IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 3, 2021, accepted December 25, 2021, date of publication January 4, 2022, date of current version January 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140378

Automated Annotator Variability Inspection for Biomedical Image Segmentation

MARCEL P. SCHILLING, TIM SCHERR, FRIEDRICH R. MÜNKE, OLIVER NEUMANN, MARK SCHUTERA, RALF MIKUT, AND MARKUS REISCHL
Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen, Germany

Corresponding author: Marcel P. Schilling (marcel.schilling@kit.edu)

This work was supported in part by the KIT Future Fields Project ''Screening Platform for Personalized Oncology (SPPO),'' in part by the HoreKa Supercomputer through the Ministry of Science, Research, and the Arts Baden-Württemberg, in part by the Federal Ministry of Education and Research, in part by the Helmholtz Association's Initiative and Networking Fund through the Helmholtz AI and the Helmholtz Imaging Platform, and in part by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

ABSTRACT Supervised deep learning approaches for automated diagnosis support require datasets annotated by experts. Intra-annotator variability of a single annotator and inter-annotator variability between annotators can affect the quality of the diagnosis support. As medical experts will always differ in annotation details, quantitative studies concerning the annotation quality are of particular interest. A consistent and noise-free annotation of large-scale datasets by, for example, dermatologists or pathologists is a current challenge. Hence, methods are needed to automatically inspect annotations in datasets. In this paper, we categorize annotation noise in image segmentation tasks, present methods to simulate annotation noise, and examine the impact on the segmentation quality. Two novel automated methods to identify intra-annotator and inter-annotator inconsistencies based on uncertainty-aware deep neural networks are proposed. We demonstrate the benefits of our automated inspection methods such as focused re-inspection of noisy annotations or the detection of generally different annotation styles using the biomedical ISIC 2017 Melanoma image segmentation dataset.

INDEX TERMS Artificial neural networks, automation, machine learning, segmentation, image processing.

I. INTRODUCTION

Recent research in the domain of diagnosis support focuses on Deep Learning (DL) to solve computer vision tasks. Especially in complex scenarios, Deep Neural Networks (DNNs) can outperform traditional image processing methods [1] and are extensively used in biomedical applications [2]–[8].

For the training of state-of-the-art supervised DNNs, image annotation by medical experts is crucial. Following the arguments in [1], [9], annotation is a general bottleneck in the context of DL since it is time-consuming, burdensome, expensive, and often requires subject-specific knowledge, i.e., by dermatologists or pathologists. As time is short and opinions by medical experts often differ, a challenge is the consistent annotation of datasets and the prevention of noise during annotation. Annotation noise is used as a collective term and includes partially false annotations, complete annotation errors, or biases. In contrast, in this article, noise is not used for classical stochastic deviations according to the understanding in measurement and control theory. Following Karimi *et al.* [10], there are two main reasons for annotation noise: (i) intra-annotator variability, occurring within the annotation of an individual annotator, and (ii) inter-annotator variability between different annotators. Noise can be based on different decision boundaries. For instance, one dermatologist prefers drawing oversized melanomas, whereas another dermatologist focuses on rather too small melanomas. Moreover, different hardware, e.g. properties of used monitors, qualification of annotators, spent annotation time, fatigue, or the available amount of time affect the way of doing image annotation. In case of annotation teams, multiple annotators can annotate a sample multiple times or using division of labor to reduce effort during annotation. Choosing division of labor, annotation noise can be introduced in the merged dataset in terms of different annotation styles.

The authors in [11], [12] highlight that issues concerning annotation quality can negatively affect model performance.

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

Developing DNN has been extensively studied for a variety of tasks, but, for instance, quantitative studies concerning annotation quality lack. On the one hand, inconsistent annotations may lead to problems during the training process. The interpretation of similar patterns would yield to different optimization steps. On the other hand, noisy annotations can result in a selection of sub-optimal DL models, e.g. benchmarks during validation with faulty annotations are not reliable [11]. The authors of [10], [13] argue that, especially in small-scale data scenarios like biomedical problems, the noisy annotation may significantly reduce the performance of DNNs.

Hence, we propose two automated inspection pipelines to identify intra-annotator and inter-annotator variability in image annotation for segmentation tasks utilizing uncertainty-aware DNNs. In particular, the proposed novel inspection method concerning inter-annotator variability has no need for multiple annotations per sample. Hence, the method can be used in division of labor annotation scenarios. The inspection pipeline is designed to assist domain experts, i.e. medical practitioner. We categorize annotation noise in image segmentation tasks, introduce methods to simulate those, and analyze the impact of annotation noise on DNN performance.

Our key contributions are the following: (i) a general categorization, simulation and visualization of annotation noise in image segmentation tasks through a generic synthetic dataset, (ii) novel automated methods forming inspection pipelines to identify intra-annotator and inter-annotator noise in image segmentation talks utilizing uncertainty-aware DNNs to support annotators in biomedical DL projects, and (iii) a demonstration of the proposed pipelines using the biomedical ISIC 2017 Melanoma [7] image segmentation dataset.

Related work is summarized in Section II. Material and methods are given in Section III. The results and discussion are presented in Section IV. Finally, we conclude our work in Section V.

II. RELATED WORK

A. PRELIMINARIES

A generic dataset $\mathcal{D} = (\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \dots, N\}$ combines the sample set $\mathcal{X}$ and the annotation set $\mathcal{Y}$. It is composed of N samples $\mathbf{x}_i \in \mathcal{X}$ and corresponding true, but mostly unknown, annotations $\mathbf{y}_i \in \mathcal{Y}$. In image segmentation tasks, annotations are referred to as masks. Potentially noisy annotations are denoted as $\tilde{\mathbf{y}}_i$. Inserting noise to datasets is referred to as corruption of the dataset. Hence, in this article, noisy annotations and corrupted annotations are used synonymously. The noisy dataset including corruptions is denoted by $\mathcal{D}_{\text{corr}}$. The property "clean" refers to datasets without noisy annotations. Predictions of a DNN $f_{\boldsymbol{\theta}}$ set with parameters $\boldsymbol{\theta}$ using sample $\mathbf{x}_i$ are noted as $\hat{\mathbf{y}}_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i)$. Variables related to different annotators j are denoted via $^j(\bullet)$. Different elements k belonging to the same sample i are indexed by $(\bullet)^k$.

1) NOISY ANNOTATIONS IN DATASETS

The work in [10], [14], [15] gives surveys in terms of noisy annotations in the context of DL. There are two possibilities to deal with noisy annotations within DL: (i) reducing the impact of noisy annotations while training a DNN, i.e., adapting DL architectures [16], loss functions [17], [18], importance of samples [19], data representation [20], or training procedures [2], [21], [22] to handle noisy annotations without removing them from the dataset, or (ii) detecting noisy annotations directly and correcting them. The authors of CleanNet [23] extract image-wise feature vectors and compare them to a feature vector that is representative for a respective class. Noisy annotations are detected if feature vectors show large deviations from the feature vector of a representative class. However, this method is not applicable for an image segmentation task since a direct comparison of annotations is not feasible.

Another approach is Rank Pruning presented in [24], which considers the DNN output for noise detection. Thereby, it is assumed that network outputs such as the softmax or sigmoid function, also referred to as soft annotations, can be interpreted as a probability distribution and a distinct output correlates to an assumed clean annotation. The Confident Learning framework in [25] is based on Rank Pruning as well. Though, the work of [26], [27] show that DNNs are too confident and thus output cannot be equated with prediction probability in general. DNNs with this property are often referred to as non-calibrated.

The issue of missing calibration in DNN outputs is taken up in the work of [28] dealing with classification problems. However, both methods are not directly applicable to segmentation tasks because of the pixel-wise output in segmentation tasks. Considering the issue of inter-annotator noise, repeating the annotation using multiple human annotators may serve for identifying noisy annotations. A method to consider different annotators for classification tasks is depicted in [29]. The key idea is to estimate the skill level of each annotator in form of confusion matrices. Kohl *et al.* [30] propose the architecture Probabilistic U-Net to cope with multiple noise annotations in image segmentation tasks using a variational auto-encoder.

However, the typical scenario in the context of annotation is the division of labor between annotators. Drawback w.r.t. methods utilizing multiple annotations per sample is an increased annotation effort. This procedure may not be possible in many applications due to time or cost restrictions. Further, the development towards automated annotation pipelines in different domains like presented in [8], [31], [32] requires manual inspection in cases where correct annotated data for benchmarks or evaluation is needed. In summary, the state of the art is strongly focused on classification tasks and less on image segmentation.

2) UNCERTAINTY IN DNNs

DNNs can solve complex tasks through accurate function approximation. However, the work of [33] demonstrates the vulnerability of DNNs. Adversarial attacks applying small perturbation of the input images considering physical boundary conditions, e.g. wrong classification in different view positions, can lead to totally wrong predictions with a high level of confidence. Hence, being aware of model uncertainty[1] is of importance. The work of Abdar *et al.* [35] gives an overview concerning uncertainty in DNNs.

