Deep-learning-based membranous nephropathy classification and Monte-Carlo dropout uncertainty estimation

Paulo Chagas[1], Luiz Souza[1], Izabelle Pontes[2], Rodrigo Calumby[1,3], Michele Angelo[3], Angelo Duarte[3], Washington L.C. dos-Santos[4], Luciano Oliveira[1]

[1]IVISION Lab, Universidade Federal da Bahia, Bahia, Brazil

[2]Universidade Federal da Bahia, Bahia, Brazil

[3]Universidade Estadual de Feira de Santana, Bahia, Brazil

[4]Fundação Oswaldo Cruz – Instituto Gonçalo Moniz, Bahia, Brazil

{paulo.chagas,luiz.otavio,lrebouca}@ufba.br

{rtcalumby,angeloduarte}@uefs.br

mfangelo@ecomp.uefs.br

izabelleyz@gmail.com

wluis@bahia.fiocruz.br

Abstract. *Membranous Nephropathy (MN) is one of the most common glomerular diseases that cause adult nephrotic syndrome. To assist pathologists on MN classification, we evaluated three deep-learning-based architectures, namely, ResNet-18, DenseNet and Wide-ResNet. In addition, to accomplish more reliable results, we applied Monte-Carlo Dropout for uncertainty estimation. We achieved average F1-Scores above 92% for all models, with Wide-ResNet obtaining the highest average F1-Score (93.2%). For uncertainty estimation on Wide-ResNet, the uncertainty scores showed high relation with incorrect classifications, proving that these uncertainty estimates can support pathologists on the analysis of model predictions.*

1. Introduction

A modern alternative to human inspection of pathological slides is the exploitation of Artificial Intelligence technologies for computer-aided diagnosis. In this context, computer vision methods have already been proposed for the automatic identification of glomerular findings from pathological slide images [Becker et al., 2020]. Nevertheless, given the importance of these pathological assessments and the critical impacts in real applications, the effectiveness of current models is still unsatisfactory and does not meet real-world demands. Given that large scale data acquisition for supervised learning is still an open issue in the field, the difficulty of building a large enough training corpus has imposed significant limitations to previous work.

Considering the nephropathology field, Membranous Nephropathy (MN) is a common glomerular disease, which usually is a cause of the nephrotic syndrome in adults. MN is an autoimmune glomerular disease characterized with the presence of a large amount of immune complex sediments on the epithelial cells. This characteristic

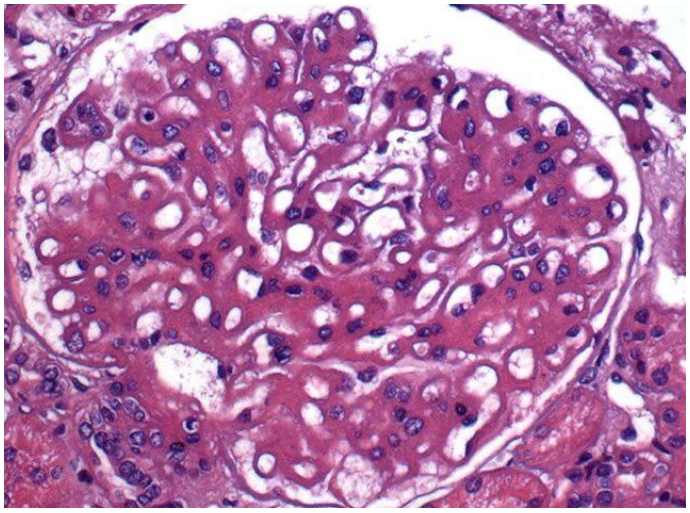

Figure 1. An example of a glomerulus with membranous nephropathy.

causes a thickening in the glomerular basement membrane, which is the main visual feature for detecting this disease. Figure 1 illustrates an example of glomerulus with MN where we can note the thickened membranes. Detecting the visual features that mark a MN disease is not an easy task, which requires experienced pathologists and even might lead to non-consensus situations. In this context, automatic classification methods can assist pathologists by becoming an useful tool in the decision making pipeline. By introducing and optimizing deep learning models, computer vision applications have significantly improved through time and its full potential for pathological analysis is still being investigated [Litjens et al., 2017].

