Toward unbounded open-set recognition to say "I don't know" for glomerular multi-lesion classification

Paulo Chagas[a], Luiz Souza[a], Rodrigo Calumby[b], Izabelle Pontes[a], Stanley Araújo[c], Angelo Duarte[b], Nathanael Pinheiro[d], Washington Santos[e], and Luciano Oliveira[a]

[a]Ivisionlab, Federal University of Bahia, Bahia, Brazil
[b]University of Feira de Santana, Bahia, Brazil
[c]Federal University of Minas Gerais, Minas Gerais, Brazil
[d]IMAGEPAT, Bahia, Brazil
[e]Fundação Oswaldo Cruz – Instituto Gonçalo Moniz, Bahia, Brazil

ABSTRACT

Glomeruli are histological structures located at the beginning of the nephrons in the kidney, having primary importance in the diagnosis of many renal diseases. Classifying glomerular lesions is time-consuming and requires experienced pathologists. Hence automatic classification methods can support pathologists in the diagnosis and decision-making scenarios. Recently most of state-of-the-art medical imaging classification methods have been based on deep-learning approaches, which are prone to return overconfident scores, even for out-of-distribution (OOD) inputs. Determining whether inputs are OOD samples is of underlying importance so as to ensure the safety and robustness of critical machine learning applications. Bearing this in mind, we propose a unified framework comprised of unbounded open-set recognition and multi-lesion glomerular classification (membranous nephropathy, glomerular hypercellularity, and glomerular sclerosis). Our proposed framework classifies the input into in- or OOD data: If the sample is an OOD image, the input is disregarded, indicating that the model "doesn't know" the class; otherwise, if the sample is classified as in-distribution, an uncertainty method based on Monte-Carlo dropout is used for multi-lesion classification. We explored an energy-based approach that allows open-set recognition without fine-tuning the in-distribution weights to specific OOD data. Ultimately, our results suggest that uncertainty estimation methods (Monte-Carlo dropout, test-time data augmentation, and ensemble) combined with energy scores slightly improved our open-set recognition for in-out classification. Our results also showed that this improvement was achieved without decreasing the 4-lesion classification performance, with an F1-score of 0.923. Toward an unbounded open-set glomerular multi-lesion recognition, the proposed method also kept a competitive performance.

Keywords: Glomeruli, multi-lesion classification, deep-learning, out-of-distribution detection, uncertainty estimation, energy functions.

1. INTRODUCTION

Kidney disease markers are mostly found in the glomerular structures, presenting varied and heterogeneous features. The glomerulus is a histological structure located at the beginning of the nephrons in the kidney, formed by a network of capillaries. The main function of the glomeruli is the filtration of the blood that leads to

Further author information: (Send correspondence to Luciano Oliveira)
Paulo Chagas: E-mail: paulo.chagas@ufba.br
Luiz Souza: E-mail: luiz.souza@ufba.br
Rodrigo Calumby: E-mail: rtcalumby@uefs.br
Izabelle Pontes: E-mail: izabelle.rocha@ufba.br
Stanley Araújo: E-mail: stanleyaa@gmail.com
Angelo Duarte: E-mail: angeloduarte@uefs.br
Nathanael Pinheiro: E-mail: nathan.pinheiro@me.com
Washington Santos: E-mail: washington.santos@fiocruz.br
Luciano Oliveira: E-mail: lrebouca@ufba.br

urine production.[1] As a primary filtering structure, the glomerulus is vulnerable to many types of lesions, leading to various primary and systemic diseases. Although lesion recognition is time-consuming and requires experienced pathologists, who quite frequently do not reach a broad consensus,[2] this is an essential step to the diagnosis and treatment of many renal diseases.[3] In this context, considering recent advances in artificial intelligence methods and computer-aided-diagnosis applied to medical imaging, automatic recognition of glomerular lesions emerges as a promising alternative to aid pathologists in kidney diseases diagnose.

Deep-learning approaches, more specifically convolutional neural networks (CNN), have achieved state-of-the-art results in several medical imaging tasks,[4] including glomerular specific[5] and multi-lesion[6] recognition. The main shortcoming of these neural architectures is that CNN predictions are usually based on softmax scores, which are prone to return overconfident results.[7] This situation is cumbersome mainly when the target image belongs to a different domain (*e.g.*, not a glomerulus), to an unknown class of the same domain (*e.g.*, glomerulus with an unknown lesion), or even when the classifier just makes a confident, but incorrect prediction. These situations where the unknown classes have no constraints, *i.e.*, when we do not bound the inputs to glomerular domain or non-glomerular domain, are called unbounded open-set recognition.[8] Since we do not want to separate whether the image is a glomerulus with a novel lesion or a non-glomerulus, we can consider our problem as an unbounded recognition: We only care about classifying the images into in- or out-of-distribution (OOD*).

