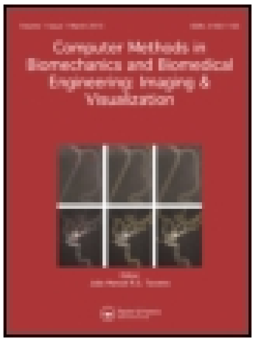

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tciv20

Uncertainty-aware membranous nephropathy classification: A Monte-Carlo dropout approach to detect how certain is the model

Paulo Chagas, Luiz Souza, Izabelle Pontes, Rodrigo Calumby, Michele Angelo, Angelo Duarte, Washington Lc-Dos Santos & Luciano Oliveira

To cite this article: Paulo Chagas, Luiz Souza, Izabelle Pontes, Rodrigo Calumby, Michele Angelo, Angelo Duarte, Washington Lc-Dos Santos & Luciano Oliveira (2022): Uncertainty-aware membranous nephropathy classification: A Monte-Carlo dropout approach to detect how certain is the model, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, DOI: 10.1080/21681163.2022.2029573

To link to this article: https://doi.org/10.1080/21681163.2022.2029573

Published online: 04 Feb 2022.

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
https://doi.org/10.1080/21681163.2022.2029573

Uncertainty-aware membranous nephropathy classification: A Monte-Carlo dropout approach to detect how certain is the model

Paulo Chagas[a], Luiz Souza[a], Izabelle Pontes[b], Rodrigo Calumby[c], Michele Angelo[c], Angelo Duarte[c], Washington Lc-Dos Santos[b] and Luciano Oliveira[a]

[a]IvisionLab, Universidade Federal Da Bahia, Bahia, Brazil; [b]Fundação Oswaldo Cruz, Instituto Gonçalo Moniz, Bahia, Brazil; [c]Departamento de Tecnologia, Universidade Estadual de Feira de Santana, Bahia, Brazil

ABSTRACT

Membranous nephropathy (MN) is among the most common glomerular diseases that cause nephrotic syndrome in adults. To aid pathologists on performing the MN classification task, we proposed here a pipeline consisted of two steps. Firstly, we assessed four deep-learning-based architectures, namely, ResNet-18, MobileNet, DenseNet, and Wide-ResNet. To achieve more reliable predictions, we adopted and extensively evaluated a Monte-Carlo dropout approach for uncertainty estimation. Using a 10-fold cross-validation setup, all models achieved average F1-scores above 92%, where the highest average value of 93.2% was obtained by using Wide-ResNet. Regarding uncertainty estimation with Wide-ResNet, high uncertainty scores were more associated with erroneous predictions, demonstrating that our approach can assist pathologists in interpreting the predictions with high reliability. We show that uncertainty-based thresholds for decision referral can greatly improve classification performance, increasing the accuracy up to 96%. Finally, we investigated how the uncertainty scores relate to complexity scores defined by pathologists.

ARTICLE HISTORY
Received 1 October 2021
Accepted 11 January 2022

KEYWORDS
Membranous nephropathy; deep learning; uncertainty estimation

1. Introduction

Membranous nephropathy (MN) is a common autoimmune glomerular disease, usually associated with the cause of nephrotic syndrome in adults. The main characteristic of MN is the large quantity of immune complex sediments on the epithelial cells, visually indicated by a thickening in the glomerular basement membrane. Figure 1 depicts a normal glomerulus (left) and another one with MN (right), where we can identify the presence of thickened membranes (boundaries of white areas inside the right glomerulus). Distinguishing these visual characteristics is not a trivial task, demanding trained pathologists that do not always reach a consensus. This way, automatic classification approaches can aid pathologists in the decision-making pipeline. By developing deep learning models, computer vision applications have significantly advanced through time and its full potential for diagnostic-driven studies is still being investigated (Lit- jens et al. 2017).

Towards MN classification, the results in the literature (Chen et al. 2020; Uchino et al. 2020) have mostly relied on limited and highly unbalanced data sets, which hinder the development of reliable models. In terms of model variability, just a few deep learning architectures have been evaluated: we highlight the U-Net for glomeruli segmentation Chen et al. 2020), and a few other traditional convolutional neural networks (CNN) for classification, e.g. InceptionV3 (Uchino et al. 2020) and ResNet (Chen et al. 2020).

In the literature, the proposals for MN classification have focused exclusively on label-only classification, presenting no supplementary data for the pathologist decision-making pipeline (Chen et al. 2020; Uchino et al. 2020). Begoli et al. (2019) emphasise the need to estimate a reliable uncertainty score for medical imaging applications, stating that a proper uncertainty score can help in both research and real-life problems in the medical field. An optimal uncertainty metric should correlate to erroneous predictions, indicating that a high uncertainty score leads to a highly 'confused' model about its prediction.

Uncertainty information could be helpful when specialists interpret the model output: *How much should the pathologists trust the input and its respective prediction?* Hence we propose here an evaluation pipeline composed of two tasks: i) evaluating deep learning architectures for MN classification, and ii) performing and investigating an uncertainty estimation approach for supportive information using the best architecture from task i). We adopted Monte-Carlo dropout (Gal and Ghahramani 2016) for uncertainty estimation because it does not require extensive modifications in the baseline architectures. Also, Monte-Carlo dropout showed noteworthy results in several works in medical imaging field (better described in Section 1.1). Inspired by the work of Combalia et al. (2020), which combines Monte-Carlo dropout and test-time data augmentation (TTA), we extensively evaluated Monte-Carlo with and without TTA along with different parameters. We also examine how our uncertainty scores correlated with erroneous classifications, and, more specifically, how these scores were distributed for each class. In addition, we evaluated whether the uncertainty scores can be used with a threshold to drop an 'uncertain' image and consequently improve the classification

CONTACT Paulo Chagas paulo.chagas@ufba.br IvisionLab, Universidade Federal Da Bahia, Bahia, Brazil

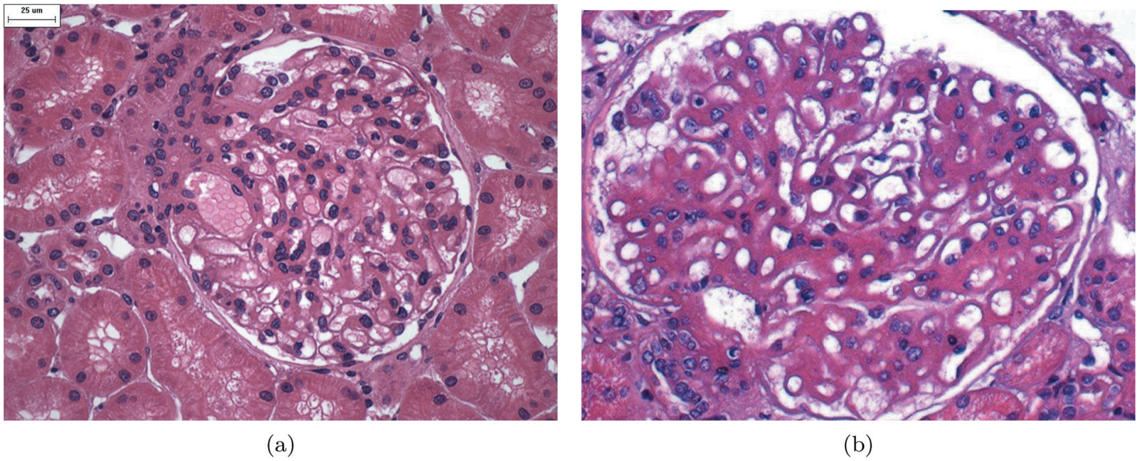

Figure 1. Example of a) a glomerulus with no lesion, and b) a glomerulus with membranous nephropathy.

metrics. Lastly, we investigate the following question: *'Are high uncertainty images "harder" for pathologists to diagnose?'*. Trying to answer this question, we collected complexity scores (ranging from 1 to 5, representing from 'no complexity' up to 'highest complexity') from five pathologists using the ten highest and ten lowest uncertainty images. Our goal is to explore the correlation between uncertainty scores and pathologist-defined complexity scores.

