Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification

Federico Pollastri*, Juan Maroñas†, Federico Bolelli*, Giulia Ligabue‡
Roberto Paredes†, Riccardo Magistroni‡, Costantino Grana*

*Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia, Italy
Email: {*name.surname*}@unimore.it

†PRHLT Research Center
Universitat Politècnica de València, Spain
Email: {*jmaronas,rparedes*}@prhlt.upv.es

‡Dipartimento Chirurgico, Medico, Odontoiatrico e di Scienze Morfologiche con Interesse Trapiantologico, Oncologico e di Medicina Rigenerativa – Università degli Studi di Modena e Reggio Emilia, Italy
Email: {*name.surname*}@unimore.it

***Abstract*—With this work we tackle immunofluorescence classification in renal biopsy, employing state-of-the-art Convolutional Neural Networks. In this setting, the aim of the probabilistic model is to assist an expert practitioner towards identifying the location pattern of antibody deposits within a glomerulus. Since modern neural networks often provide overconfident outputs, we stress the importance of having a reliable prediction, demonstrating that Temperature Scaling (TS), a recently introduced re-calibration technique, can be successfully applied to immunofluorescence classification in renal biopsy. Experimental results demonstrate that the designed model yields good accuracy on the specific task, and that TS is able to provide reliable probabilities, which are highly valuable for such a task given the low inter-rater agreement.**

I. Introduction

Immunofluorescence is a powerful technique for light microscopy that makes use of fluorescent-labeled antibodies to detect specific target antigens. It is widely used in both scientific research and clinical laboratories [1]. As a matter of fact, it represents a step of the diagnostic pipeline needed to address a correct renal histopathological diagnosis. Some specific renal diseases such as IgA Nephropathy, Membranous Glomerulonephritis and the anti-GBM glomerulonephritis, can be virtually diagnosed using only the result of immunofluorescence. The intensity and patterns of the deposits for each applied antibody must be analyzed and evaluated by an operator with strong experience in the field.

This work aims at introducing an automated tool to aid professionals in this time-consuming and highly variable evaluation process, by focusing on the identification of patterns of different antibody deposits using state-of-the-art Convolutional Neural Networks (CNNs). These models have outperformed other strategies in several computer vision tasks such as semantic segmentation [2], [3], [4], object detection [5], [6], object classification [7], [8], [9], video captioning [10], [11], [12], [13], and many others. Moreover, CNNs are nowadays the cornerstone of medical image analysis: they have been obtaining state-of-the-art results in skin lesion analysis [14], [15], [16], [17], brain and lung tumor detection [18], [19], and countless other medical imaging tasks [20], [21]. The most attractive aspect of CNNs is their ability to learn features directly from input image data, with no need of hand-crafted features, regardless of the input image acquisition techniques (*e.g.* dermoscopy, MRI, microscopy, ultrasound).

Since our main goal is to assist an expert practitioner towards making a meticulous analysis, predictions need to be intelligible, *i.e.*, we need the model to output reliable probabilities distributions and guarantee optimal support to the final decision. However, the scores provided by modern

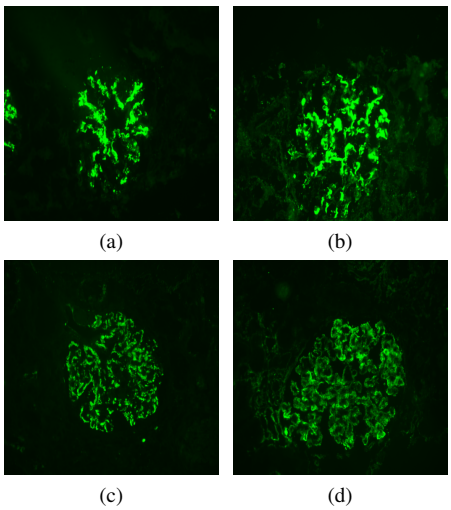


Fig. 1. Example of *mesangial* (a), (b) and *parietal* patterns (c), (d). The mesangial pattern gets the name from the heavy presence of deposits in the *mesangium* of the glomerulus. The parietal pattern can be divided in different subcategories, but the deposits are always well distributed all across the whole glomerulus and with spatial continuity.

Juan Maroñas is supported by grant FPI-UPV associated to the DeepHealth Project, grant agreement No 825111 and by the Spanish National Ministry of Education through grant RTI2018-098091-B-I00.

CNNs cannot be correctly interpreted, as the likelihoods do not represent the proportion of real outcomes. In fact, in consideration of the overconfidence of neural networks, out of 1000 images classified as disease with probability of 0.90, it cannot be assumed that 900 of them belong to that class. Proper confidence values provide valuable information to establish trustworthiness to the reader [22]. Following this observation, previous works on the field of medical diagnosis show that it is possible to apply Generalized Additive Models (GAM), such as logistic regression, to predict pneumonia risk [23], showing good accuracy performance and correct interpretability. However, these kind of models compromise accuracy in comparison to modern state-of-the-art CNNs, making the gap bigger as long as the dimensionality of the input features increases. For instance, while the features used in [23] are just 46 dimensional, in other tasks we deal with dimensionalities that are four orders of magnitude larger. Moreover, and contrary to pneumonia prediction, our inputs are images, which again motivates the use of CNNs.

