Fuzzy Uncertainty-based Out-of-Distribution Detection Algorithm for Semantic Segmentation

Qiao Lin*, Xin Chen, Chao Chen, Direnc Pekaslan, Jonathan M. Garibaldi
School of Computer Science, University of Nottingham
Email:{*qiao.lin,@nottingham.ac.uk}

Abstract—**Deep learning models have achieved high performance in numerous semantic segmentation tasks. However, when the input data at test time do not resemble the training data, deep learning models can not handle them properly and will probably produce poor results. Therefore, it is important to design algorithms for deep learning models to reliably detect out-of-distribution (OOD) data. In this paper, we propose a novel fuzzy-uncertainty-based method to detect OOD samples for semantic segmentation. Firstly, to capture both data and model uncertainties, test-time augmentation and Monte Carlo dropout are applied to a ready-trained image segmentation model for generating multiple predicted instances of a given test image. Then interval fuzzy sets are generated from these multiple predictions to describe the captured uncertainty via distance transform operators. Finally, an image-level uncertainty score, which is calculated from the generated interval fuzzy sets, is used to indicate if it is an OOD sample. Experiments on testing three OOD test sets on a skin lesion segmentation model show that our proposed method achieved significantly higher classification accuracy in detecting OOD samples than three other state-of-the-art uncertainty-based algorithms.**

I. Introduction

DEEP convolutional neural networks (CNNs) based methods have achieved high performance in numerous semantic image segmentation tasks, but most methods assume that the training and test data are sampled from the same underlying distribution. However, the data used in the test time potentially consists of anomalous data that are out of the training data distribution. In this case, the semantic segmentation models tend to generate unsatisfactory segmentation results without informing the user, since there are no ground truth masks to assist the user in evaluating segmentation quality. In some scenarios, such as self-driving and clinical applications, the error-tolerance rate is considerably low. A tiny segmentation error may result in an unexpected car accident or a medical negligence. Thus, reliable and robust image segmentation models with the ability to detect out-of-distribution (OOD) test cases are highly desirable in practical applications.

Based on the literature, there are mainly three categories of OOD detection methods. 1) Distance-based methods calculate the distances from the test image to each of the training images [1], [2]. Each image is represented by its feature maps learned from the semantic segmentation model. 2) Softmax-based methods use the maximum value of softmax function outputs as the OOD measure [3]. 3) Learning-based methods normally introduce auxiliary models to assist the OOD detection [4]. However, in practice, these methods have their limitations. For example, distance-based methods are time-consuming as they need to iterate through all the training samples in each OOD detection. Softmax-based methods only take pixel-wise OOD detection into consideration and highly rely on the pixel-wise confidence values. Moreover, learning-based approaches require larger training datasets than the other methods, due to the large number of learnable parameters.

Different from the above methods, we explore the use of image-level uncertainty as a high level and reliable measure to detect OOD. The calculation of image-level uncertainty relies only on the output of the segmentation model [5], which is easy to implement and no learning process is required. When the uncertainty is higher than a certain value, it means the input image is potentially out of distribution and requires a manual examination. Roy et al.[6] measured the image-level uncertainty using the variance over the mean of all foreground pixels based on multiple predicted images that are produced by the image segmentation model (denoted as CV method). Mehrtash et al. [7] firstly calculated the pixel-wise uncertainty and then use the mean uncertainty value of the target region to represent the image-level uncertainty (APE). Both the CV and APE methods treated the low uncertainty and the high uncertainty areas equally without acknowledge that certain object regions are more important than other parts. For instance, the segmentation around the object boundary is more challenging and important than other regions. Therefore, our previous work [5] proposed a distance-to-boundary-aware method (denoted as FIU-SQ) to quantify the image-level uncertainty based on type-1 and general type-2 fuzzy sets.

On the other hand, interval fuzzy sets are also powerful to describe and calculate different kinds of uncertainties compared to type-1 fuzzy sets, and have low computational complexity in comparison to general type-2 fuzzy sets [8]. Ananthi et al. [9] employed interval fuzzy sets to remove the uncertainty in a noisy image. Xu et al. [10] leveraged interval fuzzy sets to handle the higher-order uncertainty of remote sensing image data. However, limited research works focus on the application of applying interval fuzzy sets to assess image segmentation uncertainty and then detect OOD samples.