A DNN can be denoted as a function $f_{\boldsymbol{\theta}} : \mathbf{x} \mapsto \hat{\mathbf{y}}$ mapping a sample $\mathbf{x}$ to a prediction $\hat{\mathbf{y}}$. The function is trained on a dataset $\mathcal{D}$ to obtain parameters $\boldsymbol{\theta}$. Considering Bayesian inference described in [35], [36], a prediction of a neural network can be described in a probabilistic way by

$$p(\hat{\mathbf{y}} \mid \mathbf{x}, \mathcal{D}) = \int p(\hat{\mathbf{y}} \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \, \mathrm{d}\boldsymbol{\theta} \tag{1}$$

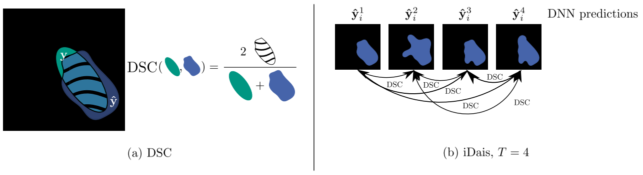
denoting $p(\boldsymbol{\theta} \mid \mathcal{D})$ as posterior distribution, which normally cannot be computed. Dropout is presented in [37] as a regularization technique which randomly deactivates neurons in order to avoid overfitting during training. The distribution of dropout $q(\boldsymbol{\theta})$ can be used to approximate the posterior. Following the proof in [36], [38] and taking T Monte Carlo samples into account, the probabilistic output of a DNN in Equation 1 simplifies to the approximation

$$p(\hat{\mathbf{y}} \mid \mathbf{x}, \mathcal{D}) \approx \int p(\hat{\mathbf{y}} \mid \mathbf{x}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{T} \sum_{t=1}^{T} p(\hat{\mathbf{y}} \mid \mathbf{x}, \boldsymbol{\theta}_t). \tag{2}$$

This idea is often referred to as Monte Carlo Dropout (MCD). Unlike dropout layers used for regularization, in MCD, dropout is activated during the application. Monte Carlo samples are drawn to estimate the output prediction distribution $p(\hat{\mathbf{y}} \mid \mathbf{x}, \mathcal{D})$ via T different parametrized networks $f_{\boldsymbol{\theta}_1}, \ldots, f_{\boldsymbol{\theta}_T}$. An alternative method for uncertainty estimation is Deep Ensemble presented in [39] which uses ensembles of DNNs to generate several predictions per sample. MCD is similar to an ensemble approach, but the DNN structure remains the same and no multiple training runs are needed. In contrast, test-time augmentation [40], [41] compares predictions of augmented images to estimate the uncertainty.

To evaluate the segmentation quality, the Dice-Sørensen coefficient $\mathrm{DSC}\big(\mathbf{y}_i, \hat{\mathbf{y}}_i\big)$ [42] can be utilized as metric comparing ground-truth segmentation mask $\mathbf{y}_i \in \mathcal{Y}$ and prediction $\hat{\mathbf{y}}_i$ of a sample $\mathbf{x}_i \in \mathcal{X}$. For T Monte Carlo predictions $\{\hat{\mathbf{y}}_i^1, \ldots, \hat{\mathbf{y}}_i^T\}$, the by Roy *et al.* [43] proposed *inverted Dice agreement in sample*

$$\begin{aligned} &\mathrm{iDAIS}(\mathbf{x}_i, f_{\boldsymbol{\theta}}, T) \\ &\quad = 1 - \frac{2}{T(T-1)} \sum_{j=1}^{T} \sum_{k=j+1}^{T} \mathrm{DSC}\big(f_{\boldsymbol{\theta}_j}(\mathbf{x}_i), f_{\boldsymbol{\theta}_k}(\mathbf{x}_i)\big) \end{aligned} \tag{3}$$



(a) DSC

(b) iDais, $T = 4$

FIGURE 1. Visualization of Dice-Sørensen coefficient and inverted Dice agreement. For the Dice-Sørensen coefficient DSC (a) the area of intersection w.r.t. two segmentation masks y and ŷ is compared to their union. In contrast, the inverted Dice agreement in sample iDAIS is based on the comparison of multiple DNN predictions $\{\hat{\mathbf{y}}_i^1, \ldots, \hat{\mathbf{y}}_i^T\}$.

can serve as metric to describe image-wise model uncertainty of $f_{\boldsymbol{\theta}}$ w.r.t. sample $\mathbf{x}_i \in \mathcal{X}$ in segmentation tasks. The metric iDAIS $\in [0, 1]$ indicates larger uncertainty of a prediction with increasing value. Using MCD in combination with Equation 3, $\mathrm{iDAIS}(\mathbf{x}_i, f_{\boldsymbol{\theta}}, T)$ estimates an inverted prediction quality score. Figure 1 visualizes both metrics. A detailed motivation of using iDAIS based on [38], [43], [44] is given in the supplementary.

The main challenges can be summarized in: (i) there is no categorization of annotation noise and its impact on DNN performance in case of image segmentation, (ii) there is a lack of methods to inspect datasets concerning noisy annotations in image segmentation tasks, (iii) DNN inspection approaches neglect the missing calibration to correct over-confident DNNs, and (iv) methods to address inter-annotator variability without requiring more than one annotation per sample are missing.

III. MATERIALS AND METHODS

A. ANNOTATION NOISE CATEGORIZATION AND SIMULATION

We subdivide annotation noise into systematic and random. Table 1 summarizes the annotation noise including category ("systematic" or "random"), description, and simulation method. In Figure 2 examples for each type are given.

The noise type of oversized masks may occur randomly or systematically. This type can be simulated with a morphological dilation operation. Analogously, an undersized mask also occurs systematically or randomly and can be simulated with a morphological erosion operation.

Holes can occur randomly in masks due to missing areas while filling segments. However, there are applications where holes in masks occur but are intended. Consequently, general post-processing in terms of hole filling cannot be done by default. Considering simulation methods, cutting out random ellipses in segments is a way to simulate this noise type.

Similar to holes, jitter occurs randomly during annotation. Unintentional clicking can be mentioned as an example in practical applications. As a simulation method for this noise type, randomly inserting ellipses in areas where initially no mask is present in the clean dataset can be named.

Besides, missing annotation can occur randomly when forgetting to annotate a sample in total. For instance, clicking through samples inside the used annotation tool as well as intentionally overstepping a sample, e.g. if an instance is not

[1] The term uncertainty is used to quantify the amount of error given in a DNN prediction. The error needs to be quantified without ground truth during DNN inference. Consequently, it should characterize model failure and lack of generalization. Uncertainty quantification is closely related to out-of-distribution detection. For instance, a current research question is whether methods for uncertainty quantification can be used for out-of-distribution detection [34].

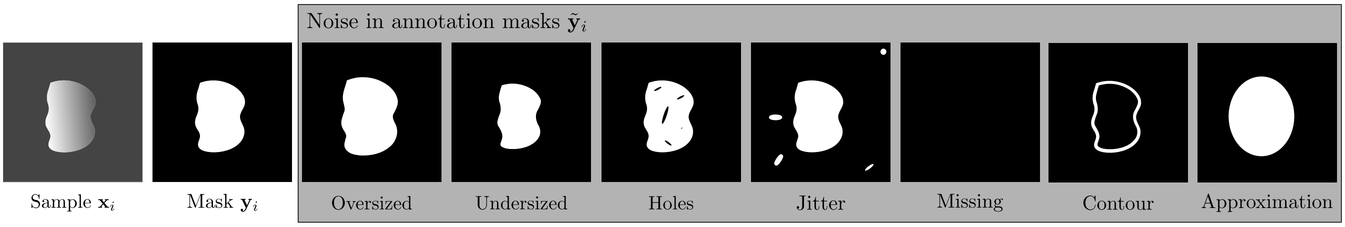


FIGURE 2. Synthetic image segmentation dataset. Presentation of sample $\mathbf{x}_i$, ground-truth mask $\mathbf{y}_i$, and occurring annotation noise $\tilde{\mathbf{y}}_i$ in segmentation masks.

TABLE 1. Overview annotation noise types. Comparison of category, description, and the operations to simulate them.

Type	Category	Description	Simulation
Oversized	Systematic/random	Mask is drawn too large	Morphological dilation
Undersized	Systematic/random	Mask is drawn too small	Morphological erosion
Holes	Random	Unintended holes in a segment	Insert holes via random oriented/sized ellipses into mask segments
Jitter	Random	Unintended segments besides mask	Insert random oriented/sized ellipses in areas apart from mask segments
Missing	Random	Mask is missing	Delete mask
Contour	Random	Mask not filled	Insert contour line with defined pixel width
Approximation	Random	Mask approximated with ellipse	Insert ellipse with same centroid and expansion as the original mask

found or the annotation task is deemed to be too difficult, are situations where this problem appears. Though, a frame without a segment can occur in a dataset, too. Consequently, it cannot be said in general that missing annotation is noise in every case.

Some annotation tools offer functionality in terms of only drawing a contour. This contour is filled when a loop closure is formed. However, if there is no exact loop closure, only the contour remains. Simulation of this random noise type can be done via inserting the contour line of a mask segment.

Further, annotating the exact contour is burdensome. The noise type approximation takes this into account. It can be simulated through inserting circular respectively elliptical masks with the same centroids compared to the original masks.

Details w.r.t. simulation can be found in the supplementary. Combinations of noise types are out of scope of this work. Systematic deviations mainly occur between different annotators (inter-annotator variability). In contrast, random noise appears during the annotation procedure of an annotator (intra-annotator variability).

To evaluate our proposed inspection pipelines, we simulate the introduced noise types. Thereby, the presented simulations are subdivided into the two use-cases, namely intra- and inter-annotator variability inspection.