To deal with the unreliable probabilities provided by modern CNNs, we apply two different re-calibration techniques: Platt Scaling (PS) [24] and Temperature Scaling (TS) [22]. By doing so, we are able to combine CNNs with an optimal decision rule, which is mandatory in critical applications as we will discuss in Section IV. To our knowledge, the latter has not yet been successfully applied to real problems. The main contributions of the paper can be thus resumed as follows:

1) This is a pioneering work that introduces for the first time the use of CNNs for immunofluorescence classification of renal biopsy.
2) We apply modern state-of-the-art re-calibration techniques to this specific task, demonstrating the importance of having reliable probabilities to support clinical decisions.
3) An exhaustive quantitative evaluation of the proposed approach is presented. Together with the qualitative analysis provided by expert nephrologists, it validates the effectiveness of our proposal.

Our final goal is to provide clinicians with a valuable tool for supporting the renal biopsy immunofluorescence image analysis.

The rest of the paper is organized as follows. In Section II a description of the existing approaches for model calibration is provided. Section III, describes the dataset used to train the proposed architecture, which is introduced in Section IV. Experimental results are then presented in Section V and discussed in Section VI. Finally, in Section VII conclusion are drawn.

II. Related Work

The first work that showed the badly calibrated probabilities of CNN is found in [22], where different classical re-calibration techniques are compared. Among all of them temperature scaling raised as the best performing technique, concluding that simple techniques should be employed for re-calibration of CNNs. Lately, [25] has shown how more complex re-calibration techniques can improve calibration if uncertainty is correctly incorporated. In that paper the authors explored the use of Bayesian Neural Networks for the purpose. In this work we focus on comparing whether simple state-of-the-art re-calibration techniques can be employed to the task tackled by this paper, establishing a performance baseline. Thus, future work will be focused on checking whether more complex techniques can be successfully applied on the same medical application.

Recently, Mozafari *et al.* proposed the Attended Temperature Scaling, a variant of TS designed to be used in scenarios where the validation set is small, or contains noisy-labeled samples [26]. On the other hand, [27] has studied how pre-training affects calibration, robustness and uncertainty quantification. The work has been then extended in [28], where self-supervised scenarios are explored. Additionally, in [29] the authors measured performance on calibration and uncertainty quantification of several techniques under dataset shift.

Many papers that study how data augmentation strategies affect the calibration, uncertainty quantification and robustness has been published in the last years. In [30], for instance, has been measured the robustness and calibration of *Mixup* training [31] showing improved results over a baseline model. However, [32] performed a deep analysis on Mixup trained models, showing that it does not always improve calibration and proposing a loss function to deal with the problem. Hence, it is not clear whether data augmentation strategies can calibrate by design.

Following related strategies, [33] proposed *On-Manifold Adversarial Data Augmentation*, which attempts to generate challenging examples by following an on-manifold adversarial attack path in the latent space of a generative model. More recently, [34] proposed *Augmix*, a technique that is build on top of Mixup and provide good results both on uncertainty quantification and robustness. Network Ensembles are another promising line of research [35].

III. Mesangial and Parietal Patterns

A. The Task

The location within a glomerulus of potential deposits is extremely meaningful during the medical diagnosis. We thus focus on recognizing the two most common and relevant deposit location patterns: *mesangial* and *parietal*, shown in Fig. 1. This two distinctive patterns emerge when an antibody gets attached to a precise type of cell inside the glomerulus, respectively mesangial and parietal cells. The two investigated location patterns are not mutually exclusive, each image can present both, only one, or neither of the two. As a matter of fact, the location pattern recognition problem could be treated as two separated binary tasks, or as a single task with four different classes, one for each combination of the predictions of the two location patterns. We consider the fact that there is no theoretical relation between the presence (or absence) of the two patterns and chose to treat this problem as two completely decoupled binary prediction tasks. Moreover, studying the two patterns jointly would introduce the issue of training the CNN

to recognize that making a wrong prediction over both of the patterns is worse than making a wrong prediction over one pattern, while correctly classifying the other one. This aspect would not be reflected by the most common loss functions employed in the neural network training.

Unlike many other computer vision tasks such as object detection, the analysis of immunofluorescence images requires a specific background and, as proved by a low inter-rater agreement, in many cases it remains ambiguous even for expert practitioners. Indeed, in Section VI we measure the Cohen's kappa coefficient from the opinions of four expert pathologists, and find it below 0.6 for both of our tasks. Since the presence or absence of a deposit location pattern is not as categorical as the presence or absence of a natural object (*e.g.* a dog or a human), the most common image classification approaches are not well suited to face the considered task. As a matter of fact, binary predictions would be an extremely underwhelming tool for immunofluorescence image analysis. On the other hand, continuous scores are a good representation of the actual opinion of an expert pathologist. In other words, rigorous likelihood scores are extremely more functional than a plain category assignment, since they can be interpreted by a specialist during the analysis of biopsied tissues. However, as discussed in Section IV, neural networks are a great tool for binary classification but they do not naturally provide accurate likelihood scores.