The main contributions of this work are listed as follows:

- The experiments were carried out by using a large amount of MN images, allowing a more reliable evaluation of our classification pipeline;
- a diverse set of deep learning architectures was used, achieving top results for all of them;
- thorough investigation of our proposed uncertainty estimation experiments with Monte-Carlo dropout, considering different parameters and the impact of using uncertainty thresholds; and
- analysis of the relationship between uncertainty scores and pathologists-defined complexity scores.

1.1. Related work

Uchino et al. (2020) carried out an extensive evaluation of deep learning models for classification of several glomerular lesions, including MN. They used class-specific binary models for each lesion by fine-tuning an InceptionV3 (Szegedy et al. 2016) network. Particularly for MN, among an amount of 3841 images, there were only 167 MN images. Even though their results showed high performance for some lesions, lower performance was achieved for the MN cases ($AUC = 0.816 \pm 0.034$ and $AUC = 0.734 \pm 0.011$ for PAS-stained and PAM-stained images, respectively). This decrease in performance might be due to the weakly representation of MN compared to the entire data set. Our work seeks to tackle this issue by using a more balanced data set.

Chen et al. (2020) introduced SPIKE-NET, a two-phase deep learning approach for MN recognition. Their proposed pipeline consists of an initial segmentation followed by classification. U-Net (Ronneberger et al. 2015) is used for glomeruli segmentation, while ResNet (He et al. 2016) is used (since it outperformed AlexNet (Krizhevsky 2014) and VGG16 (Simonyan and Zisserman 2014)) for the classification task. More specifically, SPIKE-NET achieved an accuracy of 94.44% against 92.86% and 91.27% from UNet-VGG16 and UNet-AlexNet, respectively. Although achieving considerably high results, their experiments were also conducted in a limited data set with 1,267 glomeruli (653 with MN and 614 normal), where only 126 glomerulus images were used in a single final assessment.

Considering uncertainty estimation, as our proposal focuses on a Monte-Carlo dropout approach, we highlight some works that applied this method to medical imaging. Leibig et al. (2017) introduced a deep learning model for diabetic retinopathy (DR) classification from fundus images. The authors used a custom sequential CNN and a VGG-inspired (Simonyan and Zisserman 2014) architecture. They used Monte-Carlo dropout for uncertainty estimation, adopting AUC, variance, and entropy as evaluation metrics. Leibig et al. (2017) accomplished not only competitive results for DR classification but also reliable uncertainty measures. This reliability is validated by the direct relation between high uncertainty scores and erroneous predictions.

Similarly, Laves et al. (2019) proposed comparing two different Monte-Carlo dropout configurations with a variational inference method for optical coherence tomographies (OCT) condition classification. These two variations consisted of one adding dropout before the last classification layer and another adding dropout after each convolutional block. They adopted ResNet-18 as baseline model and compared it with the other three variations using the ResNet-18 backbone. They concluded that adding dropout after each convolutional block can lead to more model noise and lower results. Their conclusion justifies our choice of using dropout layers only before the last convolutional layer. They used variance as uncertainty metric, and, as the work of Leibig et al. (2017), high variance scores correlated to incorrect classifications.

Our uncertainty estimation approach was mostly inspired by the work of Combalia et al. (2020), which combined Monte-Carlo dropout and test-time data augmentation for skin lesion classification. Our work differs from theirs on the following aspects:

We evaluated different architectures before the uncertainty estimation phase (they used only Efficient-Net-B0 (Tan and Le 2019)), we assessed Monte-Carlo dropout with and without test-time data augmentation, also evaluating different parameters.

Diving into the nephropathology field, Cicalese et al. (2020) introduced a deep-learning-based classification of kidney-level lupus nephritis with later uncertainty estimation. DenseNet (Huang et al. 2017) was adopted as CNN backbone and Monte-Carlo dropout was used for uncertainty estimation. Their pipeline was composed of a glomerular-level and a kidney-section-level classification, achieving competitive results for both types. Entropy was adopted as uncertainty score and, as occurred in previously cited works, high scores correlated to erroneous classifications. This correlation consistently occurring in several works justifies our proposal of using Monte-Carlo dropout for uncertainty estimation on the MN classification task.

2. Materials and methods

2.1. Data description and preparation

Our data set is composed of 4,682 images of human glomerulus labelled considering the following classes: Isolated membranous nephropathy (I-MN), mixed membranous nephropathy (M-MN), hypercellularity, glomerular sclerosis (cited here as sclerosis), and images with no lesion (cited here as normal). The following criteria were used: I-MN represents glomerulus that has MN characteristics only; conversely, the M-MN cases have MN properties and other lesions involved. The images were collected from the digital histological image library of the Gonçalo Moniz Institute (FIOCRUZ) and properly dissociated from their respective patient information to avoid identification. The tissue samples were fixed in Bouin's fixative or formalin–acetic acid–alcohol, included in paraffin. Haematoxylin and Eosin (H&E) were used to stain sections of 2–3 μm. All images were obtained using an Olympus QColor 3 digital camera connected to a Nikon E600 optical microscope (applying 200 × magnification). From each section, pathologically relevant regions were individually cropped and annotated by pathologists for diagnoses. The final data set was created considering only the cropped images that included at least one glomerulus.

Since that differentiation between I-MN and M-MN is usually difficult to be done using only visual features, we decided to cluster isolated and mixed MN images into a single group called 'membranous'. A typical training approach would be to use a membranous × no-lesion setup. Still, other lesions not related to MN may appear in real case situations. As the data set also included images with hypercellularity and sclerosis (both without MN), we grouped these classes into a class named 'other lesion'. Therefore, our final 3-class configuration and class distribution can be summarised as follows:

- **Membranous**: glomeruli with isolated (712 images) and mixed MN (1354 images), thus comprising 2066 images;
- **Other lesions**: glomeruli with hypercellularity (1237 images) or sclerosis (510 images), thus comprising 1747 images;
- **Normal**: glomeruli with no lesion (869 images).

Figure 2 illustrates examples of the approached lesions, and for comparison, a non-lesioned glomerulus is displayed in Figure 1. These pictures exhibit the challenge of differentiating the glomerular classes, even though some of these classes are grouped into broader labels.