While OOD detection avoids providing a lesion label to subjects other than a glomerulus, or even a forced misprediction for a glomerulus with an unknown lesion, robust glomerular multi-lesion recognition methods should be able to reject unknown objects by considering a response for both in- and OOD images. By accounting for all these challenges, and including an explicit "I don't know" answer, a multi-lesion classifier turns into an open-set recognition tool.[9] This is precisely the context for building robust glomerular lesion recognition methods, considering for instance the numerous types of glomerular lesions and multiple constraints in terms of small amount of labeled data, unsatisfactory recognition effectiveness for specific lesions, and finally lack of prediction uncertainty.

Considering open-set recognition tasks, some methods include expensive training procedures to enhance the generalization of models for OOD data or to deal with known-unknown classes.[8] In an unbounded open-set context, such approach becomes infeasible as for instance it is not actually possible to account for all unknown classes. Aiming to reduce this expensive process of fine-tuning to specific OOD data, we propose a open-set recognition approach that is only trained on the glomerular multi-lesion data set. While previous work focus on closed-set single-lesion or multi-lesion recognition, also susceptible to overconfident predictions, we propose moving to an unbounded open-set glomerular multi-lesion recognition. For that, we conceived an effective integrated framework that accounts for prediction uncertainty with an energy-based method and also exploits energy distributions to deal with OOD data and unknown glomerular lesions. We hypothesize that an energy-based method (see Section 3.1) combined with uncertainty estimation approaches (see Section 3.2) allow better in- and OOD separation. To assess our hypothesis validation, we defined a multi-lesion scenario, including membranous nephropathy, glomerular hypercellularity, and glomerular sclerosis. In summary, the main contributions of this work are twofold: (i) An integrated framework to simultaneously handle multi-lesion glomerular classification and open-set recognition; (ii) we demonstrate that combining energy scores with specific uncertainty estimation methods can improve OOD detection, compared with using energy scores only.

2. RELATED WORK

Glomerular lesion classification has already been studied by few works.[10,11] Just as there are different medical taxonomies for glomerular classification, the previous works approach different types of lesions. Barros *et al.*[5] proposed a glomerular hypercellularity recognition method based on feature extraction via classical image processing algorithms and k-nearest neighbors. Their work used images stained with hematoxylin-eosin (H&E) and periodic acid–Schiff (PAS) from a set of biopsy slides provided by Gonçalo Moniz Institute (FIOCRUZ). Using the same data set, Chagas *et al.*[6] applied a custom CNN combined with support vector machines for hypercellularity classification, but also performing a new annotation assessing the following sub-types: endocapillary hypercellularity, mesangial hypercellularity, and both lesions.

*It is also called out-of-domain.

Closer to a wider multi-lesion scenario, Zeng *et al.*[10] developed a classification pipeline for PAS-stained images considering the following lesions: glomerular hypercellularity, global sclerosis, segmental sclerosis and crescents. The proposed approach relied in different deep-learning-based architectures for different tasks: U-NET for glomerulus segmentation; and a custom pipeline for lesion classification combining DenseNet-121, LSTM-CGNet and V-Net. Considering an even larger set of lesions, Uchino *et al.*[11] proposed a deep-learning-based classification method for: global sclerosis, segmental sclerosis, endocapillary proliferation, mesangial matrix accumulation, mesangial cell proliferation, crescents and membranous nephropathy. Uchino *et al.*'s approach[11] was based on a simple training of a binary classifier for each lesion, using InceptionV3[12] as convolutional backbone. Even though their work aggregated more lesions to the task, they assessed the models over an extremely unbalanced data set, which resulted in a non reliable performance for some lesions. Both Zeng *et al.*'s[10] and Uchino *et al.*'s[11] works perform multi-lesion glomerular classification considering predictions from a single neural network model. As aforementioned, these predictions are usually based on softmax scores and tend to be overconfident even to OOD data. To the best of our knowledge, no study was conducted explicitly considering open-set recognition on glomerulus data sets. For a closely related task, Cicalese *et al.*[13] and Chagas *et al.*[14] tackled the problem of uncertainty estimation on glomerular lesion classification. Although uncertainty estimation and open-set recognition are not the same task, they are related in a higher level of abstraction, since both tasks also aim to approach the problem of overconfident outputs from neural networks. Cicalese *et al.*'s work[13] developed a Lupus level classification framework on PAS-stained images, considering uncertainty estimation using Monte-Carlo Dropout combined with a DenseNet-121[15] architecture. Chagas's work[14] focused on membranous nephropathy classification of H&E-stained images, presenting a comparison between Resnet-18, DenseNet-121 and Wide-ResNet architectures, also including uncertainty estimation with Monte-Carlo dropout. Both works applied test-time data augmentation for model prediction, increasing robustness and capturing both model and data uncertainty.[16]