B. The Dataset

In order to tackle renal biopsy immunofluorescence image analysis, we gathered a dataset composed of histological images of renal biopsies, which were captured on a fluorescence microscope (BX41 with U-RFL-T, Olympus) by a digital camera (XC30, Firmware version 4.0.2, Olympus) controlled by a dedicated software (cellB software, Olympus). The pictures are stored as one channel images with a resolution of $1\,040 \times 772$ pixels, in 12 bit uncompressed Tagged Image File Format (TIFF) [36].

Although each image shows deposits on a single glomerulus, annotations are inferred by medical reports, which were given by experts on account of the analysis of several glomeruli obtained through the biopsy. As a result of this label-deduction process, ground truth annotations contain minor inaccuracies. In consideration of the fact that the patterns are not affected in any way by the choice of antibody used during immunofluorescence, images are merged regardless of the antibody employed, leading to a database composed of $10\,979$ samples. The dataset contains many more samples that do not present the investigated patterns (negatives) than samples that do present such patterns (positive), and is therefore imbalanced. Indeed, of the $10\,979$ total images, $3\,249$ exhibit the parietal pattern, $2\,104$ exhibit the mesangial one. A total of $1\,097$ samples exhibit both of the investigated patterns.

In order to consider the two tasks as independent of one another, two different splits of the dataset are created in order to obtain, for each task, a $9\,479$ samples training set, a 500 samples validation set, and a $1\,000$ samples test set. Positive samples are over-represented in the test set, whereas the original proportions are preserved in the validation set.

IV. Probabilistic Framework

Neural networks are function approximators that combine linear and non-linear operations, which are designed to be distributed across modern hardware architectures. The key property of these models is that they make a hierarchical representation of the input, starting from low level features represented in the early layers, to more complex abstract features in the final ones. These features can be then used to take decisions.

In our classification scenario, where the goal is to assign a class label t to a set of input images x, neural networks are typically trained by maximizing an unbiased stochastic estimate of the likelihood or posterior w.r.t. a set of parameters θ given a set of N observations $\mathcal{O} = \{(x^i, t^i)\}_{i=1}^N$. In this context, the neural network models a k-Categorical class conditional probability distribution $p(t_k|x, \theta)$, where k is the total number of classes, in this case $k = 2$. Once the training criteria is optimized, we recover the optimal parameters $\hat{\theta}$ and use them to make predictions over new unseen images.

When dealing with unbalanced dataset, it should be noticed that a point-estimate selection of the optimal $\hat{\theta}$ might lead to an undesired behaviour where the most unrepresented class is ignored, and the CNN just learns to classify all the samples to the most represented class. This can be avoided by subtracting the empirical prior, something which can be done through a weighted cross entropy loss using the inverse prior probability. To justify this claim we have to consider the Bayes rule, from which we know that

$$p(t^k|x) \propto p(x|t^k)p(t^k). \tag{1}$$

Thus, by maximizing $p(t^k|x)$ we implicitly maximize $p(x|t^k)p(t^k)$. As a consequence, the cross entropy loss can be scaled by $1/p(t^k)$ so that our model learns $p(x|t^k)$ for each of the classes. This can be viewed as learning the posterior distribution assuming equal prior distributions for all classes. With this approach we force the model to learn a representation of data based on x and not based on the class proportion.

On the other hand, the key point of using neural networks in the context of medical diagnosis is to assist an expert practitioner towards the final decision, not replace it. Thus, we now discuss how the probabilistic information provided by the neural network can be used to assist an expert practitioner in an optimal manner and how expert knowledge can be combined with the information provided by the model. This will motivate the necessity of having reliable probability distributions.

A. Probabilistic Models For Optimal Decision Making

In critical applications different decisions can have extremely different consequences. For instance, in the medical context, it is different to decide towards action α_1: *the patient has a disease*, than towards α_2: *the patient does not have a disease*. In the latter, an incorrect diagnosis can have drastic consequences. For that reason, if we are going to use a

probabilistic framework to assist the decision make by an expert practitioner, we must be able to incorporate expert knowledge in the best possible way.

A probabilistic binary classifier decides towards action α_i by selecting the action that minimizes the Bayes Risk denoted by $\mathtt{R}(\alpha_i|x)$. This rule can be defined in the following way:

$$\begin{aligned} \mathtt{R}(\alpha_i|x) &= \sum_{k=1}^{2} \lambda_{ik} \cdot p(t_k|x,\hat{\theta}) \\ \alpha_i &= \underset{i\in\{1,2\}}{\operatorname{argmin}}\, \mathtt{R}(\alpha_i|x) \end{aligned} \tag{2}$$

where λ_{ik} is the loss incurred when deciding class t_i if the ground truth is t_k. It is well known that this rule provides optimal performance if our model probability $p(t|x,\hat{\theta})$ recovers the data distribution $p(t|x)$ which, in general, cannot be guaranteed. However, the better our model approximates this distribution, the closer to the optimal error we will be.