In this paper, we propose to combine interval fuzzy sets and deep learning based method for OOD detection in medical image segmentation. Firstly, various image augmentation operators e.g. rotation, scale, flipping, etc. are applied to the input test image to generate several augmented images, which are then sent into a ready-trained image segmentation model. Both test-time augmentation and Monte Carlo dropout are adopted to generate a set of predicted outputs that correspond to the augmented input images for capturing the data and model uncertainties. Next, interval fuzzy sets are applied to

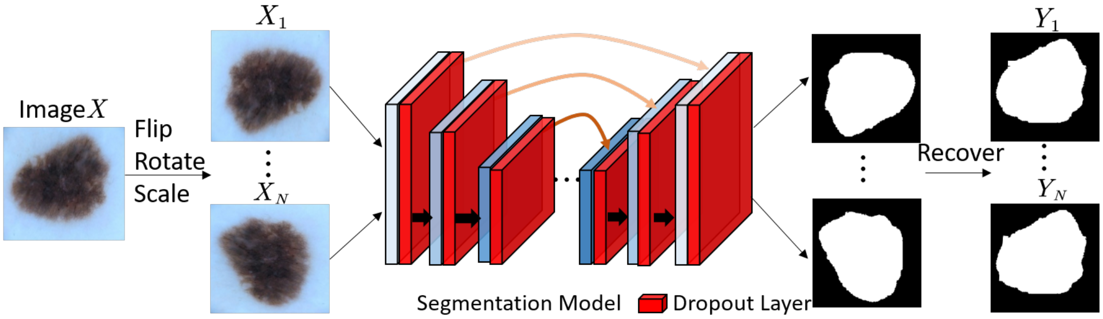


Fig. 1: The pipeline to generate multiple predicted images

describe the set of predicted images by considering pixel-to-boundary distance measures. Finally, interval-fuzzy-sets-based uncertainty is calculated to classify the in-distribution and OOD data. The key contributions of this paper are as below.

- We extend our previous study [5] and introduce interval fuzzy sets to calculate the segmentation uncertainty.
- A novel OOD detection algorithm is proposed using interval fuzzy sets and distance transform operators.
- We evaluated our method using a public skin lesion dataset as the training set and three other datasets (lung X-ray, Nuclei and skin lesion with added artificial noise) as the OOD test sets. The results show that our method is more powerful to detect OOD samples while maintaining higher classification accuracy compared to other state-of-the-art uncertainty-based methods.

The remainder of this paper is structured as follows: Section II introduces deep learning based image segmentation. Section III describes the detailed procedure of our proposed interval-fuzzy-sets-uncertainty-based OOD detection algorithm (FIU-OD). Section IV presents the datasets, implementation details, experimental design and results. Conclusions are drawn in Section V.

II. Background

Image segmentation is the process of dividing a digital image into various image regions, the goal of which is to reduce the complexity of the image and improve the efficiency of image analysis. Recently, with the development of deep learning technologies and the increasing amount of training data, convolutional neural networks (CNNs) and transformer modules have been widely used to construct segmentation models and then segment images via supervised learning. The deep-learning-based segmentation process is pixel-wise classification and normally called semantic segmentation. In 2015, Long et al. [11] firstly designed a pixel-to-pixel and end-to-end semantic segmentation model, known as Fully Convolutional Networks (FCN), based on multiple convolutional layers. Following the introduction of FCN, various improved methods e.g. Unet [12], SegNet [13], Deeplab [14], were developed and achieved excellent performance. With the advent of the transformer module [15] in 2017, many researchers adopted the transformer module to replace CNNs and achieved outstanding performance such as Swin transformer [16]. However, transformer-based methods are more time-consuming and require a larger amount of data for model training in comparison to CNN-based methods.

Note that our developed OOD detection algorithm can be considered as an add-on module to existing segmentation models, which is not highly dependent on the model architecture (as long as drop-out layers are included). In this paper, due to the popularity of Unet [12] in medical image segmentation over other CNN-based models, we adopt the Unet model as our backbone segmentation model.