Particularly for MN classification, the already reported experimental results in the literature [Uchino et al., 2020, Chen et al., 2020] have generally relied on extremely limited and highly unbalanced data sets, which impinges the construction of effective generalized predictive models. Such constrained experimental settings avoid a rigorous validation of the models and reduce the extent and reliability of the findings. In terms of the underlying learning infrastructure, just a few deep networks have been assessed, notably the U-Net (for glomeruli segmentation [Chen et al., 2020]) and just a few mainstream convolutional and residual networks such as InceptionV3 [Uchino et al., 2020] and ResNet [Chen et al., 2020].

Beyond this, the few initiatives on MN identification, besides achieving low effectiveness and using limited data, have only focused on image-level label-only classification, providing no additional supportive information for the pathologist decision making process. Begoli et al. [2019] highlight the importance of estimating a reliable uncertainty score for medical imaging assessment, stating that this estimation can benefit both research and practical applications in the medical domain. An ideal uncertainty metric should be related with erroneous predictions, leading to an interpretation that a high uncertainty prediction indicates a ”confused” model. This information could be useful when specialists analyze the model predictions. Therefore, in this work we evaluate deep learning architectures for MN classification, as well as we perform an uncertainty estimation for supportive information. Among several uncertainty estimation methods, we opted for Monte-Carlo Dropout [Gal and Ghahramani, 2016] approach, due to its implementation

that needs no large changes in the baseline models.

The main contributions of this work are listed as follows:

- The experimental analysis is performed over an unprecedentedly large Membranous Nephropathy image collection, allowing a reasonable assessment of the target nephropathy task;
- the effectiveness assessment is performed for a diverse set of deep network architectures; and
- the exploration and validation of an uncertainty estimation technique in the context of nephropathy identification and its impact to overall recognition effectiveness.

2. Related Work

Considering the absence of comprehensive models for the identification of many of the major glomerular pathological findings, Uchino et al. [2020] conducted a broad assessment of a deep learning network for pathology identification. The authors trained independent binary models for each pathology using a fine-tuning approach over the InceptionV3 [Szegedy et al., 2016] network. Specifically for MN, the dataset included 167 MN cases out of the total 3481 cases (including seven types of findings). Although the experimental results demonstrated high performance for some findings (*e.g.*, $AUC = 0.983 \pm 0.001$ / $AUC = 0.986 \pm 0.001$ with Periodic acid–Schiff (PAS) and Periodic acid-methenamine silver (PAM) stainings for global sclerosis), for MN it achieved a lower performance with $AUC = 0.816 \pm 0.034$ / $AUC = 0.734 \pm 0.011$ (PAS/PAM). Notice that MN was also weakly represented in the experimental data set, corresponding to 1.5% / 2.1% (PAS/PAM) and 2.3% / 4.6% (PAS/PAM) of the training and testing data, respectively.

Chen et al. [2020] proposes the SPIKE-NET, a simple two-phase process for MN identification combining convolutional and residual deep network architectures. The first phase regards the glomeruli segmentation and the second phase comprises the lesion classification. For the glomeruli segmentation, SPIKE-NET relies on the well known U-Net CNN-based network [Ronneberger et al., 2015]. For the classification phase, the authors proposed using the ResNet residual network [He et al., 2016], which achieved superior classification accuracy when compared to AlexNet [Krizhevsky, 2014] and VGG16 [Simonyan and Zisserman, 2014] networks. Specifically, SPIKE-NET achieved 94.44% of MN identification accuracy against 92.86% and 91.27% of UNet-VGG16 and UNet-AlexNet, respectively. The authors highlighted a 98.26% effectiveness in terms of recall, which means a low rate of missed diagnosis. Although such effectiveness may be considered significantly high, the whole experiments were conducted in a limited dataset with 1,267 glomeruli (653 with MN and 614 normal). In fact, a single assessment round was performed with only 126 glomeruli images used for the final evaluation of the optimized models, which hardens further conclusions.