In order to illustrate how expert knowledge can be incorporated, we rewrite this rule. We decide α_2 if:

$$p(x|t_2) \cdot p(t_2) \cdot \lambda_{12} > p(x|t_1) \cdot p(t_1) \cdot \lambda_{21} \tag{3}$$

where we are assuming that there is no loss associated in correctly classifying a sample, *i.e.* $\lambda_{ii} = 0$. Expert knowledge can now be incorporated through λ_{ik}. For example, if the risk of deciding towards action 2 (our patient is sane) is higher than deciding towards class 1 (our patient is not sane), then an expert practitioner can set $\lambda_{21} \gg \lambda_{12}$. If an automatic system assigns moderated probabilities, but slightly higher for action α_2, a proper choice of λ_{ik} can change our final decision to be α_1. This will lead an expert practitioner to perform new medical tests to deeply analyze that particular patient, given that the automatic system, and probably the expert, is not confident enough on the decision to be taken.

Following the previous example, it should be noted that when the probability provided by the model is overconfident the expert knowledge incorporated through λ_{21} could not change the final decision taken, with potential drastic consequences. This means that in critical applications, it is important to provide the correct class as much as to provide a probability distribution that actually reflects the ground truth.

B. Model Calibration

As already introduced, to provide optimal performance we need the model probability $p(t|x,\hat{\theta})$ to resemble the true unknown distribution, *i.e.* we need reliable probability distributions. To analyze this reliability, we need to measure two different concepts: accuracy and calibration [37]. The accuracy reflects if our final choice on the action to take is the correct one, while the calibration measures how informative are the confidence scores used to make these choices.

In a classification scenario, the calibration can be interpreted as the agreement between the probabilities assigned by a model and the distribution that characterizes the data. In other words, if our probabilistic classifier assigns class t_1 with probability 0.6 to a set of samples, then we expect that 60% of these samples actually belong to class t_1. If that happens, our model is perfectly calibrated because the confidences provided are reflecting the true proportion of samples in the distribution. In this work we measure calibration by taking an unbiased estimate of the Expected Calibration Error (ECE). This is done by partitioning the probabilistic space in M equally spaced bins and then computing the metric:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{4}$$

where B_m represent the set of samples that lie in bin m, $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the accuracy and average confidence of that bin. This metric is basically a weighted average of the intuitive description we have given for calibration. ECE is 0 if the accuracy is equal to the average confidence obtained on the samples contained in every bin. For a wider description see [22], [38].

It is important to remark that the two metrics, calibration and accuracy, are decoupled. This means that we can have perfect calibration and useless accuracy, as for example in a prior classifier. Thus, our model must provide sharp and calibrated probability distribution, *i.e.* models that correctly classify the data with a reliable confidence. The ideal case would be to classify correctly all the samples assigning them 1.0 confidence, that would be the case of a data distribution totally separable. On the other hand, if the data overlaps, our model must incorporate this uncertainty in the decision by setting adequate final confidences.

In theory, to achieve optimal accuracy and calibration we need to train the model by minimizing a proper scoring rule [39], [35]. Examples of these are the Negative-Log-Likelihood (NLL) or the Brier Score (BS). In our set-up, the CNNs are trained to minimize the NLL theoretically providing good accuracy and calibration performance. However, this is far from being true when using deep CNNs.

C. Convolutional Neural Networks

Convolutional Neural Networks are the current state-of-the-art in computer vision applications. They are designed to provide translation and scale invariance and detect local and global patterns. In this scenario, an input image x is mapped to a probability distribution over the different classes $p(t|x,\hat{\theta})$.

A recent work [22] has shown that state-of-the-art CNN architectures are badly calibrated in general, although they are trained to minimize a proper scoring rule. They show that these models provide overconfident predictions, which means that they assign high confidence without presenting such a high accuracy. Thus, if we want to use a deep CNN to parameterize $p(t|x,\theta)$ we need techniques to calibrate the output, in order to provide optimal decision performance, as discussed above. In the literature we can find *decoupled* techniques which take as input the logit or pre-softmax from an already trained CNN to train a re-calibration mapping, or *coupled* techniques, that aim at directly calibrating the model.

On the other hand, in scenarios where the input distribution configured by the training images is complex, or the number

of training samples is scarce, a common practice is to use pre-trained CNNs. Firstly, the CNN is trained on a complex and large image dataset. In this step, the network learns relevant features that describe images in general. Afterwards, the CNN is fine-tuned with the samples from our training distribution. In other words, instead of training our model starting from a random choice of the parameters, we start from a model that already makes a good representation of images in general.