There are several works on OOD detection, but most of them focuses on fine-tuning pretrained models with specific OOD data.[17,18] Among them, we explored the method proposed by Liu *et al.*,[19] which consists of using energy scores instead of softmax scores to distinguish in- and OOD data. Their approach allows one to compute energy scores directly from a pretrained model, without the need to fine-tuning to OOD samples. Nevertheless, their report shows that fine-tuning can lead to a better separation between in- and OOD data. Considering this limitation, we decided to adopt their energy-based approach and investigate the combination with an uncertainty estimation method to better separate in- and OOD data without the extra cost of retraining the model. The proposed method is based on Monte-Carlo dropout, test-time data augmentation and ensemble-based classification for uncertainty estimation. These uncertainty methods introduce data and model randomness, which we believe can widen the difference between in- and OOD of the energy scores.

3. UNBOUNDED OPEN-SET RECOGNITION FOR GLOMERULAR MULTI-LESION CLASSIFICATION

We propose a unified framework for open-set recognition on glomerular multi-lesion classification. The open-set recognition task can be defined as unbounded because we do not differentiate the types of OOD data, *i.e.*, we do not consider prior information about the unknown classes. This way, our proposed framework groups novel glomerular lesions and out-of-domain data into the same OOD class. Figure 1 lays out our proposed framework while detailing the following tasks approached in this work. For the first task, we propose a novel method by combing energy-based scores (see Section 3.1) and uncertainty estimation approaches (see Section 3.2). We adopted Monte-Carlo dropout, test-time data augmentation, and ensemble-based model for uncertainty estimation. These uncertainty-based approaches rely in adding randomness on both data and model, inspiring us to bring the hypothesis that this variability can widen the in- and OOD energy scores. For the latter task, we used the traditional softmax scores to find the most probable lesion. Since the adopted uncertainty methods are based on sampling images/weights N times, both energy and softmax scores are averaged for these N predictions and for the M models from the ensemble.

3.1 Energy-based out-of-distribution detection

To achieve robust predictions, a model should be able to return trustworthy scores for different input domains. If we have a model trained on images of glomeruli, that model should indicate if a new input belongs to a different

