Out of distribution detection for intra-operative functional imaging

Tim J. Adler[1,2](✉), Leonardo Ayala[1], Lynton Ardizzone[3], Hannes G. Kenngott[4], Anant Vemuri[1], Beat P. Müller-Stich[4], Carsten Rother[3], Ullrich Köthe[3], and Lena Maier-Hein[1](✉)

[1] Division Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Heidelberg, DE
{t.adler,l.maier-hein}@dkfz.de

[2] Faculty of Mathematics and Computer Science, Heidelberg University, DE

[3] Visual Learning Lab, Heidelberg University, DE

[4] Division of Minimally-invasive Surgery of the Department of General Surgery, Heidelberg University, DE

Abstract. Multispectral optical imaging is becoming a key tool in the operating room. Recent research has shown that machine learning algorithms can be used to convert pixel-wise reflectance measurements to tissue parameters, such as oxygenation. However, the accuracy of these algorithms can only be guaranteed if the spectra acquired during surgery match the ones seen during training. It is therefore of great interest to detect so-called *out of distribution* (OoD) spectra to prevent the algorithm from presenting spurious results. In this paper we present an information theory based approach to OoD detection based on the *widely applicable information criterion* (WAIC). Our work builds upon recent methodology related to *invertible neural networks* (INN). Specifically, we make use of an ensemble of INNs as we need their tractable Jacobians in order to compute the WAIC. Comprehensive experiments with *in silico*, and *in vivo* multispectral imaging data indicate that our approach is well-suited for OoD detection. Our method could thus be an important step towards reliable functional imaging in the operating room.

1 Introduction

The most commonly applied approach to computer aided surgery (CAS) relies on fusing pre-operative medical images with the current patient anatomy for augmented reality guidance. While this approach is well-suited for displaying subsurface structures detected in pre-operative images, such as tumors or vessels, a main bottleneck is the fact that it cannot account for tissue dynamics; live monitoring of perfusion, for example, is not possible with an approach that relies on 'offline images'. To address this shortcoming, recent research has focused on intra-operative functional imaging using biophotonics techniques. In this context, multispectral optical imaging is evolving as a key tool. Previous work has shown that machine learning algorithms can be used to convert pixel-wise reflectance measurements to tissue parameters, such as oxygenation [13,14]. These methods learn to infer tissue parameters via training samples providing a spectrum and the correct corresponding tissue parameter(s) (supervised learning). However, the accuracy of these algorithms is heavily effected by aleatoric and epistimic uncertainties [6].

In this paper, we argue for a multi-stage process for uncertainty handling as illustrated in Fig. 1. (1) To investigate whether the input is sufficiently close to the training data, *out of distribution* (OoD) detection is performed. (2) If the input is regarded as valid, the corresponding functional tissue parameters are computed, and a full posterior probability distribution is provided as output for each tissue parameter. As the second part of this pipeline has already been addressed in a recent publication [1, 2], we will focus on the first part. The following sections present and validate our proposed approach to OoD detection.

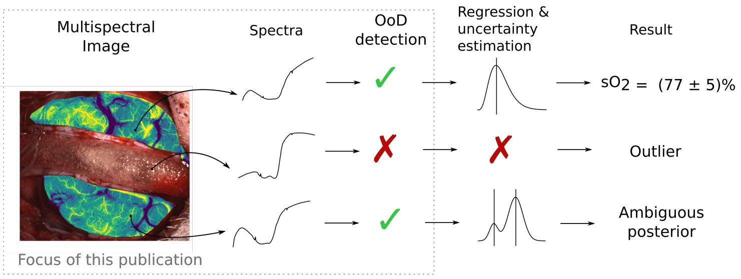


Fig. 1: Proposed multi-stage process for uncertainty handling in multispectral image analysis. To investigate whether the input (here: a spectrum) is sufficiently close to the training data, *out of distribution* (OoD) detection is performed. If the input is regarded as valid, the corresponding functional tissue parameters are computed. To address the potential inherent ambiguity of the problem a full posterior probability distribution rather than a point estimate is provided for each tissue parameter (here: blood oxygenation).