1) INTRA-ANNOTATOR VARIABILITY SIMULATION

The simulation process concerning intra-annotator variability to create a noisy dataset $\mathcal{D}_{\text{corr}}$ is presented in Figure 3. First, a random shuffling is executed. Subsequently, the initially clean dataset $\mathcal{D}$ is split into two parts ${}^{\text{A}}\mathcal{D}$ and ${}^{\text{B}}\mathcal{D}$ controlled by the corruption ratio parameter $\gamma \in [0, 1]$. Split ${}^{\text{A}}\mathcal{D}$ is corrupted using the introduced noise types. It is not guaranteed that all images within a dataset have the same sizes. Being faced with images of different sizes, a first resize operation is performed per segmentation mask to ensure comparable results which are independent of the initial image size. Images $\mathbf{x}_i$ remain unchanged to prevent the loss of information. Details of resizing parameters are given in the supplementary. Afterwards, corruptions are inserted and the corrupted annotation mask is resized to its original size. The corrupted dataset $\mathcal{D}_{\text{corr}}$ is a union of the corrupted split ${}^{\text{A}}\mathcal{D}_{\text{corr}}$ and the unmodified dataset split ${}^{\text{B}}\mathcal{D}$. All resize operations use nearest-neighbor interpolation to preserve class information.

2) INTER-ANNOTATOR VARIABILITY SIMULATION

Figure 4 presents our proposed method to simulate inter-annotator noise. The clean dataset $\mathcal{D}$ is split into two parts. Here, however, the parts are equal-sized. After inserting individual corruptions A,B, corrupted datasets ${}^{\text{A}}\mathcal{D}_{\text{corr}}$,${}^{\text{B}}\mathcal{D}_{\text{corr}}$ are obtained to simulate inter-annotator annotation noise. It should be noted that the case of only one present corruption is also conceivable and thus will be investigated in the experiments, too.

B. ANNOTATION INSPECTION PIPELINES

1) INTRA-ANNOTATOR VARIABILITY INSPECTION

One objective of our proposal is obtaining an inspection pipeline that should be capable of identifying noisy annotated images in a dataset that is annotated by a single annotator. The general concept of our method to detect intra-annotator noise is a hybrid approach depicted in Figure 5.

A DNN f_{θ} intended for the actual image segmentation task is trained on the noisy dataset $\mathcal{D}_{\text{corr}}$ (cf. Figure 3). Due to the random nature of intra-annotator noise, it is assumed that noisy annotations are not dominating in a dataset (corruption rate $\gamma < 0.5$). Hence, the DNN can

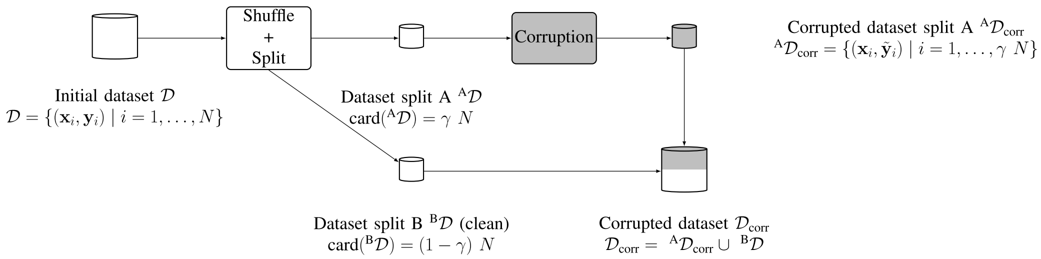

Shuffle
+
Split
Corruption
Initial dataset $\mathcal{D}$
$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N\}$
Dataset split A ${}^{\mathrm{A}}\mathcal{D}$
$\mathrm{card}({}^{\mathrm{A}}\mathcal{D}) = \gamma\ N$
Corrupted dataset split A ${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}}$
${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid i = 1, \ldots, \gamma\ N\}$
Dataset split B ${}^{\mathrm{B}}\mathcal{D}$ (clean)
$\mathrm{card}({}^{\mathrm{B}}\mathcal{D}) = (1-\gamma)\ N$
Corrupted dataset $\mathcal{D}_{\mathrm{corr}}$
$\mathcal{D}_{\mathrm{corr}} = {}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}} \cup {}^{\mathrm{B}}\mathcal{D}$

FIGURE 3. **Intra-annotator variability simulation. A noisy dataset $\mathcal{D}_{\mathrm{corr}}$ is generated via inserting corruption to a clean dataset $\mathcal{D}$ controlled by parameter γ which describes the corruption ratio. Thereby, corruptions are inserted to the split ${}^{\mathrm{A}}\mathcal{D}$ leading to ${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}}$ whereas ${}^{\mathrm{B}}\mathcal{D}$ remains unchanged.**

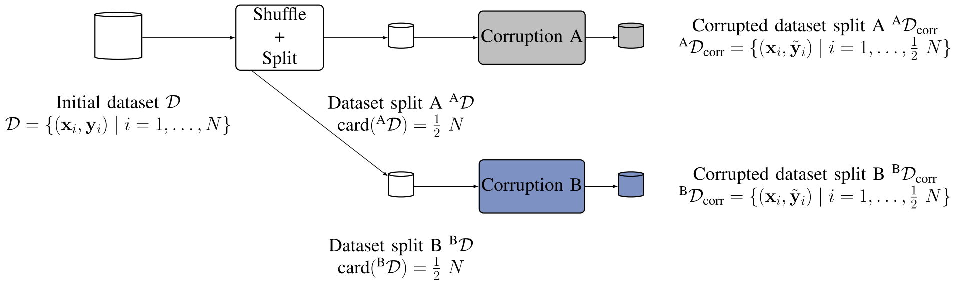

Shuffle
+
Split
Corruption A
Corruption B
Initial dataset $\mathcal{D}$
$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N\}$
Dataset split A ${}^{\mathrm{A}}\mathcal{D}$
$\mathrm{card}({}^{\mathrm{A}}\mathcal{D}) = \frac{1}{2}\ N$
Corrupted dataset split A ${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}}$
${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid i = 1, \ldots, \frac{1}{2}\ N\}$
Corrupted dataset split B ${}^{\mathrm{B}}\mathcal{D}_{\mathrm{corr}}$
${}^{\mathrm{B}}\mathcal{D}_{\mathrm{corr}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \mid i = 1, \ldots, \frac{1}{2}\ N\}$
Dataset split B ${}^{\mathrm{B}}\mathcal{D}$
$\mathrm{card}({}^{\mathrm{B}}\mathcal{D}) = \frac{1}{2}\ N$

FIGURE 4. **Inter-annotator variability simulation with two independent annotators. A clean dataset $\mathcal{D}$ is split in two equal-sized parts to insert corruptions A,B to generate noisy datasets ${}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}}$,${}^{\mathrm{B}}\mathcal{D}_{\mathrm{corr}}$.**

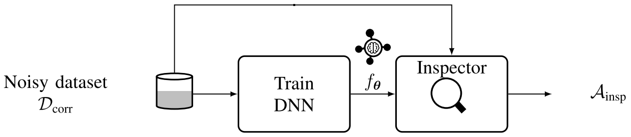


FIGURE 5. **Inspection pipeline intra-annotator variability. The noisy dataset $\mathcal{D}_{\mathrm{corr}}$ is used to obtain an uncertainty-aware DNN f_θ. The inspector uses the DNN to automatically predict an ordered list $\mathcal{A}_{\mathrm{insp}}$ which can be used for re-inspection of potentially noisy annotations.**

focus on the majority of correct annotations to learn helpful representations. A comparison of the DNN prediction $\hat{\mathbf{y}}_i$ and the potentially faulty annotations $\tilde{\mathbf{y}}_i$ is possible. The amount of matching between prediction $\hat{\mathbf{y}}_i$ and annotation $\tilde{\mathbf{y}}_i$ can be described quantitatively utilizing $\mathrm{DSC}(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i)$. However, this metric may be misleading for samples $\mathbf{x}_i$ that are difficult to learn and where the DNN fails. The metric DSC states a large discrepancy between prediction and annotation due to DNN failure, but the annotation needs not be noisy. Therefore, we propose an extension to consider uncertainty in our approach for detecting potentially noisy annotations. We adapt the DNN to take the uncertainty of the predictions into account. Thereby, we use the previously introduced MCD in Equation 2 since it can be easily integrated into an existing architecture by inserting dropout layers. Due to the focus on the application of estimating model uncertainty, investigations regarding other described uncertainty estimation approaches (Deep Ensembles or test-time augmentation) are not subject of this work. Being faced with image segmentation tasks, we suggest using $\mathrm{iDAIS}(\mathbf{x}_i, f_\theta, T)$ introduced in Equation 3 to evaluate image-wise uncertainty of DNN f_θ w.r.t. sample $\mathbf{x}_i$. The annotation inspection should focus on samples that show both, a large matching error of prediction and annotation in terms of $\mathrm{DSC}(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i)$ as well as low uncertainty values considering $\mathrm{iDAIS}(\mathbf{x}_i, f_\theta, T)$ indicating confidence in the DNN predictions. Hence, this yields to the sample strategy `get_next_inspection`

$$(\mathbf{x}^*, \mathbf{y}^*) = \underset{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}_{\mathrm{corr}}}{\operatorname{argmin}} \Big((1-\alpha)\mathrm{DSC}(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i) + \alpha\ \mathrm{iDAIS}(\mathbf{x}_i, f_\theta, T) \Big) \tag{4}$$

to obtain the most critical sample/mask pair $(\mathbf{x}^*, \mathbf{y}^*)$ for inspection. The uncertainty weighting parameter $\alpha \in [0, 1]$ in Equation 4 controls the ratio between prediction to annotation matching and uncertainty. More information w.r.t. parameter α is shown in the supplementary. Algorithm 1 summarizes the whole inspection pipeline to filter out potentially noisy annotated images. The outcome of Algorithm 1 is an ordered list denoted as $\mathcal{A}_{\mathrm{insp}}$, which orders all samples in terms of their inspection necessity. The inspection necessity results from position in $\mathcal{A}_{\mathrm{insp}}$. Tuples of annotation and sample at the head of list $\mathcal{A}_{\mathrm{insp}}$ should be inspected first since they are potentially wrong (high mismatch between annotation and DNN prediction, confident DNN prediction/low prediction uncertainty).

