Article

Glomerulus Classification and Detection Based on Convolutional Neural Networks †

Jaime Gallego [1], Anibal Pedraza [1] , Samuel Lopez [1] , Georg Steiner [2], Lucia Gonzalez [3], Arvydas Laurinavicius [4] and Gloria Bueno [1,*]

1   Electrical Engineering Department, University of Castilla La Mancha, Ciudad Real 13071, Spain;
    Jaime.Gallego@uclm.es (J.G.); Anibal.Pedraza@uclm.es (A.P.); Samuel.Lopez@uclm.es (S.L.)
2   TissueGnostics GmbH, Vienna 1020, Austria; georg.steiner@tissuegnostics.com
3   Hospital General Universitario, Ciudad Real 13005, Spain; lmgonzalez@sescam.jccm.es
4   Vilnius University Hospital Santariskes Clinics and Vilnius University, Vilnius 08406, Lithuania;
    Arvydas.Laurinavicius@vpc.lt
*   Correspondence: Gloria.Bueno@uclm.es
†   This paper is an extended version of our paper published in: Pedraza A., Gallego J., Lopez S., Gonzalez L.,
    Laurinavicius A., Bueno G. (2017) Glomerulus Classification with Convolutional Neural Networks.
    In: Valdés Hernández M., González-Castro V. (eds) Medical Image Understanding and Analysis. MIUA 2017.
    Communications in Computer and Information Science, vol. 723. Springer, Cham.

Abstract: Glomerulus classification and detection in kidney tissue segments are key processes in nephropathology used for the correct diagnosis of the diseases. In this paper, we deal with the challenge of automating Glomerulus classification and detection from digitized kidney slide segments using a deep learning framework. The proposed method applies Convolutional Neural Networks (CNNs) between two classes: Glomerulus and Non-Glomerulus, to detect the image segments belonging to Glomerulus regions. We configure the CNN with the public pre-trained AlexNet model and adapt it to our system by learning from Glomerulus and Non-Glomerulus regions extracted from training slides. Once the model is trained, labeling is performed by applying the CNN classification to the image blocks under analysis. The results of the method indicate that this technique is suitable for correct Glomerulus detection in Whole Slide Images (WSI), showing robustness while reducing false positive and false negative detections.

Keywords: Glomerulus classification; Glomerulus detection; digital pathology; Convolutional Neural Networks

1. Introduction

The whole-slide scanner allows the creation of high-resolution images from tissue slides. This technology has created the possibility to apply image processing and machine learning techniques to these Virtual Slides (VS) to automate and/or enhance traditional pathology processes in what is known as Digital Pathology. This current trend is becoming a necessary tool for pathologists, and it is used to perform a fast and robust diagnosis from high-resolution images or, in some cases, to completely automate quantitative evaluations. Bueno et al. give a general survey of digital Pathology trends in [1].

Deep Learning techniques based on Neural Networks represent an important area in the Digital Pathology field. These machine learning methods, such as Convolutional Neural Networks (CNN), are becoming central in VS analysis since they give promising results in difficult tasks like region detection and classification [2]. CNN allows a relatively simple configuration thanks to its unsupervised feature generation, the availability of larger datasets to train the models, and the possibility to import

pre-trained models to partially set-up the neural networks. The major drawback of these techniques is the huge amount of computational power necessary to correctly train the networks in a reasonable time. Madabushi et al. [3] show machine learning methods applied to different Digital Pathology challenges.

In this paper, we will focus on Glomerulus classification based on deep learning in the context of digital pathology applied to kidney tissue slides. Glomerulus classification is central in nephropathology field to get a correct medical diagnosis. These networks of capillaries serve as the first stage in the filtering process of the blood that retains higher molecular weight proteins in blood circulation in the formation of urine. Therefore, any pathological change associated to them inside a tissue section, like the number of Glomeruli or their sizes, is essential to detect diseases from the patients.

Glomeruli structures visualized in a VS present high variability in terms of size, shape and color. The reasons of this high variability in samples are multiple: the relative position of the Glomerulus inside the renal section, the heterogeneity in immunohistochemistry staining, or the presence of internal biological processes.

- Size and shape: In a healthy kidney before sectioning, Glomeruli present a spherical shape with fixed size (diameter ranges between 350 and 100 μm), but its aspect can change due to the presence of medical diseases. For instance: Glomeruli can present a swell aspect under hypertension [4] or diabetes [5] conditions. After sectioning, the presence of pathologies affect the appearance inside the VS. Besides, the different Glomeruli sizes observed could vary depending on where the cross-section was taken with respect to each Glomerulus sphere.
- Color: In our configuration, we use PAS (Periodic Acid Schiff) stain in tissue sections, which gives a purple-magenta color to the slides. The amount of stain present in each slide will determine the color intensity of the segments under analysis. Since this process is not perfect, each slide can present different intensities. Moreover, the presence of medical diseases can vary the amount of stain present in the Glomeruli under study.

All this variability related to the Glomerulus appearance, and in general to VS images, makes it difficult to automate the classification using a digital pathology framework. Therefore, any proposal devoted to this task must be robust to these aspects.

Previous Work

Many authors have been working in digital pathology developing techniques to automate quantitative analysis tasks. Methods devoted to classify and detect Glomeruli in the nephropathology field have become of main importance in diseases diagnosis and research. The classification methods of the literature can be divided into two groups: handcrafted features approaches and unsupervised features approaches.

Handcrafted features: These classification methods try to find specific measurable attributes in the image that can be utilized to correctly identify the structures. In [6], the authors propose to use a genetic algorithm for edge patching by using a Canny edge detector to get discontinuous edges of Glomerulus and its final detection and classification. The main problem of this technique is that borders detected by means of Canny edge detector are not robust and is prone to errors under color intensity variations. In [7,8], the authors use Rectangular-Histogram of Gradients (R-HOG [9]) as feature vector to classify Glomeruli. The main problem of these techniques is the fact that R-HOG has rigid block division that results into considerable number of false positives. In order to improve the R-HOG framework, Kato et al. [10] suggest the Segmental HOG (S-HOG) as potential candidate descriptor for Glomerulus detection, where the block division is not rigid, and uses nine discretized oriented gradients and Support Vector Machine (SVM [11]) as a supervised learning classifier. In [12], an analysis of renal microscopic images is applied by using a detection of borders improved with the Convex Hull algorithm ([13]). This method detects and classifies Glomeruli for the measurement of both diameter and Bowman's Space thickness. In [14], the authors propose an approach for Glomeruli detection

based on the ellipse shape of Glomeruli to obtain candidate regions. These regions will be classified in a final step with a trained SVM classifier. The method uses a color normalization step to deal with variations that appear in the stain process. Brandon Ginley et al. [15] present a method to identify glomerular boundaries based on Gabor filtering, Gaussian blurring, statistical F-testing and distance transform. In [16], the authors propose a computer-assisted topological analysis to detect tubules and interstitium, glomeruli, and peritubular capillaries in kidney samples. The method is based on double immunostaining of both capillaries and an immune cell population, so that the decomposition of colors allows identification of the nuclei of all cells present in the biopsy specimen, of the cytoplasm of stained immune cells, of capillary walls, and the lumens of capillaries and Glomeruli. Coupling decomposed colors allow precise identification of the regions of interest, that is, the stained immune cells, capillaries, and Glomeruli. Simon et al. [17], designed a method for the segmentation of glomeruli using an adaptation of the local binary patterns (LBP) image feature vector to train a support vector machine (SVM) model. Finally, as an example of Glomeruli classification in 3D dataset, Zhang et al. [18] propose a Glomerulus detection and classification using 3D Magnetic Resonance Imaging (MRI) by means of Hessian based Difference of Gaussians (HDoG) detector.

Unsupervised features: These techniques avoid the necessity to define handcrafted features to perform the classification. Deep Learning methods such as CNNs belong to this group. They learn the best features for the problem, so that the scientists are only responsible of providing large representative datasets, as well as choosing the network architecture. Kothari et al. [19] perform a thorough review of how the application of this kind of techniques is currently the most suitable one, pointing at the challenges and future opportunities that are involved.

