A hybrid of deep and textural features to differentiate glomerulosclerosis and minimal change disease from glomerulus biopsy images

Justino D. Santos[a,b,*], Rodrigo M. S. Veras[a], Romuere R. V. Silva[c], Nayze L. S. Aldeman[a], Flávio H. D. Araújo[c], Angelo A. Duarte[d], João Manuel R.S. Tavares[e]

[a]*Departamento de Computação, Universidade Federal do Piauí, Teresina, Brasil*
[b]*Instituto Federal de Educação, Ciência e Tecnologia do Piauí, São Raimundo Nonato, Brasil*
[c]*Curso de Bacharelado em Sistemas de Informação, Universidade Federal do Piauí, Picos, Brasil*
[d]*Departamento de Tecnologia, Universidade Estadual de Feira de Santana, Feira de Santana, Brasil*
[e]*Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal*

Abstract

The minimal change disease (MCD) and glomerulosclerosis (GS) are two common kidney diseases. Unless adequately treated, these diseases leads to chronic kidney diseases. Accurate differentiation of these two diseases is of paramount importance as their methods of treatment and prognoses are different. Thus, this article propose a method capable of differentiating MCD from GS in glomerulus biopsies images based on a new hybrid deep and texture feature space. We conducted an extensive study to determine the best set of features for image representation. Our feature extraction methodology, which includes Haraliks and geostatistics texture descriptors and pre-trained CNNs, resulted in 13,476 characteristics. We then used mutual information to order the elements by importance and select the best set for differentiating MSC from GS using the random forest classifier. The proposed method achieved an accuracy of 90.3% and a Kappa index of 80.5%. Representation of glomerulus biopsy images with a hybrid of deep and textural features facilitates the accurate differentiation of GS and MCD.

Keywords: Deep learning, Feature Extraction, Feature Selection, Image Analysis, Image Classification;

*Corresponding author
Email addresses: justinoduarte@gmail.com (Justino D. Santos), rveras@ufpi.edu.br

1. Introduction

Glomerulopathies are kidney diseases with different histopathological subtypes. Microscopic evaluation is crucial for their diagnosis since it provides prognostic data and guidance for treatment. In Brazil, for example, glomerulopathies are among the leading causes of end-stage kidney disease (ESKD) and account for 11% of patients on dialysis [1]. On the other hand, nearly 125,000 people in the USA started treatment for ESKD in 2016 [2].

Nephrotic syndrome is one of the primary forms of the glomerular disease, and when symptoms are persistent, it is associated with a progression to chronic kidney disease (CKD). A publication by [2] from the US Department of Health & Human Services reports that 15% of US adults (37 million people) are estimated to have CKD.

Several histological abnormalities may lead to the development of the nephrotic syndrome. Common causes of idiopathic nephrotic syndrome are minimal change disease (MCD) and glomerulosclerosis (GS). In children, MCD is the cause of nephrotic syndrome in 90% of patients. Simultaneously, in adults, primary glomerular diseases such as GS and MCD cause a nephrotic syndrome in 70% of cases. When considering only the adult population, GS is the leading cause of nephrotic syndrome in several countries [3].

Therefore, it is essential to understand the differences between these two glomerulopathies (MCD and GS). From the therapeutic point of view, there are differences in the treatments concerning the attack phase duration in the case of corticosteroids, treatment response rates and prognosis [3].

Computer-aided diagnosis (CAD) systems aim to assist medical specialists by offering information that helps in diagnosis [4]. Such systems take as their input annotated image tests, blood tests, biopsy results or other forms of information, which are often available as a dataset of examples, and apply image processing and machine learning techniques to output a supplemental diagnosis, such as a classification into "healthy" and "unhealthy", or "benign" and "malignant". These systems are often employed in screening for diseases and provide a preliminary diagnosis or offer an opinion based on previously labeled examples.

This article proposes a computational approach that distinguishes glomerulus biopsy images with MCD and glomerulosclerosis (GS), based on a novel hybrid deep and texture feature space. To achieve this goal, texture descriptors and transfer learning (TL) techniques were evaluated using convolutional neural networks (CNNs), in ordr to produce hybrid descriptors that were inputted to supervised classifiers. Although this article proposes a complete system, its main focus is on describing the image to be analyzed; i.e., to define the best features set used to differentiate between MCD and GS.

(Rodrigo M. S. Veras), romuere@ufpi.edu.br (Romuere R. V. Silva), nayzealdeman@gmail.com (Nayze L. S. Aldeman), flavio86@ufpi.edu.br (Flávio H. D. Araújo), angeloduarte@uefs.br (Angelo A. Duarte), tavares@fe.up.pt (João Manuel R.S. Tavares)

This article is organized as follows. Section 2 presents related works; Section 3 describes the used materials and techniques and proposed method; Section 4 presents the obtained results and their discussion; and finally, Section 5 presents the conclusion and future work perspectives.