D. A Baseline for Renal Biopsy Immunofluorescence Images

In order to tackle our two classification tasks, we train several versions of ResNet [40], a state-of-the-art neural network introduced in 2016 that has obtained excellent results on several tasks by means of residual blocks. We further extend the experiments by employing one version of DenseNet [41], which was introduced in 2017 and enhances the concept of residual blocks, and three versions of EfficientNet, which was introduced in 2019 and aims to obtain state-of-the-art results with more shallow and efficient architecture [42]. Moreover, we aim to improve the results obtained by ResNet and DenseNet by introducing a dropout layer [43] right before the last fully connected layer of the neural network, thus improving regularization and avoiding overfitting. We do not add any regularization layer to EfficientNet, given the fact that a dropout layer is already built into the model.

The absence of balance between the positive and negative samples, which is relevant for both tasks, is handled through several techniques. First of all, we perform data augmentation by randomly flipping and rotating input images, which mitigates the need of huge amounts of training samples without altering the semantic content of an image. Moreover, neural networks are pre-trained on ImageNet [44] and then fine-tuned to minimize the weighted Cross-Entropy Loss over renal biopsy images: the loss obtained by each input image is multiplied by a weight inversely proportional to the quantity of samples belonging to the same class in the dataset. Finally, we monitor the F1-score metric and make use of the validation set to apply the early-stopping technique, which ends up forcing the last 40 of the 80 total epochs to be always ignored. During the fine-tuning of every network, the learning rate is initially set to 1^{-5} and then adjusted by the Adam optimizer [45].

E. Re-Calibration of Convolutional Neural Networks

Motivated by the observation in [22] we have measured the calibration performance of the trained models. We observed that they are very bad calibrated despite being pre-trained models [27]. For that reason we compare and discuss the performance of two simple yet effective and well-established decoupled calibration techniques [22], [24]. To our knowledge, only [26] has explored the use of re-calibration techniques for real-case scenarios involving Deep CNNs (skin-cancer detection). We build up on this work, exploring the use of re-calibration on a different medical application and proving his effectiveness. To illustrate how both calibration techniques work, consider a dataset $\mathcal{O} = \{(l_i, t_i)\}_{i=1}^{N}$ where l_i is the logit vector (pre-softmax) computed by taking an input image x and passing it through the already trained CNN.

Platt Scaling [24]: platt scaling maximizes the log-likelihood of the conditional distribution $p(t|l, W, b) = softmax(W \cdot l + b)$ w.r.t a set of parameters $W \in \mathbb{R}^{2\times 2}$ $b \in \mathbb{R}^2$ on a validation set.

Temperature Scaling [22]: temperature scaling maximizes the log-likelihood of the conditional distribution $p(t|l, T) = softmax(l/T)$ w.r.t a single parameter $T \in \mathbb{R}$ on a validation set. This parameter is applied to all the elements of the logit vector. This calibration technique does not change the accuracy of the CNN model because the transformation applied is monotonic, the softmax function does not affect the argmax of the probability vector. This is the best property of temperature scaling. The main drawback is that it is a very simple transformation that might not work in complex scenarios. However, as showed in [22] it works well in CNN for image classification.

A good property of both techniques is that, given a dataset, the optimization problem is convex. Thus, as these techniques minimize a proper scoring rule, we are guaranteeing optimal performance in terms of accuracy and calibration on the validation dataset[1]. Moreover, this optimization is not expensive and can be run in a normal CPU.

V. Experimental Results

Experimental results on the mesangial and parietal pattern recognition tasks are reported in Table I and Table II respectively. For each of the trained neural networks, the reported metrics are accuracy (Acc), precision (Prec), Recall (Rec), F1-Score (F1-S), Area Under the ROC Curve (AUC), and Expected Calibration Error (ECE). Eight different state-of-the-art pre-trained models are employed to solve the classification problem, they all rely on residual blocks and are presented in the first column of the Tables. The second column of the Tables displays the probability of dropping out each CNN unit [43].

The last part of the Tables is divided according to the two different calibration methods applied to the CNNs and described in Section IV-E. For both of them, the impact on the calibration metric ECE is reported. However, the Platt Scaling (PS) re-calibration technique has a negative impact on the final class decision, whereas the Temperature Scaling (TS) technique does not affect the discriminative power of the neural networks. Therefore, classification metrics are reported under the section platt scaling.

Experimental results show that every CNN yields a good classification accuracy on both tasks. Setting the dropout probability to 0.5 usually grants a minor boost in accuracy, although the only noticeable improvement is for DenseNet-121 in Table II, which is the only model providing an accuracy over 80% on parietal pattern recognition. Moreover, all the trained architectures achieve a valuable balance between recall

[1]In order to make TS convex the logit transformation must be $l \cdot T$, however we follow the same notation from the original authors. In fact with the original notation we can reach $T < 0$ depending on where do we initialize the parameter.