III. A Novel Fuzzy-Uncertainty-Based OOD Detection Algorithm

The proposed fuzzy-uncertainty-based OOD detection algorithm consists of four steps. Firstly, a CNN-based image semantic segmentation model is trained based on a set of training images with their corresponding ground truth masks. Secondly, test-time augmentation (TTA) and Monte Carlo dropout (MCdropout) are applied to capture the data uncertainty and model uncertainty, respectively. Then interval fuzzy sets are adopted to describe the captured uncertainty. Finally, the interval-fuzzy-sets-based uncertainty value is utilized to achieve OOD detection in the model inference process.

A. CNNs-based Semantic Segmentation Model

In this paper, Unet [12] is implemented to segment medical images. Unet is a widely-used semantic segmentation model, which is mainly designed to process 2D medical images. The architecture consists of a contractive path, an expansive path, and skip connections. The contractive path is a down-sampling process including convolutional operators, relu activation function, and max-pooling operators. This process is employed to capture image features in multiple resolutions. However, while feature information grows during this process, spatial information diminishes. To recover spatial information, the expansive path applies an up-sampling process so that the output image and the ground truth image have the same sizes. The skip connections transmit the features learned in the contractive path to the expansive path by using concatenation operators, allowing the expansive path to combine the spatial and feature information simultaneously.

B. Test-time Argumentation and Monte Carlo Dropout

Once the Unet is ready-trained using the training images, it is used for image segmentation of unseen test images.

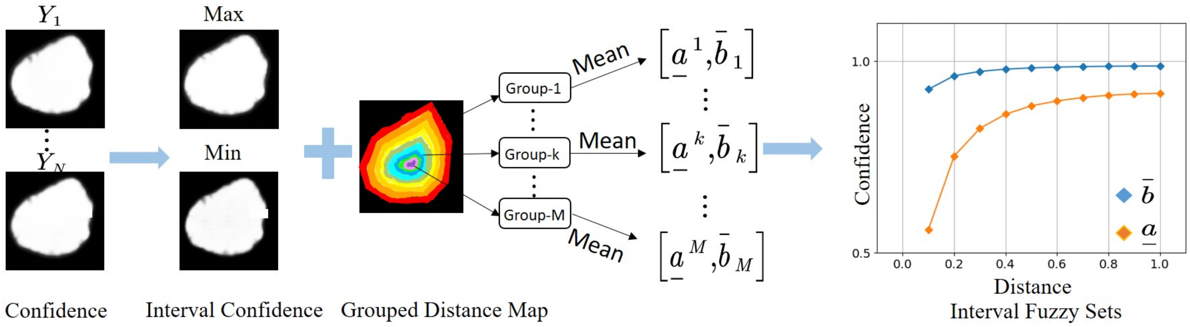


Fig. 2: The pipeline to obtain the interval fuzzy sets

During the test stage, TTA [17] and MCdropout [18] are used simultaneously to estimate the data uncertainty and model uncertainty, respectively, as illustrated in Fig. 1. For a given test image X, N variations $\{X_1, X_2, \cdots, X_N\}$ are created by using rotation, flipping, scale etc., transformation operators. Then these transformed versions of the given test image are input to the ready-trained segmentation model (Unet). Note that the dropout layers in the segmentation model are still active in the test stage for MCdropout calculation. Thus, for $\{X_1, X_2, \cdots, X_N\}$, N corresponding predicted images $\{Y_1, Y_2, \cdots, Y_N\}$ are generated and then used to calculate the segmentation uncertainty in the next step.

C. Grouped Distance Map Generation

Before calculating the segmentation uncertainty, a grouped distance map is generated to divide the segmented target region into M groups $\{p_1, p_2, \cdots, p_M\}$ based on each pixel's distance to the segmented object boundary. The reason is that the pixels having the same distance to the segmentation object boundary share the same uncertainty and different distances to the boundary imply different uncertainties. For pixels that are closer to the object boundary, the uncertainties are normally higher, because the object boundary is a critical and more challenging image region for segmentation. Thus dividing the pixels of the segmented region into different groups helps assessing the segmentation uncertainty more accurately. The detailed calculation procedure is described in Algorithm 1.

