input and/or output QC in the context of medical imaging is presented. To this end, a systematic search was performed on October 2023 using Google Scholar and PubMed to identify papers on QC for medical imaging published from 2016 (included) to October 2023. The following keywords were employed: "Quality Control", "Deep Learning", and "Medical image". Papers were included if they explored automatic QC solutions applied to medical images, resulting in a list of 95 papers selected for reading and analysis. The collected papers were further classified according to the QC method as well as to the QC domain (input, output, or both). Within each category, the frequency of each proposed framework is further reported. The resulting taxonomy is presented in Figure 1.2. In the following of the section, each QC method class is briefly described.

1.2.1 QC on the input image domain

QC on the input image domain (40.0% of the reviewed papers) aims at detecting images that have a poor resolution or present artifacts that make them unqualified

for further processing. The corresponding methods are generally evaluated in a binary classification scenario using a set of images labeled as poor or good-quality samples.

1.2.1.1 Supervised Input QC

The most straightforward and widely-used approach to automatize input QC is to train a classifier (e.g. CNN or Transformer) to separate conform and non-conform images. This requires the building of a large dataset comprising annotated examples of each category. For instance, Bottani et al. [14] propose a CNN to reject poor-quality images from a large brain T1w MRI warehouse. Images were manually labeled into good, medium and bad quality images based on different characteristics: the presence of contrast agent, motion artifact or noise, and the quality of contrast. Additionally, their model was able to detect images that *were not* brain T1w MRI. Similarly, in the QC-Automator [97] framework, authors trained two different CNNs in a supervised fashion to detect artifacts in axial and sagittal slices respectively. The models were developed for diffusion MRIs and were able to distinguish between multi-band interleaving, ghosting, susceptibility, herringbone and chemical-shift artifacts. This supervised approach was also extensively explored for quality assessment of brain MRI [53, 3, 27, 65, 135], diffusion MRI [49, 37], prostate MRI [75] and cardiac MRI [132, 112, 130, 18, 131, 79, 65]. Esses et al. [26] trained a CNN to assess the diagnostic power of liver MRI, while different studies can be found in fetal MRI making use of supervised CNNs [98, 126, 56, 61, 29, 125]. While numerous applications can be found for MR imaging due to the heterogeneity of the data, similar strategies were also investigated for radiography [19, 110, 122, 24], computed tomography (CT) [88, 78], mammography [119] and PET imaging [85].

Instead of directly predicting the quality of the input image in a binary setting (accept or reject), which may lack interpretability in the decision, other methods have been proposed to mimic the process of a human rater performing visual QC. To do so, the classification models are trained to predict a set of relevant image metrics from which a decision (accept or reject) can be derived. An illustrative example is the work of Sun et al. [110], where an input QC model (CNN) is trained to predict QC indicators for knee radiography, namely, the anteroposterior/lateral overlap ratios and flexion angle. These indicators correspond to criteria that a human rater verifies to ensure the correct positioning of the knee in the image [76]. Similarly in the context of lumbar spine radiography, authors derive an automated QC protocol based on criteria defined in radiology textbooks. From the segmentation of the spine, they control the visibility and number of key anatomical features, allowing them to reject non-conform images.

However, such supervised approaches still require access to a sufficient number of noisy data for the training of the DL model and a time-consuming annotation labeling by human experts. Thus, some supervised strategies proposed to use data augmentation to inject artifacts in noise-free data, which allows to easily

construct large databases for automatic input QC. For example, the RegQCNET [102] was developed to detect affine registration errors in brain MRI using simulated spatial transformations for training. Another illustrative example is the work of Zhang et al. [132] for missing slice detection in cardiac MRI. As the training dataset contained only good-quality images, they proposed to remove slices to simulate acquisition problems.

1.2.1.2 Unsupervised Anomaly Detection for Input QC

The Unsupervised Anomaly Detection (UAD) framework corresponds to a set of methods aiming to detect outlier images without requiring a labeled training dataset.

A general UAD methodology is to detect outlier images as deviations from a reference distribution. A typical example of such paradigm for input QC corresponds to latent-space methods, illustrated in Figure 1.3. The principle is as follows: from a trained neural network, intermediate NN activations are collected for a set of in-distribution (artifact-free) data. They are used to model the distribution of the activations of conforming images. The hypothesis is that activations corresponding to non-conform images should deviate from this distribution. At test time, the activations generated for the test sample are compared to the in-distribution distribution by computing a distance metric. A popular choice is the Mahalanobis Distance (MD), widely used for medical image analysis [36, 17, 58] or the L2 distance [22].

Alternatively, self-supervision techniques can be derived to perform QC and can be seen as a form of UAD as they do not require explicit QC labels. Briefly, self-supervision involves training a discriminative model to perform a supervised task (e.g. segmentation or classification) jointly with a fully unsupervised surrogate task (e.g. edge detection) for which the target can directly be derived from the input image, without expert annotation. The rationale is that the performance of this proxy task should be poor when confronted with non-conform input images, allowing for their detection. In Gonzalez et al. [35], authors investigate two self-supervision tasks, namely edge detection and contrastive learning, in order to detect outlier data in the context of cardiac segmentation. Contrastive learning was also further explored by Zuo et al. [136] for artifact detection in brain MRI.

1.2.1.3 Uncertainty-based Input QC

Uncertainty Quantification is increasingly being explored in AI-based medical image processing to detect model misfunctions due for instance to noisy data or learning pitfalls [57]. Traditionally, uncertainty is divided into two categories: aleatoric uncertainty, which is linked to the inherent and irreducible randomness in the input data, and epistemic uncertainty which stems from the lack of knowledge of the predictive model concerning a given input.

In particular, epistemic uncertainty is expected to be high for images that

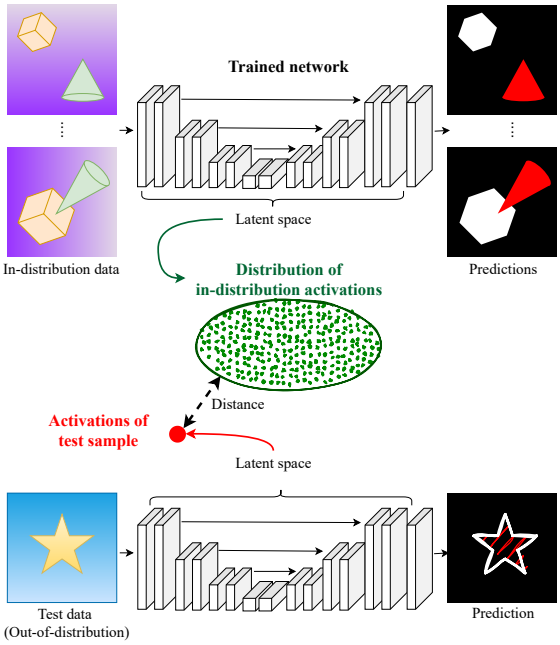

FIGURE 1.3 Illustration of the latent-space approach for out-of-distribution detection.

are different from the ones encountered during training [50]. Thus, monitoring model uncertainty could theoretically be used to detect out-of-distribution samples. This idea was explored in [71] where authors used Monte Carlo dropout [30] to estimate the uncertainty of a brain tumor segmentation model. They further used this uncertainty score to detect bad-quality MRI scans. However, uncertainty is rarely used to perform input QC and only one occurrence of this approach was identified when preparing this review.

1.2.2 QC on the output prediction domain

Evaluating the quality of a prediction is trivial when ground truth data is available. However, the task is much more challenging when no references are available. Automatic QC on the output domain aims at detecting predictions that do not meet a predefined level of quality. For segmentation tasks, one can think of a Dice threshold under which predictions should be disqualified. The metric and threshold are dependent of course on the underlying predictive task and its overall difficulty. Numerous studies have focused on this task (52.6% of the review papers) and the associated frameworks are presented in the following Section.

1.2.2.1 Uncertainty-based Output QC

In contrast to input QC, uncertainty has been more successfully explored for output QC. Pioneer work has been carried out by Roy et al. [95] who used Monte Carlo (MC) dropout to obtain a set of plausible whole-brain segmentation masks for each input image. From this set of images, they propose different metrics that consider the level of agreement between the predictions: the average Dice between samples, the coefficient of variation (CoV), the Intersection over Union between the samples, and the average uncertainty (see Figure 1.4). They showed that these estimators had an important correlation with the segmentation quality, estimated using the Dice score. These scores, easily computable from a set of segmentation masks, could then serve as proxies to estimate the performance of a prediction without ground truth. This framework was further explored and improved in several studies. From the outputs of a MC dropout model, several other proxy metrics were proposed: the Predictive Dice Coefficient [41], the Contour Quality metric [10], the Doubt score [47] or the mean uncertainty [81, 80]. Other variants replaced MC dropout by Test Time Augmentation [116, 117], Ensembling [52, 42, 94, 72, 5, 133], probabilistic models [87] or fuzzy uncertainty [62, 63], which are alternative uncertainty approaches to generate the set of plausible masks.

To further improve the output QC procedure, several works have explored the use of these uncertainty metrics as features to train a ML model to directly infer the prediction quality, in a regression setting. Ghosal et al. [34] and Hann et al. [39, 38, 40] trained linear regression models to predict the Dice directly from uncertainty estimates, for digital histopathology image segmentation and cardiac MRI segmentation, respectively. Alternatively, Arega et al. [7] used a Random Forest (RF) either in a binary classification approach (accept/reject poor segmentation) or regression (predict the Dice score) from the outputs of a MC dropout model. These approaches require building a training dataset comprising automated predictions and associated quality to allow the training of the auxiliary ML model. Such a concept is closely related to the Supervised Output QC approach presented in the next section.