2. Related works

Medical images are widely used in computer-aided diagnostic systems. Two types of tasks are commonly found in these systems: segmentation, whose purpose is to separate specific regions, and classification, whose aim is to define related groups or classes.

Among the works that use kidney images and are focused on the segmentation or identification of renal structures, especially those related to the glomerulus, one can find the work of Zhao et al. [5]. These authors proposed an automated glomerulus extraction framework based on a micrograph of the entire kidney. On the other hand, Sarder, Ginley, and Tomaszewski [6] estimated the location of the glomerulus in images of kidney biopsies. These authors developed a methodology to extract regions containing a single glomerulus and use them to segment the glomerular boundary. In [7], the authors applied an integrated approach using Gabor filtering and Gaussian blurring to label glomerular textural edges.

Recently, Rehem et al. [8] proposed a glomerulus detection method on renal histological images. The authors applied a single shot multibox detector with Inception V2 (SI2) and reached 0.88 of mAP and 0.94 of F1-score using 909 images splitted in 509 for training, 200 for validation and 200 for the test.

There are already methods aimed at locating and segmenting the glomerulus. It is in this structure that changes due to kidney diseases. A helpful step after segmenting the glomerulus is analyzing the segmented image by identifying patterns can that be used to assist the disease diagnosis. This step falls within the classification field.

Focuses on classification tasks, there are works such as the one from Barros et al. [9]. These authors proposed a computer system to detect proliferative glomerular lesions (PGL) that could differentiate them from healthy images. They used the k-nearest neighbor (KNN) algorithm to classify the input images. The accuracy achieved in their work was $88.3 \pm 3.6\%$.

Araújo et al. [10] used images of single glomeruli to detect segmental glomerulosclerosis. Three feature vectors were extracted and supplied to four classifiers: KNN, support vector machine (SVM), neural network and naive Bayes. The authors achieved an accuracy of 84.8% for hematoxylin-eosin (H&E) stained samples and of 81.3% for periodic acid–Schiff (PAS) stained samples.

Ginley et al. [11] proposed an approach to define the structural progression of human glomeruli in diabetic nephropathy. The authors segmented glomerular compartment boundaries and quantified 47 features from each glomerulus, including texture based features. They used a naive Bayes classifier on the feature set and reported that their method could distinguish pathological stages

of diabetic nephropathy; they reach sensitivity and specificity of 0.89 and 0.93, respectively.

Sheehan and Korstanje [12] developed a method for identifying and collecting quantitative data from glomeruli. The suggested approach is semi-automatic, since it requires intervention from a specialist. The authors used contrast enhancement and Gaussian blurring, followed by a size filter to identify regions of interest corresponding to glomeruli tufts. Three features were extracted: mesangial matrix expansion (MME), the number of nuclei and capillary openness, which were classified using a random forest (RF) based approach [13]. A strong correlation was reported between MME and the analyzed phenotypes.

Marsh et al. [14] described a deep learning model that identifies and classifies non-sclerosed and sclerosed glomeruli in whole-slide images of frozen biopsy sections of donor's kidneys. This differentiation is meaningful because the criterion for accepting or rejecting the donor's kidneys relies heavily on the pathologist's determination of the percentages of glomeruli that are normal and sclerotic. The proposed approach fine-tuned the VGG-16 [15] CNN using 48 whole-slide images. According to the authors, the model achieved a precision of 81.28% in the identification of non-sclerosed glomeruli. They concluded that the method outperformed another model trained on image patches of isolated glomeruli in terms of accuracy and computational cost.

Chagas et al. [16], like Barros et al. [9], also worked on PGL detection. Their proposals perform the classification into specific PGL subcategories: endocapillary, mesangial, and both. These authors built a CNN-based architecture to extract features from glomerulus images, that were supplied to an SVM classifier. In the classification task, their method achieved an accuracy of 82%.

In addition to these works, whose tasks are focused on the renal tissue, other works employ the same techniques to analyze images of different organs or tissues. For example, Sousa et al. [17] proposed a method for diagnosing glaucoma using geostatistics [18] as a texture descriptor for images, and transfer learning techniques that use medical images as input and are also widely applied in other CAD systems [19].

To the best of our knowledge, there are no datasets or previous studies that used computational methods to differentiate between GS and MCD. It is, therefore, the main contribution of the current study. Although the aforementioned works were developed with a different purpose, they were taken into account in the developing of the proposed solution.