TABLE I
PERFORMANCE FOR MESANGIAL PATTERN CLASSIFICATION.

		Uncalibrated						PS						TS
Model	**Drop**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**ECE**
DenseNet-121	0	81.00	76.70	70.90	73.70	79.00	13.19	77.50	81.00	52.30	63.50	72.50	4.96	2.31
DenseNet-121	0.5	82.20	76.50	75.70	76.10	80.90	4.19	78.80	86.90	51.2	64.40	73.30	5.27	3.00
ResNet-101	0	82.10	75.40	77.60	76.50	81.20	8.86	80.00	85.40	56.30	67.80	75.30	3.08	2.67
ResNet-101	0.5	82.10	79.20	70.90	74.80	79.90	12.64	78.80	85.00	52.80	65.10	76.30	3.77	3.06
ResNet-18	0	81.30	78.30	69.30	73.50	78.90	1.62	79.40	85.70	54.10	66.30	74.30	4.40	1.41
ResNet-18	0.5	81.90	76.40	74.90	75.60	80.50	3.35	78.50	83.60	53.10	64.90	73.40	6.33	2.96
ResNet-50	0	81.60	72.70	81.60	76.90	81.60	7.59	79.70	85.20	55.50	67.20	74.90	4.71	2.19
ResNet-50	0.5	81.70	77.30	72.50	74.80	79.90	3.62	79.60	85.90	55.20	67.20	74.90	3.83	2.58
ResNet-152	0	81.60	75.50	75.50	75.50	80.40	10.40	79.80	85.30	55.70	67.40	75.00	4.45	3.00
ResNet-152	0.5	82.10	73.80	81.10	77.30	81.90	2.22	80.00	86.90	54.90	67.30	75.00	4.53	2.29
EfficientNet-b3	0.3	78.40	72.50	68.30	70.30	76.40	12.54	77.60	82.10	51.50	63.30	72.40	4.94	3.13
EfficientNet-b4	0.4	79.60	75.20	68.00	71.40	77.30	14.54	78.40	85.00	51.50	64.10	73.00	4.78	4.00
EfficientNet-b5	0.4	79.40	75.50	66.70	70.80	76.90	13.16	76.70	81.40	49.10	61.20	71.20	7.02	5.70

TABLE II
PERFORMANCE FOR PARIETAL PATTERN CLASSIFICATION.

		Uncalibrated						PS						TS
Model	**Drop**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**ECE**
DenseNet-121	0	76.80	79.90	64.70	71.50	75.70	15.42	76.40	83.40	59.30	69.30	74.80	6.97	5.73
DenseNet-121	0.5	80.30	78.70	77.10	77.90	80.00	13.25	77.20	85.60	59.30	70.10	75.60	4.20	3.21
ResNet-101	0	77.30	75.40	73.60	74.50	77.0	17.31	76.00	83.40	58.20	68.60	74.40	4.57	3.88
ResNet-101	0.5	75.90	82.60	58.90	68.70	74.40	18.93	75.20	84.50	54.70	66.40	73.20	5.04	3.77
ResNet-18	0	75.60	76.50	66.00	70.90	74.70	15.04	75.60	82.60	58.00	68.10	74.00	4.85	4.36
ResNet-18	0.5	78.20	79.00	70.20	74.30	77.50	11.37	76.10	83.10	58.90	68.90	74.50	5.36	4.19
ResNet-50	0	76.80	82.10	62.00	70.60	75.50	17.38	75.20	86.10	53.80	66.20	73.30	5.34	3.66
ResNet-50	0.5	76.90	82.10	62.20	70.80	75.60	16.78	75.80	84.30	56.00	67.30	73.70	5.55	4.52
ResNet-152	0	77.60	81.20	65.30	72.40	76.50	18.59	76.00	84.30	57.30	68.20	74.30	4.23	4.06
ResNet-152	0.5	76.00	80.00	62.20	70.00	74.70	19.00	74.70	82.40	56.00	66.70	73.10	5.53	4.53
EfficientNet-b3	0.3	78.20	74.90	77.60	76.20	78.10	8.52	74.40	83.20	54.00	65.50	72.50	5.80	2.35
EfficientNet-b4	0.4	77.50	77.80	70.00	73.70	76.80	12.37	74.30	82.90	54.00	65.40	72.50	6.36	3.69
EfficientNet-b5	0.4	77.50	77.50	70.40	73.80	76.90	14.62	75.10	82.70	56.40	67.10	73.40	5.77	3.85

and precision, obtaining positive results in both F1-Score and AUC. The ECE is always enhanced when the temperature scaling method is applied, whereas the platt scaling approach is unable to yield good results when the model is already fairly calibrated. Moreover, temperature scaling preserves the classification capabilities of the networks, while platt scaling tends to degrade them. However, both of the techniques are able to calibrate the output of the CNN trained to recognize the parietal pattern. It is remarkable that, although we are using pre-trained models, the baseline results are rather uncalibrated. This is in contrast with the observations in [27].