2) INTER-ANNOTATOR VARIABILITY INSPECTION

A naive way is performing annotation of one sample by several annotators. Hence, comparing and averaging of all available annotations per sample is possible. However, this

Algorithm 1 Intra-Annotator Variability Detection

```
Require: $\mathcal{D}_{\text{corr}}, f_{\boldsymbol{\theta}}, \alpha, T$
Ensure: $\mathcal{A}_{\text{insp}}$
  for $i = 1, \ldots, N$ do
      $\hat{\mathbf{y}}_i \leftarrow \texttt{mean}(\{\hat{\mathbf{y}}_i^1, \ldots, \hat{\mathbf{y}}_i^T\})$   ▷ $T$ inference steps using $f_{\boldsymbol{\theta}}$ trained on $\mathcal{D}_{\text{corr}}$
      $(\mathbf{x}^*, \mathbf{y}^*) \leftarrow \texttt{get\_next\_inspection}(\mathbf{x}_i, \hat{\mathbf{y}}_i, f_{\boldsymbol{\theta}}, T, \alpha)$
      $\mathcal{A}_{\text{insp}}.\text{append}\big((\mathbf{x}^*, \mathbf{y}^*)\big)$
      $\mathcal{D}_{\text{corr}}.\text{delete}\big((\mathbf{x}^*, \mathbf{y}^*)\big)$
  end for
```

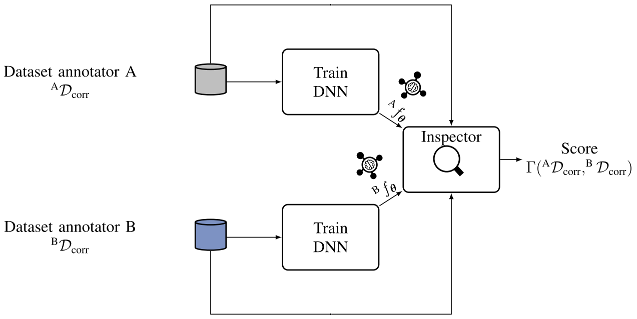

Dataset annotator A
$^{\text{A}}\mathcal{D}_{\text{corr}}$
Train DNN
$^{\text{A}}f_{\boldsymbol{\theta}}$
Inspector
Score
$\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$
$^{\text{B}}f_{\boldsymbol{\theta}}$
Dataset annotator B
$^{\text{B}}\mathcal{D}_{\text{corr}}$
Train DNN

FIGURE 6. Inspection pipeline inter-annotator variability. Datasets of annotator A $^{\text{A}}\mathcal{D}_{\text{corr}}$ and annotator B $^{\text{B}}\mathcal{D}_{\text{corr}}$ are used to train DNNs to simulate their annotation style. Subsequently, the inspector cross-checks $^{\text{A}}\mathcal{D}_{\text{corr}}$ and $^{\text{B}}\mathcal{D}_{\text{corr}}$ to predict a consistency score $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$ (cf. Algorithm 2).

leads to additional effort and is often not feasible in practical applications.

Thus, we propose the inter-annotator variability inspection pipeline depicted in Figure 6. The result of the pipeline is a predicted consistency score $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$ comparing the annotation styles of two annotators A and B. Their annotated but disjoint datasets are denoted by $^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}}$. In particular, the presented method does not rely on several annotations per sample to rate consistency between samples. The predicted consistency score $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$ can be used as an indicator whether the merging of datasets annotated by different annotators leads to a consistency gap or not. First, we propose using DNNs $^{\text{A}}f_{\boldsymbol{\theta}}$ and $^{\text{B}}f_{\boldsymbol{\theta}}$, as introduced before, to learn and simulate the individual annotation style. This part of the proposal enables a cross-check of the disjoint datasets. Algorithm 2 provides an overview of the calculations. In particular, we suggest comparing the predictions of $^{\text{A}}f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ and $^{\text{B}}f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ w.r.t. to a sample $\mathbf{x}_i$ to evaluate consistency in annotation styles. Similarly to the previously introduced intra-annotator detection, in cases where DNNs fail, the comparison in terms of $\text{DSC}\big(^{\text{A}}f_{\boldsymbol{\theta}}(\mathbf{x}_i), ^{\text{B}}f_{\boldsymbol{\theta}}(\mathbf{x}_i)\big)$ can be misleading. To avoid this unfavorable property, we take uncertainty (*use_uncertainty* $= 1$) here as well into account. There are three ways to consider uncertainty in the calculation, namely considering only one iDAIS of a prediction, a combination of both iDAIS prediction scores, or a comparison between noisy annotation $\tilde{\mathbf{y}}_i$ and predictions $\hat{\mathbf{y}}_i$ to rate prediction uncertainty. To simplify the notation, only the variant of taking one iDAIS of a prediction is shown in Algorithm 2. The other variants result from the previous descriptions analogously. We calculate a weighted average of matching scores $\text{DSC}\big(^{\text{A}}f_{\boldsymbol{\theta}}(\mathbf{x}_i), ^{\text{B}}f_{\boldsymbol{\theta}}(\mathbf{x}_i)\big)$ using weights $w_i = \Big(1 - \text{iDAIS}\big(^{\text{A/B}}\mathbf{x}_i,\ ^{\text{A/B}}f_{\boldsymbol{\theta}}, T\big)\Big)$ to finally obtain the consistency score $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$ of both annotators A/B (cf. Algorithm 2). It should be noted that our proposed detection pipeline is not limited to two annotators. A cross-wise comparison of more than two annotators would yield multiple consistency scores. A detailed description of the procedure is given in the supplementary.

Algorithm 2 Inter-Annotator Variability Detection

```
Require: $^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{A}}f_{\boldsymbol{\theta}}, ^{\text{B}}\mathcal{D}_{\text{corr}}, ^{\text{B}}f_{\boldsymbol{\theta}}$, use_uncertainty
Ensure: $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})$
  $\mathbf{L} \leftarrow [\text{A}, \text{B}]$   ▷ list transforming annotator to corresponding index notation
  $w, \Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}}) \leftarrow 0, 0$
  for $j = 1, \ldots, 2$ do
      $k, l \leftarrow \mathbf{L}[j], \mathbf{L}[3-j]$
      for $i = 1, \ldots, \text{card}(^{k}\mathcal{D}_{\text{corr}})$ do
          if use_uncertainty then
              $w_i \leftarrow \Big(1 - \text{iDAIS}\big(^{k}\mathbf{x}_i,\ ^{k}f_{\boldsymbol{\theta}}, T\big)\Big)$
          else
              $w_i \leftarrow 1$
          end if
          $\Gamma_i \leftarrow w_i\, \text{DSC}\big(^{k}f_{\boldsymbol{\theta}}(^{k}\mathbf{x}_i),\ ^{l}f_{\boldsymbol{\theta}}(^{k}\mathbf{x}_i)\big)$
          $w \leftarrow w + w_i$
          $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}}) \leftarrow \Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}}) + \Gamma_i$
      end for
  end for
  $\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}}) \leftarrow \frac{\Gamma(^{\text{A}}\mathcal{D}_{\text{corr}}, ^{\text{B}}\mathcal{D}_{\text{corr}})}{w}$
```

C. EVALUATION METHODS

1) UNCERTAINTY ESTIMATION

To evaluate the MCD uncertainty estimation in terms of iDAIS introduced in Equation 3, we consider the Pearson correlation R_{pearson}. By that, the correlation between iDAIS and mismatch of network prediction $\hat{\mathbf{y}}_i$ to ground-truth annotation $\mathbf{y}_i$ in terms of $1-\text{DSC}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$ is analyzed. Thereby, comparisons are all referred to clean annotations $\mathbf{y}_i$ since the uncertainty awareness to ground-truth is of interest. A perfect uncertainty estimation should be aware of the degree of error in predictions. Thus, $R_{\text{pearson}} \rightarrow 1$ is the optimal and desired property.

2) ANNOTATION INSPECTION

To assess intra-annotator variability, we propose to use Receiver Operating Characteristic (ROC) curve to evaluate performance of the proposed inspection pipeline. Hereby, we denote a noisy mask as a member of the positive class, correct annotations as examples of the negative class. Decision criteria used in ROC refer to the inverted sampling score presented in Equation 4, namely

$1 - \texttt{get\_next\_inspection}(\mathbf{x}_i, \hat{\mathbf{y}}_i, f_{\boldsymbol{\theta}}, T, \alpha)$. For performance evaluation, we consider Area Under Curve AUC as metric listed in [45]. The True Positive Rate is denoted with TPR (successful detection of a noisy annotation) respectively False Positive Rate using FPR (detecting a clean annotation as noisy), $\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\ \mathrm{dFPR}$ represents an integral of TPR over FPR between the range of 0 and 1. It should be remarked, that an $\mathrm{AUC} > 0.5$ indicates superior performance to random classifiers whereas $\mathrm{AUC} < 0.5$ characterizes classifiers pointing to consistent wrong inspections.