Algorithm 1 Grouped Distance Map

```
Input: $N$ predicted images $\{Y_1, Y_2, \cdots, Y_N\}$, the size of each image is $R \times C$, the number of distance groups is $M$.
Output: the grouped pixel locations $\{p_1, p_2, \cdots, p_M\}$
calculate the average predicted image $Y_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
obtain the binary image $\varLambda$ of $Y_{\text{avg}}$
apply the Euclidean distance transform algorithm [19] on $\varLambda$ to get the distance map $\varTheta$, in which each pixel value represents the minimum distance from this pixel location to the segmentation boundary
normalize the distances in $\varTheta$ into the range of $[0, 1]$
calculate the evenly distributed distance $\varepsilon$ for each group by $\varepsilon = \frac{1}{M}$
for $k$ from 1 to $M$ do
    $g_{\text{max}} = \varepsilon \times (k - 1)$
    $g_{\text{min}} = \varepsilon \times k$
    define the pixel locations group $p_k$, where the size of $p_k$ is $t = 0$
    for $i$ from 1 to $R$ do
        for $j$ from 1 to $C$ do
            if then $g_{\text{min}} \le \varTheta[i, j] \le g_{\text{max}}$
                add the location$(i, j)$ to the $p_k$ and the size of $p_k$ is updated by $t \leftarrow t + 1$
return $\{p_1, p_2, \cdots, p_M\}$
```

D. Interval-fuzzy-sets-based Uncertainty

After adopting the TTA and MCdropout, N predicted images $\{Y_1, Y_2, \cdots, Y_N\}$ are obtained. To capture the variation range of each pixel, an interval map $\varPhi$ is generated to describe the $\{Y_1, Y_2, \cdots, Y_N\}$. In $\varPhi$, each pixel value is an interval data and is represented as: $\varPhi^{[i,j]} = \left[\min\left(Y_1^{[i,j]}, Y_2^{[i,j]}, \cdots Y_N^{[i,j]}\right), \max\left(Y_1^{[i,j]}, Y_2^{[i,j]}, \cdots Y_N^{[i,j]}\right)\right]$, where i and j refer to the pixel location.

Then the interval map $\varPhi$ is divided into M groups based on each pixel's position $[i, j]$ and grouped distance map $\{p_1, p_2, \cdots, p_M\}$. If the position $[i, j]$ belongs to p_k, it means that $\varPhi^{[i,j]}$ is assigned into the k^{th} group. Thus, $\varPhi$ is represented by $(([a_1^1, b_1^1], [a_2^1, b_2^1], \cdots, [a_{\tau_1}^1, b_{\tau_1}^1]), \cdots, ([a_1^k, b_1^k], [a_2^k, b_2^k], \cdots, [a_{\tau_k}^k, b_{\tau_k}^k], \cdots, ([a_1^M, b_1^M], [a_2^M, b_2^M], \cdots, [a_{\tau_M}^M, b_{\tau_M}^M]))$ where a is the minimum value of the predicted values, b is the maximum value of the predicted values for a given pixel position. τ_k is the number of pixels in the k^{th} group. It is noted that pixel-wise information in a given image has little effect on inferring the overall segmentation quality since pixel-wise information is noisy in comparison to region-wise information [6]. Hence, in the next step, the τ_k pixel-wise intervals in the k^{th} group are aggregated into one interval to represent region-wise interval confidence. Based on the literature [20], the commonly-used interval data aggregation algorithm is Arithmetic mean.

Definition 1: Given n intervals $I_1, I_2, \cdots, I_n$, the Arithmetic mean is defined as

$$I_{Am} = \left[\frac{1}{n}\sum_{i=1}^{n} \underline{I}_i, \frac{1}{n}\sum_{i=1}^{n} \overline{I}_i\right], \tag{1}$$

where $\underline{I}_i = inf\,(I\,|I \in I_i)$ and $\overline{I}_i = sup\,(I\,|I \in I_i)$.

After applying Arithmetic mean on each group of $\varPhi$, region-wise interval confidence is obtained $\left(\left[\underline{a}^{1},\overline{b}^{1}\right],\cdots\left[\underline{a}^{k},\overline{b}^{k}\right],\cdots\left[\underline{a}^{M},\overline{b}^{M}\right]\right)$, where $\underline{a}^{k}$ is the aggregation of $\left(a_{1}^{k},a_{2}^{k},\cdots,a_{\tau_k}^{k}\right)$, and $\overline{b}^{k}$ is the aggregation of $\left(b_{1}^{k},b_{2}^{k},\cdots,b_{\tau_k}^{k}\right)$. The generated interval confidence sets can be treated as interval fuzzy sets (shown in Fig. 2). The x-axis is the pixel-to-boundary distance. The y-axis is the predicted value which refers to the degree of belonging to the segmentation object. $\left[\underline{a}^{1},\cdots,\underline{a}^{k},\cdots,\underline{a}^{M}\right]$ and $\left[\overline{b}^{1},\cdots,\overline{b}^{k},\cdots,\overline{b}^{M}\right]$ represent the lower membership function and the upper membership function, respectively.