2.2. Evaluation protocol

Figure 3 illustrates our evaluation protocol, which can be summarised into the following steps: Architecture comparison, uncertainty estimation, and qualitative evaluation. In the first step, we compared four DL architectures in a 10-fold cross-validation setup, namely, MobileNet-V2 (Sandler et al. 2018), ResNet-18 (He et al. 2016), DenseNet-121 (Huang et al. 2017), and Wide-ResNet-121 (Zagoruyko and Komodakis 2016). The main goal of this first phase is to select the architecture that achieved the highest average F1-Score. After selecting the best architecture, we added a dropout layer and started the second step, which consists of performing an uncertainty estimation using Monte-Carlo dropout. In the final step, we carried out a qualitative study by evaluating the relationship between uncertainty scores and the labelling challenge-level defined by pathologists. In the next sections, we cover each step of the proposed protocol.

2.3. On selecting the best architecture for MN classification

Since our uncertainty estimation evaluation covers experiments with different hyperparameters, we perform an initial architecture selection to analyse the parameter combinations using a single architecture. We assessed four CNN architectures: MobileNet-V2, ResNet-18, DenseNet-121, and Wide-ResNet. Given the diversity of architectures in the literature, we selected networks that bring different deep-learning novelties on architecture design. As the name suggests, ResNet-18 introduced the so-called residual block, innovating by adding skipping connections inside each block thus preventing the vanishing-gradient problem. Inception-V3 updated the inception block that, in the latter one, is composed of parallel convolutions of different kernel sizes. Expanding the concept of skipping connection, in a DenseNet architecture, each layer is connected to every forward layer, still reducing the vanishing-gradient problem and also strengthening feature propagation. Wide-ResNets are a variant of ResNets with decreased depth and increased width (introducing wide residual blocks). These wide residual blocks are capable of learning more features with no deepening of the network, achieving faster convergence during the training phase. Finally, by combining residual connections and separable convolutions, MobileNet-V2 introduces a novel architecture focused on running on devices constrained by low memory and low computation.

As we do not wish to 'waste' training data with a traditional train/validation/test split, we used a K-fold cross-validation approach for training and validation of the candidate architectures. Instead of generating fixed training and validation sets, we split the data set into K folds, interactively leaving onefold for validation and using the remaining $K-1$ folds for training. This way, we train and evaluate the model on K rounds of

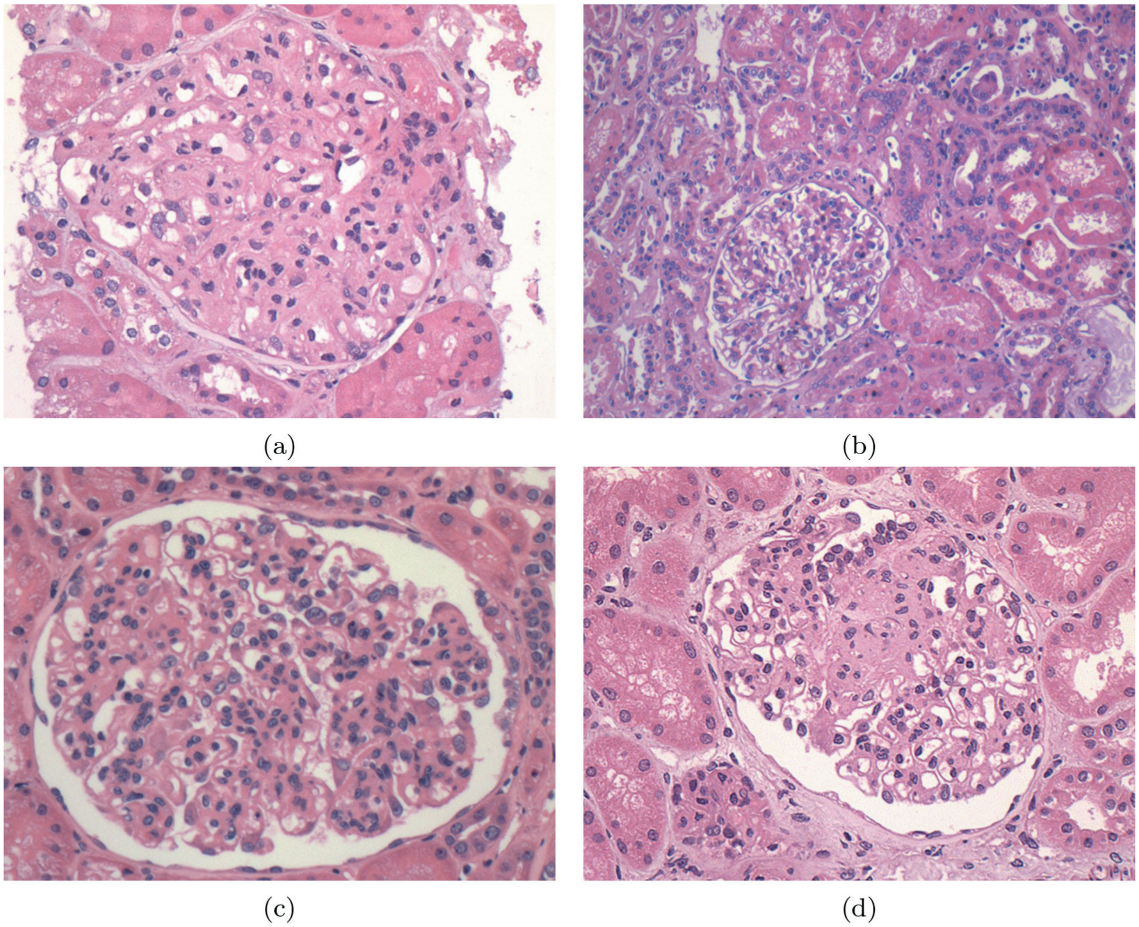

Figure 2. Example of a) a glomerulus with isolated MN, b) mixed MN, c) hypercellularity, and d) sclerosis.

different training and validation sets. Finally, as we do not have a final test set, we validated the architectures using average metrics (to be described later) across the K rounds. To avoid a large reduction on the training set, we adopted $K = 10$.

2.4. *Estimating uncertainty of deep-learning architectures*

Neural networks are known to almost always return high confidence predictions, usually softmax-based probabilities (Hein et al. 2019). Tackling this issue of yielding high confidence scores for inputs far away from the training distribution is essential for some tasks, as safety-critical systems should be aware of inputs they 'are not sure' about the prediction. Considering medical applications, avoiding overconfident predictions can aid specialists and students in the analysis of the input and model outcome. With a proper uncertainty score, the following questions arise: *'How certain is the model?', 'should its prediction be reevaluated?', 'was this class present in the training data set?'*.

Usually, one considers two types of uncertainty: Aleatory and epistemic (Kiureghian and Ditlevsen 2009). Aleatory uncertainty (also called data uncertainty) represents randomness inherent to the observed data, mostly related to the data set generation issues (such as labelling phase and domain characteristics). Conversely, the epistemic type (also called model uncertainty) captures uncertainty about the model and the generalisation over the data. Our uncertainty estimation pipeline is inspired by the work of Combalia et al. (2020), which combines Monte-Carlo dropout (Gal and Ghahramani 2016) and test-time data augmentation (Ayhan and Berens 2018) to estimate both aleatory and epistemic uncertainty.