1.2.2.2 Supervised Output QC

As for its input counterpart, supervised output QC is extremely popular. A straightforward idea for supervised output QC is to build a CNN model that takes as input both the image and automated prediction and infers the quality of the prediction, for example by estimating the Dice score [28], or by predicting a QC label (for example, high-quality or poor segmentation) [59]. Additionally to the input image and prediction, studies also incorporated uncertainty maps as input to the output QC model to enhance the Dice prediction [23, 20], while Galdran et al. [33] used the segmentation only as input to a CNN. As an alternative to training directly from the images, Jungo et al. [46] extracted radiomics from uncertainty maps in the context of brain tumor segmentation,

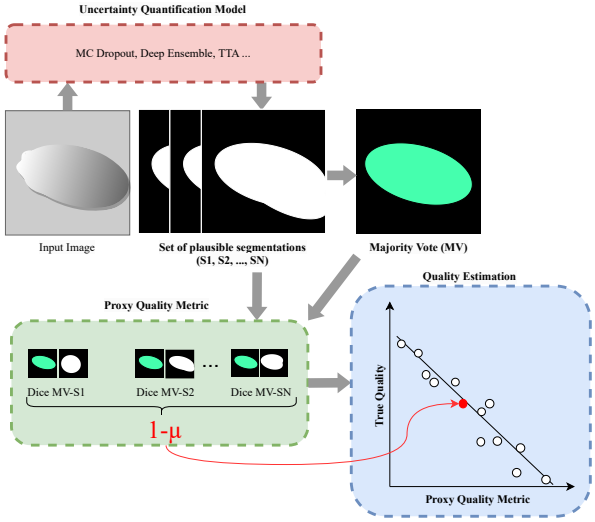

FIGURE 1.4 Illustration of output QC based on the generation of various plausible segmentation masks using an uncertainty framework.

and further trained a RF model to predict Dice scores. A similar approach is adopted by extracting a set of features from the segmentation, followed by the training of a Support Vector Machine to predict the Dice score [51]. Lastly, another approach is to extract a set of meaningful statistical and geometrical features from the segmentations and use them to train a ML classifier to detect failed analyses [2, 103, 25].

While these previous works focus on estimating the overall quality of the segmentation, other studies instead proposed to detect local errors within the segmentation [129, 99]. It is worth mentioning that while most supervised output QC studies are proposed for segmentation tasks, similar techniques can be employed to detect poor image registrations [102, 11].

However, building such a supervised output QC module requires access to an auxiliary database for which automated predictions are available. Additionally, to correctly train the model the dataset should contain a sufficient amount of poor predictions, which however are scarce when fortunately the predictive model has good performances. A potential way to alleviate this is to artificially degrade segmentation masks to mimic segmentation errors during training [33].

1.2.2.3 Unsupervised Anomaly Detection for Output QC

Output QC can be tackled in an unsupervised fashion, without explicitly training a model to distinguish between high-quality and poor predictions using annotated data. For instance, Hui et al. [44] proposed an outlier approach to output QC, by first modeling the features of valid delineations using a parametric distribution.

Then, poor segmentations are detected as outliers to this historic distribution. Alternatively, Audelan et al. [8] proposed an unsupervised Bayesian framework to provide a surrogate segmentation that smoothly follows the visible boundaries of the target ROI (e.g. lesion). The Dice coefficient between the surrogate and automated segmentation is then used as a proxy score for the real Dice. Finally, Luo et al. [66] propose to compute variograms to automatically detect erroneous ground truth landmarks in annotated image registration datasets.

Another popular unsupervised approach is the reconstruction paradigm. The concept is simple: an autoencoder is trained to reconstruct its input using conform data only. At test time, when confronted with non-conform inputs, the autoencoder is supposed to produce large reconstruction errors, allowing to detect the outlier. One idea is to mask the input image using the predicted segmentation, and train an autoencoder to recover the original input from the masked one [134, 121]. Alternatively, the autoencoder can be used to reconstruct the segmentation mask [32]. The hypothesis is that incorrect segmentations should yield to masked images that are difficult to reconstruct. Alternatively, a conditional Generative Adversarial Network (GAN) can be trained to reconstruct the image from its automated segmentation [15]. The differences between the original and synthesized images can then be used to detect segmentation errors. Alternatively, a variational autoencoder is used in [118] to learn a manifold of valid pairs of inputs and outputs, by learning to reconstruct the inputs. At test time, the predicted segmentation is projected to the learned manifold, yielding to a surrogate segmentation. A quality score can then be computed by comparing the original segmentation and the surrogate one. A similar idea is pursued in the CRISP framework making use of contrastive learning to learn a manifold of valid input-segmentation pairs [45] Finally, Nourzadeh et al. [77] used autoencoders trained to reconstruct the posterior probabilities of an ensemble on classification models to detect errors in organ delineations.

1.2.2.4 Reverse Classification Accuracy

Reverse Classification Accuracy (RCA) was first proposed by Valindria et al. [114] and further extended in follow-up studies [90, 89, 91]. The RCA strategy requires a labeled reference database comprising images and ground truth segmentation. Then, for a pair composed of a test image and a prediction whose quality we wish to evaluate, RCA proposes to build a classifier using only the test pair as training data. To do so, the automated segmentation is considered as a pseudo ground truth. The assumption is that if the automated segmentation is of good quality, the trained classifier should be able to segment at least one reference image with high performance. In contrast, if the segmentation is poor, the trained classifier should fail on all reference images. To estimate a quality score, the trained classifier is applied to all images of the reference set to obtain automated segmentation, and a metric (e.g. Dice) is computed using the available reference ground truth. Then, the highest score is used as a proxy estimation

of the true quality. In the original RCA paper, several RCA classifiers have been explored, including Atlas Forests, Deep Learning, and registration [114]. The latter was found to outperform competing approaches. Registration was then further used for RCA [91, 90]. Interestingly, RCA does not require access to a database comprising good and poor segmentation. The only assumption is that training and test data are similar.

1.2.3 Combining Input and Output QC

In Sections 1.2.1 and 1.2.2, the previously proposed frameworks to perform input and output QC were introduced. The following step is to examine whether potential gains could be obtained by addressing input and output QC in a unified way. Few works can be found in the literature in this direction (only 7.4% of the reviewed papers). A straightforward choice is to combine two different QC techniques, one for the input image and one for the prediction. For instance, Machado et al. [67] leverage a supervised CNN for input QC and RCA for prediction quality estimation. Ruijsink et al. [96] detect outlier images using a CNN, while the adequacy of the prediction is estimated using a Support Vector Machine classifier. Alternatively, Comballa et al. [21] rely on the uncertainty quantification paradigm which can be interestingly used to estimate both the confidence in the prediction and to detect outlier images, in the context of dermoscopic image classification.

To go further in associating input and output QC, the pioneering work of Shaw et al. [105] defined MRI quality with respect to the model's ability to provide correct output. Thus, if an image presents an artifact that does not prevent the correct functioning of the model, it should still be considered valid. They further proposed to enhance a segmentation model with a heteroscedastic uncertainty quantification module. It allows the quantification of the noise present in the input images (e.g. motion, blurring, or ghosting artifacts), and the uncertainty score was found to correlate strongly with the Dice coefficient, thus bridging the gap between input and output QC.

The concept of defining non-conform images with respect to the performance of the downstream task is also explored in three recent studies focusing on image segmentation [115, 124, 60]. Instead of defining OOD inputs as images presenting artifacts or missing attributes, they propose to cast OOD images as cases for which the associated segmentation is poor. This redefinition of OOD allows to take into account the generalization capability of the network, which can be robust to certain types of noise, as illustrated in Figure 1.1. On the other hand, it will also be considered as OOD, a conform (noise-free) image when poorly segmented by the model. Vasiliuk et al. [115] further proposed a new metric to evaluate the performance of an OOD detector in this setting, called the Expected Performance Drop (EPD). The principle is to determine the expected performance on clean data used as a target performance. Then, the OOD detector is used to reject samples expected to yield to poor predictions (OOD samples) at

test time. Finally, EPD computes the difference between the target performance and the empiric performance on the remaining test data points. The metric can be minimized by rejecting poorly segmented samples, while correctly segmenting the remaining ones.

For output QC, Lennartz et al. [60] proposed a proxy score called *Segmentation Distortion*, computed from the latent space of a trained segmentation model, which exhibits a strong correlation with the Dice coefficient. Authors compared their metric with MD (introduced in Section 1.2.1.2). They show that MD, which has been designed for input QC, performs poorly at predicting the segmentation quality (output QC). On the other hand, Segmentation Distortion is successful at output QC but performs poorly at detecting shifts in the acquisition device (input QC), while MD performs extremely well. Authors conclude that OOD detection is more suitable for critical applications where strong protection against silent failures is required. However, this may also consider as OOD images that are well processed by the model. Alternatively, output QC is most preferred when the goal is to identify cases that should be reviewed by a human expert.

In the following of this chapter, based on these previous studies, a general framework to encompass both input and output QC is illustrated to enhance QC decision-making.

1.3 PRACTICAL DEMONSTRATION OF A UNIFIED INPUT-OUTPUT QC FOR MEDICAL IMAGE SEGMENTATION

1.3.1 Ensemble-based input and output QC scores

In this section, a practical application of the different tools presented for input and output QC is proposed. The aim of the following experiments is to serve as a concrete illustration of how QC can be implemented to monitor simultaneously the conformity of the input image and its associated prediction. Here, the focus is on medical image segmentation due to its preponderance in the automatized QC literature.