To evaluate the inter-annotator inspection quality in terms of consistency score $\Gamma({}^{\mathrm{A}}\mathcal{D}_{\mathrm{corr}}, {}^{\mathrm{B}}\mathcal{D}_{\mathrm{corr}})$ introduced in Algorithm 2, we consider Weighted Root Mean Square Error WRMSE. Thereby, we use the respective annotation styles of annotators A and B to simulate two noisy annotation ${}^{\mathrm{A}}\tilde{\mathbf{y}}$ and ${}^{\mathrm{B}}\tilde{\mathbf{y}}$ serving as a reference. Hence, a ground-truth mismatch $\mathrm{DSC}({}^{\mathrm{A}}\tilde{\mathbf{y}}_i, {}^{\mathrm{B}}\tilde{\mathbf{y}}_i)$ can be obtained. This mismatch is compared to the predicted mismatch $\mathrm{DSC}({}^{\mathrm{A}}\hat{\mathbf{y}}_i, {}^{\mathrm{B}}\hat{\mathbf{y}}_i)$. In the case of activated uncertainty, the WRMSE of differences between $\mathrm{DSC}({}^{\mathrm{A}}\tilde{\mathbf{y}}_i, {}^{\mathrm{B}}\tilde{\mathbf{y}}_i)$ and $\mathrm{DSC}({}^{\mathrm{A}}\hat{\mathbf{y}}_i, {}^{\mathrm{B}}\hat{\mathbf{y}}_i)$ is using inverted uncertainty, like depicted in Algorithm 2, as weights w_i. Thus, it can be obtained

$$\mathrm{WRMSE} = \sqrt{\frac{\sum_{i=1}^{N} w_i \left(\mathrm{DSC}({}^{\mathrm{A}}\tilde{\mathbf{y}}_i, {}^{\mathrm{B}}\tilde{\mathbf{y}}_i) - \mathrm{DSC}({}^{\mathrm{A}}\hat{\mathbf{y}}_i, {}^{\mathrm{B}}\hat{\mathbf{y}}_i)\right)^2}{\sum_{i=1}^{N} w_i}} \tag{5}$$

as an evaluation metric. In contrast, taking uncertainty not into account leads to the Root Mean Square Error RMSE since all weights are equal to one. In case of no inspection algorithm, $\mathrm{DSC}({}^{\mathrm{A}}\tilde{\mathbf{y}}_i, {}^{\mathrm{B}}\tilde{\mathbf{y}}_i) = 1$ is assumed to enable a benchmark.

IV. RESULTS AND DISCUSSION

A. DATASET

We use the ISIC 2017 Melanoma image segmentation dataset presented at the International Skin Imaging Collaboration (ISIC) challenge 2017 [7] to evaluate our proposed inspection approaches in practical biomedical image tasks. One objective of the challenge is an automated lesion segmentation of images to assist diagnosis of skin cancer. Experts provide an annotation of dermoscopic images with binary lesion masks. The dataset is composed of 2000 training and 600 test samples with varying image sizes. For simulations of noise, the training dataset is exclusively taken into account. Figure 7 depicts samples $\mathbf{x}_i$ with corresponding segmentation masks $\mathbf{y}_i$ of the introduced dataset. It is assumed that this dataset is initially clean and all annotations are correct. Figure 8 illustrates the introduced annotation noise types (cf. Figure 2 and Table 1) using a sample of the ISIC 2017 Melanoma image segmentation dataset.

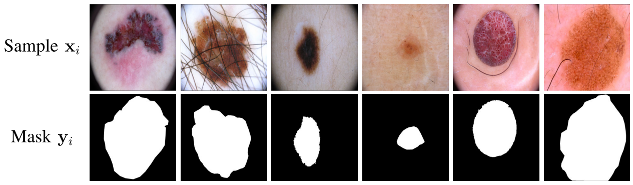


FIGURE 7. ISIC 2017 Melanoma image segmentation dataset [7]. Exemplary samples $\mathbf{x}_i$ and masks $\mathbf{y}_i$ are given.

B. ARCHITECTURE, IMPLEMENTATION, AND TRAINING PROCESS

The U-Net [46] serves as the basis for our proposed uncertainty-aware DNN architecture. We propose to extend this architecture by inserting dropout layers to enable uncertainty estimation in terms of MCD. Inspired by [43], we use $T = 20$ for uncertainty estimation. We implemented our uncertainty-aware U-Net in PyTorch Lightning [47] and used data augmentation provided by Albumentations [48]. The image processing libraries OpenCV [49] and scikit-image [50] are used for noise simulation.

The used train-validation split ratio is 80/20. The split is done randomly since the considered dataset is not sequential. An evaluation in form of k-fold cross-validation ($k = 5$) concerning robustness of evaluation scores concerning random splits shows a standard deviation $\sigma(\mathrm{DSC}) < 1.5\%$. Hence, the results are not susceptible regarding train-validation split. To reduce DNN training durations, the proposed hybrid high-performance computing/high-throughput computing concept in [51] is considered. Logging for interpretation of results is performed by Weights&Biases [52]. The proposed algorithm random search [53] is considered in order to determine hyperparameters. More details are given in the supplementary.

C. EXPERIMENTS

To evaluate our intra- and inter-annotator variability detection pipelines, we simulate noisy datasets as presented in Figure 3 and Figure 4. In the case of intra-annotator variability, we consider noise types which are members of category random (cf. Table 1) using corruption rate $\gamma \in \{0.1, 0.25, 0.5\}$. In contrast, inter-annotator noise is simulated with all combinations of systematic annotation noise (cf. Table 1), including unmodified clean splits of the initial dataset. Parameters for the corruption simulation can be found in the supplementary.

D. IMPACT OF NOISY ANNOTATIONS

To begin with, the impact of noisy annotations on DNN test performance is discussed. For simplicity, we limited our investigations to the case of inter-annotator variability simulations. Table 2 shows segmentation quality by using $\mathrm{DSC}_{\mathrm{test}}$ on the test dataset in comparison to different corruptions. It can be determined that undersized and missing masks lead to monotonic decreasing DNN performance regarding corruption rate γ. Moreover, the contour corruption reduces DNN generalization capability. This finding can be obtained since the metric $\mathrm{DSC}_{\mathrm{test}}$ in the case of noise is reduced

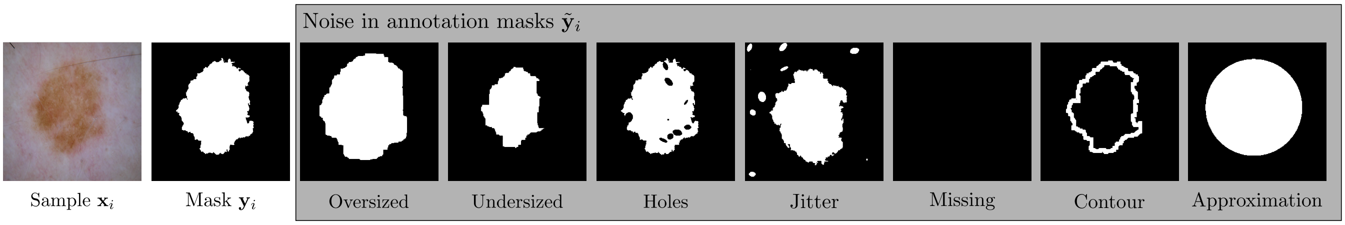


FIGURE 8. Simulated corrupted ISIC 2017 Melanoma image segmentation dataset. Sample $\mathbf{x}_i$, ground-truth mask $\mathbf{y}_i$, and occurring annotation noise $\tilde{\mathbf{y}}_i$ in segmentation masks are presented.

TABLE 2. Results concerning intra-annotator noise. Presentation of the impact of noisy annotations on test performance $\mathrm{DSC}_{\mathrm{test}}$, evaluation of uncertainty estimation using R_{pearson}, and comparison of intra-annotator inspection performance in terms of AUC considering uncertainty ($\alpha = 0.4$) respectively not considering uncertainty ($\alpha = 0.0$). The baseline w.r.t. the clean dataset $\mathcal{D}$ is $\mathrm{DSC}_{\mathrm{test}} = 78.28\%$ and $R_{\mathrm{pearson}} = 0.85$.

Type	Undersized			Oversized			Jitter			Holes			Missing			Contour			Approximation		
γ	0.10	0.25	0.50	0.10	0.25	0.50	0.10	0.25	0.50	0.10	0.25	0.50	0.10	0.25	0.50	0.10	0.25	0.50	0.10	0.25	0.50
$\mathrm{DSC}_{\mathrm{test}}$ in %	77.87	76.54	73.58	78.65	79.77	77.79	78.24	78.99	79.05	78.10	78.25	78.50	77.26	75.93	74.05	77.67	77.80	77.05	78.30	80.00	79.31
R_{pearson}	0.85	0.87	0.85	0.81	0.81	0.65	0.81	0.83	0.82	0.84	0.83	0.85	0.85	0.87	0.84	0.83	0.84	0.85	0.82	0.81	0.72
AUC ($\alpha = 0.0$)	0.882	0.864	0.773	0.824	0.717	0.219	0.708	0.696	0.695	0.729	0.721	0.711	1.000	1.000	1.000	0.963	0.956	0.948	0.820	0.770	0.390
AUC ($\alpha = 0.4$)	0.906	0.896	0.809	0.856	0.745	0.185	0.750	0.735	0.727	0.770	0.765	0.755	0.999	0.998	0.998	0.977	0.976	0.975	0.844	0.788	0.375

compared to the clean baseline result. In contrast, oversized corruption only leads to mild performance degradation in case of a corruption rate $\gamma = 0.5$. Moreover, the approximation corruption does not yield a performance degradation. This finding can be explained having a closer look at the test dataset. There are many melanomas depicted with low contrast to the surrounding area (e.g. Figure 7, column 4). Therefore, oversized masks or larger elliptical masks may help the U-Net to rather predict a mask than not predicting a mask segment. A closer look at jitter and hole noise types shows an interesting phenomena. Contrary to expectation, these corruptions appear to improve DNN generalization.