Dropout is a regularisation method usually applied at training time to avoid overfitting (Srivastava et al. 2014). This method consists in randomly dropping out weights on a given layer of the network. In this context, Monte-Carlo-based experiments use repetition through sampling to model probabilities of events constrained by random variables (Kroese et al. 2013). The idea of MCD is thus to use that randomness on the network to estimate model uncertainty on prediction. Gal and Ghahramani (2016) applied this sampling idea to the CNN context by keeping dropout activated during test time while performing multiple forward passes of a given input through the model. With a probability p of randomly dropping out weights, each forward pass results in a different set of weights, leading to different predictions. Consequently, each input x_i yields M predictions $p = \{p_{i1}, p_{i2}, \ldots, p_{iM}\}$. One common approach is to use the mean of these predictions as the final prediction y_i and the variance as the uncertainty score u_i.

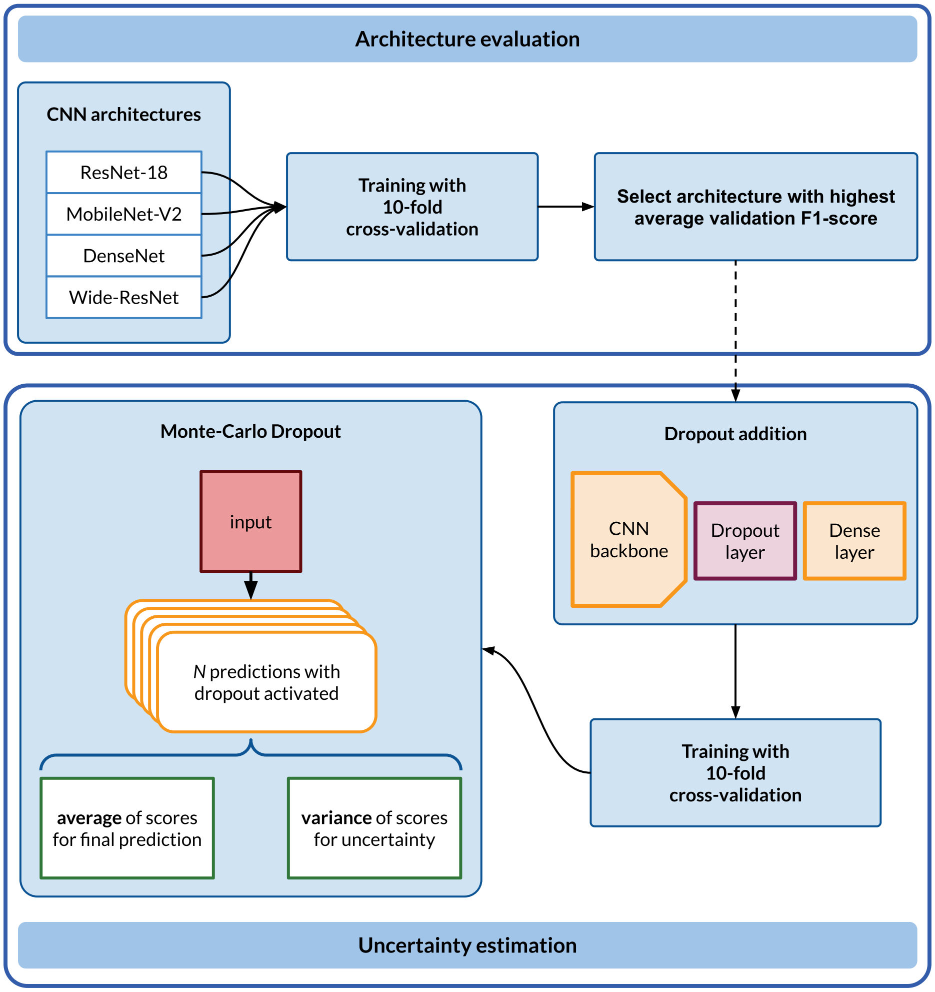
Architecture evaluation
CNN architectures
ResNet-18
MobileNet-V2
DenseNet
Wide-ResNet
Training with 10-fold cross-validation
Select architecture with highest average validation F1-score
Monte-Carlo Dropout
input
N predictions with dropout activated
average of scores for final prediction
variance of scores for uncertainty
Dropout addition
CNN backbone
Dropout layer
Dense layer
Training with 10-fold cross-validation
Uncertainty estimation

Figure 3. Proposed evaluation pipeline split into two steps: Evaluation of chosen architectures, and uncertainty estimation.

Following a similar approach, Ayhan and Berens (2018) proposed the test-time data augmentation to estimate aleatory uncertainty. Data augmentation consists in applying random transformations to the input, usually during training. The core idea is to increase the variability of the training data and to improve the generalisation. By applying these random transformations on inference for every input x_i, one can evaluate how much the network output varies with random and close samples from x_i.

Combalia et al. (2020) combined those two uncertainty estimation methods by adding random data augmentation to each forward pass on MCD. This way, they included randomness on both data and model. For MCD implementation, as Figure 3 illustrates, we selected the best architecture from the architecture evaluation phase and added a dropout layer right before the last classification layer. Then, we retrain the model considering different p (dropout probability) and M (number of forward passes) values. Since we are mostly interested in the uncertainty over the model prediction, we assessed MCD with and without TTA. Even though the experiments to assess the uncertainty methods' performance have been performed by considering the best network architecture (without loss of generality), both MCD and TTA were originally developed to work regardless of the neural network.

2.5. Metrics

A quite traditional metric for evaluating classification models is **accuracy**, which corresponds to the percentage of correct predictions among all instances. To assure the generalisation for each class, we adopted **precision, recall** and **F1-score**. Precision summarises how many positive predictions are indeed positive. Recall indicates how many positive samples were correctly classified. F1-score is the harmonic mean between precision and recall. Each metric is based upon the following definitions: True positive (TP) represents the correct predictions of the positive class, true negative (TN) represents the correct predictions of the negative class, false positive (FP) represents the incorrect predictions of the positive class, and false negative (FN) represents the incorrect predictions of the negative class. Each metric is then given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

For uncertainty estimation, we adopted two metrics: **predictive variance** and **uncertainty ratio**. Predictive variance (also cited here as MCD variance) is estimated by taking the variance of *M* predictions for each class and then computing the mean of these variances. As MCD variance is measured for each image, we used the uncertainty ratio to evaluate a set of images. The uncertainty ratio is the ratio between the average MCD variance of incorrect predictions and the average MCD variance of correct predictions. This way, we can analyse the relation between MCD variance and correctness in predictions.

3. Experiments and results

3.1. Implementation details

All architectures were modelled, trained, and validated using the Pytorch framework (Paszke et al. 2019). To achieve faster and better convergence for medical imaging (Raghu et al. 2019), all models were initially loaded with weights pretrained on the ImageNet data set (Russakovsky et al. 2015) with an adapted dense layer with three nodes, respectively, to the target classes. For training parameters, we used AdamW optimiser (Loshchilov and Hutter 2017) and set 100 epochs with a batch size of 32, defining an initial learning rate of 0.0001 with step decay of factor 0.1 at every 30 epochs. To apply data augmentation during training and testing (for TTA), we included the following transformations: Random rotations within an range of 90 degrees and probability of 0.5, random horizontal and vertical flips, and random crops of 224×224 pixels after resizing the input height to 224, thus maintaining aspect ratio and matching the input size of 224×224 for all networks. All experiments were run on a machine with 8GB RAM and an NVIDIA GEFORCE GTX 1060.