The next step is to calculate the uncertainty of the interval fuzzy set. Based on the literature [21], the commonly-used interval fuzzy sets uncertainty formula is

$$U_{\mathrm{IFS}}\left(\tilde{A}\right)=\frac{1}{M}\sum_{k=1}^{M}\left[\overline{\mu}_{\tilde{A}}\left(x_k\right)-\underline{\mu}_{\tilde{A}}\left(x_k\right)\right] \tag{2}$$

where $\overline{\mu}_{\tilde{A}}\left(x_k\right)$ is $\overline{b}^{k}$, $\underline{\mu}_{\tilde{A}}\left(x_k\right)$ is $\underline{a}^{k}$ in our case. M is the number of membership values. This formula uses the average difference between the upper membership function and the lower membership function to represent the uncertainty of the generated interval fuzzy sets. Then the uncertainty value is applied to detect OOD by finding a threshold, which is determined using an in-distribution test set.

IV. Evaluation

In this section, the performance of our proposed FIU-OD algorithm is evaluated using public medical image datasets. The training (in-distribution) dataset is skin lesion and the OOD datasets are nuclei, lung and the skin lesion images with added Gaussian noise. The goal of the experiments is to discriminate the in-distribution and OOD datasets by finding a threshold of uncertainty value.

A. Datasets

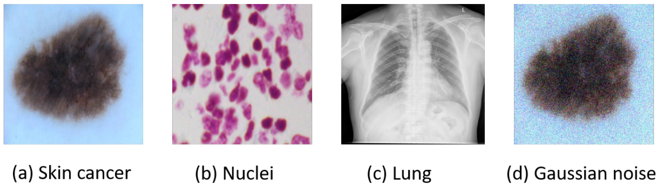
In-distribution dataset:

- Skin cancer (SK): this dataset is publicly available in the ISIC-2018 challenge (https://challenge2018.isic-archive.com/) that aims to segment skin lesions. It consists of 2594 raw dermoscopic images and their corresponding ground truth images. Each image only has one skin lesion region and it is a binary segmentation task.

Out-of-distribution datasets:

- Nuclei: unlike the SK dataset, the nuclei dataset has more than one object to be segmented. Moreover, the objects present various sizes and shapes. This dataset is obtained from the 2018 Data Science Bowl (https://www.kaggle.com/c/data-science-bowl-2018-/data). 140 images were used as OOD samples to test our method.
- Lung: this public dataset [22] contains lung x-ray images that all patients share similar shapes of lung regions to be segmented. 140 images were used as the OOD samples to test our method.
- Gaussian noise: this is not a strictly OOD dataset, which is generated from the SK dataset by adding Gaussian noises to each of the SK images. The mean and variance of the Gaussian noise were set to 0 and 30. The same 519 test images as the SK dataset were used for testing.

Fig. 3 shows the example images of in-distribution and out-of-distribution datasets.



(a) Skin cancer (b) Nuclei (c) Lung (d) Gaussian noise

Fig. 3: (a) is from the in-distribution skin lesion dataset. (b), (c) and (d) are from the out-of-distribution datasets of nuclei, lung and SK with noise respectively.

B. Implementation Detail

In this paper, Unet was used as the segmentation model for all experiments. The pixel size for the input and output of Unet is 256×256, which means that all images of different datasets were resized to the same size. This model has five layers in both the contractive (encoding) and expansive (decoding) paths. In the encoding stage, the number of kernels for the convolutional operators was 32 (1^{st} layer), 64 (2^{nd} layer), 128 (3^{rd} layer), 256 (4^{th} layer), and 512 (5^{th} layer). The kernel size of each layer was 3×3 with stride of 1, and maxpooling was adopted to down-sample the feature maps. During the decoding stage, de-convolution operations with the kernel size of 2×2 were adopted to up-sample the feature maps.