E. UNCERTAINTY QUANTIFICATION

Before using uncertainty in our proposed inspection approaches, we evaluate uncertainty in terms of the introduced correlation metric R_{pearson}. Table 2 illustrates results in cases of the clean dataset and corruptions. We demonstrate that the estimated uncertainty in terms of iDAIS correlates with the error $1 - \mathrm{DSC}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$ in cases of the clean dataset and the noisy datasets. Hence, we are able to be aware of failures in DNN predictions. Excluding two simulation categories (undersized mask and approximation with corruption rate $\gamma = 0.5$, discussed later), we achieve correlation coefficients $R_{\mathrm{pearson}} \geq 0.81$. Hence, the benefit of the proposed uncertainty estimation can be demonstrated. Figure 9 represents the correlation of uncertainty and DNN prediction errors in the case of the clean dataset showing all 2000 samples. For visualization purposes, a linear line fit presents the correlation as well. Furthermore, Figure 10 gives a visualization of uncertainty prediction of different samples. Thereby, clean annotation $\mathbf{y}_i$, predicted mask $\hat{\mathbf{y}}_i$, and uncertainty respectively quantitative metrics in terms of iDAIS and DSC are presented for different samples $\mathbf{x}_i$. With this, an accurate and an inaccurate prediction are opposed. The visual evaluation as well as the quantitative data support that uncertainty can be used to detect DNN failures assuming correct labels within the evaluation metric.

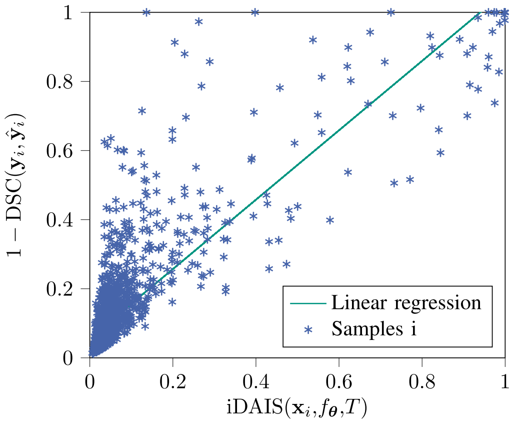


FIGURE 9. Evaluation of uncertainty estimation. Comparison of prediction error $1 - \mathrm{DSC}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$ to $\mathrm{iDAIS}(\mathbf{x}_i, f_\theta, T)$ of clean training dataset samples i and linear line fit.

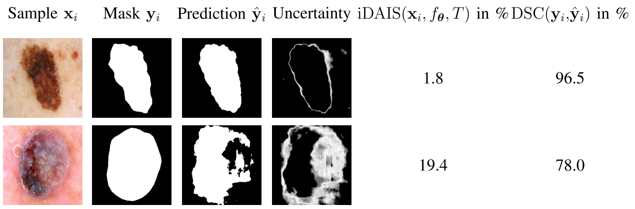

Sample $\mathbf{x}_i$
Mask $\mathbf{y}_i$
Prediction $\hat{\mathbf{y}}_i$
Uncertainty
$\mathrm{iDAIS}(\mathbf{x}_i, f_{\boldsymbol{\theta}}, T)$ in %
$\mathrm{DSC}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$ in %
1.8
96.5
19.4
78.0

FIGURE 10. Uncertainty estimation. Visualization of different samples $\mathbf{x}_i$, ground-truth mask $\mathbf{y}_i$, predicted mask $\hat{\mathbf{y}}$, and uncertainty (increasing with gray-level) respectively quantitative metrics in terms of $\mathrm{iDAIS}(\mathbf{x}_i, f_\theta, T)$ and $\mathrm{DSC}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$.

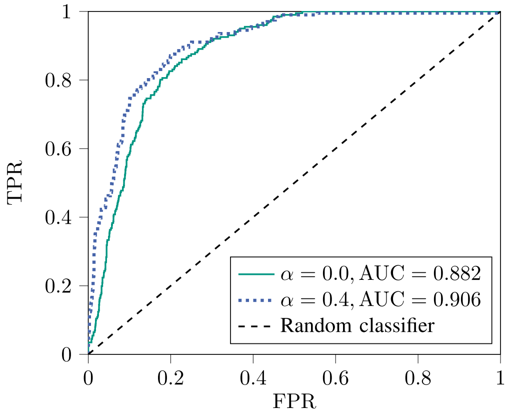
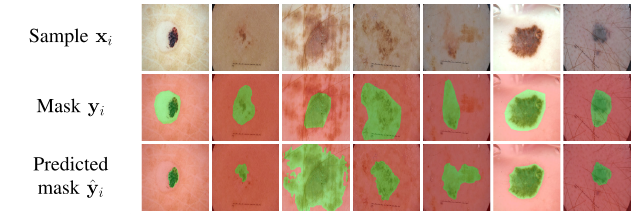


FIGURE 11. ROC curve. Opposing TPR and FPR in case of undersized mask ($\gamma = 0.1$) comparing a random classifier to the proposed annotation intra-annotator inspection considering ($\alpha = 0.4$) and not considering ($\alpha = 0.0$) uncertainty.



iDAIS($\mathbf{x}_i, f_{\boldsymbol{\theta}}, T$) in %	0.6	0.6	5.1	4.0	3.5	1.4	3.6
DSC($\mathbf{y}_i$,$\hat{\mathbf{y}}_i$) in %	37.5	38.5	36.5	56.5	65.0	84.2	70.3

FIGURE 12. Inspection of the clean dataset $\mathcal{D}$. Exemplary samples $\mathbf{x}_i$ of the clean dataset suggested by the automated annotation inspection pipeline to re-inspect. A comparison of annotated mask $\mathbf{y}_i$ and predicted mask $\hat{\mathbf{y}}_i$ by means of uncertainty-aware U-Net respectively quantitative metrics in terms of iDAIS($\mathbf{x}_i, f_{\theta}, T$) and DSC($\mathbf{y}_i$,$\hat{\mathbf{y}}_i$) is given.

F. ANNOTATION NOISE INSPECTION

In the following, it will be analyzed to what extent the two proposed inspection methods are suitable to detect intra-annotator and inter-annotator noise.

1) INTRA-ANNOTATOR VARIABILITY DETECTION

EVALUATION ON THE SIMULATED DATASET $\mathcal{D}_{corr}$

To evaluate the proposed intra-annotator variability detection pipeline, AUC is represented in Table 2 for the simulated annotation noise. Defining parameter $\alpha = 0.4$ leads to the overall best results. A parameter study concerning the impact of α on ROC curves or AUC is given in Supplementary. First, it can be determined that we obtain AUC ≥ 0.727 considering uncertainty if the exception cases are excluded. Uncertainty awareness ($\alpha = 0.4$) leads to performance improvement in almost every corruption case. A small exceptional case is the noise type of missing annotation. Hereby, uncertainty leads to a maximum degradation of 0.002 since the approach without uncertainty completely solved this detection task. This degradation is attributed to iDAIS since uncertainty estimation is not correct for all samples (cf. Table 2, $R_{\text{pearson}} < 1$). Though, this effect can be neglected compared to the performance boost in the other noise type cases. Moreover, a tendency of decreasing the performance in the form of AUC with an increasing corruption rate γ can be obtained (cf. Table 2). This result is understandable since, as presented in the previous section, in most cases, more noise leads to a less performing DNN. Hence, inspection is more difficult in cases of a large corruption ratio γ. Two corner cases occur for the noise simulation types oversized mask and approximation considering a corruption rate of $\gamma = 0.5$. The DNN U-Net converges to be in favor of predicting noise type instead of correct annotations. This claim is supported by the performance criteria AUC < 0.5, which indicates a misleading classification in general. Therefore, meeting our made assumption of $\gamma < 0.5$ is important to ensure meaningful functionality of the proposal to detect intra-annotator variability. In all other cases, as depicted in Table 2, the performance metric AUC > 0.5 indicates that the proposed inspection pipeline is able to detect noisy annotations. Figure 11 depicts an exemplary ROC curve in the case of undersized masks ($\gamma = 0.1$) and illustrates the superiority of the uncertainty awareness ($\alpha = 0.4$) within the inspection pipeline.

a: EVALUATION ON THE CLEAN DATASET $\mathcal{D}$

Up to this point, we assumed that the ISIC 2017 Melanoma image segmentation dataset contains only clean annotations. We used our inspection pipeline on the clean dataset as well. Exemplary samples $\mathbf{x}_i$ suggested for re-inspection by Algorithm 1 are displayed in Figure 12 comparing dataset annotations $\mathbf{y}_i$ and predictions $\hat{\mathbf{y}}_i$.