Regarding the application of these techniques to specific problems, some of the most interesting works that have been published focus on different medical areas: Wang et al. [20] use a cascaded approach for mitosis detection in breast cancer that combines a CNN model and handcrafted features (morphology, color and texture features). In [21], the authors deal with mitotic figure classification combining manually designed nuclear features with learned features extracted by CNN. Xu et al. [22] explore the use of very deep CNNs to automatically classify diabetic retinopathy by configuring a network with 18 layers. Havaei et al. [23] apply CNN to automatic brain tumor segmentation in MRI images that exploits both local features as well as more global contextual features simultaneously, in a CNN cascade configuration.

Table 1 shows a summary of the different experiments performed in the publications mentioned above, specifying the kind of problem, network architecture (and any particularity if necessary to be mentioned), the dataset used and the obtained results. The revision of every publication stated before is out of the scope of this article and therefore we focus on those works related to CNN.

Some recent publications focus on unsupervised Glomerulus detection by means of Deep Learning. In [24], the authors evaluate the detection of glomerular structures in WSIs stained with multiple histochemical staining using HOG feature classifier and CNN with a specific model. They compare the performance of both methods for each stain, and the combination of the detection in consecutive sections stained with different histochemical staining. Their results showed that HOG detector perform between 10% and 20% worse than CNN detector. Besides, their method using CNN detector with PAS samples performed on average 66.38% F1-score. Combining consecutive sections, after WSI registration, the final score reaches 78%. The main drawback of this proposal is the low F1-score using only the PAS WSI and the complexity of combining consecutive sections. The training method, which does not take into account the differences between stains, and the CNN analysis carried out to accept a region as a Glomerulus are partially the cause of these inferior results. More recently, Gadermayr et al. [25] deal with Glomerulus detection and segmentation using CNN cascades where first, a network is used to detect regions of interest, and next, a U-Net network is applied to achieve the Glomerulus segmentation. The resulting detection rate is 91%, which is higher than previous proposals, but no color normalization is applied to the data, which led us to consider miss-detection under different conditions like PAS samples belonging to different laboratories.

Table 1. Previous work experiments and results.

Year [Reference]	Application	Dataset	Features	Classification	Performance
2013 [21]	Mitotic figures	5 WSIs with 35 ROIs and 226 samples	Handcrafted (Colour, texture, shape) + CNN (LeNet based)	SVM	0.659 (F-score)
2014 [20]	Mitotic figures	35 HPFs	Handcrafted (Morphology, Intensity, Texture) + Custom CNN (2CV + FC)	Ensemble	0.735 (F-score)
2016 [2]	Nuclei segmentation	141 ROIs with 12,000 samples	AlexNet based CNN	Softmax	0.83 (F-score)
	Epithelium segmentation	42 ROIs with 1,735 samples	AlexNet based CNN	Softmax	0.83 (F-score)
	Tubule segmentation	85 ROIs with 795 samples	AlexNet based CNN	Softmax	0.84 (F-score)
	Lymphocyte detection	100 ROIs with 3064 samples	AlexNet based CNN	Softmax	0.9 (F-score)
	Mitosis detection	311 ROIs with 550 samples	AlexNet based CNN	Softmax	0.53 (F-score)
	Invasive ductal carcinoma	162 WSIs	AlexNet based CNN	Softmax	0.765 (F-score)
	Lymphoma classification	374 samples	AlexNet based CNN	Softmax	97.0% (Acc)
2017 [23]	Brain tumour segmentation	BRATS2013	Custom CNN (4CV)	Softmax	0.84 (Dice)
Proposed	Glomeruli classification	10,600 ROIs from 40 WSIs	CNN-AlexNet (pre-trained)	Softmax	**0.999** (F-score)
Proposed	Glomeruli classification	10,600 ROIs from 40 WSIs	CNN-AlexNet (from-scratch)	Softmax	0.992 (F-score)
Proposed	Glomeruli classification	10,600 ROIs from 40 WSIs	CNN-GoogleNet (pre-trained)	Softmax	**0.999** (F-score)
Proposed	Glomeruli classification	10,600 ROIs from 40 WSIs	CNN-GoogleNet (from-scratch)	Softmax	0.994 (F-score)

In this paper we propose two methods devoted to: Glomerulus classification and Glomerulus detection. In the Glomerulus classification proposal (Method I), we study the viability of the CNN to achieve a correct Glomerulus classification into Glomerulus/non-Glomerulus classes. In the Glomerulus detection method (Method II), we use the CNN configuration that gave the best results and performance for Glomerulus classification (Method I) to achieve the Glomeruli detection in a WSI. In this proposal, we present a method suitable for PAS samples belonging to different laboratories, thanks to color normalization process used in the data augmentation step. Our method uses a sliding window with overlapping to perform the detection, and utilizes a majority vote decision for each pixel to reduce false positive detections. With this proposal, we achieve a detection score of 93.7%, which improves upon the results obtained in reference methods.

The remainder of the paper is as follows. The Materials and Methods are explained in Section 2, where Glomerulus classification and Glomerulus detection methods are described in detail. Results for Glomerulus classification and Glomerulus detection are shown in Section 3. Finally, Section 4 is devoted to the Discussion.

2. Materials and Methods

This section explains how a Glomerulus dataset have been built from WSI and the use of Convolutional Neural Networks to develop two methods able to:

- Classify patches in two classes: Glomerulus and non-Glomerulus.
- Detect Glomerulus in WSI.

We focus in specifically designed deep models to be used with images.

2.1. Dataset

The samples used to obtain the collection were stained by PAS (Periodic Acid Schiff) staining, which is mainly used for staining structures containing a high proportion of carbohydrate macromolecules (glycogen, glycoprotein, proteoglycans), typically found in connective tissues, mucus, the glycocalyx, and basal laminae.

Many laboratories use a battery of stains for evaluating renal disease, e.g., H&E, PAS, trichrome, and silver stains. The silver stain accentuates collagenous structures, e.g., in the Glomerulus, the mesangial matrix and the glomerular basement membrane. The PAS stain also accentuates matrix and basement membrane constituents, as does the blue or green component on the trichrome stain. In certain circumstances the trichrome stain demonstrates immune deposits as fuchsinophilic (red) structures.

Regarding the utility, the PAS stain is used to identify pathological disorders in which weakness or improper functioning of basal membranes are present. Kidney glomerular diseases are an example of this kind of pathologies [26].

This study involved digital slides from the AIDPATH kidney database [27]. It is composed of kidney tissue cohorts from four different institutions and pathology labs around Europe, including Spain, Germany, Italy and Lithuania. The dataset is made of samples extracted by expert pathologists from incisional biopsies, with the objective to detect potential renal pathologies. It comprises a collection of 108 PAS WSIs. The slides were prepared serially, using Formalin-Fixed Paraffin-Embedded (FFPE) tissue sections with 4 μm of thickness. They were obtained using the systematic pathological process defined by the highest quality standards [28].

2.1.1. Slide Digitation

Tissue samples from biopsies were digitized with a scanner Aperio ScanScope XT (Leica Biosystems GmbH, Nussloch, Germany) at 20× and 40× with no compression, into *SVS* format (*SVS* is a semi-proprietary file format consisting on a single-file pyramidal tiled *TIFF*, which can be

opened with Aperio ImageScope software (Leica Biosystems GmbH, Nussloch, Germany. The version used in this research is: 12.1.0.5029), and by ImageJ or Fiji via the Bio-Formats plugin).

To reduce the time and effort required to manipulate the digital slides, create the ground truth annotations and the computational processing, we worked at 20× magnification. Therefore, the WSI acquired at 20× magnification have been analyzed in full resolution, while slides acquired at 40× magnification have been down-sampled by a factor of two. Depending on the size of tissue fixed on the slide, digital slides range between 20 k × 10 k pixels to 70 k × 50 k pixels of resolution with sizes from 0.5 GB to 2 GB.