For this proof-of-concept, the emphasis is on QC methods that are i) easy to implement, ii) do not require the training of a dedicated model dedicated to QC, and iii) do not require the building of a labeled dataset of good/bad quality for inputs and outputs, which is cumbersome to obtain for most applications. Instead, QC metrics are seamlessly obtained from an ensemble of DL segmentation models. More formally, let's consider an ensemble of K encoder-decoder segmentation models (e.g. the U-Net [93]). From this ensemble, two QC estimates are computed: the *Globally Normalized Mahalanobis Distance* (GNMD) [17] for input QC, and the *Ensemble Dice Agreement* for output QC, respectively. Both scores are introduced below.

1.3.1.1 Input QC using the Globally Normalized Mahalanobis Distance

The GNMD score is a latent-space approach for input QC [17]. In practice, the GNMD score can be computed from the latent space of a single trained model. The final input-level QC score will correspond to the average score across the K models.

We consider the k-th model of the ensemble, composed of L convolution layers $C_1, ..., C_N$. To compute the GNMD for layer C_i with M_i filters, the initial step is to model the output of the layer by a multivariate Gaussian distribution with mean $\mu_i \in \mathcal{R}^{M_i}$ and covariance matrix $\Sigma_i \in \mathcal{R}^{M_i \times M_i}$. To do so, the feature maps $\phi_{C_i}(x)$ are collected at layer C_i on a set of in-distribution (noise-free) images $\mathcal{S}$. The resulting maps are then reduced over the spatial dimensions:

$$\left\{A^i(x) : x \in \mathcal{S}\right\} \text{ with } A^i(x) = \frac{1}{H}\frac{1}{W}\sum_{h=1}^{H}\sum_{w=1}^{W}\phi_{C_i(x)}(h,w) \tag{1.1}$$

From this set of reduced activations, the layer mean μ_i, layer variance σ_i, and covariance matrix Σ_i are estimated. Then, for a test image x_T, the MD for layer C_i is obtained using:

$$\begin{aligned}\tilde{A}^i(x_T) &= (A^i(x_T) - \mu_i)/\sigma_i \\ \Sigma_i &= \text{Cov}(\tilde{A}^i(x_T)) \\ MD^i(x_T) &= \sqrt{\tilde{A}^i(x_T)^T \Sigma_i^{-1} \tilde{A}^i(x_T)}\end{aligned}$$

To obtain the GNMD for the k-th model, this process is repeated for each convolution layer, independently. As MD scales with the number of filters M, the layer scores are aggregated using a weighted average:

$$GNMD_k(x_T) = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{M_i}MD^i(x_T) \tag{1.2}$$

In the proposed setting, the GNMD is computed for each model of the ensemble, separately. Then the ensemble's GNMD is taken as the average of the individual GNMDs. It is supposed to yield high values for images that are out-of-distribution, as their representations in the latent space of the models are expected to deviate from the learned in-distribution distribution.

1.3.1.2 Output QC using the Ensemble Prediction Agreement

The inter-predictions agreement is a popular proxy metric to estimate the quality of a segmentation, from a set of plausible masks [39, 132, 40, 42]. In the proposed setting, the masks are provided by the ensemble of segmentation models. Each model produces a prediction mask S_k, $k \in [1, ..., K]$. The first step is to compute the majority vote segmentation MV that will be used as the

final ensemble prediction. Then, the overlap between each mask and the majority vote is measured using $\text{Dice}(S_k, MV)$. The final output QC estimate, named Ensemble Predictions Agreement (EPA), corresponds to:

$$EPA = \frac{1}{K} \sum_{i=1}^{K} \text{Dice}(S_k, MV) \tag{1.3}$$

The hypothesis is that if the segmentation is of good quality, the individual segmentations should be stable between the ensemble's members, hence the EPA should be high. In contrast, if the segmentation deviates significantly from one model to the other, it is likely that the prediction is uncertain, and its overall quality should be rather low. To get a score that grows with the degree of nonconformity, $1 - EPA$ is used in practice as the output QC score.

1.3.2 Defining thresholds for input and output QC

The selected input and output QC scores produce continuous values, however, for practical QC it is desired to have binary decisions: either accept or reject. Thus, it is necessary to define thresholds on the scores above which the sample (image or prediction) will be flagged for review. To determine these thresholds, the predictions and QC scores are gathered on a clean validation labeled dataset, representative of conforming inputs and outputs. Following the intuition that non-conform inputs and outputs should be rare events [115], the thresholds for the $GNMD$ and EPA metrics are set as the 95-th percentiles of the validation QC scores.

1.3.3 Stratification of the Prediction space

The prediction space is referred to as the positioning of the test pair (image, prediction) according to their unified QC protocol. More specifically, 4 cases are possible using the proposed protocol (see Figure 1.5), listed below in increasing priority:

- **Region A - Input QC ✓ and Output QC ✓:** optimum operating regime; corresponding to the ideal setting where the image is close to the training distribution and the output prediction is estimated as accurate. It is expected that this subgroup contains the high-quality predictions of the model.
- **Region B - Input QC ✗ and Output QC ✓:** Robust operating regime; corresponding to images that may contain an anomaly (artifact), but for which the output QC is successful. This could represent images that the model is able to process even though their quality is not perfect.
- **Region C - Input QC ✓ and Output QC ✗:** Dysfunctional regime, corresponding to images that have passed the input QC, but for which the prediction is unsure (high EPA). This could represent images that are conform in terms of quality but are still poorly segmented. The reviewing priority for this

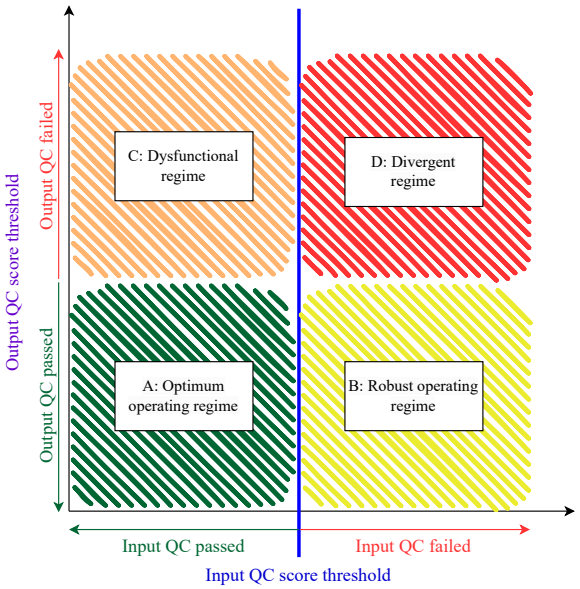

FIGURE 1.5 Proposed stratification of the prediction space using input and output QC estimates.

subgroup is high.

- **Region D - Input QC ✗ and Output QC ✗:** Divergent regime, corresponding to the worst-case scenario where both input and output QC failed. This could represent out-of-distribution images for which the prediction is highly sub-optimal. This subgroup should be reviewed with top priority.

1.3.4 Experiments

The general unified input-output QC protocol is illustrated on two tasks: binary segmentation of polyps in endoscopic images and multi-class tumor segmentation in multi-model brain MRI. For each setting, a synthetic and a real scenario are investigated, causing degradations to the input images, and/or an expected degradation on the output performance.

1.3.4.1 Task 1: binary polyp segmentation in 2D endoscopic images

For this task, a training dataset is built composed of data collected from different sources: Kvasir [84] (1000 images), ETIS-LaribPolyp [107] (196 images), CVC-ColonDB [13] (380 images) and CVC-ClinicDB [12] (612 images). This results in a set of 2188 endoscopic images with associated binary polyp mask, from which a random split is performed in 60% for training (1312 images), 20% for validation (438 images) and 20% (438 images) for in-distribution test, referred

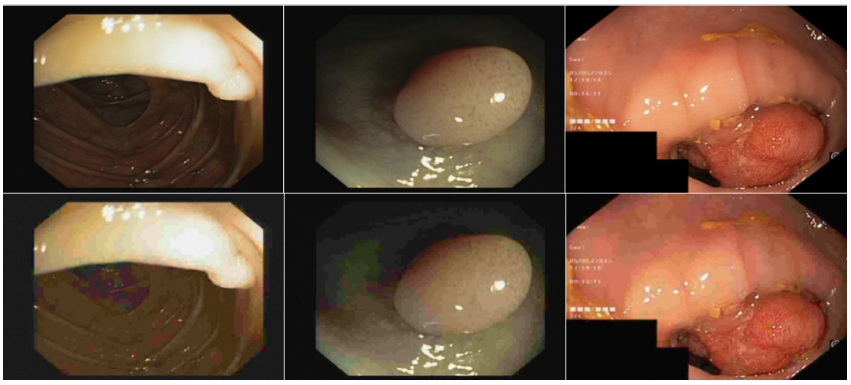

FIGURE 1.6 Samples from 2D endoscopic images. Top row: *Test ID*, initial samples. Bottom row: *Synthetic Degradation*, associated downgraded versions.

to as *Test ID* in the following. All images are resized to a resolution of 768×512.

Two degraded datasets are then explored to test the QC protocol on poor-quality samples (poor inputs and/or poor predictions). The first dataset, *Synthetic Degradation*, is an augmented version of *Test ID*. More precisely, the albumentation library [16] is employed to artificially downgrade the quality of the images by applying the following operators: $Downscale$, $ImageCompression$ and $ISONoise$.

The second scenario explores *domain-shifts* using the PolypGen dataset [4]. This dataset comprises endoscopy images from 6 different centers, exhibiting a heterogeneous population and acquired with different endoscopic systems.