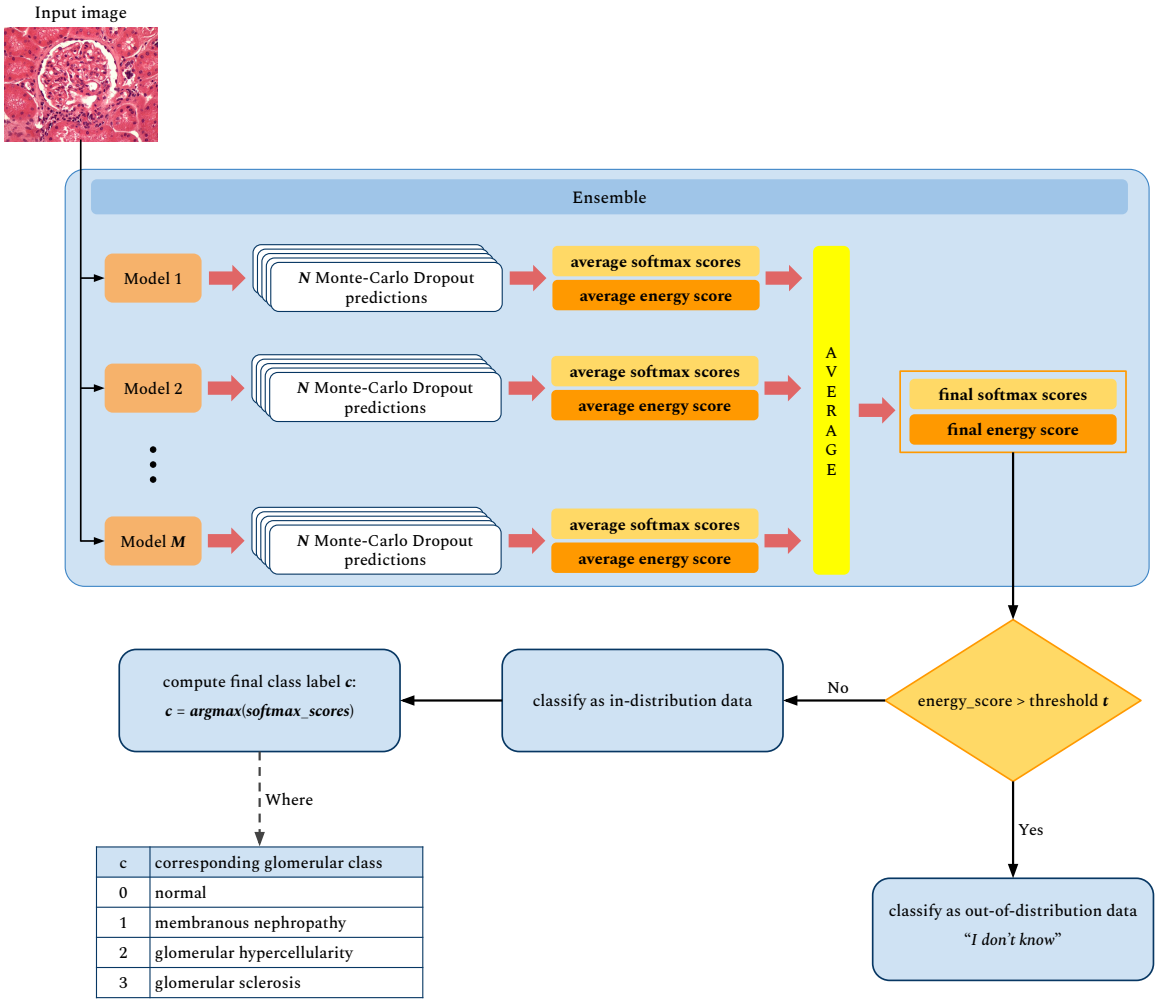


Figure 1: Proposed framework for open-set recognition on multi-lesion glomerular classification. Our framework combines uncertainty estimation methods with energy scores. For OOD detection, a threshold value is set to classify a new energy score. If the sample is classified as in-distribution, the lesion is assigned using averaged softmax scores. Alternatively, if the sample is classified as OOD, we can interpret the output as the model saying "*I don't know*", where we can disregard the sample or further evaluate it.

distribution. This task is called OOD detection and usually relies in the fine-tuning of pretrained models to detect specific OOD data. Our goal was post-processing the output of pretrained models by an energy function, without retraining the model considering specific data.[19] An energy function $E(x) : \mathbb{R}^D \rightarrow \mathbb{R}$ can be defined as a function that maps each point of x to a single scalar called *energy*.[20] The energy scores are computed from the Helmholtz free energy, given by

$$E(x; f) = -T * log(\sum_{i}^{K} e^{f_i(x)/T}),$$

where K is the number of classes, $f(x)$ is the output from the last dense layer of the neural network, and T is a temperature scaling value equal to 1 in the most cases. Using these energy scores from training and OOD data, we set an energy threshold value to classify new images as in- or OOD samples (see Fig. 1). Liu *et al.*[19] demonstrate that one can use their energy function as a scoring function for any pretrained network or as a trainable cost function to fine-tune the model. However, fine-tuning the model led to better results, which

Table 1: Summary of the data sets used in the experiments.

Glomerular lesion data set				OOD data sets	
Normal	Membranous	Hipercellularity	Sclerosis	Caltech101	BreakHis
869	2,066	1,237	510	9,146	9,109

brought us to the question of *"how to improve the energy-based OOD detection without the costly fine-tuning process?"*. This question led us to hypothesize that by adding uncertainty estimation methods, we could improve the separation of in- and OOD scores.

3.2 Uncertainty estimation combination

Our assumption is that, by introducing randomness on both data and model for uncertainty estimation, we can achieve a wider variance in the energy distributions. As we performed a K-fold cross-validation (see Section 4), we needed to train $K = M$ different models using M different training sets. For each model, we estimated the uncertainty by combining Monte-Carlo dropout[21] with test-time data augmentation.[16] Monte-Carlo dropout consists on keeping dropout layers activated during inference time, performing N predictions for each input image. The final class prediction is usually the average of these N predictions, and the uncertainty score is commonly assigned as the variance of these predictions. We inserted the dropout layer just before the last dense layer, where we evaluated different probability values (see Section 4 for details). Test-time data augmentation has a similar approach. Instead of using data augmentation only during training, we also applied it during inference time for N predictions, averaging these predictions to a final output. The combination of these two methods is illustrated in Fig. 1, where for each one of the M models random augmentation and random dropout are applied during inference, performing an averaging fusion at the end. For the multi-lesion classification, we average the softmax scores, as we aim at selecting the most probable class. Alternatively, we average the energy scores for OOD detection.