2.1.2. Ground Truth Annotation

The process of labeling was performed manually using the Aperio ImageScope tool, as shown in Figure 1. As a result, the coordinates of the different Glomerulus detected by the pathologists were available, so that the Glomerulus patches could be cropped from the images.
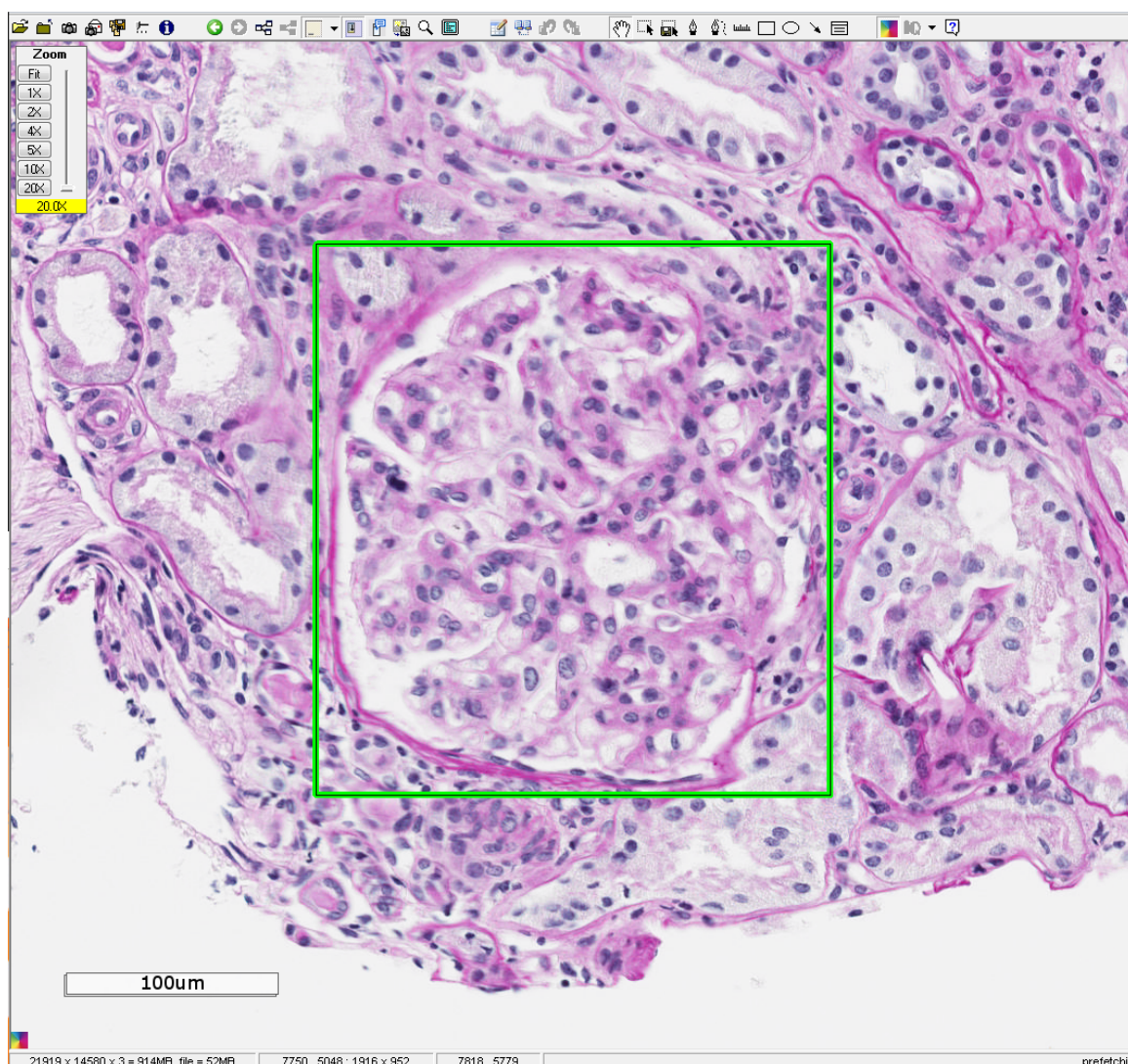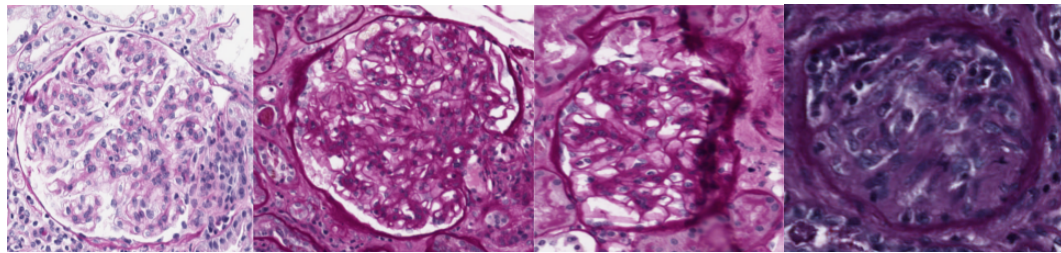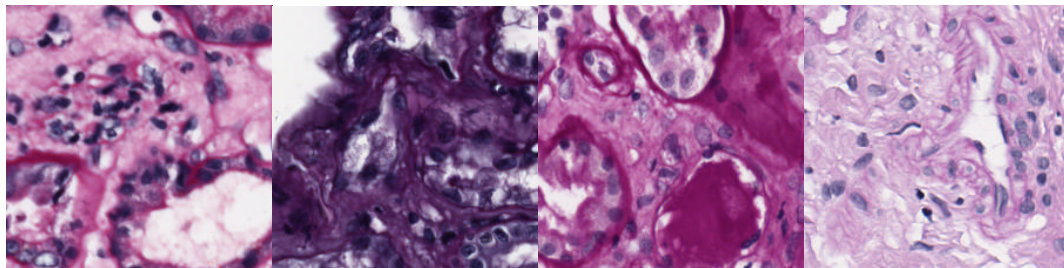

Figure 1. Glomerulus example labeled using Aperio ImageScope tool.

Following this process, 4300 Glomeruli structures were obtained. As shown in Figure 2a, these samples present high variability in terms of shape, size and texture.

(**a**) Glomerulus examples

(**b**) No Glomerulus examples

Figure 2. Examples of image patches from dataset.

2.2. Method I: Glomerulus/Non-Glomerulus Classification

The approach followed in the proposed method is based on Convolutional Neural Networks used to classify patches in Glomerulus/Non-Glomerulus classes. The objective is to be able to discern whether a patch from a WSI could be classified as a Glomerulus structure or not. Hence, the network needs to be trained with Glomeruli and Non-Glomeruli annotated regions.

2.2.1. Materials

In order to reduce the computational time in the CNN training process, we use a representative subset of the overall database to correctly check the CNN capability to classify selected patches in Glomerulus and Non-Glomerulus classes. The subset used is composed of 40 WSIs used to extract 700 Glomeruli annotations and 700 non-Glomeruli to train the CNN. Extracted Glomeruli diameters range between 350 μm and 100 μm, which is translated to blocks with sides in the range of 400 pixels to 200 pixels. Non-Glomeruli regions are extracted randomly, outside Glomeruli regions, with similar block sides. Since the 700 Glomeruli samples may seem barely enough to be fed into a Convolutional Neural Network, a data augmentation technique was agreed upon to be applied. A combination of rotations in 0°, 90°, 180° and 270° and vertical flip were performed. The final number of samples per class (Glomerulus and non-Glomerulus) was established in 5300. In order to fed the CNN model, the samples are resized to the required size of 227 × 227 pixels. Figure 2 shows some examples of patches from both classes.

2.2.2. CNN Training

Two different approaches have been considered. First, a technique called fine-tuning has been used. That is, taking a previously trained model that has proven reliable (with good general features) and apply that to our own dataset, modifying the weights to be adapted to this problem. The model used was a pre-trained AlexNet model from the ImageNet challenge [29]. The architecture of this network is composed of eight layers: five convolutional and three fully-connected. Given that, it was possible to achieve good results with only a few epochs/iterations. After that, the same network was employed without any previous training, what is called to train a network "from-scratch". In order

to increase the variability of architectures the GoogleNet ([30]) architecture was employed also with both techniques.