1.3.4.2 Task 2: multi-class tumor segmentation in multi-modal 3D brain MRI

For this task, the large-scale BraTS 2023 dataset [9, 73] is employed to train models that take as inputs 4 MRI sequences (T1w, T2w, T1w with contrast agent, and T2w FLAIR) and segment tumor tissues into 3 classes: necrotic tissue, edematous and enhancing tumor. The open source adult population dataset comprises 1133 patients, randomly split into 60% for training (679 images), 20% for validation (227 images), and 20% for in-distribution testing (227 images) (referred to *Test ID* in the following).

Then, a *Synthetic Degradation* dataset specially tailored for MRI data is created. More specifically, the TorchIO library [83] is employed to downgrade the quality of *Test ID*. The following operators are applied: *RandomAnisotropy* (in each of the 3 axes), and *RandomMotion*. Illustrations of this process are provided in Figure 1.7.

Additionally, the BraTS 2023 dataset also includes various auxiliary datasets to explore domain-shift robustness. To mimic shifts in the population demography, the BraTS-Africa dataset [1] is employed, comprising 60 cases exhibiting

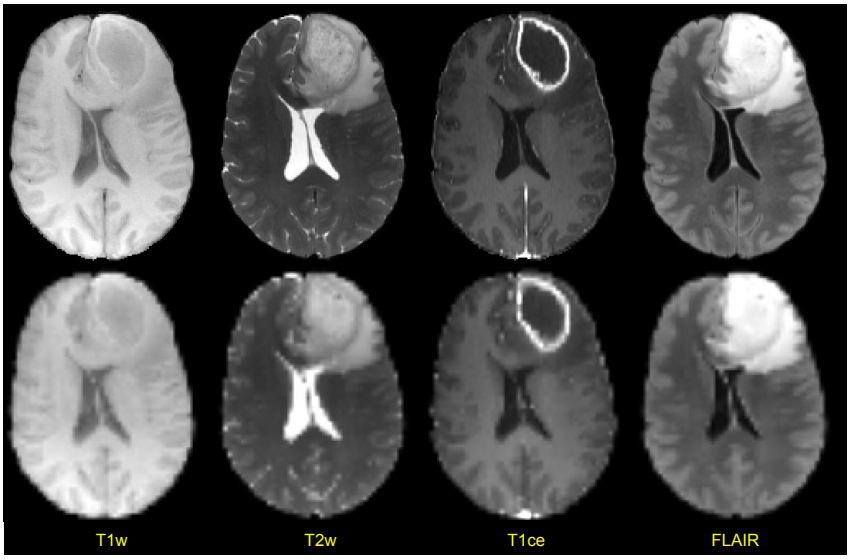

FIGURE 1.7 Samples from the tumor test datasets. Top row: axial views of a patient from *Test ID*, bottom row: axial views of the corresponding degraded patient in *Synthetic Degradation*.

lower MRI quality and more advanced stages of the disease as well as the BraTS-Pediatric dataset [48] comprising 99 cases. To mimic shifts in the observed pathology, the BraTS 2023 Metastasis dataset [74] is used (238 cases) as well as 250 cases from the BraTS 2023 Intracranial Meningioma dataset [54]. These last two datasets contain patients with a type of brain tumor unseen during training, as the Adult dataset only includes glioblastoma. All images are provided pre-processed, including brain extraction and registration of the MRI sequences to a common anatomical template.

1.3.4.3 Evaluation Metrics

To evaluate the quality of a segmentation $\hat{Y}$ with respect to a ground truth annotation Y, two metrics are used. First, the Dice coefficient, which is a popular volumetric overlap score in medical image segmentation defined as:

$$Dice(\hat{Y}, Y) = \frac{2 \times (\hat{Y} \cap Y)}{\hat{Y} \cup Y} \tag{1.4}$$

However, the Dice score is known to be biased toward large segmented volumes [68]. Thus, the Surface Dice (SD) is also used to measure the overlap of two surfaces instead of the overlap of two volumes. SD computes the proportion of the segmentation boundary correctly identified. A point of boundary is considered as correct if the closest distance to the ground truth boundary is

smaller than or equal to a user-level tolerance threshold, defining the acceptable deviation in pixels (or voxels). SD is defined as:

$$SD(\hat{Y}, Y) = \frac{|\mathcal{D}'_Y| + |\mathcal{D}'_{\hat{Y}}|}{|\mathcal{D}_Y| + |\mathcal{D}_{\hat{Y}}|} \tag{1.5}$$

where $\mathcal{D}_Y$ and $\mathcal{D}_{\hat{Y}}$ are the sets of nearest neighbors from the predicted boundary towards the reference and vice versa. $\mathcal{D}'_Y$ and $\mathcal{D}'_{\hat{Y}}$ are the subsets of distances that are smaller than or equal to the tolerance threshold τ:

$$\mathcal{D}'_Y = \{d \in \mathcal{D}_Y | d \leq \tau\} \tag{1.6}$$

For the standard Dice and SD, the metric is computed for each foreground class independently (thus excluding the background class), and the final score is the average of the classes scores. For SD, a tolerance threshold of 3 pixels/voxels is set.

1.3.4.4 Implementation Details

For both tasks, a segmentation ensemble is obtained by aggregating 5 U-Nets independently trained with the same conditions. The implementation of the base network is obtained from the Monai library[2]. Models are trained using a combination of the Dice loss and cross-entropy loss until the Dice score on the validation dataset seizes to improve for 20 epochs. The Adam optimizer is employed with a learning rate of 2×10^{-4}, with a batch size of 8 for the polyp models and a batch size of 1 for brain tumors.

1.3.5 Results

Tables 1.1 and 1.2 present the segmentation metrics on each test dataset and on each of the 4 identified QC regions, for polyps and brain tumors, respectively. Additionally, Figure 1.8 presents the result of the stratification of test datapoints according to their input and output QC estimates. Three visualizations of the same scatter plot are provided: one showing the source dataset of each test point, and the two others that show the true quality of the predictions estimated using Dice and SD scores, respectively.

First, in terms of segmentation performance, both the polyp and brain tumor ensemble achieve high segmentation quality on in-distribution data (Test ID). Then, for the perturbed datasets exhibiting degraded image quality (Synthetic Degradation, BraTS Africa), domain shifts (PolypGen), population shift (BraTS Pediatric), or target shifts (Brats Metastases and Meningioma), the quality of the predictions degrades. For the polyp models, it can be observed that this

2. `https://docs.monai.io/en/stable/_modules/monai/networks/nets/basic_unet.html`

TABLE 1.1 Ensemble segmentation performance on each test dataset for polyp segmentation in endoscopy images. A: Optimal regime, B: Robust regime, C: Dysfunctional regime, D: Divergent regime.

Dataset	N samples	Dice	Surface Dice
Test ID	438	0.87 ± 0.15	0.64 ± 0.25
Synthetic Degradation	438	0.78 ± 0.22	0.48 ± 0.68
PolypGen-Center 1	251	0.80 ± 0.21	0.50 ± 0.26
PolypGen-Center 2	270	0.74 ± 0.26	0.49 ± 0.29
PolypGen-Center 3	456	0.84 ± 0.17	0.61 ± 0.24
PolypGen-Center 4	146	0.56 ± 0.30	0.29 ± 0.22
PolypGen-Center 5	206	0.53 ± 0.34	0.32 ± 0.26
PolypGen-Center 6	83	0.76 ± 0.27	0.52 ± 0.27
Total	2288	-	-
QC Region	**N samples**	**Dice**	**Surface Dice**
A	1333 (58.26%)	0.88 ± 0.13	0.63 ± 0.23
B	358 (15.65%)	0.80 ± 0.23	0.54 ± 0.27
C	225 (9.83%)	0.60 ± 0.23	0.28 ± 0.18
D	372 (16.26%)	0.46 ± 0.29	0.22 ± 0.18

TABLE 1.2 Ensemble segmentation performance on each test dataset for tumor segmentation in brain MRI.A: Optimal regime, B: Robust regime, C: Dysfunctional regime, D: Divergent regime.

Dataset	N samples	Dice	Surface Dice
Test ID	227	0.84 ± 0.13	0.91 ± 0.14
Synthetic Degradation	227	0.78 ± 0.13	0.89 ± 0.15
BraTS-Africa	60	0.74 ± 0.19	0.81 ± 0.20
BraTS-Pediatric	99	0.39 ± 0.23	0.41 ± 0.25
BraTS-Metastases	238	0.55 ± 0.29	0.64 ± 0.33
BraTS-Meningioma	250	0.66 ± 0.34	0.54 ± 0.31
Total	1101	-	-
QC Region	**N samples**	**Dice**	**Surface Dice**
A	736 (66.85%)	0.78 ± 0.17	0.84 ± 0.19
B	85 (7.72%)	0.70 ± 0.18	0.76 ± 0.23
C	204 (18.53%)	0.45 ± 0.35	0.38 ± 0.29
D	76 (6.90%)	0.32 ± 0.33	0.28 ± 0.27

degradation is very heterogeneous depending on the source center. For instance, data from center 3 is segmented with a performance close to the one achieved on Test ID data, while data from center 5 is very poorly segmented. Similarly, for the brain models, the performance on BraTS Africa remains acceptable while the performance collapses on pediatric data.

For both experiments, the general input/output QC strategy allows a stratification of the predictions in 4 different sub-regions (A, B, C and D in Figure 1.8). Predictions in Region A (Optimal operation regime, success of input and output

QC) contain the best quality predictions, with an average performance close to the one achieved on Test ID. Then, B, C, and D regions exhibit performances that are increasingly decreasing, with D containing the worst segmentation.