2 Methods

While we are not aware of any previous work in OoD detection in the field of optical imaging, the topic has gained increasing interest in the machine learning community. To implement the proposed multi-stage process for uncertainty handling in multispectral image analysis (Fig. 1), we build our method upon the work by Choi et al. [3] who proposed the *widely applicable information criterion* (WAIC) as a means to measure the closeness of a new sample to the training distribution. The advantage of this method lies in the fact that it outperforms many other ensemble based unsupervised learning methods while still being easily computable. An unsupervised approach is integral to the method as it is not feasible to generate enough labeled negative samples to train a discriminator between in- and outliers [3, 10]. The challenge in applying WAIC is the fact that it is an ensemble based method leading to the necessity of training a model multiple times. Depending on the data dimensions this can become prohibitively expensive both in terms of time and hardware requirements. In this work, we use invertible neural networks (INN) [2] to estimate WAIC on multispectral endoscopic imaging data.

In this section, we briefly revisit WAIC [3] and give an intuition for this quantity (Section 2.1), present the INN architecture as an integral ingredient to apply WAIC in the surgical domain (Section 2.2) and describe our experimental validation (Section 2.3).

2.1 Principle of WAIC

In the original contribution [12], WAIC was defined as

$$\mathrm{WAIC}(x) = \mathrm{Var}_{\Theta}[\log p(x \mid \Theta)] - \mathbb{E}_{\Theta}[\log p(x \mid \Theta)], \tag{1}$$

where $\mathrm{WAIC}(x)$ quantifies the proximity of a sample x to the distribution of the training data X^{tr}, and Θ is distributed according to $p(\Theta \mid X^{\mathrm{tr}})$. In a very recent publication [3] it was suggested to use WAIC as a means for OoD in the setting of neural networks.[1] The variance term in equation (1) measures 'how certain' the posterior distribution $p(\cdot \mid \Theta)$ is about a sample x, the heuristic being that it should be more certain about samples that are close to what it has seen before. The expectation term in equation (1) is used for normalization. The idea is that if the expectation of $\log p(x \mid \Theta)$ is high then the spread measured by the variance might also be larger without really measuring internal uncertainty of the model. Hence, it is subtracted to account for this effect.

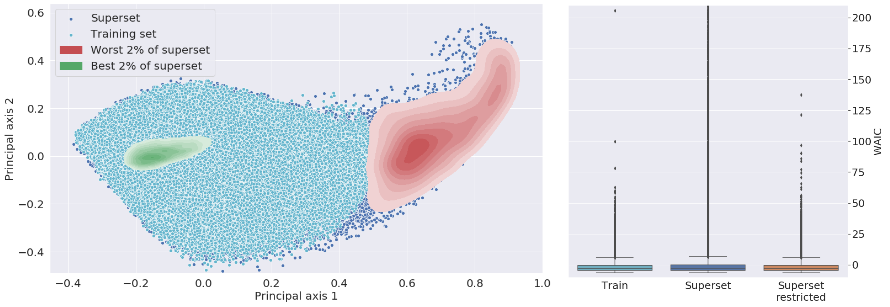
2.2 WAIC computation with INNs

WAIC only works for parametrized models. To meet this precondition, we use a deep neural network to encode the spectra X in a latent space Z following an analytically tractable distribution, which we chose to be a multivariate standard Gaussian. Let $f_{\Theta}\colon X \subset \mathbb{R}^n \to Z \subset \mathbb{R}^n$ denote the the neural network with parameters Θ. Then we can use the change of variable formula to compute the log-likelihood $\log p(x \mid \Theta)$ for a spectrum x as

$$\log p(x \mid \Theta) = -\frac{1}{2}\|f_{\Theta}(x)\|_2^2 - \frac{n}{2}\log(2\pi) + \log|\det Jf_{\Theta}(x)|, \tag{2}$$

where Jf_{Θ} denotes its Jacobian [11]. Equation (2) shows that it is mandatory for the log-Jacobi determinant of our network to be efficiently computable. One established architecture is the one of *normalizing flows* originally introduced in [4] and refined in [2] under the name of *invertible neural networks* (INN). For each of the experiments described in the next section, we trained an ensemble of INNs to estimate $p(\Theta \mid X^{\mathrm{tr}})$. Each network consisted of 10 layers of so called coupling blocks (see [4]) each followed by a permutation layer. Each coupling block consisted of a 3 layer fully connected network with ReLU activation functions. The networks were trained using Maximum-Likelihood training, i. e. by minimizing the loss $L(x) = -\log p(x \mid \Theta)$ as given in Equation (2) using the Adam optimizer [7].

[1] Please note that the sign convention of WAIC of Choi and Watanabe are opposite. We chose Watanabe's definition.