The training process was carried out using two different sets of parameters, depending on whether the network was pre-trained or not. For the pre-trained network the learning rate is started with an initial of 0.001, decreasing it with drop factor of 0.1 with a period of eight. For back propagation, the Stochastic Gradient Descent was used, establishing 0.004 for L2-Regularization. With the previous parameters, training reaches the best results with just 10 epochs (after that number the loss value and accuracy do not improve). It takes around 9 min to perform the training process for AlexNet and 25 min for GoogleNet (i.e., time for training one-fold in cross-validation), using the GPU NVIDIA GTX 960 Ti with 6 GB of VRAM. On the other hand, training from-scratch requires slightly different parameters to obtain the best convergence in terms of accuracy and loss values. For these experiments, and initial value of 0.01 was employed for the learning rate decreasing it with drop factor of 0.1 with a period of 16. For back propagation, the Stochastic Gradient Descent was used, establishing 0.004 for L2-Regularization. Due to a higher number of epochs, it takes 17 min for AlexNet and 33 min for GoogleNet to perform the training process with this technique and parameters, using the same hardware as described above.

2.2.3. Validation

To perform the experiments, 10-fold cross-validation has been applied, with nine folds for training and the remaining one in each iteration for testing. Hence, joining the results from test folds, the whole dataset is validated.

In order to measure the performance, the *F*-score (also known as *F*-measure) has been chosen. This is one of the most commonly used metrics to obtain the performance of a binary classification problem, such as Glomeruli, nuclei, or tumor detection, in which, for a given sample, is determined whether it belongs to the specified class or not. T. Fawcett performs a comprehensive review of this topic in [31]. Here, the most important concepts are summarized.

Equation (1) states the general expression that rules the *F*-score, in terms of True positive (TP), True negative (TN), *Type 1* error (False positive, FP) and *Type 2* error (False negative, FN), of a given confusion matrix:

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{TP}}{(1 + \beta^2) \cdot \text{TP} + \beta^2 \cdot \text{FN} + \text{FP}} \tag{1}$$

This expression can be stated in terms of Precision (*P*, defined in Equation (2)) and Recall (*R*, defined in Equation (3)), as shown in Equation (4)

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R} \tag{4}$$

The *F*-score can be understood as a weighted measure of false positives and negatives, varying the balance between them with the β value, so that it is given β times more importance to recall than precision. Some common values to β are 0.5, 1 and 2. For this problem the selected value is 1, so the statement employed to calculate the measure in our experiments is the one shown in Equation (5).

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{5}$$

2.3. Method II: Glomerulus Detection in WSI

As said in previous sections, Glomerulus can present different sizes, shapes and textures. We have seen in Method I that this variability can be learnt in order to achieve a correct classification of complete Glomerulus regions. Glomerulus detection supposes a more complicated challenge, since the mentioned variability causes the necessity to deal with block size issue and high confusion with other similar structures.

The proposed method is based on the use of Convolutional Neural Networks classification to detect the Glomerulus regions present in the WSIs. It allows to automate time consuming tasks like Glomeruli counting or Glomeruli distribution analysis inside a WSI. In this proposal, we use CNN AlexNet model to classify the image blocks. In order to deal with Glomeruli variation, we use a fixed block size of 227×227 pixels, which is the block size used by the AlexNet model, to train and classify the tissue regions of the WSI. The training and detection processes are explained in detail in the following sections.

2.3.1. Materials

In order to allow the Glomeruli detection from a full WSI, more Glomerulus samples, than the ones used in Method I, are necessary to train the model. We use the complete data set composed of 108 WSI, thus using the 4300 annotated Glomerulus. A total of 98 WSI, with around 4000 Glomeruli, are used for training. The remaining 10 WSI are used for testing with 275 Glomeruli.

2.3.2. CNN Training

Since we are going to classify image blocks of 227×227 pixels, we need to adapt the Glomeruli data to this specific size. Hence, we split the 4000 Glomeruli, extracted at $20\times$ magnification, into blocks of 227×227 pixels. With this splitting process, we achieve 5000 Glomerulus blocks. Note that we split the Glomerulus in this block size and we reject remainder regions with smaller size. Figure 3a shows a graphical representation of this process, where Glomerulus regions smaller than this size are not used in the training process.

Regarding non-Glomerulus class, we extract blocks of 227×227 pixels from tissue regions not belonging to Glomerulus annotated areas. Due to the high variability that these tissue areas present, we decide to strengthen the learning of this class by using a higher number of blocks than Glomerulus class. In our implementation, we use 12,300 non-Glomerulus blocks. Figure 3b displays a graphical representation of this block extraction.
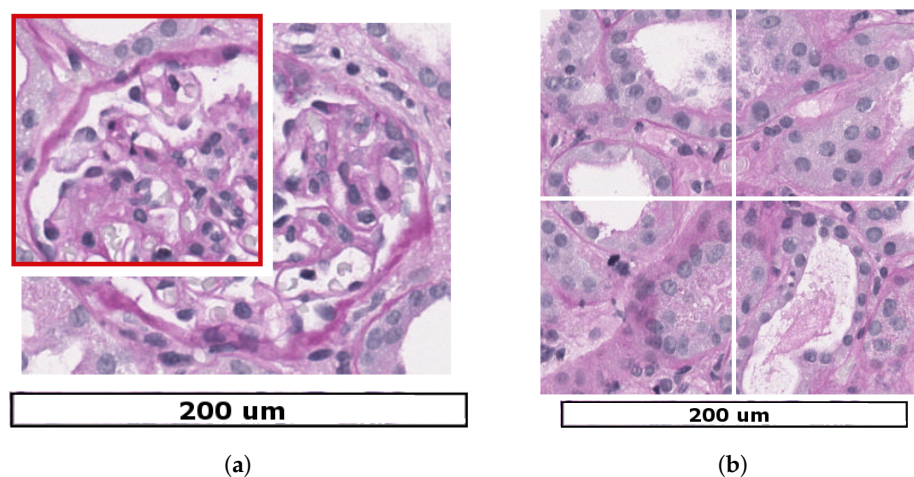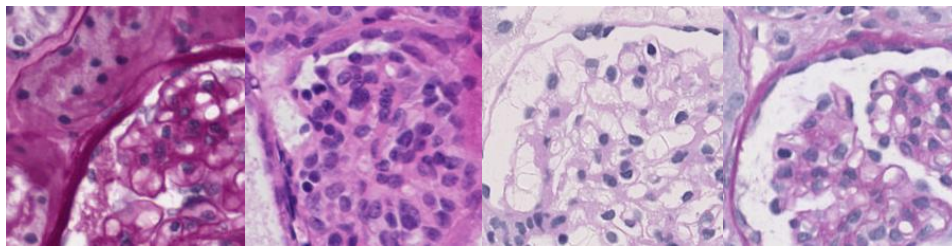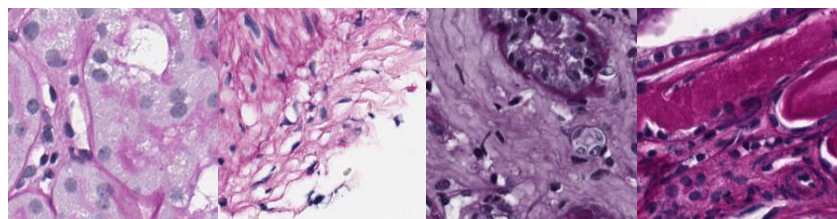

Figure 3. Graphical representation of blocks generation for Glomerulus and non-Glomerulus classes with 227×227 pixels size. (**a**) Representation of Glomerulus block extraction; (**b**) Representation of non-Glomerulus blocks.

Figure 4 shows an example of Glomerulus and non-Glomerulus blocks used to train the model.


(**a**) Example of Glomerulus blocks.

(**b**) Example of non-Glomerulus blocks.

Figure 4. Example of 227×227 pixels training blocks for Glomerulus and non-Glomerulus classes.

Data augmentation techniques are also applied to this method. A combination of rotations in $0°$, $90°$, $180°$ and $270°$ and vertical flip were performed. Besides, color modification has been performed, in order to enhance the CNN robustness in front of the staining color variability present between laboratories. We apply this color modification by means of Reinhard's method [32], which has proved to work accurately in previous publications devoted to digital pathology [33,34]. We select 10 distinct color patterns, to use as reference samples, corresponding to the most representative staining variations observed in the dataset.