Medical imaging applications require associated uncertainty scores. These needs have led several works to apply uncertainty estimation approaches on different medical domains. Since our work focus on Monte-Carlo Dropout estimation (detailed in Section 3), we cite some works that apply this approach for medical imaging. Leibig et al. [2017] proposed a deep-learning-based approach to detect diabetic retinopathy (DR) from fundus images. They used a custom sequential convolutional neural network and a VGG-inspired [Simonyan and Zisserman, 2014] network. For uncertainty estimation, Leibig

et al. [2017] used Monte-Carlo Dropout, adopting AUC, variance, and entropy as evaluation metrics. They achieved competitive results for DR classification, as well as reliable uncertainty estimates via Monte-Carlo Dropout. This reliability is confirmed by the results revealing the relation between high uncertainties and incorrect classifications.

Laves et al. [2019] have a similar proposal by comparing Monte-Carlo Dropout with a variational inference approach considering an optical coherence tomographies (OCT) condition classification. An interesting point is they compare two variations of Monte-Carlo Dropout. One by adding dropout right before the last full-connected layer, and another one by adding dropout after each residual block. They used ResNet-18 as baseline model, comparing with the other three variations. Considering classification scores, they achieved competitive results with the lowest F1-Score associated with the model that used second type of dropout. Their conclusion was that adding dropout after each residual block led to higher noise and consequently lower results. This outcome motivated us to evaluate the Monte-Carlo Dropout considering dropout layers added before the last convolutional layer only. The variance of the Monte-Carlo Dropout was chosen for the uncertainty analysis. Just as the work of Leibig et al. [2017], high uncertainties were associated with incorrect predictions.

Combalia et al. [2020] combines Monte-Carlo Dropout with Test Time Data augmentation for skin lesion classification. This combination will be detailed in Section 3, because we adopt their uncertainty estimation method in our experiments. Since they use Efficient-Net-B0 [Tan and Le, 2019] only as convolutional backbone, our approach differs on evaluating different architectures before estimating uncertainties for our task.

In Nephropathology field, Cicalese et al. [2020] proposed a kidney level lupus nephritis classification with uncertainty estimation. They adopted DenseNet [Huang et al., 2017] as backbone and Monte-Carlo Dropout for uncertainty estimation. Competitive results were achieved for both glomerular-level and kidney-section-level classification. Predictive entropy was used for uncertainty analysis. As occurred in other works, high uncertainties were related with incorrect predictions, which justify our selection of Monte-Carlo Dropout for uncertainty estimation on MN classification.

3. Materials and Methods

3.1. Data set

The data set consists of 4,682 images of human glomerulus, containing images labelled as one of the following classes: Primary membranous nephropathy, secondary membranous nephropathy, hypercellularity, glomerular sclerosis (referred as sclerosis), and images with no lesion (referred as normal). The images were selected from the digital histological image library of the [anonymized for revision] and properly disconnected to the patient information to avoid identification. The tissue samples were fixed in Bouin's fixative or formalin–acetic acid–alcohol, included in paraffin. Sections of 2–3 μm were stained by Hematoxylin and Eosin (H&E). The images were captured using an Olympus QColor 3 digital camera attached to a Nikon E600 optical microscope (using 200$\times$magnification). From each section, relevant regions were cropped and labelled individually by pathologists for diagnosis purposes. The data set was built considering only the crops that contained at least one glomerulus.

Table 1. Class distribution considering individual labels and grouped classes.

Normal	Membranous		Other lesion	
	Primary	Secondary	Hypercellularity	Sclerosis
869	712	1354	1237	510
	2066		1747	

Our work focused on the membranous nephropathy lesion, which was collected in primary and secondary variations. Since for now we are not interested in the differentiation between the MN types, primary and secondary MN images were grouped into a single group called "membranous". In fact, that differentiation is frequently hard to be made considering only visual features, but in practice, the following criteria can be used: glomeruli with primary membranous have MN characteristics only; diversely, glomeruli with secondary membranous have MN characteristics and other lesions involved. A common validation approach would be to train the models in a membranous$\times$no-lesion setup. However, in real case scenarios, other lesions not related to MN may show up. Since the data set also contained glomerular images with hypercellularity and sclerosis, we grouped these images into a class called "other lesions". Therefore, the final class configuration can be summarized as follows (see Table 1 for detailed class distribution):

- **Membranous:** glomeruli with any lesion combination that includes MN;
- **Other lesions:** glomeruli with hypercellularity or sclerosis, but no MN;
- **Normal:** glomeruli with no lesion.