(a) (b)

Fig. 2: In silico validation. (a) The WAIC method trained on the training set (turquoise points, here projected onto the first two principal components (PCA) of the complete training set $X_{\mathrm{SC}}^{\mathrm{tr}}$) was applied to the superset (here: blue points). The 2 % percentile of best and worst superset spectra according the WAIC value are shown as a kernel density estimation in green (representing *in distribution* samples) and red (representing *out of distribution* samples) respectively. (b) The WAIC distribution of the training set, superset and restricted superset is shown (superset boxplot truncated).

2.3 Experiments

The purpose of our experiments was to validate our approach to OoD detection *in silico* (Section 2.3), and to present *in vivo* use cases (Section 2.3).

In silico quantitative validation In our simulation framework, a multispectral imaging pixel is generated from a 8-valued vector $\mathbf{t}_i$ of tissue properties, which are assumed to be relevant for the image formation process. Plausible tissue samples $\mathbf{t}_i$, are drawn from a layered tissue model as proposed in [14]. The framework was used to generate a data set X_{raw}, consisting of 550,000 high resolution spectra and corresponding ground truth tissue properties. It was split in a training $X_{\mathrm{raw}}^{\mathrm{tr}}$ and test set $X_{\mathrm{raw}}^{\mathrm{te}}$, comprising 500,000 and 50,000 spectra respectively. For the *in silico* quantitative validation we converted the (high resolution) spectra of the simulated data sets to plausible camera measurements using the filter response functions of the 8-band Pixelteq SpectroCam. We use a subscript (here: *SC* for SpectroCam) to refer to the data set X_{raw} after it was adapted to a certain camera. $X_{\mathrm{SC}}^{\mathrm{tr}}$ was split into a small training set $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ and a *superset* $X_{\mathrm{SC}}^{\mathrm{sup}}$, such that the support of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ lay within the support of $X_{\mathrm{SC}}^{\mathrm{sup}}$ and $X_{\mathrm{SC}}^{\mathrm{sup}}$ consisted of a cluster of data points outside of the support of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$, as illustrated in Fig. 2 (a). This led to a split of $X_{\mathrm{SC}}^{\mathrm{tr,d}}$ of approximately 49% of $X_{\mathrm{SC}}^{\mathrm{tr}}$ and $X_{\mathrm{SC}}^{\mathrm{sup}}$ of approximately 51% of $X_{\mathrm{SC}}^{\mathrm{tr}}$. An ensemble of five INNs was trained on $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ and the WAIC value was evaluated on $X_{\mathrm{SC}}^{\mathrm{sup}}$. We defined $X_{\mathrm{SC}}^{\mathrm{sup,r}}$ as the *reduced* data set of $X_{\mathrm{SC}}^{\mathrm{sup}}$ lying in the support of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$. We then investigated

(1) whether the WAIC distribution of the $X_{\mathrm{SC}}^{\mathrm{sup,r}}$ matches that of the $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ and whether (2) the part of $X_{\mathrm{SC}}^{\mathrm{sup}}$ not in the support of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ was correctly classified as outliers by our method.

In vivo application While the goal of the previous experiment was to confirm the validity of our approach in an *in silico* setting, the purpose of the *in vivo* experiments were to showcase applications in which the OoD detection could be useful.

Anomaly/Novelty detection: Detecting (parts of) a multispectral image in which the spectra do not closely match the training data distribution can be useful for many reasons. Possible applications include the detection of abnormal tissue or of artifical objects (e. g. instruments). To investigate this aspect, we used the complete $X_{\mathrm{SC}}^{\mathrm{tr}}$ to train an ensemble of five INNs. As *in vivo* test data, we acquired endoscopic images of porcine organs which we classified as *organs lying in the simulation domain* X^{iD} and *organs not lying in the simulation domain* X^{oD}. These spectra were acquired using a Pixelteq SpectroCam on a 30° Stortz laparascope with a a Stortz Xenon light source (Storz D-light P 201337 20). We classified liver, spleen, abdominal wall, diaphragm and bowl as in domain organs as hemoglobin can be assumed to be the main absorber in these. In contrast, we classified gallbladder as an out of domain organ, since bile is a notable absorber but has not been considered in our simulation framework. With this, X^{iD} and X^{oD} consisted of 50000 spectra and 10000 spectra respectively. Our hypothesis was that the WAIC values of X^{iD} should be much lower than those for X^{oD}. For reference, we also compared the resulting WAIC distributions to that of the simulated test data $X_{\mathrm{SC}}^{\mathrm{te}}$.