3.2. Architecture evaluation

The first phase of experiments consists in evaluating the candidate architectures in a cross-validation setup. Table 1 presents the results of each architecture for the classification metrics defined earlier. The four architectures achieved competitive results, presenting average metrics above 92% with low standard deviations. The models returned similar values for all metrics, indicating that the networks also achieved a good generalisation per class. As expected, Wide-ResNet returned the highest average F1-Score (in bold) with the closest values across all metrics, showing robustness among different training and validation sets. Also, Wide-ResNet achieved the lowest standard deviation for all metrics, showing more stable results across all folds.

Another noteworthy point is the performance of MobileNet, which, despite being the most lightweight model, still outperformed ResNet-18. Table 2 brings the number of trainable parameters of each architecture, where we can observe that MobileNet is almost five times less costly than ResNet-18. In addition, MobileNet achieved results relatively close to the other architectures, still being approximately three times less costly than DenseNet, and thirty-three times less costly than Wide-ResNet. In this work, our uncertainty estimation experiments are performed regarding only the best architecture, namely Wide-ResNet. However, considering a performance-efficiency trade-off, MobileNet would be the right choice.

Despite a high F1-Score *per si* indicates great learning per class, we also computed the confusion matrix for Wide-ResNet to better visualise intraclass predictions. For an aggregated view, we analysed the confusion matrix sum over the 10 validation folds. The confusion matrix is summarised in Table 3, where the rows represent ground-truth labels, and the columns represent the predicted classes. Thus, the main diagonal (highlighted in purple) indicates the correct predictions. Besides, we calculated the F1-score for each class, where competitive scores for all classes were achieved.

3.3. Uncertainty estimation

The second phase of our evaluation pipeline can be defined as estimating uncertainty scores considering two methods: MCD and TTA. As pointed in Section 2.4, we evaluated MCD with and without TTA. To assess how MCD parameters impact model predictions, the experiments were performed considering $p = \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and $M = \{10, 50, 100\}$. With Wide-ResNet selected as the best architecture, we added a dropout layer between the Wide-ResNet backbone and the last classification layer. For every combination of *p*, *M*, and the presence/absence of TTA, we retrained the architecture starting from the pre-trained ImageNet weights. The results of each parameter combination can be found in Table 4. As occurred in Section 3.2, the average metrics also showed similar values for each combination. However, a greater variation can be observed if we consider all parameter combinations. The F1-scores ranged from 91.7% to 93.5%, where the top value is

Table 1. Comparative results of MobileNet, ResNet-18, DenseNet and Wide-ResNet deep-learning-based architectures.

Model	μAccuracy	μF1-score	μPrecision	μRecall
MobileNet	0.926(±0.008)	0.924(±0.008)	0.924(±0.008)	0.925(±0.008)
ResNet-18	0.924(±0.010)	0.922(±0.008)	0.922(±0.010)	0.922(±0.009)
DenseNet	0.932(±0.011)	0.930(±0.011)	0.928(±0.012)	0.933(±0.013)
Wide-ResNet	0.937(±0.007)	0.936(±0.007)	0.936(±0.008)	0.936(±0.008)

Table 2. Number of trainable parameters of the candidate architectures.

Model	# trainable parameters
MobileNet	2.227.715
ResNet-18	11.178.051
DenseNet	6.956.931
Wide-ResNet	66.840.387

Table 3. Confusion matrix sum for Wide-ResNet predictions over the 10-fold cross-validation.

	N	M	O	F1-score
Normal (N)	808	49	12	0.93
Membranous (M)	49	1948	69	0.94
Other lesion (O)	19	95	1633	0.94

below but still close to the former no-uncertainty Wide-ResNet (93.6%). Taking the cases without TTA, the *M* values had little to no impact on the average metrics. Also, the absence of TTA returned lower uncertainty ratios for almost all cases compared to the cases with TTA, maybe due to the reduction of randomness on prediction. Nevertheless, without the noise introduced by TTA, higher scores were achieved for average metrics. Regarding the probability values, the highest average metrics were achieved using *p*=0.4, which might be a good trade-off between an excessive amount of weights dropped (which could harm the generalisation) and too little weights dropped (that could introduce less regularisation). Interestingly, the highest uncertainty ratio was 9.252, which was achieved with *p*=0.3. We cannot conclude that lower dropout probabilities always led to greater uncertainty ratios because *p*=0.7 also returned a competitive score of 9.150.

For further investigation, we considered only the model trained with TTA applied, *p*=0.3, and *M*=100. the highest uncertainty ratio (9.252), it is possible to say that the average uncertainty of incorrect classification was approximately nine times higher than the average uncertainty of correct predictions, showing a relation between MCD variance and erroneous predictions. For a better understanding of how the MCD variance is distributed, Figure 4 illustrates area charts for each class combination. We can note that MCD variance of correct predictions are clustered in a region close to the minimum variance, while the incorrect predictions have a wider distribution over areas with higher uncertainty scores. Since higher MCD variances are more likely to represent incorrect classifications, we can set a threshold value to remove or re-evaluate a given sample. To

Table 4. Results of Wide-ResNet with MCD considering different parameters over 10 validation folds. The *p* parameter refers to the dropout probability, TTA refers to the presence (✓) or absence (✗) of test-time data augmentation, and *M* indicates the number of forward passes of the input over the network.