For all the segmentation tasks, the parameters of Unet were initialized from a uniform distribution. Unet was trained with cross-entropy loss. During the training process, early stopping was applied to avoid the over-fitting issue. Adam was the optimization algorithm with an initial learning rate of 0.0001 to update the parameters for Unet. Once the segmentation model is ready-trained, it is then used for uncertainty estimation and OOD detection for any given test image.

C. Experimental Design

We compared our method to three other uncertainty-based methods, namely CV [6], APE [7] and FIU-SQ [5]:

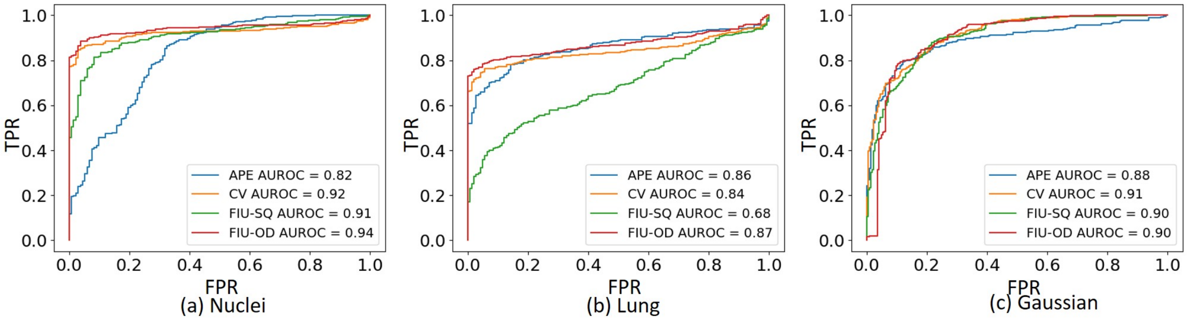
1) CV: After obtaining N predicted images $\{Y_1,Y_2,\cdots,Y_N\}$ by TTA and MCdropout, the region-wise uncertainty is calculated by the standard deviation σ over the mean μ:

$$U_{CV}=\frac{\sigma\left\{Y_1^S,Y_2^S,\cdots Y_N^S\right\}}{\mu\left\{Y_1^S,Y_2^S,\cdots Y_N^S\right\}} \tag{3}$$

where S is the segmented target region.

2) APE: Firstly, the pixel-wise uncertainty is calculated as

$$U_p=-\frac{1}{N}\sum_{i=1}^{N}p_i\left(x\right)\log\left(p_i\left(x\right)\right) \tag{4}$$



(a) Nuclei (b) Lung (c) Gaussian

Fig. 4: The ROC curves for three OOD detection algorithms: FIU-OD, CV, FIU-SQ and APE

where N is the number of the predicted images generated by TTA and MCdropout. $p(x)$ is the pixel-wise predicted value. Then the region-wise uncertainty is calculated as:

$$U_{APE} = \frac{1}{|S|} \sum_{i \in S} U_p^i, \tag{5}$$

where S is the segmented target region.

3) FIU-SQ: The following fuzzy entropy formula is chosen to calculate the uncertainty using FIU-SQ method.

$$U\left(A\right) = 1 - \frac{\left[\sum_{i=1}^{M} \left|2\mu_A\left(x_i\right) - 1\right|^2\right]^{\frac{1}{2}}}{M^{\frac{1}{2}}}, \tag{6}$$

where μ is the fuzzy membership function and M is the number of the elements in the defined fuzzy sets. A more detailed explanation of the formula (6) is presented in our previous work [5].

For method evaluation, each of the OOD datasets was added to the in-distribution SK test set to build a new test set. Then all methods were executed to generate an uncertainty value for each of the test images. If the segmentation uncertainty value is smaller than a threshold, it is classified as an in-distribution image (positive), otherwise an OOD image (negative). The higher the classification accuracy in discriminating in-distribution and OOD the better the algorithms' performance. The threshold of the uncertainty value is determined by fixing the true positive rate (TPR) at an acceptable level (application dependent) using the in-distribution test set. The area under the receiver operating characteristic curve (AUROC), classification accuracy and false positive rate (FPR) are used as the evaluation metrics for method comparison.