Interestingly, even though conform in-distribution data are under-represented in our evaluation protocol (438 / 2288 samples for polyps, and 227 / 1101 for brain tumors), the majority of samples are still located in Region A (58.26% of samples for polyps, and 66.26% for brain tumors). This indicates that the segmentation models are able to generalize to some extent to noisy data. The dysfunctional (C) and divergent regimes (D) contain together approximately one-quarter of the data points (26.10% for polyps and 25.43% for brain tumors). This correspond to cases that should be absolutely reviewed by the user.

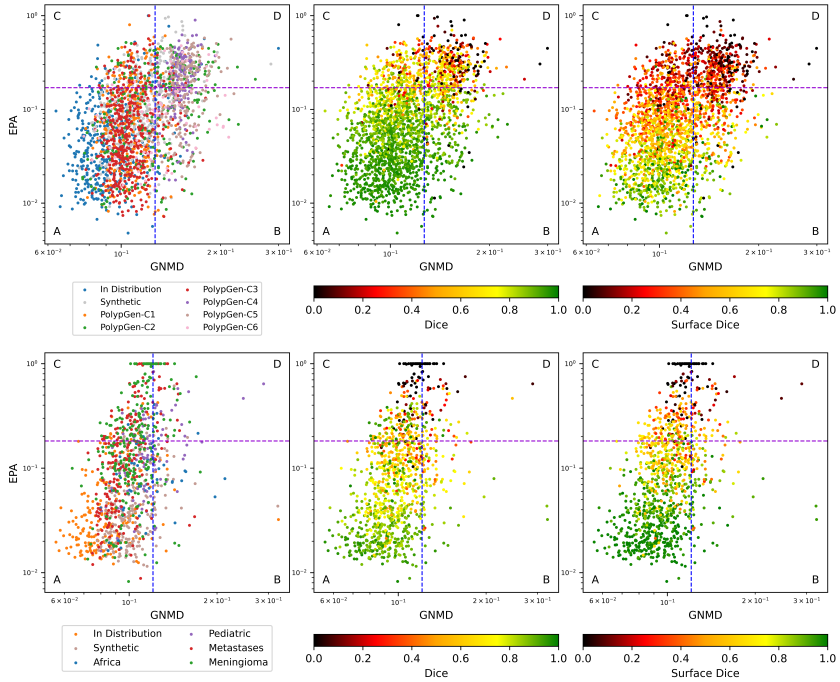

FIGURE 1.8 Stratification of the polyp (top row) and brain tumor (bottom row) prediction spaces in 4 regions according to the input QC scores (x-axis) and output QC scores (y-axis). Left: prediction distribution according to the source dataset. Center and Right: prediction distribution according to the Dice and Surface Dice scores, respectively. The vertical blue dashed line indicates the threshold for the GNMD values, and the horizontal purple dashed line is the threshold for the EPA scores.

1.4 CONCLUSION

In this chapter, the different techniques proposed in the literature to tackle QC in the context of ML-based medical image analysis have been discussed. While

most proposed methods have been proposed to either tackle input QC or output QC, there is a growing interest regarding the merging of both notions. In this direction, Uncertainty Quantification appears as a natural paradigm, allowing the assessment of both data (aleatoric) and model (epistemic) uncertainty. The most popular uncertainty approaches, including Monte Carlo Dropout [30] and Deep Ensemble [55], are easy to implement and are thus a scalable and general solution for QC. However, in practice, it appears that uncertainty is mostly investigated for output QC. This can be explained by the fact that uncertainty estimates have been shown to be poorly calibrated on non-conform inputs, thus they are generally not reliable for input QC [100, 113]. Other techniques address QC more directly, such as the supervised techniques, which were widely investigated for both input and output QC. Their increased performance is at the cost of a dedicated training stage making use of a dataset comprising examples of both poor and good-quality input and/or predictions.

After reviewing the main paradigms for QC, a practical illustration of a unified input/output QC strategy was detailed. This method relies on 2 features: one for estimating the conformity of the input data and one for estimating the quality of the output segmentation. Both metrics are easily obtained from an ensemble of DL models and do not require the training of auxiliary models dedicated to QC. By jointly considering input and output QC, it is possible to stratify the space of predictions into 4 sub-regions. First, the optimal regime contains the best quality predictions. Second, the robust regime which contains samples that the model is able to process smoothly, despite they are far from the training distribution. Third, the dysfunctional regime corresponds to images that are considered as conform, yet the quality of the prediction is doubtful. Finally, the divergent regime corresponds to the case where both input and QC failed and thus encompass the worst quality predictions. The corresponding experiments on medical image segmentation serve as a demonstration of the usefulness of a unified QC strategy, providing additional information to the user and helping prioritize the cases to review.

Bibliography

[1] Maruf Adewole, Jeffrey D Rudie, Anu Gbdamosi, Oluyemisi Toyobo, Confidence Raymond, Dong Zhang, Olubukola Omidiji, Rachel Akinola, Mohammad Abba Suwaid, Adaobi Emegoakor, et al. "The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa)". In: *ArXiv* (2023).

[2] Xenia Alba, Karim Lekadir, Marco Pereanez, Pau Medrano-Gracia, Alistair A Young, and Alejandro F Frangi. "Automatic initialization and quality control of large-scale cardiac MRI segmentations". In: *Medical image analysis* 43 (2018), pp. 129–141.

[3] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. "Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank". In: *Neuroimage* 166 (2018), pp. 400–424.

[4] Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. "A multi-centre polyp detection and segmentation dataset for generalisability assessment". In: *Scientific Data* 10.1 (2023), p. 75.

[5] Natália Alves, Joeran S Bosma, Kiran V Venkadesh, Colin Jacobs, Zaigham Saghir, Maarten de Rooij, John Hermans, and Henkjan Huisman. "Prediction variability to identify reduced AI performance in cancer diagnosis at MRI and CT". In: *Radiology* 308.3 (2023), e230275.

[6] Kyriakos D Apostolidis and George A Papakostas. "A survey on adversarial deep learning robustness in medical image analysis". In: *Electronics* 10.17 (2021), p. 2132.

[7] Tewodros Weldebirhan Arega, Stéphanie Bricq, François Legrand, Alexis Jacquier, Alain Lalande, and Fabrice Meriaudeau. "Automatic uncertainty-based quality controlled T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using Bayesian vision transformer". In: *Medical image analysis* 86 (2023), p. 102773.

[8] Benoıt Audelan and Hervé Delingette. "Unsupervised quality control of image segmentation based on Bayesian learning". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (2019), pp. 21–29.

[9] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification". In: *arXiv preprint arXiv:2107.02314* (2021).

[10] Anjali Balagopal, Dan Nguyen, Howard Morgan, Yaochung Weng, Michael Dohopolski, Mu-Han Lin, Azar Sadeghnejad Barkousaraie, Yesenia Gonzalez, Aurelie Garant, Neil Desai, et al. "A deep learning-based framework for segmenting invisible clinical target volumes with estimated uncertainties for post-operative prostate cancer radiotherapy". In: *Medical image analysis* 72 (2021), p. 102101.

[11] S Bannister, D Page, T Standen, A Dunne, J Rawling, CJ Birch-Sykes, MZ Wilson, S Holloway, J McClelland, and Y Peters. "Deep neural networks for quality assurance of image registration". In: *Medical Imaging with Deep Learning: MIDL 2019* (2019).

[12] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.

[13] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. "Towards automatic polyp detection with a polyp appearance model". In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182.

[14] Simona Bottani, Ninon Burgos, Aurélien Maire, Adam Wild, Sebastian Ströer, Didier Dormont, and Olivier Colliot. "Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse". In: *Medical Image Analysis* 75 (2022), p. 102219.

[15] Irene Brusini, Daniel Ferreira Padilla, José Barroso, Ingmar Skoog, Örjan Smedby, Eric Westman, and Chunliang Wang. "A deep learning-based pipeline for error detection and quality control of brain MRI segmentation results". In: *Medical Imaging with Deep Learning* (2020).

[16] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. "Albumentations: fast and flexible image augmentations". In: *Information* 11.2 (2020), p. 125.

[17] Erdi Calli, Bram Van Ginneken, Ecem Sogancioglu, and Keelin Murphy. "FRODO: An In-Depth Analysis of a System to Reject Outlier Samples From a Trained Neural Network". In: *IEEE Transactions on Medical Imaging* 42.4 (2022), pp. 971–981.

[18] Emily Chan, Ciaran O'Hanlon, Carlota Asegurado Marquez, Marwenie Petalcorin, Jorge Mariscal-Harana, Haotian Gu, Raymond J Kim, Robert M Judd, Phil Chowienczyk, Julia A Schnabel, et al. "Automated Quality Controlled Analysis of 2D Phase Contrast Cardiovascular Magnetic Resonance Imaging". In: (2022), pp. 101–111.

[19] Xiao Chen, Qingshan Deng, Qiang Wang, Xinmiao Liu, Lei Chen, Jinjin Liu, Shuangquan Li, Meihao Wang, and Guoquan Cao. "Image Quality Control in Lumbar Spine Radiography Using Enhanced U-Net Neural Networks". In: *Frontiers in Public Health* 10 (2022), p. 891766.

[20] Xinyuan Chen, Kuo Men, Bo Chen, Yu Tang, Tao Zhang, Shulian Wang, Yexiong Li, and Jianrong Dai. "CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy". In: *Frontiers in Oncology* 10 (2020), p. 524.

[21] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. "Uncertainty estimation in deep neural networks for dermoscopic image classification". In: (2020), pp. 744–745.

[22] Matilde Costa, Sofia C Pereira, João Pedrosa, Ana Maria Mendonça, and Aurélio Campilho. "Deep Feature-Based Automated Chest Radiography Compliance Assessment". In: *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)* (2023), pp. 64–67.