Figure 5a depicts the selected blocks used as reference in our implementation. The number of reference blocks should be defined according to the color variability of our database. The more variability the higher the color modifications. Hence, using color normalization the method is invariant to differences in PAS staining related to different intensities and tonalities. Figure 5b shows the color transformations generated in a block through Reinhard's method. From the observation, color normalization allows for adapting blocks to distinct stain tonalities. After color modification, the final number of blocks for Glomerulus class is 150,000, meanwhile the number of blocks for non-Glomerulus class is 370,000. Therefore, we decide to use unbalanced training set for Glomerulus and Non-Glomerulus classes. In detection, this configuration is appropriate since non-Glomerulus regions comprise several structures that present high similarity to Glomerulus blocks. Figure 6 shows an example of three of the main regions present in the background: Tubuli, Interstitium and blood vessels. Therefore, due to this variability we need more data to allow the network for learning correctly the characteristics of each non-Glomerulus structure.

According to the results obtained in Method I (Section 3.1, Table 3), pre-trained AlexNet and GoogleNet models give the best results in terms of classification. Since AlexNet model performs better in terms of computational cost, this model is used in this proposal in order to enhance results and speed up the training process of the model. We use fine-tuning technique in order to adapt the model to our training data with only few epochs/iterations. The set of parameters used, as well as the hardware and GPU configuration is the same as defined in Method I for the pre-trained Model (Section 2.2.2).
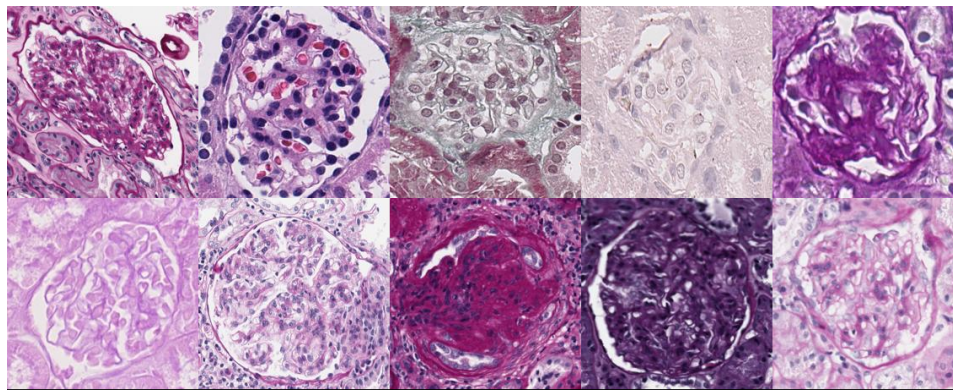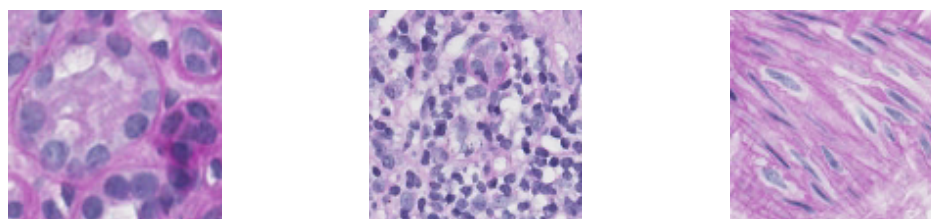
(**a**) The 10 blocks used as reference colors in color normalization process.
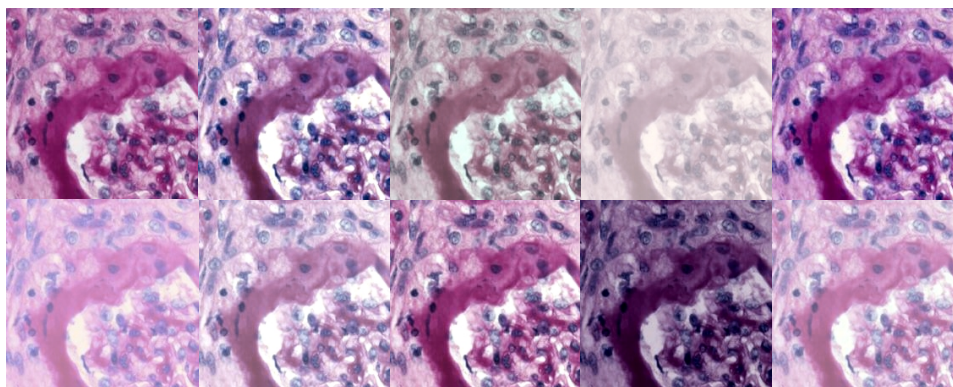
(**b**) Effect of the color normalization process in one block by using the 10 reference blocks.

Figure 5. Color normalization by means of Reinhard's method.

(**a**) Tubuli.　　　　　　　　　(**b**) Interstitium.　　　　　　　　(**c**) Blood vessel.

Figure 6. Example of non-Glomerulus regions belonging to tubuli, interstitium and blood vessels structures.

2.3.3. Test Validation

We test the proposed Glomeruli detection by analyzing the 10 WSI devoted for testing. Firstly, the WSI under analysis is processed to segment tissue regions. Background color thresholding is applied to perform this segmentation. Only in these tissue regions, we apply a sliding window of 227×227 pixels with 80% overlapping in both, horizontal and vertical axis. This overlapping is necessary in order to get redundancy in the block detections. Thus, for each one of the blocks, CNN classifier is applied to obtain the block classification in Glomerulus/non-Glomerulus detection.

Once all the WSI blocks are classified, we achieve the Glomeruli detection by performing a pixel-wise analysis. Hence, we classify each pixel as Glomerulus if the pixel belongs, at least, to nine blocks classified as Glomerulus by the CNN, out of 10 different blocks that overlap each pixel. Thanks to this analysis, we reduce the false positive detections that appear in the CNN classification, adding robustness to the detection process. Figure 7 is an example of the overlapping map computed in a WSI region. We can observe how overlapping value below nine (blue, violet and green color),

increases the false positive detections outside the Glomerulus region. Besides, all decisions below nine increases the over detection in the boundaries of the Glomerulus. Decisions equal to nine achieve the detection of the Glomerulus region while reducing over detection in the boundaries of the Glomerulus. None of these values present False Negative detections. Detections equal to 10 give similar values than redundancy of nine in terms of False positive detections but present some False Negative detections, which make the decrease of the performance. Table 2 shows Precision and Recall metrics (Equations (2) and (3)) of the detection process obtained in our database after evaluating with different redundancy values. As we can see, redundancy of nine gives the best False positive ratio while keeping 0 false negative detections.
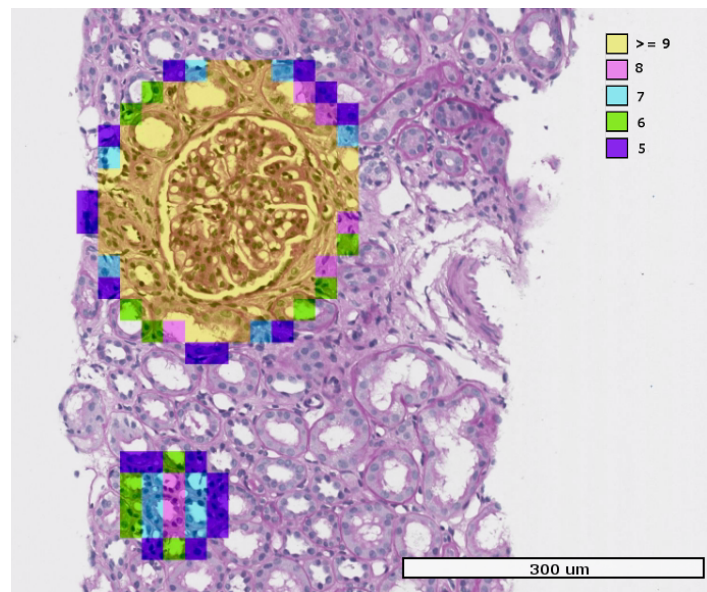

Figure 7. Redundancy map obtained in pixel classification. Displayed from overlapping 5 to overlapping ≥9.

Table 2. Precision and Recall in Glomerulus detection according to Redundancy decision value.

Redundancy Value	Precision	Recall
10	0.881	0.989
9	**0.881**	**1**
8	0.804	1
7	0.781	1
6	0.759	1
5	0.739	1

Glomerulus regions in the WSI will be detected by segmenting all the pixels that fulfill this condition.