D. Experimental Results

The receiver operating characteristic (ROC) curves of the four methods on the three test sets are shown in Fig. 4. The AUROC values are reported in TABLE I. It is seen that our proposed FIU-OD algorithm performs better than all other methods on the nuclei and lung datasets with statistical significance (measured by Wilcoxon signed rank test with $p < 0.05$). While for the SK dataset with Gaussian noise, CV has better classification ability but no statistical significance to our method.

Furthermore, the distribution of the uncertainty values calculated by each method for the in-distribution SK test set and three OOD test sets are shown in Fig. 5. It is seen that our FIU-OD method (Fig.5 (a)) can better separate the in-distribution test images (blue line) from other OOD test images. The threshold that applies to the uncertainty value for OOD detection is determined by fixing the TPR at an acceptable level based on the in-distribution test set. We used 80% TPR as an example in our experiment for the skin lesion detection problem, which can be set differently dependent on applications. Based on 80% TPR, the thresholds for FIU-OD, CV, APE, and FIU-SQ are 0.473, 0.244, 0.086, and 0.325 respectively (black dash lines in Fig. 5). Then based on these threshold values, the classification accuracy of each method and FPR are calculated and reported in TABLE I. For the lung and nuclei datasets, our method has the highest classification accuracy and lowest FPR with statistical significance compared to other methods. For the SK dataset with Gaussian noise, APE has the highest accuracy but with no statistical significance to our method. It is observed that all methods performed similarly well for the SK with noise dataset, because this is an artificially corrupted in-distribution dataset that is relatively easy to handle. Note that current findings of the proposed OOD algorithm are only based on comparisons with three other OOD methods. To draw more solid conclusions, extra experiments and comparisons with other approaches may be needed.

V. Conclusions

This paper proposes an interval-fuzzy-sets-uncertainty-based OOD detection algorithm, which can be used to infer segmentation quality without access to the ground truth segmentation masks. This method firstly generates several predicted images to capture the data and model uncertainties by TTA and MCdropout. Then interval fuzzy sets are applied to quantify the captured uncertainty. Finally, a threshold is determined using TPR of the in-distribution test set to classify OOD data and in-distribution data. Experimental results show that our proposed FIU-OD method has better classification accuracy and lower FPR than three other state-of-the-art uncertainty-based methods. In the future work, more comprehensive studies could be performed with other (e.g. 3D) datasets and OOD approaches.

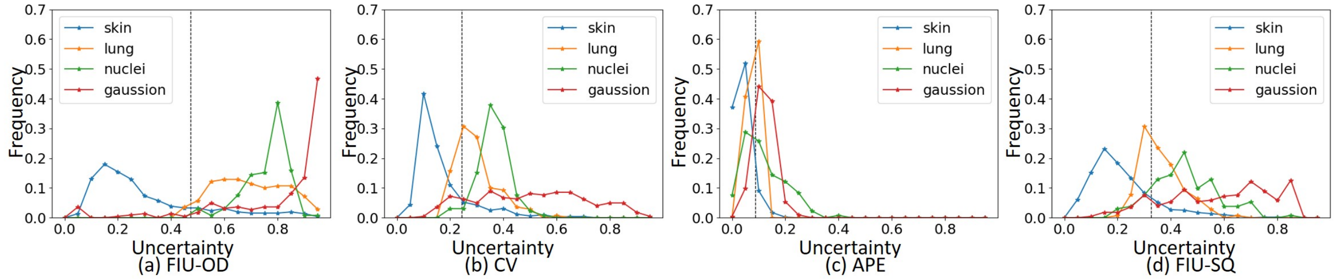


Fig. 5: The distribution of uncertainty values for (a) FIU-OD, (b) CV, (c) APE, and (d) FIU-SQ. The blue line is the in-distribution dataset. The green, yellow and red lines are OOD datasets. The black line is the selected threshold for each method.

TABLE I: Experimental results for FIU-OD, CV, APE and FIU-SQ in three OOD datsaets: Nuclei, Lung and Gaussian Noise. Mean ± standard deviation values are reported for all the evaluation measures. Wilcoxon signed rank test was used as statistical test. $*$ represents FIU-OD and CV are significantly different with $p < 0.05$; $\sharp$ represents FIU-OD and APE are significantly different with $p < 0.05$; $\diamond$ represents FIU-OD and FIU-SQ are significantly different with $p < 0.05$.