3.2. Evaluation pipeline

The proposed evaluation pipeline is shown in Figure 2. The process can be split into two steps: evaluation of selected deep learning architectures for MN classification; and uncertainty estimation. The first step consists of training and validation of the following network architectures selected from the literature: ResNet-18, DenseNet, and Wide-ResNet ([He et al., 2016, Huang et al., 2017, Zagoruyko and Komodakis, 2016] respectively). This validation phase allowed the selection of the architecture that achieved the highest effectiveness in terms of F1-Score.

The second step starts by introducing the MC dropout to the best architecture and retraining this updated network. With dropout activated, we can estimate uncertainty by performing Monte-Carlo samples from the trained network. The details of the training and validation procedures and the uncertainty estimation process are presented next.

3.2.1. Training and validation

For training and validation of the candidate architectures, a K-fold cross-validation approach was used. This approach consists of splitting the data set into K folds and interactively evaluate the models by using $K-1$ folds for training and the 1 left for validation. Therefore, on each iteration we train and evaluate the model using different training and validation sets. This way, we validate the architectures by considering the average performance on K rounds. In order to avoid a large reduction on the training set, the cross-validation was performed adopting $K = 10$.

3.2.2. CNN architectures and training procedure

In the experiments, three CNN architectures were assessed: ResNet-18, DenseNet, and Wide-ResNet. Each architecture rely on a different approach in the learning process. ResNet-18 introduced the residual blocks, which main novelty was the skip connections to prevent the vanishing gradient problem. DenseNets expand the skip connections by connecting each layer to every other forward layer. Finally, Wide-ResNet is a ResNet variant with decreased depth and increased width, which allows the learning of more features without increasing the depth of the network, leading to a faster convergence on training.

The Pytorch framework [Paszke et al., 2019] was used for modeling, training and evaluating the models. Since fine-tuning pretrained models leads to faster and better convergence for medical imaging [Raghu et al., 2019], all models were initialized with weights pretrained on the ImageNet data set [Russakovsky et al., 2015] with an adjusted softmax layer with three neurons respective to the target classes. The networks were trained across 100 epochs with a batch size of 32, setting a learning rate schedule with step decay of factor 0.1 at every 30 epochs. We experimented the values of 0.1, 0.001, and 0.0001 for the initial learning rate, and 0.0001 was the top performing configuration. All training procedures used AdamW optimizer [Loshchilov and Hutter, 2017] running on a machine with 8GB RAM and an NVIDIA GEFORCE GTX 1060. In order to increase input variability, we adopted online image augmentation by applying pre-defined random transformations. These transformations include: Random rotation within an angle range of 90 degrees and probability of 0.5; random horizontal and vertical flip; random crop of size 224×224 after resizing the input height to 224, thus keeping aspect ratio and matching the input size of 224×224 for all networks.

3.2.3. Classification metrics

A quite common metric for evaluation of classification models is **Accuracy**, which corresponds to the ratio between the number of correct predictions and the number of instances.