3. Materials and methods

This section presents a solution capable of differentiating between kidney biopsy images with GS or MCD. We performed experiments using texture features, such as Haralick's features [20], and geostatistics [18], and pre-trained CNNs, mainly VGG-16, VGG-19 [15], Xception [21] and ResNet50 [22].

The following sections describe the proposed solution and the involved techniques, the metrics adopted to assess the solution and the used image datasets.

3.1. Proposed method

The proposed method has four main steps (Figure 1): the **pre-processing** step receives the original input image and then applies size adjustments and uses the local binary patterns (LBP) [23] representation. The second step, **feature extraction**, receives the processed image as input and produces a set of features describing the image in the form of a numerical feature vector. The extracted vector is then inputted into the **feature selection** step, where the most relevant features are selected in order to be used with machine learning algorithms in the **classification** step.

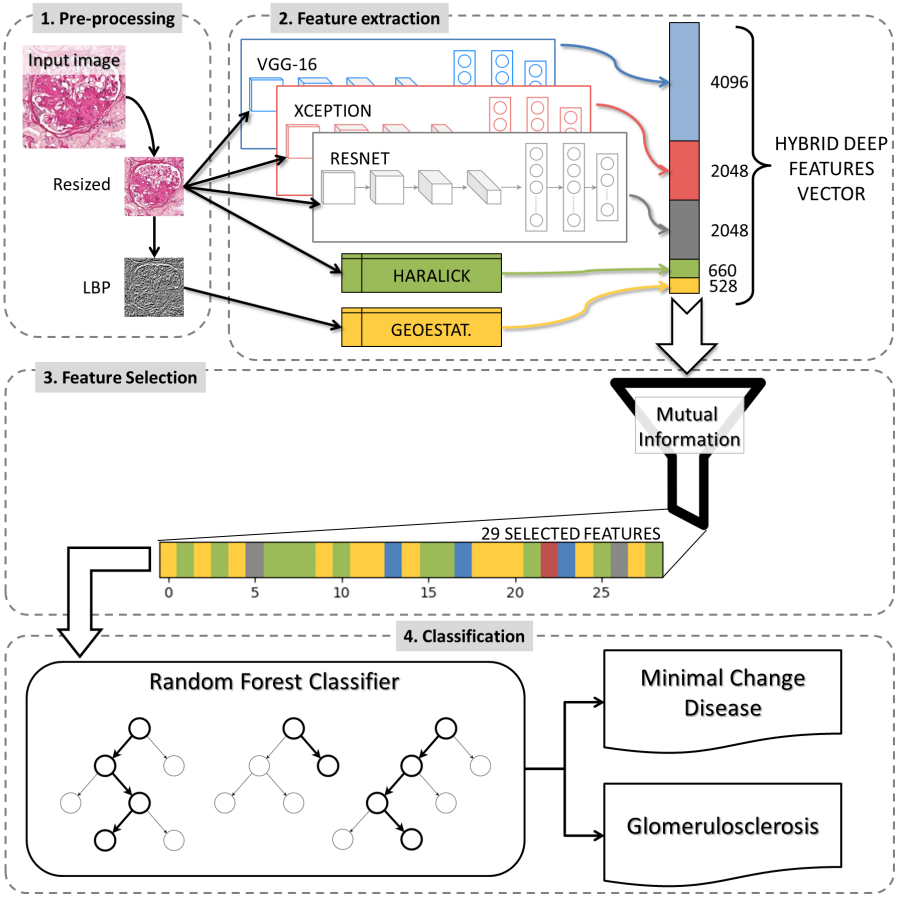

Figure 1: Flowchart of the proposed method of automatic differentiation between MCD and GS in biopsy images.

The following sections detail the steps of the proposed method and the techniques evaluated in each of them.

3.1.1. Pre-processing

In the pre-processing step, the input images are resized to the respective CNNs default input dimensions: 224×224 for VGG-16, VGG-19 and ResNet50, and 299×299 for Xception.

For extracting texture features according to the Haralick's method, the input images are also scaled to 299×299. The same is performed in the case of geostatistics; however, the resized image is represented in the form of local binary patterns (LBP) for each color channel.

Image texture is one of the visual characteristics observed by pathologists to make their diagnosis. Therefore, the proposed method generates a texture representation by processing the input image using LBP. Other works also use texture representations in order to compute geostatistical functions from medical images [17] [24].

The computation of the LBPs is depicted in Figure 2. This approach is based on the neighborhood of the central pixel of the used processing window (Figure 2 (a)); then, neighbors with values greater than or equal to the central pixel are mapped (Figure 2 (b)), and to each neighbor i is assigned a weight of 2^i (Figure 2 (c)). The value of the LBP is then the sum of the weights relative to the mapped pixels (Figure 2 (d)). After applying this process to all the ipnut image pixels, the computed values create a new image with the same size as the original one. Figures 2 (e) and (f) show a glomerulus input image and its LBP representation, respectively.