p	TTA	M	μ Accuracy	μ F1-score	μ Precision	μ Recall	Unc. ratio
0.3	✗	10	0.933(±0.008)	0.931(±0.007)	0.930(±0.008)	0.933(±0.008)	7.060
	✗	50	0.934(±0.009)	0.932(±0.008)	0.931(±0.009)	0.934(±0.008)	6.583
	✗	100	0.933(±0.009)	0.932(±0.008)	0.931(±0.009)	0.933(±0.009)	6.811
	✓	10	0.921(±0.012)	0.921(±0.011)	0.920(±0.011)	0.922(±0.012)	8.481
	✓	50	0.925(±0.008)	0.923(±0.008)	0.922(±0.008)	0.925(±0.010)	9.008
	✓	100	0.921(±0.015)	0.920(±0.014)	0.919(±0.014)	0.921(±0.014)	9.252
0.4	✗	10	0.937(±0.008)	0.935(±0.009)	0.935(±0.009)	0.936(±0.011)	6.736
	✗	50	0.936(±0.009)	0.935(±0.009)	0.934(±0.010)	0.936(±0.011)	6.995
	✗	100	0.937(±0.009)	0.935(±0.009)	0.935(±0.009)	0.936(±0.011)	6.584
	✓	10	0.927(±0.011)	0.925(±0.012)	0.923(±0.011)	0.927(±0.014)	8.319
	✓	50	0.925(±0.006)	0.923(±0.005)	0.922(±0.007)	0.926(±0.007)	7.935
	✓	100	0.923(±0.009)	0.922(±0.009)	0.920(±0.010)	0.924(±0.011)	8.482
0.5	✗	10	0.933(±0.011)	0.931(±0.010)	0.931(±0.010)	0.932(±0.013)	7.725
	✗	50	0.933(±0.011)	0.931(±0.010)	0.931(±0.010)	0.931(±0.013)	6.801
	✗	100	0.933(±0.012)	0.931(±0.011)	0.932(±0.010)	0.932(±0.023)	7.006
	✓	10	0.924(±0.008)	0.924(±0.008)	0.923(±0.008)	0.924(±0.009)	8.306
	✓	50	0.921(±0.007)	0.920(±0.006)	0.919(±0.008)	0.921(±0.009)	9.117
	✓	100	0.927(±0.009)	0.926(±0.009)	0.925(±0.009)	0.927(±0.011)	8.578
0.6	✗	10	0.931(±0.011)	0.928(±0.011)	0.927(±0.011)	0.930(±0.012)	7.842
	✗	50	0.932(±0.011)	0.929(±0.011)	0.928(±0.011)	0.931(±0.012)	7.746
	✗	100	0.932(±0.011)	0.929(±0.011)	0.928(±0.011)	0.930(±0.013)	7.796
	✓	10	0.928(±0.011)	0.927(±0.011)	0.926(±0.009)	0.927(±0.014)	7.690
	✓	50	0.924(±0.013)	0.922(±0.014)	0.920(±0.016)	0.925(±0.014)	8.947
	✓	100	0.926(±0.011)	0.924(±0.011)	0.922(±0.010)	0.927(±0.013)	7.663
0.7	✗	10	0.931(±0.011)	0.928(±0.011)	0.929(±0.009)	0.928(±0.015)	6.720
	✗	50	0.930(±0.011)	0.927(±0.011)	0.928(±0.009)	0.927(±0.015)	7.208
	✗	100	0.931(±0.011)	0.928(±0.010)	0.929(±0.008)	0.928(±0.014)	6.710
	✓	10	0.923(±0.013)	0.923(±0.012)	0.922(±0.012)	0.923(±0.014)	7.811
	✓	50	0.920(±0.011)	0.918(±0.010)	0.918(±0.010)	0.919(±0.013)	8.652
	✓	100	0.920(±0.009)	0.917(±0.010)	0.916(±0.010)	0.919(±0.011)	9.150

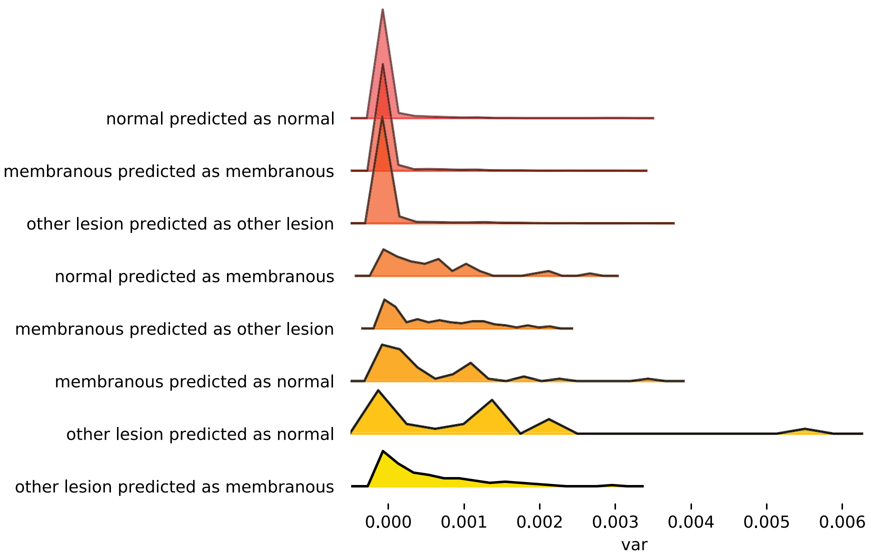

Figure 4. Distribution of variance score based on real/predicted class combinations.

assess whether a threshold can improve the model performance, we evaluated several thresholds values ranging from the minimum to the max variance. For each threshold value *t*: 1) we drop samples that received an uncertainty score greater than *t*, 2) we recompute the accuracy with the new set of samples. Figure 5 illustrates a monotonic increase in accuracy as MCD variance thresholds decrease, where we can see the accuracy improve up to 96%.

Even though the models were trained considering a unique membranous class, we had the annotations for isolated MN and MN combined with other lesions. We expected mixed-MN samples would be misclassified as 'other lesion' more frequently than the isolated MN cases. Since mixed-MN images can have any other lesion besides MN, there might be images visually similar to hypercellularity or sclerosis (both present in the 'other lesion' class). What was expected indeed occurred: Among all isolated MN cases, 3.65% were misclassified as 'other lesion'; while among all mixed-MN cases, 6.72% were misclassified as 'other lesion'. It is necessary thus to investigate if higher uncertainty scores represent higher complexity for pathologists. To tackle this issue, we selected the ten highest uncertainty images and the ten lowest uncertainty images. To five pathologists were asked to accomplish two tasks:

- Classify each image into one of the following classes: isolated MN, mixed-MN, other lesion, no lesion;
- ranging from 0 (no complexity) to 5 (highest complexity), which complexity would he/she give to the classification task?

Table 5 summarises the predictions and average complexity scores for each image. The top part (first ten rows) represents the lowest uncertainty images. If we group I-MN and M-MN cases, there is an agreement of 8/10 between the ground-truth and the majority voting considering all five pathologists. The

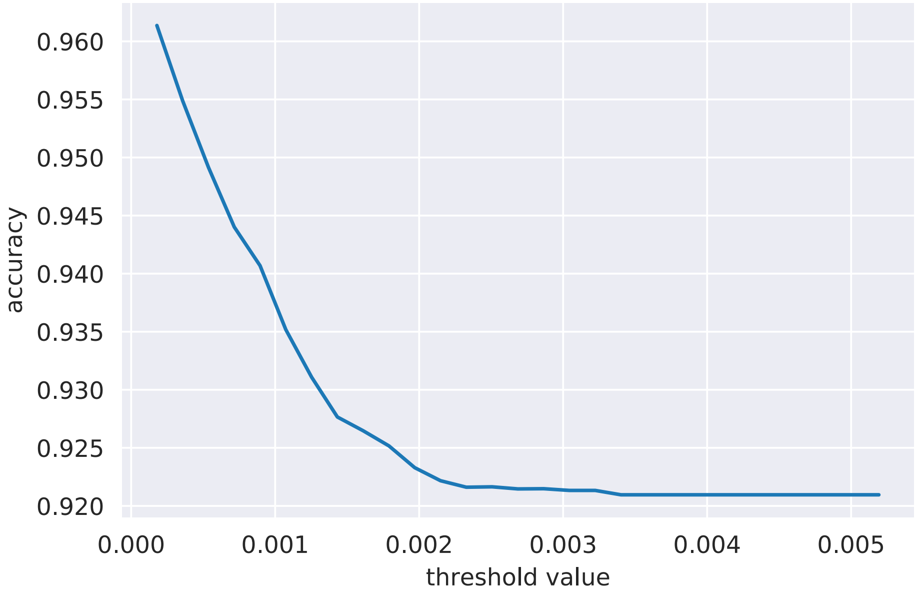

Figure 5. Improvement in accuracy using a decision approach based on MCD variance threshold.