Detection of scene changes: Intra-operative image modalities often rely on a careful calibration of the device. When recovering blood oxygenation from multispectral measurements, for example, the regressor is typically trained with the light source that is used during test time. To investigate whether WAIC is applicable to detect illumination changes (which would substantially harm the method and render the estimation results invalid), we adapted X_{raw} to a xiQ XIMEA (Muenster, Germany) $SNm4 \times 4$ mosaic camera consisting of 16 bands assuming a Wolf LED light source (Wolf Endolight LED 2.2). We trained an ensemble of five INNs on $X_{\mathrm{Xim}}^{\mathrm{tr}}$. Furthermore, we recorded 200 512 × 272-pixel images of the lip of a healthy human volunteer using the xiQ XIMEA camera and a 30° Stortz laparascope (cf. Fig. 4 (b)). At around image 80 we switched the endoscope from a Stortz Xenon light source (Storz D-light P 201337 20) to a Wolf LED light source (Wolf Endolight LED 2.2). Based on the hypothesis that the switch in light source would be detected by WAIC analysis, we computed the WAIC time series for the region of interest depicted in Fig. 4 (a).

3 Results

In silico validation The distribution of the reduced training data $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ and the superset $X_{\mathrm{SC}}^{\mathrm{sup}}$ (projected to the first two principal components) can be found in Fig. 2 (a). It can be seen that the samples with poor WAIC values (red)

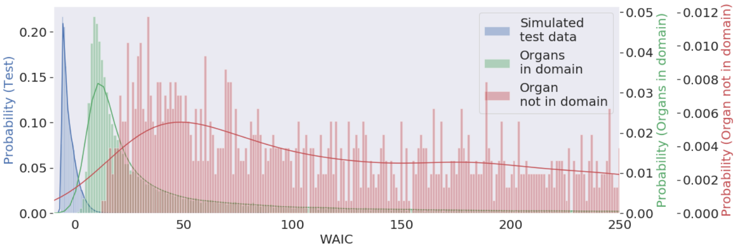


Fig. 3: Histogram of the WAIC values for the simulated test set, and *in vivo* multispectral measurements of organs that do (green) or do not (red) match the model assumptions based on which the training data was generated (tails truncated). Please note the different scales for the three distributions.

concentrate in the superset part not contained in $X_{\mathrm{SC}}^{\mathrm{tr,s}}$, whereas samples with low WAIC values (green) are contained in the interior of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$. Fig. 2 (b) shows a comparison between the WAIC distributions of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$, $X_{\mathrm{SC}}^{\mathrm{sup}}$ and the restricted superset $X_{\mathrm{SC}}^{\mathrm{sup,r}}$. The data sets $X_{\mathrm{SC}}^{\mathrm{tr,s}}$ and $X_{\mathrm{SC}}^{\mathrm{sup,r}}$ are in excellent agreement. The superset $X_{\mathrm{SC}}^{\mathrm{sup}}$ only differs in the regard that there are far more outliers, which can be accounted to the data points outside of $X_{\mathrm{SC}}^{\mathrm{tr,s}}$.

Application The WAIC distribution for the test data $X_{\mathrm{SC}}^{\mathrm{te}}$, the *in domain organs* X^{iD} and the *out of domain organ* X^{oD} can be found in Fig. 3. The distribution of the test data is by far the sharpest with a maximum aposterior probability (MAP) at -4.9. The distribution closely matches that of the training data (MAP = -4.9). The distribution of X^{iD} also possesses a sharp maximum, however with a far heavier tail. The MAP estimate yields 9.3 which indicates a still existent domain gap between our simulation domain and the organ domain. The distribution of X^{oD} is very noisy and has an even heavier tail than X^{iD}. The MAP lies at 34. This indicates that our WAIC estimate is suitable to distinguish between in domain tissue and out of domain tissue. Similarly, Fig. 4 illustrates that the change in illumination as performed in the *detection of scence changes* experiment results in a drastic change of WAIC values.

4 Discussion

The accuracy of machine learning-based regression methods in multispectral imaging crucially depend on whether the spectra acquired during surgery match the ones seen during training. Although initial steps with respect to uncertainty estimation and compensation have been taken in the field of optical imaging [1, 2, 5, 8, 9, 15], we are, to our knowledge, the first to address the problem of OoD detection to prevent algorithms from presenting spurious results.