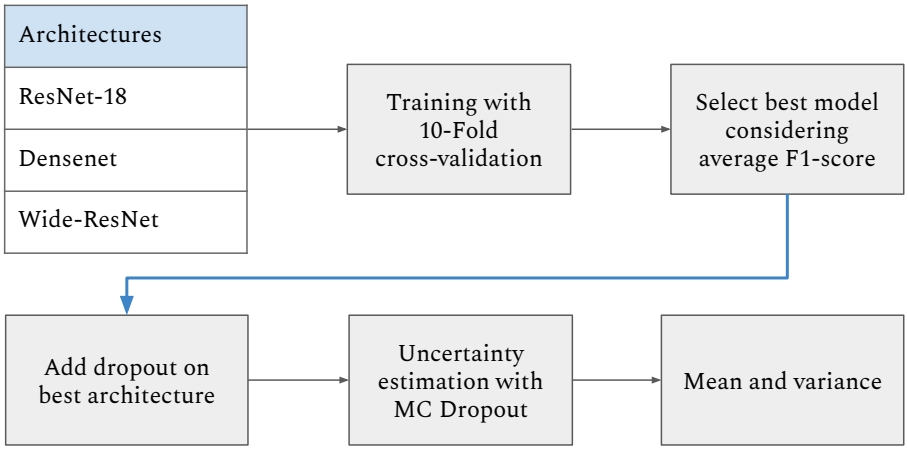

Figure 2. Proposed evaluation pipeline split into two steps: Evaluation of chosen architectures, and uncertainty estimation.

However, this metric can lead to a biased analysis when some class is predominant or under-represented. Since we have a slightly unbalanced data set, additional evaluation measures were adopted to ensure the performance for each class is taken into account. The **Precision** metric summarizes how much positive predictions are actually positive. The **Recall** metric summarizes how much positive examples were correctly classified as positive. By combining these two metrics, we used the **F1-score**, which is computed by taking the harmonic mean between precision and recall. Considering that there are some similarities between secondary membranous and the "other lesions" class, we computed the confusion matrix of the overall best model.

3.2.4. Uncertainty estimation

We followed the uncertainty estimation approach described by Combalia et al. [2020]. This approach combines Monte-Carlo Dropout [Gal and Ghahramani, 2016] and Test-Time Data Augmentation [Ayhan and Berens, 2018] to estimate both aleatoric and epistemic uncertainty. Aleatoric uncertainty represents noise inherent to the observed data, mostly related to the labelling process and the challenges of the domain. Conversely, the epistemic type captures uncertainty about the model and the generalization process.

After selecting the architecture with the best average F1-Score on the 10-fold cross-validation, a dropout layer was introduced previous to the last fully-connected layer. Hence, the network was retrained with a dropout rate of 0.5. This means that for each forward pass, there was a chance of 50% of turning off the neurons connected in the dropout layer. Dropout helps preventing overfitting and can be used to estimate uncertainty via Monte-Carlo sampling. During test time, we keep dropout activated and for each input x we perform M forward passes. Each forward pass results in a different set of activated neurons as well as a different prediction score. So, each image x_i yields M predictions $p = \{p_1, p_2, \ldots, p_M\}$. The mean of these predictions is the final prediction y_i, and the variance is interpreted as the uncertainty score u_i. Additionally, random data augmentation was applied during test time for each input x_i at each of the M forward passes, thus achieving combined aleatoric and epistemic uncertainty.

4. Results and discussion

Table 2 shows the results of the chosen architectures on our proposed MN classification. For each architecture, average metrics and respective standard deviations are displayed. Overall, all architectures achieved competitive results, displaying average metrics above 92% with low standard deviation. As expected, Wide-ResNet achieved the highest average F1-Score (in bold) with similar values for all the other metrics, showing robustness in the results. Also, Wide-ResNet achieved the lowest standard deviation for all metrics, showing more stable results across all folds.

Even though a high F1-Score indicates a good learning per class, we computed the confusion matrix for Wide-ResNet to check whether the class imbalance was prioritizing one class in the learning process. In addition, we can use the confusion matrix to check the most common misclassification cases. To get an aggregated view, we summed the confusion matrices across all 10 folds. Figure 3 illustrates the confusion matrix sum using a heatmap approach to highlight the highest values. The correct classifications are located

Table 2. Comparative results of the ResNet-18, DenseNet and Wide-ResNet deep-learning-based architectures.