[23] Terrance DeVries and Graham W Taylor. "Leveraging uncertainty estimates for predicting segmentation quality". In: *arXiv preprint arXiv:1807.00502* (2018).

[24] AA Dovganich, AV Khvostikov, AS Krylov, and LE Parolina. "Automatic quality control in lung X-ray imaging with deep learning". In: *Computational Mathematics and Modeling* 32 (2021), pp. 276–285.

[25] Jingwei Duan, Mark E Bernard, James R Castle, Xue Feng, Chi Wang, Mark C Kenamond, and Quan Chen. "Contouring quality assurance methodology based on multiple geometric features against deep learning auto-segmentation". In: *Medical Physics* (2023).

[26] Steven J Esses, Xiaoguang Lu, Tiejun Zhao, Krishna Shanbhogue, Bari Dane, Mary Bruno, and Hersh Chandarana. "Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture". In: *Journal of Magnetic Resonance Imaging* 47.3 (2018), pp. 723–728.

[27] Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O Koyejo, Russell A Poldrack, and Krzysztof J Gorgolewski. "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites". In: *PloS one* 12.9 (2017), e0184661.

[28] Joris Fournel, Axel Bartoli, David Bendahan, Maxime Guye, Monique Bernard, Elisa Rauseo, Mohammed Y Khanji, Steffen E Petersen, Alexis Jacquier, and Badih Ghattas. "Medical image segmentation automatic quality control: A multi-dimensional approach". In: *Medical Image Analysis* 74 (2021), p. 102213.

[29] Borjan Gagoski, Junshen Xu, Paul Wighton, M Dylan Tisdall, Robert Frost, Wei-Ching Lo, Polina Golland, Andre van Der Kouwe, Elfar Adalsteinsson, and P Ellen Grant. "Automated detection and reacquisition of motion-degraded images in fetal HASTE imaging at 3 T". In: *Magnetic resonance in medicine* 87.4 (2022), pp. 1914–1922.

[30] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning* (2016), pp. 1050–1059.

[31] Francesco Galati, Sébastien Ourselin, and Maria A Zuluaga. "From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: a review". In: *Applied Sciences* 12.8 (2022), p. 3936.

[32] Francesco Galati and Maria A Zuluaga. "Efficient model monitoring for quality control in cardiac image segmentation". In: (2021), pp. 101–111.

[33] Adrian Galdran, Pedro Costa, Alessandro Bria, Teresa Araújo, Ana Maria Mendonça, and Aurélio Campilho. "A no-reference quality metric for retinal vessel tree segmentation". In: *International conference on medical image computing and computer-assisted intervention* (2018), pp. 82–90.

[34] Sambuddha Ghosal, Audrey Xie, and Pratik Shah. "Uncertainty quantified deep learning for predicting dice coefficient of digital histopathology image segmentation". In: *arXiv preprint arXiv:2109.00115* (2021).

[35] Camila Gonzalez and Anirban Mukhopadhyay. "Self-supervised Out-of-distribution Detection for Cardiac CMR Segmentation". In: Proceedings of Machine Learning Research 143 (July 2021). Ed. by Mattias Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schläfer, and Floris Ernst, pp. 205–218.

[36] Camila González, Karol Gotkowski, Moritz Fuchs, Andreas Bucher, Armin Dadras, Ricarda Fischbach, Isabel Jasmin Kaltenborn, and Anirban Mukhopadhyay. "Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation". In: *Medical image analysis* 82 (2022), p. 102596.

[37] Mark S Graham, Ivana Drobnjak, and Hui Zhang. "A supervised learning approach for diffusion MRI quality control with minimal training data". In: *NeuroImage* 178 (2018), pp. 668–676.

[38] Evan Hann, Luca Biasiolli, Qiang Zhang, Iulia A Popescu, Konrad Werys, Elena Lukaschuk, Valentina Carapella, Jose M Paiva, Nay Aung, Jennifer J Rayner, et al. "Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 750–758.

[39] Evan Hann, Ricardo A Gonzales, Iulia A Popescu, Qiang Zhang, Vanessa M Ferreira, and Stefan K Piechnik. "Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets". In: *Annual Conference on Medical Image Understanding and Analysis* (2021), pp. 280–293.

[40] Evan Hann, Iulia A Popescu, Qiang Zhang, Ricardo A Gonzales, Ahmet Barutçu, Stefan Neubauer, Vanessa M Ferreira, and Stefan K Piechnik. "Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping". In: *Medical image analysis* 71 (2021), p. 102029.

[41] Yuta Hiasa, Yoshito Otake, Masaki Takao, Takeshi Ogawa, Nobuhiko Sugano, and Yoshinobu Sato. "Automated muscle segmentation from clinical CT using Bayesian U-Net for personalized musculoskeletal modeling". In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1030–1040.

[42] Katharina Hoebel, Vincent Andrearczyk, Andrew Beers, Jay Patel, Ken Chang, Adrien Depeursinge, Henning Müller, and Jayashree Kalpathy-Cramer. "An exploration of uncertainty information for segmentation quality assessment". In: *Medical Imaging 2020: Image Processing* 11313 (2020), pp. 381–390.

[43] Ivalina Hristova, Ronald Boellaard, Paul Galette, Lalitha K Shankar, Yan Liu, Sigrid Stroobants, Otto S Hoekstra, and Wim JG Oyen. "Guidelines for quality control of PET/CT scans in a multicenter clinical study". In: *EJNMMI physics* 4 (2017), pp. 1–15.

[44] Cheukkai B Hui, Hamidreza Nourzadeh, William T Watkins, Daniel M Trifiletti, Clayton E Alonso, Sunil W Dutta, and Jeffrey V Siebers. "Quality assurance tool for organ at risk delineation in radiation therapy using a parametric statistical approach". In: *Medical physics* 45.5 (2018), pp. 2089–2096.

[45] Thierry Judge, Olivier Bernard, Mihaela Porumb, Agisilaos Chartsias, Arian Beqiri, and Pierre-Marc Jodoin. "Crisp-reliable uncertainty estimation for medical image segmentation". In: *International Confer-*

ence on Medical Image Computing and Computer-Assisted Intervention. Springer. 2022, pp. 492–502.

[46] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. "Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation". In: *Frontiers in neuroscience* 14 (2020), p. 282.

[47] Alain Jungo, Raphael Meier, Ekin Ermis, Evelyn Herrmann, and Mauricio Reyes. "Uncertainty-driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation". In: *Medical Imaging with Deep Learning* (2022).

[48] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. "The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs)". In: *ArXiv* (2023).

[49] Christopher Kelly, Max Pietsch, Serena Counsell, and J-Donald Tournier. "Transfer learning and convolutional neural net fusion for motion artefact detection". In: *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine, Honolulu, Hawaii* 3523 (2017).

[50] Alex Kendall and Yarin Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017).

[51] Timo Kohlberger, Vivek Singh, Chris Alvino, Claus Bahlmann, and Leo Grady. "Evaluating segmentation error without ground truth". In: (2012), pp. 528–536.

[52] Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. "Layer ensembles: A single-pass uncertainty estimation in deep learning for segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), pp. 514–524.

[53] Thomas Küstner, Annika Liebgott, Lukas Mauch, Petros Martirosian, Fabian Bamberg, Konstantin Nikolaou, Bin Yang, Fritz Schick, and Sergios Gatidis. "Automated reference-free detection of motion artifacts in magnetic resonance images". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 31 (2018), pp. 243–256.

[54] Dominic LaBella, Maruf Adewole, Michelle Alonso-Basanta, Talissa Altes, Syed Muhammad Anwar, Ujjwal Baid, Timothy Bergquist, Radhika Bhalerao, Sully Chen, Verena Chung, Gian-Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, Devon Godfrey, Fathi Hilal, Ariana Familiar, Keyvan Farahani, Juan Eugenio Iglesias, Zhifan Jiang, Elaine

Johanson, Anahita Fathi Kazerooni, Collin Kent, John Kirkpatrick, Florian Kofler, Koen Van Leemput, Hongwei Bran Li, Xinyang Liu, Aria Mahtabfar, Shan McBurney-Lin, Ryan McLean, Zeke Meier, Ahmed W Moawad, John Mongan, Pierre Nedelec, Maxence Pajot, Marie Piraud, Arif Rashid, Zachary Reitman, Russell Takeshi Shinohara, Yury Velichko, Chunhao Wang, Pranav Warman, Walter Wiggins, Mariam Aboian, Jake Albrecht, Udunna Anazodo, Spyridon Bakas, Adam Flanders, Anastasia Janas, Goldey Khanna, Marius George Linguraru, Bjoern Menze, Ayman Nada, Andreas M Rauschecker, Jeff Rudie, Nourel Hoda Tahon, Javier Villanueva-Meyer, Benedikt Wiestler, and Evan Calabrese. *The ASNR-MICCAI Brain Tumor Segmentation (BraTS) Challenge 2023: Intracranial Meningioma*. 2023. arXiv: 2305.07642 [cs.CV].

[55] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).

[56] Sayeri Lala, Nalini Singh, Borjan Gagoski, Esra Turk, P Ellen Grant, Polina Golland, and Elfar Adalsteinsson. "A deep learning approach for image quality assessment of fetal brain MRI". In: *Proceedings of the 27th Annual Meeting of ISMRM, Montréal, Québec, Canada* (2019), p. 839.

[57] Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis". In: *Artificial Intelligence in Medicine* 150 (2024), p. 102830. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2024.102830. URL: https://www.sciencedirect.com/science/article/pii/S0933365724000721.