Table 5. Pathologists' predictions and average complexity scores (μC) for the ten lowest uncertainty images (images 1–10) and ten highest uncertainty images (images 11–20). We represent each pathologist as Px, with *X* ranging from 1 up to 5. For a better understanding, we also included the ground-truth labels (GT). The following abbreviations are used: Membranous nephropathy as MN; isolated membranous nephropathy as I-MN; mixed membranous nephropathy as M-MN; other lesion as O; normal as N.

Img	P_1	P_2	P_3	P_4	P_5	GT	μ C
1	M-MN	I-MN	M-MN	M-MN	M-MN	MN	2.4
2	I-MN	M-MN	M-MN	O	O	MN	3
3	I-MN	O	I-MN	O	N	MN	2.6
4	O	O	M-MN	O	O	O	2.8
5	O	O	O	O	O	O	2.8
6	O	M-MN	O	O	O	O	2.8
7	O	O	M-MN	M-MN	O	O	2.8
8	O	O	M-MN	O	O	O	3.2
9	I-MN	I-MN	I-MN	I-MN	I-MN	MN	3.2
10	O	O	O	O	O	O	3.4
11	N	N	N	I-MN	N	MN	2.4
12	I-MN	O	I-MN	M-MN	N	MN	2.2
13	O	O	M-MN	O	M-MN	O	3.2
14	O	O	M-MN	O	O	MN	3.6
15	N	O	N	O	N	N	2.4
16	N	O	O	O	O	N	3
17	M-MN	O	O	O	O	O	2.2
18	I-MN	I-MN	I-MN	I-MN	M-MN	MN	3
19	O	N	N	N	N	N	2.8
20	M-MN	O	O	O	N	O	3.4

bottom part (last ten rows) represents the highest uncertainty images. By grouping I-MN and M-MN cases, there is an agreement of 7/10 between the ground-truth and the majority voting. We can highlight that there are only three images where the pathologists reached a consensus (images 5, 9 and 10): Two cases labelled as 'other lesion' and one sample labelled as I-MN. It is noteworthy that these three images belong to the lowest uncertainty group, and their respective majority voting classes match the GT labels. Overall, Table 5 indicates how unlikely it is to reach a consensus, where we have to consider different backgrounds, levels of experience, and expertise among the pathologists. Finally, we computed the mean of average complexity for both top- and bottom-10 groups. Although the lowest uncertainty group achieved the lowest mean of average complexity (2.82 compared to 2.9), further experiments need to be done to prove that MCD variance indicates complexity for specialists.

4. Concluding remarks

Recognising MN is an important and challenging task, which could be supported by automatic classification methods and proper uncertainty scores. We developed an evaluation pipeline consisted of two tasks: Architecture evaluation and uncertainty estimation. For the first phase, we compared ResNet-18, MobileNet, DenseNet, and Wide-ResNet in a 10-fold cross-validation setup. We obtained average F1-Scores above 92% for all models, where the highest average F1-score was achieved by Wide-ResNet (93.6%). Also, it is noteworthy that MobileNet achieved competitive scores even with much fewer parameters, becoming a feasible choice for mobile applications or other performance-constrained hardware configurations. The second phase consists in estimating and evaluating uncertainty for the best architecture (Wide-ResNet) . We applied Monte-Carlo dropout and assessed several experiments combining different values for parameters p (dropout probability), M (number of forward passes), and TTA (presence or absence of test-time data augmentation). The modified Wide-ResNet achieved slightly lower results, with F1-Scores ranging from 91.7% to 93.5%. Also, we could achieve an uncertainty ratio of 9.252, which indicates a relation between MCD variance and correctness on predictions. Regarding the evaluated parameters, p=0.4 returned the best average metrics and a competitive uncertainty score (8.482). We could conclude that M has more impact when used with TTA, and the use of TTA must consider that the data noise might increase the uncertainty ratio as the average metrics slightly decrease. Another experiment aimed to use thresholds based on MCD variance to drop samples and evaluate the impact of these dropped samples on accuracy. We showed that our uncertainty scores correlated with correctness so that decreasing threshold values led to increasing accuracy scores. In a real-world scenario, the images surpassing a specified threshold, beyond being dropped, could be re-evaluated, re-labelled, or highlighted for further inspection. Lastly, our investigation on the correlation between MCD variance and pathologists-defined complexity scores did not show a great discrepancy among the highest and lowest uncertainty groups. This way, we cannot conclude that there is a correlation, so further experiments with more images and uncertainty scores need to be made. Although we could not prove the correlation, the pathologist's predictions showed how challenging it is to diagnose a glomerular lesion and reach a consensus-based data set. For future work, we plan to evaluate other uncertainty estimation methods such as Variational Inference (VI) (Posch et al. 2019) or deep ensembles (Lakshminarayanan et al. 2016). In addition, finding an optimal threshold value is an important task, especially considering real-world scenarios. Novel lesions need to be assessed as well, and we plan to develop a robust approach considering more glomerular lesions in an open-set classification setup.

Note

1. https://pathospotter.bahia.fiocruz.br/pathospotterhome/

Notes

This work was conducted following the resolution No. 466/12 of the Brazilian National Health Council. To preserve confidentiality, the images (including those shown in this paper) were separated from other patient's data. No data presented herein allows patient identification. All the procedures were approved by the Ethics Committee for Research Involving Human Subjects of the Gonçalo Moniz Institute from the Oswaldo Cruz Foundation (CPqGM/FIOCRUZ), Protocols No. 188/09 and No. 1,817,574.

Acknowledgments

The authors would like to thank Dr. Henrique M. S. Proença (Hospital do Rim e Hipertensão, São Paulo, São Paulo, Brazil), Dr. Karla Lais Pêgas (Hospital Moinhos de Vento e Hospital Conceição, Porto Alegre, Rio Grande do Sul, Brazil), Dr. Nathanael Pinheiro Jr. (IMAGEPAT, Salvador, Bahia, Brazil), and Dr. Nayze L. S. Aldeman (Universidade Federal do Piauí, Teresina, Piauí, Brazil) for their technical assistance in the complexity scores collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

All authors belong to the Pathospotter project,[1] which is partially sponsored by Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB), grant No. TO P0008/15 and TO-SUS0031/2018 and Inova Fiocruz - Innovative ideas. Paulo Chagas and Luiz Souza have scholarships from FAPESB, grants TO-BOL0344/2018 and TO-BOL0660/2018, respectively. Washington Santos and Luciano Oliveira have research scholarships from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants 306779/2017 and 307550/2018-4, respectively.

Notes on contributors

Paulo Chagas is a doctoral student at the Graduate Program in Computer Science at the Federal University of Bahia, Brazil. He received his Bsc. and MSc. in Computer Science from Federal University of Pará, Brazil, in 2016 and 2018, respectively. He is interested in Data Mining, Machine Learning and Computer Vision, with focus on image clustering for medical imaging applications.