Model	μAccuracy	μF1-score	μPrecision	μRecall
ResNet-18	$0.924(\pm 0.010)$	$0.922(\pm 0.008)$	$0.922(\pm 0.010)$	$0.922(\pm 0.009)$
DenseNet	$0.932(\pm 0.011)$	$0.930(\pm 0.011)$	$0.928(\pm 0.012)$	$0.933(\pm 0.013)$
Wide-ResNet	**0.937($\pm$ 0.007)**	**0.936($\pm$ 0.007)**	**0.936($\pm$ 0.008)**	**0.936($\pm$ 0.008)**

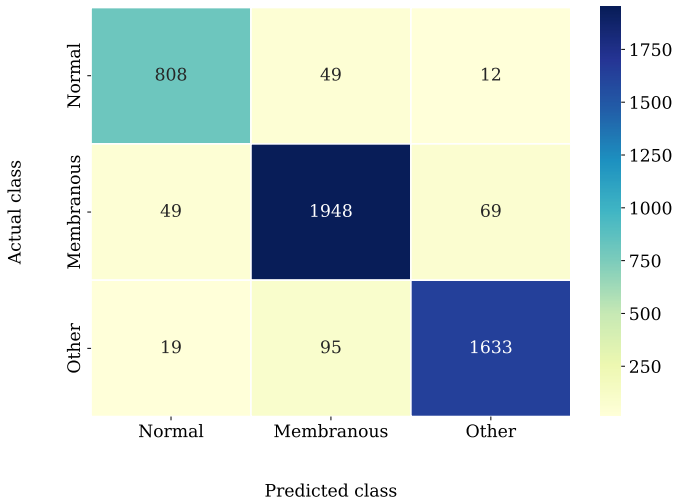

Figure 3. Wide-ResNet confusion matrix sum over the 10-fold cross-validation with a heatmap visualization.

in the main diagonal, where the ground truth labels match the predicted labels. We can note that the main diagonal is indeed the highlighted region from the heatmap, ratifying the results from Table 2. Among the regions out of the main diagonal, we selected the two highest values. Interestingly, the 2 most common misclassifications were: 95 "other lesion" images predicted as membranous; and 69 membranous images predicted as "other lesion". This behaviour was quite expected, since secondary membranous images could have visual features that are also present on the images from "other lesion" label.

For uncertainty estimation, we applied Monte-Carlo Dropout using $M = 100$ forward passes. After adding the dropout layer, we trained the model initialized with ImageNet weights and computed the same metrics in a 10-fold cross-validation. For better comparison, we refer to this model variation as D-Wide-ResNet. Table 3 summarizes the average scores of the D-Wide-ResNet over the 10 folds. The results are below, but still close to the performance of the other models, with a lower standard deviation. We used variance as the main metric on our study of uncertainty for MN classification. To

Table 3. Results of D-Wide-ResNet with $M = 100$ forward passes over 10 validation folds.

Model	μAccuracy	μF1-score	μPrecision	μRecall
D-Wide-ResNet	$0.922(\pm 0.006)$	$0.921(\pm 0.006)$	$0.920(\pm 0.006)$	$0.922(\pm 0.008)$

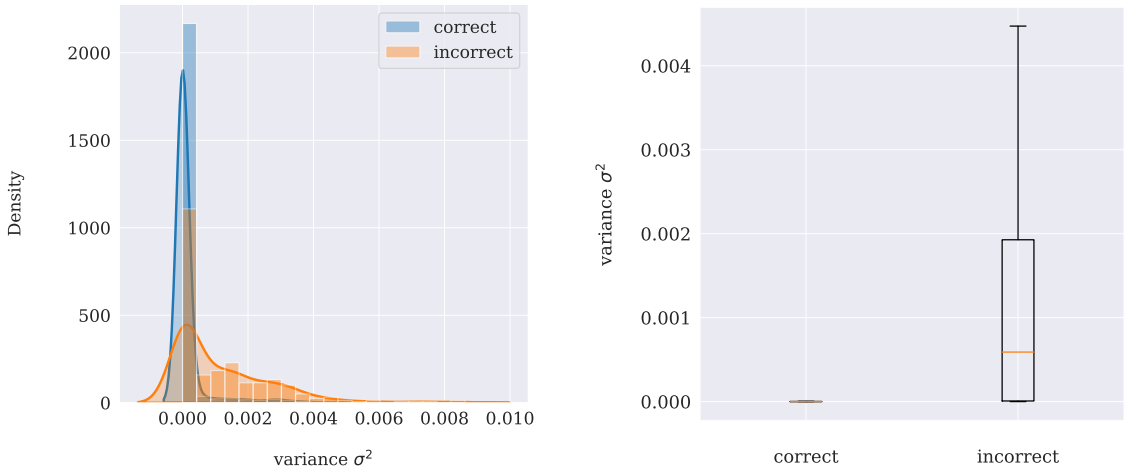

Figure 4. Uncertainty scores visualizations grouped into correct and incorrect predictions. Left: Normalized density histogram with a Gaussian kernel density estimation. Right: Boxplots of uncertainty scores.