[58] Benjamin Lambert, Florence Forbes, Senan Doyle, and Michel Dojat. "Multi-layer Aggregation as a key to feature-based OOD detection". In: *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (2023), pp. 104–114.

[59] Ho Hin Lee, Yucheng Tang, Olivia Tang, Yuchen Xu, Yunqiang Chen, Dashan Gao, Shizhong Han, Riqiang Gao, Michael R Savona, Richard G Abramson, et al. "Semi-supervised multi-organ segmentation through quality assurance supervision". In: *Medical Imaging 2020: Image Processing*. Vol. 11313. SPIE. 2020, pp. 363–369.

[60] Jonathan Lennartz and Thomas Schultz. "Segmentation Distortion: Quantifying Segmentation Uncertainty Under Domain Shift via the Effects of Anomalous Activations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), pp. 316–325.

[61] Lufan Liao, Xin Zhang, Fenqiang Zhao, Tao Zhong, Yuchen Pei, Xiangmin Xu, Li Wang, He Zhang, Dinggang Shen, and Gang Li. "Joint image quality assessment and brain extraction of fetal MRI using deep learning". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23* (2020), pp. 415–424.

[62] Qiao Lin, Xin Chen, Chao Chen, and Jonathan M Garibaldi. "A Novel Quality Control Algorithm for Medical Image Segmentation Based on Fuzzy Uncertainty". In: *IEEE Transactions on Fuzzy Systems* (2022).

[63] Qiao Lin, Xin Chen, Chao Chen, and Jonathan M Garibaldi. "Quality quantification in deep convolutional neural networks for skin lesion segmentation using fuzzy uncertainty measurement". In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2022, pp. 1–8.

[64] Zhexing Liu, Yi Wang, Guido Gerig, Sylvain Gouttard, Ran Tao, Thomas Fletcher, and Martin Styner. "Quality control of diffusion weighted images". In: *Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications* 7628 (2010), pp. 137–145.

[65] Benedikt Lorch, Ghislain Vaillant, Christian Baumgartner, Wenjia Bai, Daniel Rueckert, and Andreas Maier. "Automated detection of motion artefacts in MR imaging using decision forests". In: *Journal of medical engineering* 2017 (2017).

[66] Jie Luo, Guangshen Ma, Nazim Haouchine, Zhe Xu, Yixin Wang, Tina Kapur, Lipeng Ning, William M Wells III, and Sarah Frisken. "On the dataset quality control for image registration evaluation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 36–45.

[67] Inês Machado, Esther Puyol-Antón, Kerstin Hammernik, Gastão Cruz, Devran Ugurlu, Bram Ruijsink, Miguel Castelo-Branco, Alistair Young, Claudia Prieto, Julia A Schnabel, et al. "Quality-aware cine cardiac MRI reconstruction and analysis from undersampled k-space data". In: (2021), pp. 12–20.

[68] Lena Maier-Hein, Bjoern Menze, et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation". In: *arXiv. org* 2206.01653 (2022).

[69] José V Manjón and Pierrick Coupé. "volBrain: an online MRI brain volumetry system". In: *Frontiers in neuroinformatics* 10 (2016), p. 30.

[70] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. "The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study". In: *Medical Image Analysis* 66 (2020), p. 101714.

[71] Patrick McClure, Nao Rho, John A Lee, Jakub R Kaczmarzyk, Charles Y Zheng, Satrajit S Ghosh, Dylan M Nielson, Adam G Thomas, Peter Bandettini, and Francisco Pereira. "Knowing what you know in brain segmentation using Bayesian deep neural networks". In: *Frontiers in neuroinformatics* 13 (2019), p. 67.

[72] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation". In: *IEEE transactions on medical imaging* 39.12 (2020), pp. 3868–3878.

[73] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[74] Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Leon Jekel, Kiril Krantchev, Harrison Moy, Rachit Saluja, Klara Osenberg, Klara Wilms, et al. "The Brain Tumor Segmentation (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI". In: *ArXiv* (2023).

[75] Michael C Muelly, Paul B Stoddard, and Shreyas S Vasanwala. "Automated quality control of mr images using deep convolutional neural networks". In: *Proceedings of international society for magnetic resonance in medicine (ISMRM), Honolulu, USA* (2017), pp. 1–3.

[76] Gregory M Mundis, Behrooz A Akbarnia, and Frank M Phillips. "Adult deformity correction through minimally invasive lateral approach techniques". In: *Spine* 35.26S (2010), S312–S321.

[77] Hamidreza Nourzadeh, Cheukkai Hui, Mahmoud Ahmad, Nasrin Sadeghzadehyazdi, William T Watkins, Sunil W Dutta, Clayton E Alonso, Daniel M Trifiletti, and Jeffrey V Siebers. "Knowledge-based quality control of organ delineations in radiation therapy". In: *Medical physics* 49.3 (2022), pp. 1368–1381.

[78] Katri Nousiainen, Teemu Mäkelä, Anneli Piilonen, and Juha I Peltonen. "Automating chest radiograph imaging quality control". In: *Physica Medica* 83 (2021), pp. 138–145.

[79] Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, James R Clough, Gastao Cruz, Aurelien Bustin, Claudia Prieto, Rene Botnar, Daniel Rueckert, Julia A Schnabel, et al. "Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning". In: *Medical image analysis* 55 (2019), pp. 136–147.

[80] José Ignacio Orlando, Philipp Seeböck, Hrvoje Bogunović, Sophie Klimscha, Christoph Grechenig, Sebastian Waldstein, Bianca S Gerendas, and Ursula Schmidt-Erfurth. "U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1441–1445.

[81] Huitong Pan, Yushan Feng, Quan Chen, Craig Meyer, and Xue Feng. "Prostate segmentation from 3d mri using a two-stage model and variable-input based uncertainty measure". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 468–471.

[82] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. "Generalizability vs. robustness: investigating medical imaging networks using adversarial examples". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I* (2018), pp. 493–501.

[83] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106236.

[84] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection". In: *Proceedings of the 8th ACM on Multimedia Systems Conference* (2017), pp. 164–169.

[85] Antonella D Pontoriero, Giovanna Nordio, Rubaida Easmin, Alessio Giacomel, Barbara Santangelo, Sameer Jahuar, Ilaria Bonoldi, Maria Rogdaki, Federico Turkheimer, Oliver Howes, et al. "Automated data quality control in FDOPA brain PET imaging using deep learning". In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106239.

[86] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification". In: *Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2* (2020), pp. 74–83.

[87] Esther Puyol-Antón, Bram Ruijsink, Christian F Baumgartner, Pier-Giorgio Masci, Matthew Sinclair, Ender Konukoglu, Reza Razavi, and Andrew P King. "Automated quantification of myocardial tissue characteristics from native T 1 mapping using neural networks with uncertainty-based quality-control". In: *Journal of Cardiovascular Magnetic Resonance* 22 (2020), pp. 1–15.

[88] Anthony P Reeves, Yiting Xie, and Shuang Liu. "Automated image quality assessment for chest CT scans". In: *Medical physics* 45.2 (2018), pp. 561–578.

[89] Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V Valindria, Mihir M Sanghvi, Nay Aung, José M Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, et al. "Real-time prediction of segmentation quality". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11* (2018), pp. 578–585.

[90] Robert Robinson, Vanya V Valindria, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M Sanghvi, Nay Aung, José Miguel Paiva, Filip Zemrak, et al. "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study". In: *Journal of Cardiovascular Magnetic Resonance* 21.1 (2019), pp. 1–14.

[91] Robert Robinson, Vanya V Valindria, Wenjia Bai, Hideaki Suzuki, Paul M Matthews, Chris Page, Daniel Rueckert, and Ben Glocker. "Automatic quality control of cardiac MRI segmentation in large-scale population imaging". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20* (2017), pp. 720–727.

[92] David Rodriguez, Tapsya Nayak, Yidong Chen, Ram Krishnan, and Yufei Huang. "On the role of deep learning model complexity in adversarial robustness for medical images". In: *BMC Medical Informatics and Decision Making* 22.2 (2022), pp. 1–15.

[93] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (2015), pp. 234–241.

[94] Sarahi Rosas-Gonzalez, Taibou Birgui-Sekou, Moncef Hidane, Ilyess Zemmoura, and Clovis Tauber. "Asymmetric ensemble of asymmetric U-net models for brain tumor segmentation with uncertainty estimation". In: *Frontiers in Neurology* 12 (2021), p. 609646.

[95] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control". In: *NeuroImage* 195 (2019), pp. 11–22.

[96] Bram Ruijsink, Esther Puyol-Antón, Ilkay Oksuz, Matthew Sinclair, Wenjia Bai, Julia A Schnabel, Reza Razavi, and Andrew P King. "Fully automated, quality-controlled cardiac analysis from CMR: validation and large-scale application to characterize cardiac function". In: *Cardiovascular Imaging* 13.3 (2020), pp. 684–695.

[97] Zahra Riahi Samani, Jacob Antony Alappatt, Drew Parker, Abdol Aziz Ould Ismail, and Ragini Verma. "QC-Automator: Deep learning-based automated quality control for diffusion mr images". In: *Frontiers in neuroscience* 13 (2020), p. 1456.

[98] Thomas Sanchez, Oscar Esteban, Yvan Gomez, Elisenda Eixarch, and Meritxell Bach Cuadra. "FetMRQC: Automated Quality Control for Fetal Brain MRI". In: *Perinatal, Preterm and Paediatric Image Analysis* (2023). Ed. by Daphna Link-Sourani, Esra Abaci Turk, Christopher Macgowan, Jana Hutter, Andrew Melbourne, and Roxane Licandro, pp. 3–16.