As commented before, the Glomerulus is a 3D structure with spherical shape. When the slide cuts the Glomerulus close to the sphere poles, it presents a small shape in the WSI. Despite the proposal is also robust detecting these regions, we follow the same criterion as the pathologists: ignore them as non-informative. Therefore, we can apply a final filtering process according to the pathologists' size restriction, where only the Glomeruli with a minimum size are considered for their evaluation, i.e., >100 μm of diameter. Hence, we filter detected regions that present smaller area than the threshold established by medical doctors: 200×200 pixels, thus removing the spurious detections from our classification without losing true positive detections.

3. Results

3.1. Method I, Glomerulus Classification Results

In this research, we have studied four different configurations for the CNN model: AlexNet model pre-trained, AlexNet model from-scratch, GoogleNet model from-scratch and GoogleNet model pre-trained. Once the training process has been performed for the selected models and techniques, the test folds are used to build the confusion matrix. The detailed results of every experiment that has been carried out are shown in Table 3.

Table 3. Experiment results.

Architecture	Technique	*F*-Score
AlexNet	pre-trained	0.999
AlexNet	from-scratch	0.992
GoogleNet	pre-trained	0.999
GoogleNet	from-scratch	0.994

As it is stated, the pre-trained models are more reliable than from-scratch. Apart from the evinced performance, fine-tuning models have some advantages also in terms of computational requirements and time spent for training, as stated in Section 2.2.2. In consequence, the pretrained AlexNet model has been chosen as the main architecture for the rest of experiments also due to less layer complexity, which is translated in faster execution time. For this experiment, the confusion matrix is shown in Table 4.

**Table 4.** Testing confusion matrix for pre-trained AlexNet.

	Glomerulus	Non-Glomerulus
Glomerulus	5295	5
Non Glomerulus	0	5300

On average, the accuracy for this approach in the 10-fold cross-validation is 99.95%, with a standard deviation of 0.06.

As the performance of the pre-trained CNNs is nearly 100%, it becomes interesting to examine some patches that have been wrongly classified, in order to extract useful information about the features of those samples. The specific five Glomerulus samples that have been detected as non-Glomerulus by pre-trained AlexNet are shown in Figure 8a. The GoogleNet pre-trained model also misclassify five patches, which are shown in Figure 8b.

As it is observed for both models, four out of five samples belong to the same image, since these examples are related with the data augmentation that was performed to the dataset. This is interesting in terms of how the data augmentation can introduce orientation invariance to the model, as it is observed, whether a sample is learned as Glomerulus (or non-Glomerulus in this case), this behavior remains the same for every variation of the same image. According to the pathologists that have examined the wrongly classified Glomeruli, the Glomeruli are rather deformed and not sufficient for human evaluation. A pathologist would rather chose to ignore them as non-informative if there are other Glomeruli with correct appearance to examine.

Also, it is observed that the pre-trained approach slightly improves the performance of the model (and the computational time employed), but it is stated that from-scratch deep CNN models are also suitable to be used in this problem with a reduced dataset as the one that was employed.
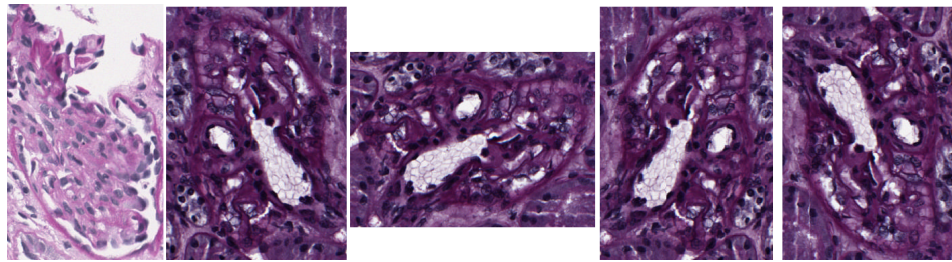
(**a**) Patches that have been classified as non-Glomerulus by pre-trained AlexNet
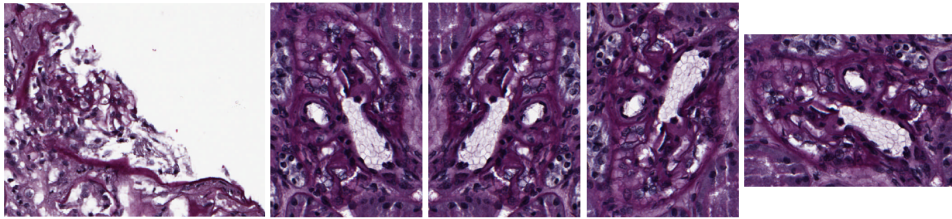
(**b**) Patches that have been classified as non-Glomerulus by pre-trained GoogleNet

Figure 8. Glomeruli that have been classified as non-Glomerulus by pre-trained AlexNet and GoogleNet models.

3.2. Method II, Glomerulus Detection Results

Once we have demonstrated that CNN performs correctly when classifying patches between Glomerulus and non-Glomerulus classes, we make now a complete evaluation of the proposed study for automatic WSI Glomerulus detection. Qualitative and quantitative evaluation have been carried out by analyzing 10 WSI devoted for testing.

In terms of qualitative results, Figure 9 shows the results obtained in two different WSI, created in different laboratories. As we can observe, the proposed method achieves the detection of all the Glomeruli present in the displayed tissue regions, even detecting the broken Glomerulus that appears in the border of the tissue in Figure 9a. This is thanks to the color normalization applied during data augmentation process. The technique is also robust when the staining process present so much variability as we can see in both samples.

Figure 10 shows some examples of the detections that we have obtained. We can observe three examples of correctly detected Glomeruli (Figure 10a), and one example of a false positive detection (Figure 10b). We can see in Figure 10a that although there is a high variability between samples, the Glomeruli are correctly detected. On the other hand, Figure 10b shows how the similarity of some non-Glomeruli regions can lead to false positive detections under determined circumstances.

Besides, we can observe in the results shown in Figure 10, the block effect that appears in the detected Glomerulus regions. This block effect appears because of the overlapping between sliding windows, which we established at 80% of the window size (227×227 pixels) in horizontal and vertical axis. Therefore, the tissue area is divided in regions of 45×45 pixels (remainder 20%) that share the same 10 windows. Since we are using redundancy where only pixels belonging to regions with nine or higher out of 10 positive classifications are classified as Glomerulus region, we obtain this block effect of 45×45 pixels between:

- Glomerulus regions (nine or ten out of ten positive detections)
- Non-Glomerulus regions (eight or less out of ten positive detections).

Hence, regions with different number of positive classifications will appear in blocks of 45×45 pixels, resulting this "Block effect" in the resultant detection.

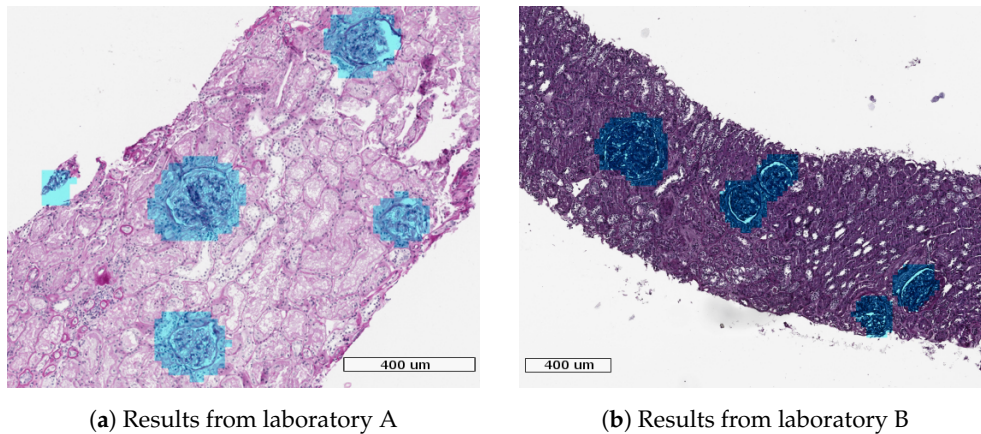(**a**) Results from laboratory A　　　　　　　　(**b**) Results from laboratory B

Figure 9. Results obtained in two WSI (Whole Slide Images) created in different laboratories. Regions detected as Glomerulus appear in blue color.