Luiz Souza is a doctoral student in Mechatronics from the Federal University of Bahia, Brazil. He received his MSc in Computer Science from the Federal University of Bahia, Brazil, in 2016. He has participated as a researcher in intelligent vision research lab in the field of Computer Vision and Image Pattern Recognition, mainly focused in image segmentation and renal pathology.

Izabelle Rocha Pontes is a undergraduate medical student at the Federal University of Bahia. She holds a undergraduate research scholarship from the Brazilian Research Council (CNPq) since 2018, to work on the PathoSpotter Project. She was one of the nominees for the Jessé Accioly Award given to student with the most outstanding research work.

Rodrigo Tripodi Calumby holds BSc., MSc., and PhD. in Computer Science. He is a tenured professor at the University of Feira de Santana (Brazil) and a visiting researcher at the Federal University of Bahia (Brazil). His main research interests include information retrieval, machine learning, and applications.

Michele Fúlvia Angelo is a professor of the Department of Exact Sciences at State University of Feira de Santana (UEFS), Brazil. She received the M.Sc. and the Ph.D. in Electrical Engineering from the University of São Paulo (USP), Brazil, in 2001 and 2007, respectively. She has experience in Computer Vision and Pattern Recognition in Images, working mainly in the following areas: medical and dental imaging, object detection and classification.

Angelo Duarte received the B.E. degree in electrical engineering in 1987, and the Ms.C. degree in 2000, both from the Federal University of Bahia (UFBA), Salvador, Brazil, and the Ph.D. degree in computer science from the Autonomous University of Barcelona (UAB) Barcelona, Spain, in 2007. Since 2009 he joined the Department of Technology, State University of Feira de Santana (UEFS), as an Assistant Professor, and in 2019 became a Professor. He is also the head of High-Performance Computing Lab at UEFS. His current research interests include computer vision, computational pathology and high-performance computing.

Washington LC dos-Santos is a M.D., Ph.D., Pathologist working for more than 20 years in diagnostic nephropathology. He is head of the Molecular and Structural Pathology of the Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Brazil.

Luciano Oliveira is a professor of the Department of Computer Science, Institute of Computing, and the head of the Intelligent Vision Research Lab at Federal University of Bahia, Brazil. He received his BSc in Computer Science and MSc in Mechatronics degrees from Federal University of Bahia, Brazil, in 1997 and 2005, respectively. In 2010, he received his PhD in Electrical and Computer Engineering at Coimbra University, Portugal. He has been working in several research projects in the fields of Computer Vision and Image Pattern Recognition with focus on Smart Cities, Biometric Systems, Biomedicine, and Robotics.

References

Ayhan MS, and Berens P. 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: International Conference on Medical Imaging with Deep Learning; Amsterdam.

Begoli E, Bhattacharya T, Kusnezov D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence. 1(1):20–23. doi:10.1038/s42256-018-0004-1.

Chen Y, Li M, Hao F, Han W, Niu D, and Wang C. 2020. Classification of glomerular spikes using convolutional neural network. In: Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare; Taiyuan China. p. 254–258.

Cicalese PA, Mobiny A, Shahmoradi Z, Yi X, Mohan C, and Van Nguyen H. 2020. Kidney level lupus nephritis classification using uncertainty guided Bayesian convolutional neural networks; IEEE Journal of Biomedical and Health Informatics. 25. p. 315–24.

Combalia M, Hueto F, Puig S, Malvehy J, Vilaplana V. 2020. Uncertainty estimation in deep neural networks for dermoscopic image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. p. 744–745.

Gal Y, Ghahramani Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machine Learning; (Proceedings of Machine Learning Research; vol. 48); 2022 2022 Jun; New York, New York, USA. PMLR. p. 1050–1059.

He K, Zhang X, Ren S, and Sun J. 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; Las Vegas, NV, USA. p. 770–778.

Hein M, Andriushchenko M, and Bitterwolf J. 2019. 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 (CVPR) Workshops; June; Long Beach, CA, USA.

Huang G, Liu Z, Van Der Maaten L, and Weinberger KQ. 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; Honolulu, HI, USA. p. 4700–4708.

Kiureghian AD, Ditlevsen O. 2009. Aleatory or epistemic? Does it matter? Structural Safety. 31(2):105–112. Risk Acceptance and Risk Communication; Available from. https://www.sciencedirect.com/science/article/pii/S0167473008000556

Krizhevsky A. 2014. One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:14045997. Available from: https://arxiv.org/abs/1404.5997

Kroese DP, Taimre T, Botev ZI. 2013. Handbook of monte carlo methods. Vol. 706. John Wiley & Sons.

Lakshminarayanan B, Pritzel A, Blundell C. 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:161201474. Available from: https://arxiv.org/abs/1612.01474

Laves MH, Ihler S, Ortmaier T. 2019. Uncertainty quantification in computer-aided diagnosis: make your model say “i don’t know” for ambiguous cases. In: International Conference on Medical Imaging with Deep Learning – Extended Abstract Track; 08–10 Jul; London, United Kingdom.

Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep. 7(1):1–14. doi:10.1038/s41598-017-17876-z.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI. 2017. A survey on deep learning in medical image analysis. Med Image Anal Available from. 42:60–88. https://www.sciencedirect.com/science/article/pii/S1361841517301135

Loshchilov I, Hutter F. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. Available from: https://arxiv.org/abs/1711.05101

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, and Antiga L, et al. 2019. Pytorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, D Alché-buc F, Fox E, and Garnett R, editors. Advances in neural information processing systems 32; Vancouver, Canada: Curran Associates, Inc. p. 8024–8035.

Posch K, Steinbrener J, Pilz J. 2019. Variational inference to measure model uncertainty in deep neural networks. arXiv preprint arXiv:190210189. Available from: https://arxiv.org/abs/1902.10189

Raghu M, Zhang C, Kleinberg J, Bengio S. 2019. Transfusion: understanding transfer learning for medical imaging. arXiv preprint arXiv, 190207208. Available from: https://arxiv.org/abs/1902.07208

Ronneberger O, Fischer P, and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention; Munich, Germany: Springer. p. 234–241.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. 2015. Imagenet large scale visual recognition challenge. Int J Comput Vis. 115(3):211–252. doi:10.1007/s11263-015-0816-y

Sandler M, Howard A, Zhu M, Zhmoginov A, and Chen LC. 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; Salt Lake City, UT, USA. p. 4510–4520.

Simonyan K, Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. Available from: https://arxiv.org/abs/1409.1556

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 15(56):1929–1958. Available from http://jmlr.org/papers/v15/srivastava14a.html

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z. 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; Las Vegas, NV, USA. p. 2818–2826.

Tan M, and Le Q. 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning; Long Beach, CA, USA. PMLR. p. 6105–6114.

Uchino E, Suzuki K, Sato N, Kojima R, Tamada Y, Hiragi S, Yokoi H, Yugami N, Minamiguchi S, Haga H, et al. 2020. Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach. Int J Med Inform. 141:104231. doi:10.1016/j.ijmedinf.2020.104231.

Zagoruyko S, Komodakis N. 2016. Wide residual networks. arXiv preprint arXiv:160507146. Available from: https://arxiv.org/abs/1605.07146