It can be shown that detected samples on the clean dataset, in any case, are complex to interpret and discussion with a dermatologist may be advantageous. Primarily, it seems that some of the annotated masks are more coarse than the U-Net predictions. An evaluation of the discussion of these samples with an dermatologist will be part of further investigations.

b: POTENTIAL W.r.t. INTERPRETABILITY AND ACTIVE LEARNING

The main objective of our proposed methods is detecting noisy annotations. However, an adaptation of the sampling function represented in Algorithm 1 may spawn another functionality for practitioners such as physicians. Using

$$(\mathbf{x}^*, \mathbf{y}^*) = \underset{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\in\mathcal{D}_{\text{corr}}}{\operatorname{argmax}} \Big((1-\alpha)\big(1-\text{DSC}(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i)\big) + \alpha \,\text{iDAIS}(\mathbf{x}_i, f_{\boldsymbol{\theta}}, T)\Big) \quad (6)$$

as a sampling strategy helps to determine and order samples that are both uncertain and hard to predict for the DNN. Hence, it can help in terms of interpretability to show the expert which samples are currently hard to learn. Furthermore, inspired by active learning approaches, the expert can use the knowledge for data acquisition and expand the dataset in order to improve DNN performance. Figure 13 gives an overview of samples $\mathbf{x}_i$ with large uncertainty in iDAIS and a large mismatch in DSC. We manually

TABLE 3. Results of inter-annotator variability inspection. Comparison of annotation consistency metric $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr})$ and quality assessment RMSE/WRMSE in cases of (no) inspection w.r.t. different dataset annotation styles ${}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}$. No inspection assumes $\mathrm{DSC}({}^{A}\tilde{\mathbf{y}}_i, {}^{B}\tilde{\mathbf{y}}_i) = 1$. Same annotations styles are denoted with ▨.

	Annotation style ${}^{B}\mathcal{D}_{corr}$		
Annotation style ${}^{A}\mathcal{D}_{corr}$	Oversized	Clean	Undersized
Oversized	0.931	0.766	0.553
Clean	0.760	0.917	0.665
Undersized	0.521	0.665	0.878

(a) Consistency score $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr})$

	Annotation style ${}^{B}\mathcal{D}_{corr}$		
Annotation style ${}^{A}\mathcal{D}_{corr}$	Oversized	Clean	Undersized
Oversized	0.000	0.237	0.518
Clean	0.237	0.000	0.385
Undersized	0.517	0.384	0.000

(b) RMSE (no inspection)

	Annotation style ${}^{B}\mathcal{D}_{corr}$		
Annotation style ${}^{A}\mathcal{D}_{corr}$	Oversized	Clean	Undersized
Oversized	0.000	0.230	0.490
Clean	0.230	0.000	0.348
Undersized	0.492	0.353	0.000

(c) WRMSE (no inspection)

	Annotation style ${}^{B}\mathcal{D}_{corr}$		
Annotation style ${}^{A}\mathcal{D}_{corr}$	Oversized	Clean	Undersized
Oversized	0.176	0.125	0.119
Clean	0.139	0.209	0.149
Undersized	0.115	0.140	0.288

(d) RMSE (inspection)

	Annotation style ${}^{B}\mathcal{D}_{corr}$		
Annotation style ${}^{A}\mathcal{D}_{corr}$	Oversized	Clean	Undersized
Oversized	0.118	0.087	0.098
Clean	0.092	0.141	0.110
Undersized	0.095	0.108	0.168

(e) WRMSE (inspection)

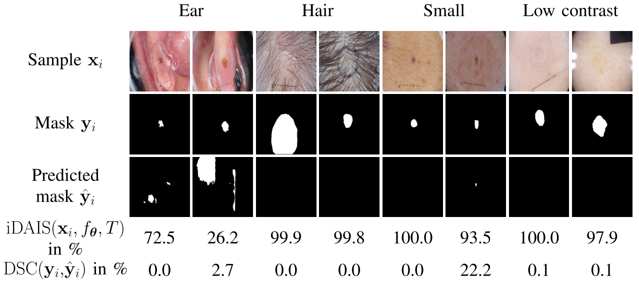


FIGURE 13. Interpretability and active learning. Exemplary samples $\mathbf{x}_i$, annotated masks $\mathbf{y}_i$, and predicted mask $\hat{\mathbf{y}}_i$ of clean dataset $\mathcal{D}$ with the property of high uncertainty as well as a large gap between annotation respectively DNN prediction represented in image clusters. Quantitative metrics in terms of iDAIS($\mathbf{x}_i, f_\theta, T$) and DSC($\mathbf{y}_i$,$\hat{\mathbf{y}}_i$) are presented.

clustered the samples obtained by the acquisition function in Equation 6. On the one hand, ear images and images including hairs seem to be hard cases. On the other hand, examples with small masks or low contrast to surrounding areas are hard to predict. Hence, expanding the dataset with more images associated with the enumerated clusters may help to improve performance.

2) INTER-ANNOTATOR VARIABILITY DETECTION

Table 3 gives an overview w.r.t the previously introduced metrics. The proposed consistency score $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr})$ is presented in Table 3a. In the case of no uncertainty consideration, Table 3b and Table 3d compare the results of the introduced error measure RMSE w.r.t. no inspection and activated inspection. Besides, the metric WRMSE given in Equation 5 to evaluate uncertainty awareness within the inter-annotator inspection approach is presented in Table 3c for the case of no inspection and in Table 3e considering inspection. As introduced in the method description, the selection of the used uncertainty score in the intra-annotator variability inspection pipeline can be done in different ways. However, using iDAIS of the prediction where sample $\mathbf{x}_i$ belonged to during training, leads to the best overall results in terms of WRMSE. This finding is comprehensible since the corresponding DNN is already familiar to the sample during training. To simplify the results, only the result of this selection are given in Table 3. First, it can be demonstrated that all tables are within a close approximation symmetric what meets the expectation that the interchangeability law holds when comparing the simulated inter-annotator variability. Taking consistency score $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr})$ depicted in Table 3 into account, a clear pattern emerges. First, similar annotation styles (represented on the diagonal and denoted by ▨) are predicted as most similar ($\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}) \in [0.878, 0.931]$). Second, different annotation styles (all entries not part of the diagonal) show lower predicted similarity scores. Furthermore, the introduced quality assessment metric correlates with the amount of difference in annotation styles. For instance, a comparison of undersized masks with oversized masks ($\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}) = 0.521/0.553$) to undersized masks with normal annotation style ($\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}) = 0.665/0.665$), shows lower scores in the first compared pair. An empirical choice for a threshold in which a re-inspection before the merging of datasets should be done can be the case of $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}) \leq 0.8$ since a perfect matching $\Gamma({}^{A}\mathcal{D}_{corr}, {}^{B}\mathcal{D}_{corr}) = 1.0$ will not be possible due to limited DNN performance. In general, the performance on the validation dataset can be used as suitable indicator for the selection of the threshold. The proposed inter-annotator variability is superior to no inspection except for diagonal elements indicated by the metrics RMSE and WRMSE. However, the exception in the case of diagonal elements is comprehensible since the DNN performance will not be able to predict the exact annotations depicted in the dataset.

The metrics RMSE and WRMSE show the similar trends and only small numeric deviation. Hence, both metrics can be used as indicators to measure errors. Besides, comparing RMSE to WRMSE, the benefits of considering uncertainty in the inspection pipeline become clear since errors can be reduced. Uncertainty awareness helps to reduce errors in the case of activated inspection. The improvement can be denoted as an interval ranging from 0.02 to 0.12 (cf. Table 3d and Table 3e).

V. CONCLUSION

Deep Learning (DL) enables a robust and accurate segmentation in biomedical imaging. The consistent and noise-free annotation of large amounts of data is a current challenge for annotators such as dermatologists or pathologist. We present a categorization of common annotation noise types. Furthermore, corresponding methods to simulate annotation noise are introduced. We demonstrate a degeneration of Deep Neural Network (DNN) performance in the cases of missing annotations, varying sizes of segmentation masks (oversized or undersized), and contour corruption.

However, the introduced noise types holes and jitter do not follow the expectation of reduced image segmentation performance. This finding is worthy of discussion and yields to the question of whether particular noise in segmentation masks can be beneficial. State-of-the-art approaches summarized as data augmentation focus on manipulating samples and do not manipulate segmentation masks in an uncoupled fashion. Further research needs to be conducted on whether unconventional segmentation mask augmentation may help to improve DNN generalization.

We proposed two automated inspection pipelines to inspect datasets w.r.t. intra-annotator and inter-annotator variability. The presented methods can be used in practical biomedical DL projects to evaluate annotation quality and to support medical practitioner during annotation. After finishing annotation, users, i.e., physicians, can focus on the most problematic annotations using our intra-inspection pipeline instead of re-inspecting all annotations. A novel inspection method of inter-annotator variability with no need for multiple annotations per sample allows detecting different annotation styles. This method is particularly useful for projects where image annotation is done via division of labor, i.e., each annotator is responsible for annotating only a subset of the entire dataset. Hence, in the case of varying annotation styles, we can raise warnings and thus show merge conflicts directly.

Using the biomedical ISIC 2017 Melanoma image segmentation dataset, we show the benefits of the proposed pipelines. Intra-annotator variability can be obtained in cases where noise is not dominant within a dataset. The corruption rate needs to be less than a half compared to the total number of samples. Besides, inter-annotator variability can be assessed in the case of deviating annotation styles. Especially, integrating uncertainty awareness improved both detection pipelines. Further, samples with a need for discussion concerning correctness are obtained. The possibility to show users difficult and uncertain samples offers potential regarding data acquisition to further improve DNN performance.