confirm whether the variance has relation with erroneous predictions, we gathered two sets: variances of correct predictions and variances of incorrect predictions over the 10 validation folds. Average uncertainty of erroneous predictions was 0.0012, which was eight times higher than the average uncertainty of correct predictions. Figure 4 illustrates relative frequencies (left) and boxplots (right) for uncertainty of correct and incorrect predictions. To illustrate the distribution of uncertainty, we plotted the normalized density histograms with a gaussian kernel density estimation to better represent the uncertainty areas of correct and incorrect predictions. The correct predictions are clustered in the region with lower uncertainty. On the other hand, incorrect predictions have a spread distribution over regions with higher uncertainty scores. The boxplots ratify that the correct predictions have lower and more concentrated uncertainties, just as incorrect predictions have higher and more disperse uncertainties.

5. Conclusion

Membranous Nephropathy (MN) is an autoimmune disease with high risk and pathology relevance. Identifying this lesion is an important and challenging task, which could be assisted with automatic classification methods. In this work, we investigated deep-learning-based architectures for MN classification considering three classes: membranous, other lesions, and no lesion. By evaluating ResNet-18, DenseNet and Wide-ResNet in a 10-fold cross-validation setup, we achieved top results with all architectures, obtaining average F1-Scores above 92% for all models. Among the chosen architectures, Wide-ResNet achieved the highest average F1-Score (93.6%). Besides evaluating MN classification, we modeled the prediction uncertainty from Wide-ResNet. Uncertainty estimation is important to achieve more reliable results, specially in a high risk domain as medical imaging. We used Monte-Carlo Dropout to estimate uncertainty scores based on the variance of Monte-Carlo samples. Although the modified Wide-ResNet achieved slightly lower results (F1-Score of 92.1%), the uncertainty estimates showed high relation with misclassifications where the average uncertainty of incorrect predictions was eight

times higher than the average uncertainty of correct predictions.

For future work, we will evaluate whether the high-uncertainty images are hard to specialists' diagnose. In addition, Monte-Carlo dropout should be evaluated by adding dropout layers inside the backbone, such as after each convolutional layer or after each convolutional block. Ultimately, we plan to comparing other uncertainty estimation methods with our current results on Monte-Carlo dropout.

Acknowledgments

PathoSpotter 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.

References

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *International Conference on Medical Imaging with Deep Learning*, 2018.

Jan U Becker, David Mayerich, Meghana Padmanabhan, Jonathan Barratt, Angela Ernst, Peter Boor, Pietro A Cicalese, Chandra Mohan, Hien V Nguyen, and Badrinath Roysam. Artificial intelligence and machine learning in nephropathology. *Kidney International*, 98(1):65–75, 2020.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Yilin Chen, Ming Li, Fang Hao, Weixia Han, Dan Niu, and Chen Wang. Classification of glomerular spikes using convolutional neural network. In *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*, pages 254–258, 2020.

Pietro A Cicalese, Aryan Mobiny, Zahed Shahmoradi, Xiongfeng Yi, Chandra Mohan, and Hien Van Nguyen. Kidney level lupus nephritis classification using uncertainty guided bayesian convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 2020.

Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 744–745, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

Max-Heinrich Laves, Sontje Ihler, and Tobias Ortmaier. 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*, London, United Kingdom, 08–10 Jul 2019.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1):1–14, 2017.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

Eiichiro Uchino, Kanata Suzuki, Noriaki Sato, Ryosuke Kojima, Yoshinori Tamada, Shusuke Hiragi, Hideki Yokoi, Nobuhiro Yugami, Sachiko Minamiguchi, Hironori Haga, et al. Classification of glomerular pathological findings using deep learning

and nephrologist–ai collective intelligence approach. *International Journal of Medical Informatics*, 141:104231, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.