As the cross-validation returns M models trained on the glomerulus data set, we used these models as a M-model ensemble (see Fig. 1). Ensemble approaches consist of training different models (either differing from random weight initialization or different training data) and performing a final prediction reducing all predictions (usually) by averaging them. By adopting ensemble of multiple models we hope to increase reliability and validity of the final output.[22]

4. EXPERIMENTAL ANALYSIS

4.1 Data set

We used anonymized images selected from the digital histological image library of FIOCRUZ, resulting in 4,682 H&E-stained images of human glomerulus with the following lesions: membranous nephropathy, glomerular hypercellularity, glomerular sclerosis, and images with no lesion (referred as "normal"). For more data acquisition details, refer to Chagas *et al.*[14] For open-set recognition, we used Caltech101[23] and BreakHis[24] data sets. Caltech101 is a data set of object pictures belonging to 101 categories, mostly used for object detection. BreakHis is a breast cancer histopathological image classification data set, containing images of breast tumor tissue collected using different magnifying factors (40X, 100X, 200X, and 400X). We chose these two sets to investigate OOD detection in two types of input data: broad object images and histology images. We wanted to evaluate how the energy scores of a "completely different" data set (Caltech) and a "quite related" one (BreakHis) associated with the glomerular data set. Despite the BreakHis belongs to a different domain, this data set contains H&E-stained images, which have color distributions quite similar to the glomerular data set. Also, to make the BreakHis images "closer" to the training set, we sampled images considering only the same magnifying factor used on the acquisition of glomerular images. Table 1 presents the class distribution of the glomerular multi-lesion data set and the two OOD data sets.

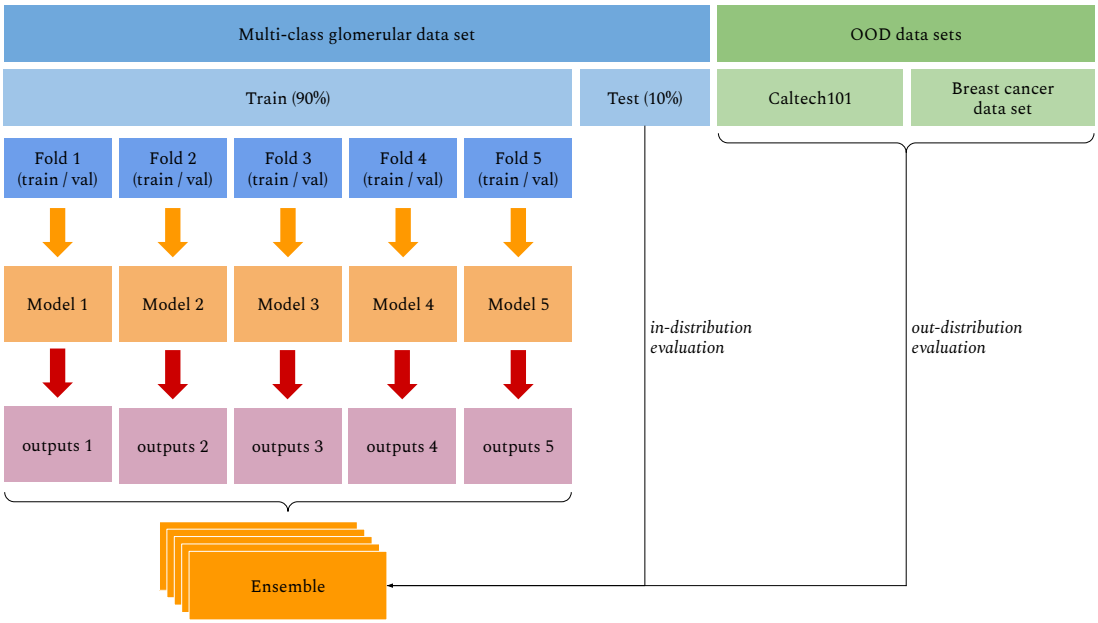


Figure 2: Ensemble-based evaluation protocol. Orange arrows represent the selection of the best model of each fold considering F1-score on the validation set. Red arrows represent model predictions, which were evaluated with and without Monte-Carlo dropout.

4.2 Evaluation protocol

The proposed evaluation protocol is illustrated in Fig. 2. The protocol was developed to consider two tasks: 1) Multi-lesion glomerular classification, and 2) OOD evaluation. For the first one, we use the multi-class glomerular data set for training, validation and testing. We separated 10% of the entire data set for testing, leaving the rest to carry on a 5-fold cross-validation. Then, we used the best model of each validation fold considering the F1-score criterion to perform a final prediction using a 5-model ensemble. For the second task, we used the 5-model ensemble, while computed the energy-based scores to achieve a measurement that can be used to distinguish in- and OOD data. For faster evaluation and to decrease the difference in size between the glomerular data set and the OOD data sets, we sampled 450 images from each OOD data set, which is a number close to the glomerular test set size (469).

Evaluation metrics: For the multi-lesion glomerular classification, we measured the following metrics: F1-score, precision and recall. For the OOD detection task, since the task becomes a binary classification (*in* or *out*), we measured the area under the receiver operating characteristic curve (AUROC), false positive rate (FPR) and false negative rate (FNR).