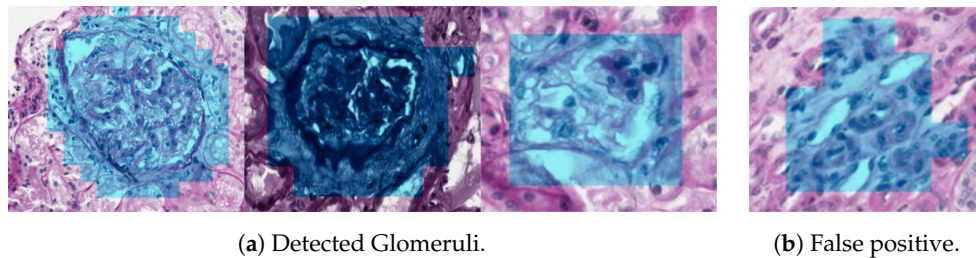(**a**) Detected Glomeruli.　　　　　　　　　　　　(**b**) False positive.

Figure 10. Example of correct and false positive detections. Regions detected as Glomerulus appear in blue color.

Regarding the quantitative results, analogously to Method I, we have computed Precision, Recall and F_1 metrics at object-wise level, with the number of Glomeruli detected in the 10 WSI under evaluation. These 10 WSI contain 275 Glomeruli. Our method has detected all the 275 Glomeruli correctly, with 37 False Positive detections, and zero False Negative detections. Table 5 shows the numerical values obtained. As we can see, our method presents a precision value of 0.881 and a recall value of 1, which supports the validity of our proposal. We can also observe that working in the reduction of the false positive detections will be necessary to enhance the robustness of this method.

Table 5. Method II: Glomerulus detection. Quantitative evaluation.

Metric	Value
Precision	0.881
Recall	1
F_1	0.937

In order to highlight the benefits of the color normalization, we have computed the results without using this modification in the training process. As expected, working with samples from different laboratories and thus, distinct stains, results in a decrease of the detection performance. With this implementation we have obtained $F_1 = 0.885$, thus 5% lower than the results obtained with color normalization. The different result obtained between both image augmentation processes will depend on the number of samples used for training and the color variability that these samples present. The higher the number of samples and the color variability, the better the model will adapt to color variations.

4. Discussion

Rather than customized architectures with few layers, the work that has been presented has relied on deep models specifically designed to be used with images, applying different techniques to determine whether a pre-trained model can significantly improve the results. Therefore, models that have been designed for large-scale image classification problems can also be used with smaller datasets to specific tasks, such as Glomeruli classification, in this case.

For this kind of models, it is suitable to state the performance of training by representing the learned features. One the most typical method is the visualization of the filters from the first convolutional layers.

For pre-trained models, fine-tuning training hyperparameters prevents from changing the first layers, since the process is focused on the last ones. For this reason, the features are basically the same that the models had learned in the original training process with the ImageNet dataset, as shown in Figure 11.
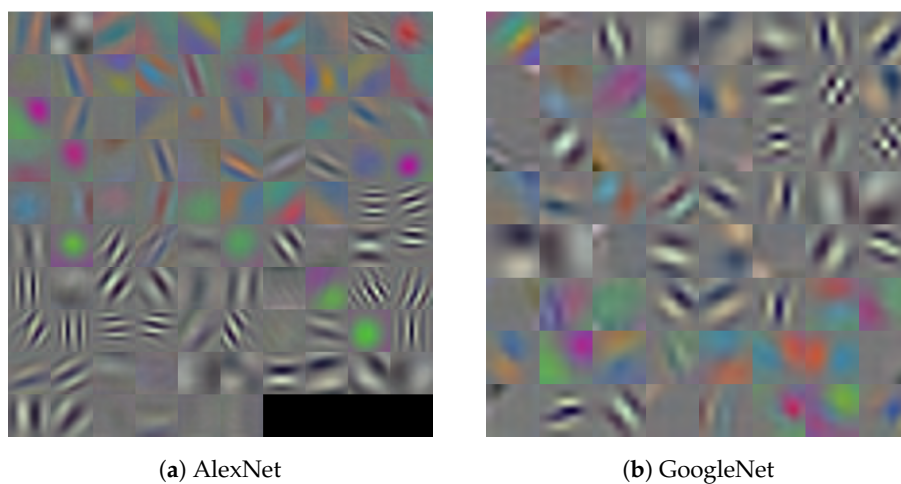

(**a**) AlexNet (**b**) GoogleNet

Figure 11. Filters learned by pre-trained models in the first convolutional layer.

On the other hand, the models learned from-scratch have not achieved features suitable enough on these layers, with some noise on their filters. This is due to lack of images to train such deep models with the dataset that was available. These filters are shown in Figure 12.
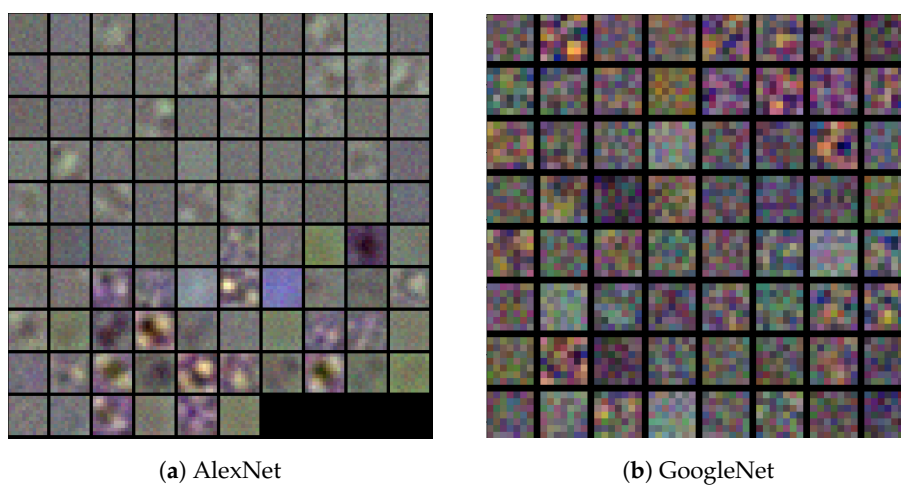

(**a**) AlexNet (**b**) GoogleNet

Figure 12. Filters learned by from-scratch models in the first convolutional layer.

Besides, one the most outstanding methods for CNNs interpretation has been used to get an idea of the concept of glomerulus and non-glomerulus that these models have. This is the Deep Dream image generation, which is able to build artificial images that performs maximum activation on a given channel. When it is applied to a classification layer, it is possible to observe the ideal representation of the different classes that the model has been trained to discern. Figure 13 shows the results obtained in the pretrained AlexNet model for Glomerulus and Non-Glomerlus class. It is stated that the model has learned the typical structure of the glomeruli, their outer shape and internal components. However, the Non-Glomerulus representation shows an undefined structure with dots related to typical individual cells.
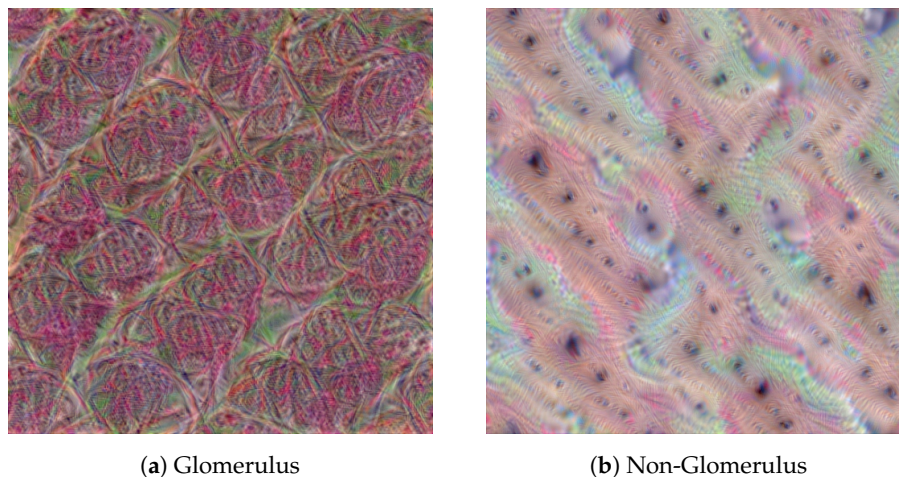

(**a**) Glomerulus (**b**) Non-Glomerulus

Figure 13. DeepDream concept class visualization for pre-trained AlexNet.