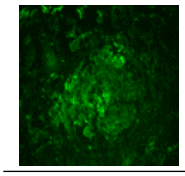
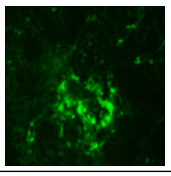
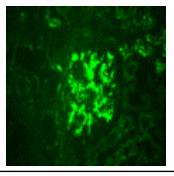
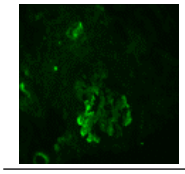
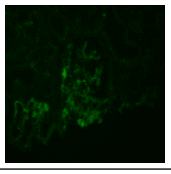
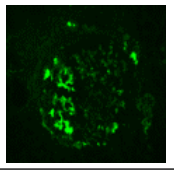
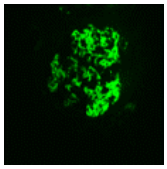
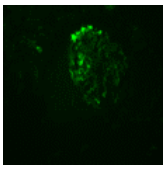
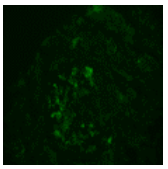
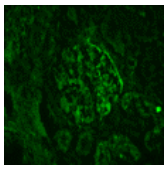
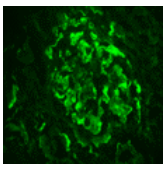
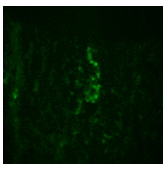
VI. DISCUSSION

Table III depicts the impact that calibration methods have on the task. The first two columns present images that DenseNet-121, with dropout probability set to 0, does not classify correctly. Differences between calibrated and uncalibrated values confirm that the neural network overconfidence is mitigated, which is undoubtedly helpful in the case of misclassified samples. The last column displays images where the mesangial pattern is correctly recognized by the CNN. Our approach proves to be helpful also in this case, when the neural network does not output a wrong prediction: as the calibrated predicted probability decreases, so does the clarity of the investigated pattern. We asked an expert practitioners to provide likelihood scores of the investigated pattern. The values presented by the human expert in the last column of Table III are very close to the calibrated output values of the CNN: when tested on a significant subset of the dataset, re-calibrating the CNNs output reduced the Mean Absolute Error by $\sim 5\%$.

Calibrated values are a plausible representation of the opinion of a wide selection of expert pathologists, with balanced scores and very few errors predicted with high confidence.

TABLE III

VISUALIZATION OF TEMPERATURE SCALING EFFECTIVENESS OVER DENSENET-121 WITH NO DROPOUT FOR THE MESANGIAL PATTERN RECOGNITION TASK. EACH COLUMN OF IMAGES IS IDENTIFIED BY THE CNN PREDICTION WITH RESPECT TO THE GROUND TRUTH ANNOTATION.

GT: yes **Pred: no**	***Calib*** **Uncalib**	**GT: no** **Pred: yes**	***Calib*** **Uncalib**	**GT: yes** **Pred: yes**	***Calib*** **Uncalib** ***Human***
	*0.830* 0.992		*0.781* 0.980		*0.965* 0.999 *1.000*
	*0.771* 0.977		*0.774* 0.964		*0.771* 0.977 *0.400*
	*0.571* 0.707		*0.572* 0.711		*0.658* 0.883 *0.600*
	*0.562* 0.684		*0.560* 0.679		*0.558* 0.673 *0.400*

Indeed, Table IV displays Cohen's kappa coefficients obtained through an inter-rater agreement study conducted among 3 expert pathologists and the ground truth (a fourth expert pathologist), across 40 images. Results clearly demonstrate that different practitioners tend to have diverse opinions on images, with wider uncertainties shown for more ambiguous samples. This outcome stresses once again that, in this sort of environment, delivering calibrated probabilities scores is highly more useful than just providing binary classification results. As a matter of fact, uncalibrated predictions present high probabilities for every input image, regardless of the clarity of the pattern. CNNs overconfidence introduces a wide gap between their output and results obtained by trained practitioners.

VII. CONCLUSION

This paper proposed a stable architecture for the analysis of the renal biopsy immunofluorescence images. State-of-the-art residual CNNs were employed to obtain an accuracy over 80% in the recognition of the two fundamental deposit location patterns: mesangial and parietal. Moreover, by means of the temperature scaling technique, the output of the neural network was exploited to retrieve consistent probability scores for each input image. To the best of our knowledge, this is the first work, alongside [26], that successfully explore the use of temperature scaling in real medical applications.

TABLE IV

AGREEMENT BETWEEN HUMAN EVALUATORS (THREE DIFFERENT EXPERT PATHOLOGISTS *P1*, *P2*, AND *P3*) AND GROUND TRUTH *GT* CALCULATED FOR BOTH MESANGIAL (A) AND PARIETAL (B) PATTERNS USING THE COHEN'S KAPPA.