4.3 Training procedure

The first part of our evaluation protocol is training a deep network model in a 5-fold cross-validation setup, considering a test set for final evaluation. As deep-learning-based architectures have achieved state-of-the-art results in several medical imaging classification tasks,[4] we decided to use a CNN backbone for multi-lesion glomerular classification. Our goal here is not to select the best architecture for the problem, but instead we want to investigate the usage of a general CNN backbone combined with uncertainty estimation methods and energy scores. In this context, we adopted the Wide-ResNet architecture[15] as CNN backbone, which is a ResNet[25] variant with reduced depth and increased width. We used the Pytorch framework[26] for training and evaluating the models. All models were pretrained on ImageNet[27] data set for faster convergence, updating the final layer to four neurons respective to the four classes tackled in this work. AdamW optimizer[28] was used with a initial learning rate of 0.0001 with decay of 0.1 at every 30 epochs, considering a total of 100 epochs. All

Table 2: Comparative results of different Wide-ResNet configurations for multi-lesion glomerular classification. These results refer to the weighted average metrics use to assess multi-classification performance.

Method	F1-score	Precision	Recall
Without Monte-Carlo Dropout	0.923	0.923	0.923
With Monte-Carlo Dropout ($p = 0.2$)	0.919	0.919	0.919
With Monte-Carlo Dropout ($p = 0.5$)	0.923	0.923	0.923

Table 3: Confusion matrix for Wide-ResNet predictions on test set with Monte-Carlo Dropout ($p = 0.5$). The rows represent ground-truth classes, and the columns represent predicted classes.

	N	M	H	S	F1-score
Normal (N)	81	4	0	2	0.915
Membranous (M)	7	195	3	2	0.944
Hypercellularity (H)	2	2	115	5	0.935
Sclerosis (S)	0	5	4	42	0.824

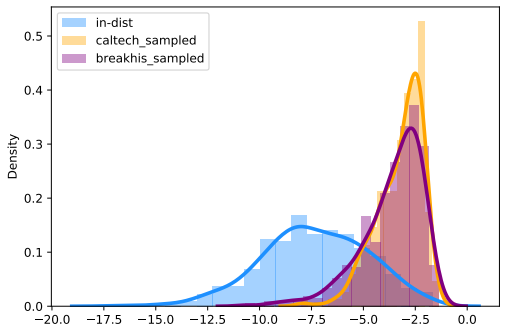
experiments were executed on a machine with 8GB RAM and an NVIDIA GEFORCE GTX 1060. For training and uncertainty estimation we used several random transformations at each batch for image augmentation. These transformations include: resizing the smallest dimension of each image to 224 pixels followed by a random crop of size 224×224 (thus keeping original aspect ratio); color transformations such as random contrast, random gamma, and random brightness; noise addiction with gaussian noise, affine transformations, and random white squares with size of 10% of image height (224 pixels).

4.4 Results and discussion

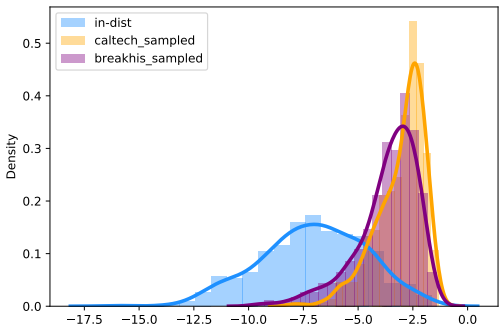
Table 2 shows the classification measures for the multi-lesion glomerular classification. We compared the Wide-ResNet with or without Monte-Carlo dropout. For the experiment without Monte-Carlo dropout, we computed the final scores by averaging the predictions from the five models. For the Monte-Carlo dropout cases, we used 50 predictions for each one of five models, and evaluated the dropout probability values of 0.2 and 0.5. We can note that all results are quite close considering the selected measures, without a clear predominance of a network. To better visualize inter-class predictions, we also present the confusion matrix in Table 3. Also, we computed the F1-score for each class, which showed high scores for all classes. As expected, the framework performed worse for glomerular sclerosis, as it is the most underrepresented class and is known to have some morphological features similar to membranous nephropathy. This similarity can occur due to a sub-type of membranous nephropathy that can include other lesions. Further studies need to be performed to address this confusion between membranous nephropathy and other lesions in general.

For OOD detection, we computed the energy scores for in- and OOD, considering the proposed Wide-ResNet variations. To compute the evaluation measures, we analyze these energy scores and define a threshold value. Figure 3 presents the energy histograms of in- and OOD, highlighting the fact that the OOD curves are quite close, with some minor variations of density in some areas. Another noteworthy behaviour is that Caltech101 and BreakHis energy distributions are very similar, almost overlapping. Considering that those data sets are from very different domains, and BreakHis is also a data set of histology images, the plots indicate a robust grouping of OOD data.