OOD Datasets	Methods	AUROC(%) ↑	Accuracy(%)↑	FPR(%) ↓
Nuclei	FIU-OD	$\mathbf{94.20 \pm 0.04}$ $*\sharp\diamond$	$\mathbf{84.00 \pm 0.07}$ $\sharp\diamond$	$\mathbf{0.00 \pm 0.00}$ $*\sharp\diamond$
	CV	92.31 ± 0.07	83.53 ± 0.25	2.28 ± 0.94
	APE	82.24 ± 0.02	77.84 ± 0.33	30.31 ± 0.72
	FIU-SQ	90.91 ± 0.01	82.30 ± 0.59	8.34 ± 0.94
Lung	FIU-OD	$\mathbf{87.33 \pm 0.13}$ $*\sharp\diamond$	$\mathbf{83.89 \pm 0.36}$ $*\sharp\diamond$	$\mathbf{1.43 \pm 0.89}$ $*\sharp\diamond$
	CV	84.39 ± 0.14	81.91 ± 1.11	10.72 ± 1.68
	APE	85.75 ± 0.06	80.24 ± 1.27	18.58 ± 2.10
	FIU-SQ	68.01 ± 0.01	80.09 ± 1.12	19.29 ± 2.10
Gaussian	FIU-OD	90.29 ± 0.83	83.64 ± 0.41	7.66 ± 1.80
	CV	$\mathbf{91.39 \pm 0.06}$	83.10 ± 0.52	9.46 ± 1.95
	APE	88.54 ± 0.17	$\mathbf{84.05 \pm 0.06}$	$\mathbf{6.31 \pm 0.37}$
	FIU-SQ	90.23 ± 0.09	82.02 ± 0.35	13.07 ± 0.27

References

[1] G. Golub and C. Van Loan, “Matrix computations 4th edition the johns hopkins university press,” *Baltimore, MD*, 2013.

[2] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” *Advances in neural information processing systems*, vol. 31, 2018.

[3] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” *arXiv preprint arXiv:1706.02690*, 2017.

[4] Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. L. Yuille, “Synthesize then compare: Detecting failures and anomalies for semantic segmentation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 145–161.

[5] Q. Lin, X. Chen, C. Chen, and J. M. Garibaldi, “A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty,” *IEEE Transactions on Fuzzy Systems*, DOI: 10.1109/TFUZZ.2022.3228332.

[6] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, “Inherent brain segmentation quality control from fully convnet monte carlo sampling,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 664–672.

[7] K. Hoebel, V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. Müller, and J. Kalpathy-Cramer, “An exploration of uncertainty information for segmentation quality assessment,” in *Medical Imaging 2020: Image Processing*, vol. 11313. International Society for Optics and Photonics, 2020, p. 113131K.

[8] K. T. Atanassov, “Interval valued intuitionistic fuzzy sets,” in *Intuitionistic fuzzy sets*. Springer, 1999, pp. 139–177.

[9] V. Ananthi and P. Balasubramaniam, “A new image denoising method using interval-valued intuitionistic fuzzy sets for the removal of impulse noise,” *Signal Processing*, vol. 121, pp. 81–93, 2016.

[10] J. Xu, G. Feng, T. Zhao, X. Sun, and M. Zhu, “Remote sensing image classification based on semi-supervised adaptive interval type-2 fuzzy c-means algorithm,” *Computers & geosciences*, vol. 131, pp. 132–143, 2019.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[17] 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, 2019.

[18] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[19] P. F. Felzenszwalb and D. P. Huttenlocher, “Distance transforms of sampled functions,” *Theory of computing*, vol. 8, no. 1, pp. 415–428, 2012.

[20] S. Ferson, V. KREINOVICK, L. Ginzburg, and F. SENTZ, “Constructing probability boxes and dempster-shafer structures,” Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia ..., Tech. Rep., 2003.

[21] D. Wu and J. M. Mendel, “Uncertainty measures for interval type-2 fuzzy sets,” *Information sciences*, vol. 177, no. 23, pp. 5378–5393, 2007.

[22] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, *et al.*, “Automatic tuberculosis screening using chest radiographs,” *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 233–245, 2013.