As the visualization techniques state, the combination of good initial features and proper fine-tuning of the last classification layers seems to be the most accurate method to train the best model for the data that was available, and, in consequence, pre-trained models perform better in this study.

We have demonstrated that CNN can be used for Glomeruli classification and its detection from full WSI, by using a specific framework based on overlapped sliding window. We have seen in the results, that AlexNet pre-trained model is the configuration that best suits the classification problem, and how its application to Glomeruli detection issue gives correct and promising results, which can be used in hospitals and laboratories to assist the pathologists in their daily tasks.

Regarding future work, the objective is to reduce false positive ratio in Glomeruli detection, in order to obtain even more robustness in the WSI Glomeruli analysis, and to achieve a correct Glomeruli segmentation independent from the block size analysis.

Author Contributions: Jaime Gallego, Aníbal Pedraza and Gloria Bueno conceived, designed the experiments and wrote the paper; Jaime Gallego and Aníbal Pedraza performed the experiments; Jaime Gallego, Samuel Lopez and Georg Steinner created the database; Lucia Gonzalez and Arvydas Laurinavicius reagents/materials/data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WSI   Whole Slide Image
CNN   Convolutional Neural Network
PAS   Periodic Acid Schiff

References

1. Bueno, G.; Fernandez-Carrobles, M.M.; Deniz, O.; Garcia-Rojo, M. New trends of emerging technologies in digital pathology. *Pathobiology* **2016**, *83*, 61–69.

2. Janowczyk, A.; Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **2016**, *7*, doi:10.4103/2153-3539.186902.

3. Madabhushi, A.; Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **2016**, *33*, 170–175.

4. Hughson, M.D.; Puelles, V.G.; Hoy, W.E.; Douglas-Denton, R.N.; Mott, S.A.; Bertran, J.F. Hypertension, glomerular hypertrophy and nephrosclerosis: The effect of race. *Nephrol. Dial. Transplant.* **2014**, *29*, 1399–1409.

5. Rasch, R.; Lauszus, F.; Thomsen, J.S.; Flyvbjerg, A. Glomerular structural changes in pregnant, diabetic, and pregnant—Diabetic rats. *Apmis* **2005**, *113*, 465–472.

6. Ma, J.; Jun, Z.; Jinglu, H. Glomerulus extraction by using genetic algorithm for edge patching. In Proceedings of the 2009 IEEE Congress on Evolutionary Computation, Trondheim, Norway, 18–21 May 2009.

7. Hirohashi, Y.; Relator, R.; Kakimoto, T.; Saito, R.; Horai, Y.; Fukunari, A.; Kato, T. Automated quantitative image analysis of glomerular desmin immunostaining as a sensitive injury marker in spontaneously diabetic torii rats. *J. Biomed. Image Process.* **2014**, *1*, 20–28.

8. Kakimoto, T.; Okada, K.; Fujitaka, K.; Nishio, M.; Kato, T.; Fukunari, A.; Utsumi, H. Quantitative analysis of markers of podocyte injury in the rat puromycin aminonucleoside nephropathy model. *Exp. Toxicol. Pathol.* **2015**, *67*, 171–177.

9. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1.

10. Kato, T.; Relator, R.; Ngouv, H.; Hirohashi, Y.; Takaki, O.; Kakimoto, T.; Okada, K. Segmental HOG: New descriptor for glomerulus detection in kidney microscopy image. *BMC Bioinform.* **2015**, *16*, 316.

11. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992.

12. Kotyk, T.; Dey, N.; Ashour, A.S.; Balas-Timar, D.; Chakraborty, S.; Ashour, A.S.; Tavares, J.M.R. Measurement of glomerulus diameter and Bowman's space width of renal albino rats. *Comput. Methods Programs Biomed.* **2016**, *126*, 143–153.

13. Graham, R.L. An efficient algorithm for determining the convex hull of a finite planar set. *Inf. Process. Lett.* **1972**, *1*, 132–133.

14. Marée, R.; Dallongeville, S.; Olivo-Marin, J.C.; Meas-Yedid, V. An approach for detection of glomeruli in multisite digital pathology. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; pp. 1033–1036.

15. Ginley, B.; Tomaszewski, J.E.; Yacoub, R.; Chen, F.; Sarder, P. Unsupervised labeling of glomerular boundaries using Gabor filters and statistical testing in renal histology. *J. Med. Imaging* **2017**, *4*, 021102.

16. Sicard, A.; Meas-Yedid, V.; Rabeyrin, M.; Koenig, A.; Ducreux, S.; Dijoud, F.; Hervieu, V.; Badet, L.; Morelon, E.; Olivo-Marin, J.C. Computer-assisted topological analysis of renal allograft inflammation adds to risk evaluation at diagnosis of humoral rejection. *Kidney Int.* **2017**, *92*, 214–226.

17. Simon, O.; Yacoub, R.; Jain, S.; Sarder, P. Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images. *arXiv* **2017**, arXiv:1709.01998.

18. Zhang, M.; Wu, T.; Bennett, K.M. A novel Hessian based algorithm for rat kidney glomerulus detection in 3D MRI. In Proceedings of the SPIE Medical Imaging. International Society for Optics and Photonics, Orlando, FL, USA, 21–26 February 2015.

19. Kothari, S.; Phan, J.H.; Stokes, T.H.; Wang, M.D. Pathology imaging informatics for quantitative analysis of whole-slide images. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 1099–1108.

20. Wang, H.; Cruz-Roa, A.; Basavanhally, A.; Gilmore, H.; Shih, N.; Feldman, M.; Madabhushi, A. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *J. Med. Imaging* **2014**, *1*, 034003.

21. Malon, C.D.; Cosatto, E. Classification of mitotic figures with convolutional neural networks and seeded blob features. *J. Pathol. Inform.* **2013**, *4*, 9.

22. Xu, K.; Zhu, L.; Wang, R.; Liu, C.; Zhao, Y. Automated Detection of Diabetic Retinopathy Using Deep Convolutional Neural Networks. *Med. Phys.* **2016**, *3*, 633–641.

23. Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Larochelle, H. Brain tumour segmentation with deep neural networks. *Med. Image Anal.* **2017**, *35*, 18–31.

24. Temerinac-Ott, M.; Forestier, G.; Schmitz, J.; Hermsen, M.; Bräsen, J.H.; Feuerhake, F.; Wemmert, C. Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities. In Proceedings of the 2017 10th International Symposium on Image and Signal Processing and Analysis (ISPA), Ljubljana, Slovenia, 18–20 September 2017; pp. 19–24.

25. Gadermayra, M.; Dombrowskia, A.K.; Klinkhammerb, B.M.; Boorb, P.; Merhofa, D. CNN Cascades for Segmenting Whole Slide Images of the Kidney. *arXiv* **2016**, arXiv:1708.00251.

26. Agarwal, S.K.; Sethi, S.; Dinda, A.K. Basics of kidney biopsy: A nephrologist's perspective. *Indian J. Nephrol.* **2013**, *23*, 243.

27. AIDPATH (Academia and Industry for Digital Pathology). European Project FP7 612471. Kidney database. Available online: http://aidpath.eu/ (accessed on 16 January 2018).

28. Smith, N.R.; Womack, C. A matrix approach to guide IHC-based tissue biomarker development in oncology drug discovery. *J. Pathol.* **2014**, *232*, 190–198.

29. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 3–8 December 2012; pp. 1097–1105.

30. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

31. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.

32. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41.

33. Gallego, J.; Swiderska-Chadaj, Z.; Deniz, O.; Bueno, G. Ki67 hot-spots detection on histopathological images of breast carcinoma using convolutional neural networks. In Proceedings of the Anual Congress of the Biomedical Engineering Spanish Society, Lejona, Spain, 29 November–1 December 2017.

34. Swiderska, Z.; Korzynska, A.; Markiewicz, T.; Lorent, M.; Zak, J.; Wesolowska, A.; Roszkowiak, L.; Slodkowska, J.; Grala, B. Comparison of the manual, semiautomatic, and automatic selection and leveling of hot spots in whole slide images for ki-67 quantification in meningiomas. *Anal. Cell. Pathol.* **2015**, *2015*, 498746.