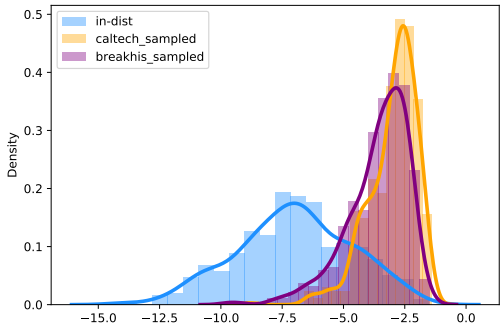
We assessed the OOD detection in two scenarios: Binary and multi-class classification. In the first scenario, the energy threshold defines the input samples as in- or OOD data. We evaluated several threshold values ranging from the minimum to the maximum energy, using a step of 0.01. For each threshold value t, if the energy score is smaller than t, our framework provides a prediction as an in-distribution data; otherwise, the image is predicted as OOD. With these predictions, we computed the AUROC to evaluate which Wide-ResNet variation better separates in- and OOD data. Table 4 summarizes the AUROC alongside to false positive and false negative rates for each Wide-ResNet configuration, showing that Wide-ResNet with Monte-Carlo Dropout ($p = 0.5$) outperformed the other variations. As expected, TPRs and TNRs followed an order behaviour similar to the AUROC for all three models. In addition, no great divergence between TPRs and TNRs was observed, indicating no dominance for in- or out-distribution classes.



(a) Without Monte-Carlo Dropout



(b) With Monte-Carlo Dropout ($p = 0.2$)



(c) With Monte-Carlo Dropout ($p = 0.5$)

Figure 3: In- and OOD energy scores of different Wide-ResNet configurations.

Table 4: Comparative results of different Wide-ResNet configurations for OOD detection. The presented results refer to measurements for binary classification.

Method	AUROC	FPR	FNR
Without Monte-Carlo Dropout	0.847	0.111	0.194
With Monte-Carlo Dropout ($p = 0.2$)	0.847	0.146	0.157
With Monte-Carlo Dropout ($p = 0.5$)	**0.854**	**0.101**	**0.189**

The results shows that a highest dropout probability allowed for a higher AUROC, which indicates a better in- and OOD separation. This behaviour corroborates with our initial assumption that introducing both data and model randomness might lead to a better OOD separation. Thus, we can conclude that the uncertainty estimation methods have potential to improve the OOD detection with energy-based scores. In addition, this improvement was achieved without decreasing the classification performance and without retraining the models to specific OOD data.

The last OOD detection scenario was the multi-class classification. In this context, instead of considering predictions of in- and OOD only, we considered the "unknown" alternative to indicate as not belonging to a known lesion class. Since the in- and OOD still have a relevant overlapping area (see Fig. 3), it is expected to see a loss in performance compared to the binary classification and the previous glomerular multi-lesion classification. Table 5 summarizes the confusion matrix for the Wide-ResNet with Monte-Carlo Dropout ($p = 0.5$) in this new scenario of unknown class for multi-class recognition. Similar to the previous experiments, glomerular sclerosis had the worst results, with the method providing an "I don't know" answer for many images. Nevertheless, the weighted average F1-score considering all classes was 0.850, which is a promising result.

5. CONCLUDING REMARKS

In this work, we proposed a unified framework for open-set recognition and multi-lesion glomerular classification. Considering our prior decision of using energy-based scores for OOD detection, we hypothesize whether uncer-

Table 5: Confusion matrix for Wide-ResNet predictions with Monte-Carlo Dropout ($p = 0.5$) for multi-lesion glomerular classification considering an *unknown* class. The rows represent ground-truth classes, and the columns represent predicted classes.

	N	M	H	S	IDK	F1-score
Normal (N)	70	2	0	0	15	0.737
Membranous (M)	3	165	2	0	37	0.819
Hypercellularity (H)	2	0	102	2	18	0.745
Sclerosis (S)	0	2	2	26	21	0.650
"I don't know" (IDK)	28	27	44	1	800	0.893

tainty estimations methods can improve the OOD detection performance. We demonstrated that our hypothesis indeed improved AUROC for in-out recognition using models pretrained on glomerular multi-lesion classification only. The main contribution resided in the improvements of OOD detection that were achieved without fine-tuning to specific OOD data. In addition, the introduction of randomness of the uncertainty estimation approaches did not result in performance loss compared with the original ensemble model without uncertainty methods. Even though our proposed framework considers the open-set binary classification as a step before multi-lesion classification (see Fig. 1), the multi-class OOD detection was important to verify what class was the most misclassified. And not surprisingly, these misclassifications were occurring in the most underrepresented class.