	GT	**P1**	**P2**
P3	0.50	0.70	0.34
P2	0.50	0.50	
P1	0.80		

(a) Mesangial

	GT	**P1**	**P2**
P3	0.40	0.60	0.60
P2	0.40	0.42	
P1	0.60		

(b) Parietal

Experimental results display both quantitatively and qualitatively the effectiveness of the proposed method, and provide a insightful visualization of the importance of calibrated probabilities. As a matter of fact, likelihood scores are an accurate representation of an expert opinion, and can thus lead to excellent medical image analysis tools.

REFERENCES

[1] I. D. Odell and D. Cook, "Immunofluorescence Techniques," *The Journal of investigative dermatology*, vol. 133, no. 1, p. e4, 2013.

[2] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-Supervised Semantic Segmentation with High- and Low-level Consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[3] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Adaptive Weighting Multi-Field-Of-View CNN for Semantic Segmentation in Pathology," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] J. Perry and A. Fernandez, "MinENet: A Dilated CNN for Semantic Segmentation of Eye Features," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[5] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] C. R. Qi, H. Su, M. Niessner *et al.*, "Volumetric and Multi-View CNNs for Object Classification on 3D Data," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object Classification Using CNN-Based Fusion of Vision and LIDAR in Autonomous Vehicle Environment," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4224–4231, 2018.

[9] A. A. M. Muzahid, W. Wan, F. Sohel, N. U. Khan, O. D. Cervantes Villagómez, and H. Ullah, "3D Object Classification Using a Volumetric Deep Neural Network: An Efficient Octree Guided Auxiliary Learning Approach," *IEEE Access*, vol. 8, pp. 23 802–23 816, 2020.

[10] F. Bolelli, L. Baraldi, and C. Grana, "A Hierarchical Quasi-Recurrent approach to Video Captioning," in *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018, pp. 162–167.

[11] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2641–2650.

[12] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-Term Feature Banks for Detailed Video Understanding," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 284–293.

[13] S. Pini, M. Cornia, F. Bolelli, L. Baraldi, and R. Cucchiara, "M-VAD Names: a Dataset for Video Captioning with Naming," *Multimedia Tools and Applications Journal*, vol. 78, pp. 14 007–14 027, 2018.

[14] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting Data with GANs to Segment Melanoma Skin Lesions," *Multimedia Tools and Applications*, vol. 79, pp. 15 575–15 592, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] L. Canalini, F. Pollastri, F. Bolelli, M. Cancilla, S. Allegretti, and C. Grana, "Skin Lesion Segmentation Ensemble with Diverse Training Strategies," in *Computer Analysis of Images and Patterns*. Springer, 2019, pp. 89–101.

[17] S. Allegretti, F. Bolelli, F. Pollastri, S. Longhitano, G. Pellacani, and C. Grana, "Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1–8.

[18] A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[19] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

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

[21] M. H. Jafari, Z. Liao, H. Girgis, M. Pesteie, R. Rohling, K. Gin, T. Tsang, and P. Abolmaesumi, "Echocardiography Segmentation by Quality Translation Using Anatomically Constrained CycleGAN," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 655–663.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1321–1330.

[23] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1721–1730.

[24] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[25] J. Maroñas, R. Paredes, and D. Ramos, "Calibration of Deep Probabilistic Models with Decoupled Bayesian Neural Networks," *Neurocomputing*, vol. 407, pp. 194–205, 2020.

[26] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné, "Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks," *arXiv preprint arXiv:1810.11586*, 2018.

[27] D. Hendrycks, K. Lee, and M. Mazeika, "Using Pre-Training Can Improve Model Robustness and Uncertainty," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97. PMLR, 2019, pp. 2712–2721.

[28] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 637–15 648.

[29] J. Snoek, Y. Ovadia, E. Fertig *et al.*, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 969–13 980.

[30] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 888–13 899.

[31] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[32] J. Maroñas, D. Ramos, and R. Paredes, "On Calibration of Mixup Training for Deep Neural Networks," 2020.

[33] K. Patel, W. Beluch, D. Zhang, M. Pfeiffer, and B. Yang, "On-manifold Adversarial Data Augmentation Improves Uncertainty Calibration," *arXiv preprint arXiv:1912.07458*, 2019.

[34] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," in *ICLR*, 2020.

[35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 6402–6413.

[36] G. Ligabue, F. Pollastri, F. Fontana *et al.*, "Evaluation of the Classification Accuracy of the Kidney Biopsy Direct Immunofluorescence through Convolutional Neural Networks," *Clinical Journal of the American Society of Nephrology*, vol. 15, no. 10, pp. 1445–1454, 2020.

[37] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing Cross-Entropy for Probabilistic Binary Classifiers," *Entropy*, vol. 20, 2018.

[38] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 04 2015.

[39] M. H. DeGroot and S. E. Fienberg, "The Comparison and Evaluation of Forecasters," *The Statistician: Journal of the Institute of statisticians*, vol. 32, pp. 12–22, 1983.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[42] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[45] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.