Part of future work may be the integration of the proposed algorithms as extensions inside state-of-the-art annotation tools [54]–[56]. The development of a Graphical User Interface (GUI) can boost the usability of the proposed methods for users in the domain medical imaging. Simulating noise in terms of mask sizes via distributions instead of taking sharp expectation values can be another part of the following experiments. Hence, this would yield to more variations in labels masks of the corresponding noise type and can be an alternative simulation approach. Besides, using the proposal for other image recognition tasks such as object detection or classification is a further objective. Thereby, DNN architecture, as well as metrics comparing prediction to ground-truth, needs to be adapted.

REFERENCES

[1] N. O. Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Wals, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision*. Cham, Switzerland: Springer, 2019, pp. 128–144.

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[3] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, and S. Wolfson, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020.

[4] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. E. Carpenter, "Nucleus segmentation across imaging experiments: The 2018 data science bowl," *Nature Methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[5] M. Böhland, L. Tharun, T. Scherr, R. Mikut, V. Hagenmeyer, L. D. R. Thompson, S. Perner, and M. Reischl, "Machine learning methods for automated classification of tumors with papillary thyroid carcinoma-like nuclei: A quantitative analysis," *PLoS ONE*, vol. 16, no. 9, pp. 1–21, 2021.

[6] K. Aggarwal, M. M. Mijwil, A.-H. Al-Mistarehi, S. Alomari, M. Gök, A. M. Z. Alaabdin, and S. H. Abdulrhman, "Has the future started? The current growth of artificial intelligence, machine learning, and deep learning," *Iraqi J. Comput. Sci. Math.*, vol. 3, no. 1, pp. 115–123, 2022.

[7] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Apr. 2018, pp. 168–172.

[8] V. Ulman, M. Maška, and C. Ortiz-de-Solorzano, "An objective comparison of cell-tracking algorithms," *Nature Methods*, vol. 14, no. 12, pp. 1141–1152, Oct. 2017.

[9] W. Chi, L. Ma, J. Wu, M. Chen, W. Lu, and X. Gu, "Deep learning-based medical image segmentation with limited labels," *Phys. Med. Biol.*, vol. 65, no. 23, Dec. 2020, Art. no. 235001.

[10] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101759.

[11] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," 2021, *arXiv:2103.14749*.

[12] S. Yu, M. Chen, E. Zhang, J. Wu, H. Yu, Z. Yang, L. Ma, X. Gu, and W. Lu, "Robustness study of noisy annotation in deep learning based medical image segmentation," *Phys. Med. Biol.*, vol. 65, no. 17, Aug. 2020, Art. no. 175007.

[13] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*.

[14] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," 2020, *arXiv:2007.08199*.

[15] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101693.

[16] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[17] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1919–1925.

[18] Y. Liu and H. Guo, "Peer loss functions: Learning from noisy labels without knowing noise rates," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6226–6236.

[19] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2233–2241.

[20] T. Scherr, K. Löffler, M. Böhland, and R. Mikut, "Cell segmentation and tracking using CNN-based distance predictions and a graph-based matching strategy," *PLoS ONE*, vol. 15, no. 12, pp. 1–22, 2020.

[21] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang, "CurriculumNet: Weakly supervised learning from large-scale web images," in *Computer Vision—ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 139–154.

[22] Z. Mirikharaji, Y. Yan, and G. Hamarneh, "Learning to segment skin lesions from noisy annotations," in *Domain Adaptation and Representation Transfer and Medical Image Learning With Less Labels and Imperfect Data*. Cham, Switzerland: Springer, 2019, pp. 207–215.

[23] K.-H. Lee, X. He, L. Zhang, and L. Yang, "CleanNet: Transfer learning for scalable image classifier training with label noise," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5447–5456.

[24] C. G. Northcutt, T. Wu, and I. L. Chuang, "Learning with confident examples: Rank pruning for robust classification with noisy labels," 2017, *arXiv:1705.01936*.

[25] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, Apr. 2021.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[27] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2902–2913.

[28] J. M. Köhler, M. Autenrieth, and W. H. Beluch, "Uncertainty based detection and relabeling of noisy image labels," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 2019, pp. 33–37.

[29] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11236–11245.

[30] S. A. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6965–6975.

[31] R. Hollandi, Á. Diósdi, G. Hollandi, N. Moshkov, and P. Horváth, "AnnotatorJ: An ImageJ plugin to ease hand annotation of cellular compartments," *Mol. Biol. Cell*, vol. 31, no. 20, pp. 2179–2186, Sep. 2020.

[32] S. Isele, M. Schilling, F. Klein, S. Saralajew, and J. Zoellner, "Radar artifact labeling framework (RALF): Method for plausible radar detections in datasets," in *Proc. 7th Int. Conf. Vehicle Technol. Intell. Transp. Syst.*, 2021, pp. 22–33.

[33] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.

[34] A. Schwaiger, P. Sinhamahapatra, J. Gansloser, and K. Roscher, "Is uncertainty quantification in deep learning sufficient for out-of-distribution detection?" in *Proc. Int. Joint Conf. Artif. Intell., Workshop Artif. Intell. Saf.*, 2020, pp. 1–8.

[35] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021.

[36] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 1050–1059.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[38] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5580–5590.

[39] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[40] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019.

[41] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Sci. Rep.*, vol. 10, no. 1, p. 5068, Dec. 2020.

[42] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[43] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, Jul. 2019.

[44] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Proc. Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds., Jul. 2018, pp. 1184–1193.

[45] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.

[46] 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*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[47] W. Falcon. (2019). *PyTorch Lightning*. Accessed: Jul. 20, 2021. [Online]. Available: https://github.com/PyTorchLightning/pytorch-lightning

[48] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.

[49] G. Bradski, "The OpenCV library," *Dr. Dobb's J. Softw. Tools Prof. Program.*, vol. 25, no. 11, pp. 120–123, 2000.

[50] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014.

[51] M. P. Schilling, O. Neumann, T. Scherr, C. Cui, A. A. Popova, P. A. Levkin, M. Götz, and M. Reischl, "A computational workflow for interdisciplinary deep learning projects utilizing bwHPC infrastructure," in *Proc. 7th bwHPC-Symp.*, Mar. 2022. [Online]. Available: https://publikationen.bibliothek.kit.edu/1000139674, doi: 10.5445/IR/1000139674.

[52] L. Biewald. (2020). *Experiment Tracking With Weights and Biases*. Accessed: Apr. 24, 2021. [Online]. Available: https://github.com/wandb

[53] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 281–305, Feb. 2012.

[54] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, and B. Hoff. (2020). *Computer Vision Annotation Tool (CVAT)*. Accessed: Jun. 21, 2021. [Online]. Available: https://github.com/openvinotoolkit/cvat

[55] M. P. Schilling, L. Rettenberger, F. Münke, H. Cui, A. A. Popova, P. A. Levkin, R. Mikut, and M. Reischl, "Label assistant: A workflow for assisted data annotation in image segmentation tasks," in *Proc. Workshop Comput. Intell.*, 2021, pp. 211–234.

[56] A. Bartschat. (2019). *Image Labeling Tool*. Accessed: Jun. 21, 2021. [Online]. Available: https://bitbucket.org/abartschat/imagelabelingtool

MARCEL P. SCHILLING received the bachelor's and master's degrees in mechanical engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Research Group "Machine Learning for High-Throughput and Mechatronics," Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include image processing, data-centric deep learning, and machine learning for high-throughput screening.

TIM SCHERR received the B.Sc. and M.Sc. degrees in physics from Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include biomedical image processing and deep learning applications.

FRIEDRICH R. MÜNKE received the bachelor's and master's degrees in mechanical engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Research Group "Machine Learning for High-Throughput and Mechatronics," Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include image processing, change detection in real world images, and machine learning for high-throughput screening.

OLIVER NEUMANN received the B.Sc. and M.Sc. degrees in computer science from the Karlsruhe Institute of Technology, Germany, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the Research Group "Machine Learning for Time Series and Images," Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include time-series forecasting, data analysis and transformation, and image processing.

MARK SCHUTERA is currently pursuing the Ph.D. degree in deep learning for autonomous driving with the Algorithms and Machine Learning Perception System Group within Research and Development, ZF Friedrichshafen AG, and the Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include deep learning, safety in autonomous driving, computer vision, and biomedical image processing.

RALF MIKUT received the Diploma degree in automatic control from the University of Technology, Dresden, Germany, in 1994, and the Ph.D. and Habilitation degrees in mechanical engineering from the University of Karlsruhe, Karlsruhe, Germany, in 1999 and 2007, respectively. Since 2011, he has been an Adjunct Professor at the Faculty of Mechanical Engineering and the Head of the Research Group "Automated Image and Data Analysis," Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His current research interests include machine learning, image processing, life science applications, and smart grids.

MARKUS REISCHL received the Dipl.-Ing. and Ph.D. degrees in mechanical engineering from the University of Karlsruhe, Karlsruhe, Germany, in 2001 and 2006, respectively. Since 2020, he has been an Adjunct Professor with the Faculty of Mechanical Engineering and is currently heading the Research Group "Machine Learning for High-Throughput and Mechatronics," Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany. His research interests include man–machine interfaces, image processing, machine learning, and data analytics.