As future work, we plan to investigate if different approaches of uncertainty estimation have the same improvement for OOD detection. Since Monte-Carlo dropout and test-time data augmentation comprises an uncertainty estimation method based on sampling and averaging, it is indeed necessary to check how other Bayesian methods, such as variational inference, might influence the energy scores distributions. Considering Monte-Carlo Dropout, we want to study how the number and parameters of dropout layers influence the energy score distribution. Even though the highest dropout probability achieved the best result, an optimal value should be found. We also plan to develop a novel method to minimize the intersection area between in- and out-of-distribution energy scores.

REFERENCES

[1] Weinstein, J. R. and Anderson, S., "The aging kidney: physiological changes," *Advances in chronic kidney disease* **17**(4), 302–307 (2010).

[2] Barisoni, L., Lafata, K. J., Hewitt, S. M., Madabhushi, A., and Balis, U. G., "Digital pathology and computational image analysis in nephropathology," *Nature Reviews Nephrology* **16**(11), 669–685 (2020).

[3] Fogo, A. B., "Approach to renal biopsy," *American Journal of Kidney Diseases* **42**(4), 826–836 (2003).

[4] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

[5] Barros, G. O., Navarro, B., Duarte, A., and Dos-Santos, W. L., "Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys," *Scientific reports* **7**(1), 1–8 (2017).

[6] Chagas, P., Souza, L., Araújo, I., Aldeman, N., Duarte, A., Angelo, M., Dos-Santos, W. L., and Oliveira, L., "Classification of glomerular hypercellularity using convolutional features and support vector machine," *Artificial intelligence in medicine* **103**, 101808 (2020).

[7] Hein, M., Andriushchenko, M., and Bitterwolf, J., "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 41–50 (2019).

[8] Roady, R., Hayes, T. L., Kemker, R., Gonzales, A., and Kanan, C., "Are open set classification methods effective on large-scale datasets?," *PLOS ONE* **15**(9), 1–18 (2020).

[9] Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E., "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(7), 1757–1772 (2013).

[10] Zeng, C., Nan, Y., Xu, F., Lei, Q., Li, F., Chen, T., Liang, S., Hou, X., Lv, B., Liang, D., et al., "Identification of glomerular lesions and intrinsic glomerular cell types in kidney diseases via deep learning," *The Journal of pathology* **252**(1), 53–64 (2020).

[11] Uchino, E., Suzuki, K., Sato, N., Kojima, R., Tamada, Y., Hiragi, S., Yokoi, H., Yugami, N., Minamiguchi, S., Haga, H., et al., "Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach," *International Journal of Medical Informatics* **141**, 104231 (2020).

[12] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2818–2826 (2016).

[13] Cicalese, P. A., Mobiny, A., Shahmoradi, Z., Yi, X., Mohan, C., and Van Nguyen, H., "Kidney level lupus nephritis classification using uncertainty guided bayesian convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics* **25**(2), 315–324 (2020).

[14] Chagas, P., Souza, L., Pontes, I., Calumby, R., Angelo, M., Duarte, A., dos Santos, W. L., and Oliveira, L., "Deep-learning-based membranous nephropathy classification and monte-carlo dropout uncertainty estimation," in [*Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*], 257–268, SBC (2021).

[15] Zagoruyko, S. and Komodakis, N., "Wide residual networks," *arXiv preprint arXiv:1605.07146* (2016).

[16] Ayhan, M. S. and Berens, P., "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in [*International Conference on Medical Imaging with Deep Learning*], (2018).

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

[18] Liang, S., Li, Y., and Srikant, R., "Enhancing the reliability of out-of-distribution image detection in neural networks," in [*International Conference on Learning Representations*], (2018).

[19] Liu, W., Wang, X., Owens, J., and Li, Y., "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[20] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F., "A tutorial on energy-based learning," *Predicting structured data* **1**(0) (2006).

[21] Gal, Y. and Ghahramani, Z., "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in [*Proceedings of The 33rd International Conference on Machine Learning*], Balcan, M. F. and Weinberger, K. Q., eds., *Proceedings of Machine Learning Research* **48**, 1050–1059, PMLR, New York, New York, USA (20–22 Jun 2016).

[22] Segebarth, D., Griebel, M., Stein, N., von Collenberg, C. R., Martin, C., Fiedler, D., Comeras, L. B., Sah, A., Schoeffler, V., Lüffe, T., et al., "On the objectivity, reliability, and validity of deep learning enabled bioimage analyses," *Elife* **9**, e59780 (2020).

[23] Fei-Fei, L., Fergus, R., and Perona, P., "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in [*2004 conference on computer vision and pattern recognition workshop*], 178–178, IEEE (2004).

[24] Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L., "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering* **63**(7), 1455–1462 (2015).

[25] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[26] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems 32*], Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., eds., 8024–8035, Curran Associates, Inc. (2019).

[27] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision* **115**(3), 211–252 (2015).

[28] Loshchilov, I. and Hutter, F., "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101* (2017).