The application to endoscopic organ data showed that our method is well-suited for anomaly detection. The in distribution organs are well separated from the out of distribution organ (gallbladder). Moreover, this experiment reveals a shortcoming of the simulation framework proposed by [14]: The large difference

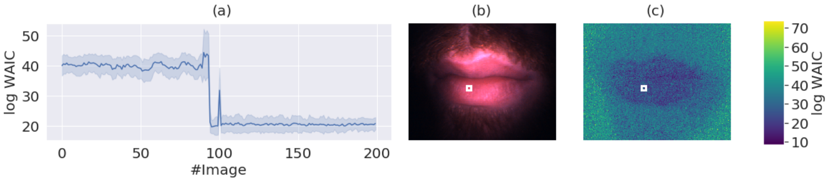


Fig. 4: Automatic detection of scene changes. (a) When a change in light source occurs, the mean log WAIC values computed for the white region of interest in (b)/(c) drop, indicating a decreasing domain gap between training and test data. (b) RGB image estimated using the 8-band measurement of human lips. (c) Corresponding WAIC values computed for the multispectral image.

in the WAIC distribution between the test set and the real data indicates a domain gap that remains to be tackled.

Our experiments with human lips show that WAIC is able to distinguish between different lighting conditions. The uncertainty prior to the change of lighting can most likely be explained by the short darkness stemming from the light source switch. The jump at image 100 was due to involuntary movement of the volunteer leading to the image being out of focus. One reason for the generally high WAIC values is the fact that melanin (a chromophore in the skin) was not simulated in the training data.

In the present implementation we used five INNs in our ensembles. According to preliminary experiments, this number is sufficient. We computed the WAIC on the data sets used for the anomality detection experiment (Section 2.3) for up to 20 ensemble members. For both the simulated test data and the in domain organs the values stabilized below $n = 10$. For the out of domain data (gallblader) the WAIC values increased throughout. This merits further investigation. However, there should be no impact on the method performance, as X^{iD} and X^{oD} were well separated.

Our findings underline the power of WAIC in the setting of medical OoD detection. However, there are still some open questions. A general downside of WAIC is its 'arbitrary units' and it is not straightforward to define a threshold for outlier detection. One approach to tackle this shortcoming would be to find a suitable normalization. Another possibility might be to just mask the worst n pixels in a certain ROI. Additionally, to this conceptual question, there are also practical limitations. The estimation of WAIC requires an *ensemble* of neural networks, which was feasible in our case, but becomes prohibitively expensive for larger input dimensions. For the future, methods for network compression might be adapted to tackle this problem.

In conclusion, this paper is the first to address the topic of OoD detection in intra-operative imaging. Due to the promising results obtained in this study, the approach proposed could not only become a valuable tool for increasing the reliability of machine learning-based regression methods but could also boost research in unsupervised intra-operative anomaly detection.

References

1. Adler, T.J., Ardizzone, L., Vemuri, A., Ayala, L., Gröhl, J., Kirchner, T., Wirkert, S., Kruse, J., Rother, C., Köthe, U., Maier-Hein, L.: Uncertainty-aware performance assessment of optical imaging modalities with invertible neural networks. International Journal of Computer Assisted Radiology and Surgery (2019)
2. Ardizzone, L., Kruse, J., Rother, C., Köthe, U.: Analyzing inverse problems with invertible neural networks. In: International Conference on Learning Representations (2019)
3. Choi, H., Jang, E., Alemi, A.A.: Waic, but why? generative ensembles for robust anomaly detection. CoRR (2018)
4. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using Real NVP. CoRR (2016)
5. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (2016)
6. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kohl, S.A.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.M.A., Rezende, D.J., Ronneberger, O.: A Probabilistic U-Net for Segmentation of Ambiguous Images (2018)
9. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports (2017)
10. Markou, M., Singh, S.: Novelty detection: a reviewpart 1: statistical approaches. Signal processing (2003)
11. Walter, R.: Real and complex analysis (1987)
12. Watanabe, S.: Algebraic geometry and statistical learning theory. Cambridge University Press (2009)
13. Wirkert, S.J., Kenngott, H., Mayer, B., Mietkowski, P., Wagner, M., Sauer, P., Clancy, N.T., Elson, D.S., Maier-Hein, L.: Robust near real-time estimation of physiological parameters from megapixel multispectral images with inverse Monte Carlo and random forest regression. International journal of computer assisted radiology and surgery (2016)
14. Wirkert, S.J., Vemuri, A.S., Kenngott, H.G., Moccia, S., Götz, M., Mayer, B.F.B., Maier-Hein, K.H., Elson, D.S., Maier-Hein, L.: Physiological Parameter Estimation from Multispectral Images Unleashed. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2017. Springer International Publishing (2017)
15. Zhu, Y., Zabaras, N.: Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. Journal of Computational Physics (2018)