[99] Jörg Sander, Bob D de Vos, and Ivana Išgum. "Automatic segmentation with detection of local segmentation failures in cardiac MRI". In: *Scientific Reports* 10.1 (2020), p. 21769.

[100] Adrian Schwaiger, Poulami Sinhamahapatra, Jens Gansloser, and Karsten Roscher. "Is uncertainty quantification in deep learning sufficient for out-of-distribution detection?" In: *Aisafety@ ijcai* 54 (2020).

[101] Euclid Seeram. "Computed Tomography-E-Book: Physical Principles, Patient Care, Clinical Applications, and Quality Control". In: (2022).

[102] Baudouin Denis de Senneville, Jose V Manjon, and Pierrick Coupé. "RegQCNET: Deep quality control for image-to-template brain MRI affine registration". In: *Physics in Medicine & Biology* 65.22 (2020), p. 225022.

[103] Reuben R. Shamir and Ze'ev Bomzon. "Evaluation of head segmentation quality for treatment planning of tumor treating fields in brain tumors". In: *ArXiv* abs/1906.11014 (2019). URL: https://api.semanticscholar.org/CorpusID:195658144.

[104] Richard Shaw, Carole Sudre, Sebastien Ourselin, and M. Jorge Cardoso. "MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty". In: *Medical Imaging with Deep Learning* (2019).

[105] Richard Shaw, Carole H Sudre, Sebastien Ourselin, M Jorge Cardoso, and Hugh G Pemberton. "A decoupled uncertainty model for mri segmentation quality estimation". In: *Journal of Machine Learning for Biomedical Imaging* (2021).

[106] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[107] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer". In: *International journal of computer assisted radiology and surgery* 9 (2014), pp. 283–293.

[108] Rebecca Smith-Bindman, Marilyn L. Kwan, Emily C. Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y. Cheng, Erin J. A. Bowles, James R. Duncan, Robert T. Greenlee, Lawrence H. Kushi, Jason D. Pole, Alanna K. Rahm, Natasha K. Stout, Sheila Weinmann, and Diana L. Miglioretti. "Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016". In: *JAMA* 322.9 (Sept. 2019), pp. 843–856.

[109] Gayathri Sreedher, Mai-Lan Ho, Mark Smith, Unni K Udayasankar, Seretha Risacher, Otto Rapalino, Mary-Louise C Greer, Andrea S Doria, and Michael S Gee. "Magnetic resonance imaging quality control, quality assurance and quality improvement". In: *Pediatric Radiology* 51 (2021), pp. 698–708.

[110] Hongbiao Sun, Wenwen Wang, Fujin He, Duanrui Wang, Xiaoqing Liu, Shaochun Xu, Baolian Zhao, Qingchu Li, Xiang Wang, Qinling Jiang, et al. "An AI-Based Image Quality Control Framework for Knee Radiographs". In: *Journal of Digital Imaging* (2023), pp. 1–12.

[111] Stephen J Swensen and C Daniel Johnson. "Radiologic quality and safety: mapping value into radiology". In: *Journal of the American College of Radiology* 2.12 (2005), pp. 992–1000.

[112] Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Antonio De Marvao, Declan P O'Regan, Stuart Cook, Ben Glocker, et al. "Learning-based quality control for cardiac MR images". In: *IEEE transactions on medical imaging* 38.5 (2018), pp. 1127–1138.

[113] Dennis Ulmer and Giovanni Cinà. "Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection". In: *Uncertainty in Artificial Intelligence* (2021), pp. 1766–1776.

[114] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. "Reverse classification accuracy: predicting segmentation performance in the absence of ground truth". In: *IEEE transactions on medical imaging* 36.8 (2017), pp. 1597–1606.

[115] Anton Vasiliuk, Daria Frolova, Mikhail Belyaev, and Boris Shirokikh. "Redesigning Out-of-Distribution Detection on 3D Medical Images". In: *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (2023), pp. 126–135.

[116] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". In: *Neurocomputing* 338 (2019), pp. 34–45.

[117] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. "Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation". In: *Frontiers in computational neuroscience* 13 (2019), p. 56.

[118] Shuo Wang, Giacomo Tarroni, Chen Qin, Yuanhan Mo, Chengliang Dai, Chen Chen, Ben Glocker, Yike Guo, Daniel Rueckert, and Wenjia Bai. "Deep generative model-based quality control for cardiac MRI segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23* (2020), pp. 88–97.

[119] Haruyuki Watanabe, Saeko Hayashi, Yohan Kondo, Eri Matsuyama, Norio Hayashi, Toshihiro Ogura, and Masayuki Shimosegawa. "Quality control system for mammographic breast positioning using deep learning". In: *Scientific Reports* 13.1 (2023), p. 7066.

[120] Brendan Williams, Nicholas Hedger, Carolyn B. McNabb, Gabriella M. K. Rossetti, and Anastasia Christakou. "Inter-rater reliability of functional MRI data quality control assessments: A standardised protocol and practical guide using pyfMRIqc". In: *Frontiers in Neuroscience* 17 (2023).

[121] Elena Williams, Sebastian Niehaus, Janis Reinelt, Alberto Merola, Paul Glad Mihai, Kersten Villringer, Konstantin Thierbach, Evelyn Medawar, Daniel Lichterfeld, Ingo Roeder, et al. "Automatic quality control framework for more reliable integration of machine learning-based image segmentation into medical workflows". In: *arXiv preprint arXiv:2112.03277* (2021).

[122] Charles E Willis, Thomas K Nishino, Jered R Wells, H Asher Ai, Joshua M Wilson, and Ehsan Samei. "Automated quality control assessment of clinical chest images". In: *Medical physics* 45.10 (2018), pp. 4377–4391.

[123] Mateusz Winder, Aleksander Jerzy Owczarek, Jerzy Chudek, Joanna Pilch-Kowalczyk, and Jan Baron. "Are We Overdoing It? Changes in Diagnostic Imaging Workload during the Years 2010–2020 including the Impact of the SARS-CoV-2 Pandemic". In: *Healthcare* 9.11 (2021).

[124] McKell Woodland, Nihil Patel, Mais Al Taie, Joshua P. Yung, Tucker J. Netherton, Ankit B. Patel, and Kristy K. Brock. "Dimensionality Reduction for Improving Out-of-Distribution Detection in Medical Image Segmentation". In: *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (2023), pp. 147–156.

[125] Lingyun Wu, Jie-Zhi Cheng, Shengli Li, Baiying Lei, Tianfu Wang, and Dong Ni. "FUIQA: fetal ultrasound image quality assessment with deep convolutional networks". In: *IEEE transactions on cybernetics* 47.5 (2017), pp. 1336–1349.

[126] Junshen Xu, Sayeri Lala, Borjan Gagoski, Esra Abaci Turk, P Ellen Grant, Polina Golland, and Elfar Adalsteinsson. "Semi-supervised learning for fetal brain MRI quality assessment with ROI consistency". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23* (2020), pp. 386–395.

[127] Zhiyun Xue, Feng Yang, Sivaramakrishnan Rajaraman, Ghada Zamzmi, and Sameer Antani. "Cross Dataset Analysis of Domain Shift in CXR Lung Region Detection". In: *Diagnostics* 13.6 (2023), p. 1068.

[128] Wenjun Yan, Yuanyuan Wang, Shengjia Gu, Lu Huang, Fuhua Yan, Liming Xia, and Qian Tao. "The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22* (2019), pp. 623–631.

[129] Fahim Ahmed Zaman, Lichun Zhang, Honghai Zhang, Milan Sonka, and Xiaodong Wu. "Segmentation quality assessment by automated detection of erroneous surface regions in medical images". In: *Computers in Biology and Medicine* 164 (2023), p. 107324.

[130] Le Zhang, Ali Gooya, Bo Dong, Rui Hua, Steffen E Petersen, Pau Medrano-Gracia, and Alejandro F Frangi. "Automated quality assessment of cardiac MR images using convolutional neural networks". In: (2016), pp. 138–145.

[131] Le Zhang, Ali Gooya, and Alejandro F Frangi. "Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets". In: (2017), pp. 61–68.

[132] Qiang Zhang, Evan Hann, Konrad Werys, Cody Wu, Iulia Popescu, Elena Lukaschuk, Ahmet Barutcu, Vanessa M Ferreira, and Stefan K Piechnik. "Deep learning with attention supervision for automated motion artefact detection in quality control of cardiac T1-mapping". In: *Artificial intelligence in medicine* 110 (2020), p. 101955.

[133] Xiaoyan Zhang, Alvaro E Ulloa Cerna, Joshua V Stough, Yida Chen, Brendan J Carry, Amro Alsaid, Sushravya Raghunath, David P VanMaanen, Brandon K Fornwalt, and Christopher M Haggerty. "Generalizability and quality control of deep learning-based 2D echocardiography segmentation models in a large clinical dataset". In: *The International Journal of Cardiovascular Imaging* 38.8 (2022), pp. 1685–1697.

[134] Leixin Zhou, Wenxiang Deng, and Xiaodong Wu. "Robust Image Segmentation Quality Assessment". In: *Medical Imaging with Deep Learning* (2020).

[135] Dženan Zukić, Anne Haley, Curtis Lisle, James Klo, Kilian M Pohl, Hans J Johnson, and Aashish Chaudhary. "Medical Image Quality Assurance using Deep Learning". In: (2022).

[136] Lianrui Zuo, Yuan Xue, Blake E Dewey, Yihao Liu, Jerry L Prince, and Aaron Carass. "A latent space for unsupervised MR image quality control via artifact assessment". In: *Medical Imaging 2023: Image Processing* 12464 (2023), pp. 278–283.