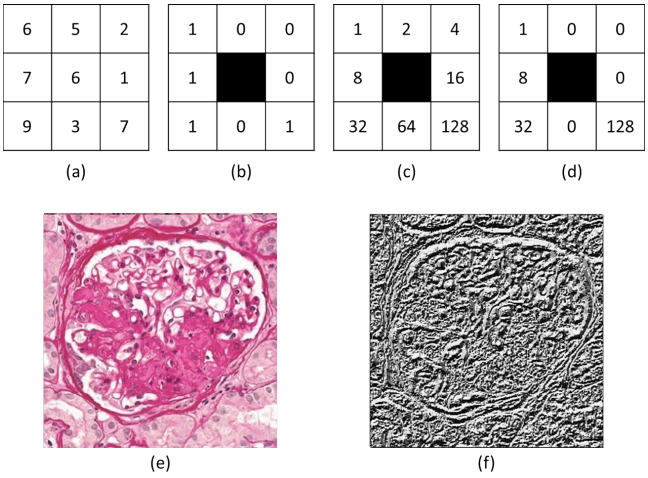

Figure 2: Calculation of LBP: a) original neighborhood values; b) mask thresholded by the central pixel; c) weights given to the corresponding pixels. d) The resulting ($b \times c$) LBP value is the sum of these values (in this example, LBP = 169); e) RGB sample input image; f) LBP from channel G of the input image.

3.1.2. Feature extraction

According to pathologists, GSF and MLD cause changes in different regions of the glomerulus. Particularly, these diseases cause the collapse of the structures, leading to adhesions and an impression of more homogeneous areas [11]. Consequently, the used description must take into account the whole image under classification. Therefore, two methods of texture characterization and deep features from four convolutional neural networks were evaluated.

Texture Features

Haralick's features are calculated based on the gray-level co-occurrence matrix (GLCM) of the input image, which is a texture descriptor that analyzes the co-occurrences between pairs of pixels and stores their relative intensities in a square matrix, with dimensions equal to the number of gray *levels*; i.e., 256 in the case of 8-bit images.The probabilities of co-occurrences ($P_{i,j}$) are calculated between two gray levels, i and j, using an angle θ (here, according to 0, 45, 90 or 135°) and a distance called the pixel pair spacing. For this purpose, 11 distance values were used: six of which are fixed (1, 2, 5, 10, 15 and 20) and five that are proportional to the image input dimensions (1.25, 2, 5, 10 and 20%).

Although there are several characteristics based on the GLCM, this experiment use the contrast (Equation 1), dissimilarity (Equation 2), homogeneity (Equation 3), angular second moment (ASM) (Equation 4), and correlation (Equation 5). The proposed method extracts texture feature using the three channels from the RGB image. Thus, the Haralick's vector have 660 attributes for each image:

$$contrast = \sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2, \tag{1}$$

$$dissimilarity = \sum_{i,j=0}^{levels-1} P_{i,j}|i-j|^2, \tag{2}$$

$$homogeneity = \sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2}, \tag{3}$$

$$ASM = \sum_{i,j=0}^{levels-1} P_{i,j}^2, \tag{4}$$

$$correlation = \sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right], \tag{5}$$

where:

$$\mu_i = \sum_{i,j=0}^{levels-1} i * P_{i,j}, \tag{6}$$

$$\mu_j = \sum_{i,j=0}^{levels-1} j * P_{i,j}, \tag{7}$$

$$\sigma_i^2 = \sum_{i,j=0}^{levels-1} P_{i,j}(i - \mu_i)^2, \tag{8}$$

and,

$$\sigma_j^2 = \sum_{i,j=0}^{levels-1} P_{i,j}(j - \mu_j)^2. \tag{9}$$

Characteristics based on the GLCM are traditionally applied for texture description based on pixel intensities. On the other hand, geostatistics takes into account, in addition to intensity, the spatial position of the pixels.

Geostatistics are statistics about a population with a known address; i.e., coordinates. The fundamental theory of geostatistics is based on the assumption that, on average, samples that are near to each other in time and space are more similar than those that are distant [18].

This work uses four geostatistical functions: semivariogram (Equation 10), semimadogram (Equation 11), covariogram (Equation 12) and correlogram (Equation 15). These functions take into account the strengths of the associations between responses as a function of distance and possibly direction [24], and can describe the texture of a given image through the degree of spatial association between its spatially referenced pixels as [17]:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (x_i - y_i)^2, \tag{10}$$

where h is the distance vector between origin values x_i and extremity values y_i, and $N(h)$ is the number of pairs in distance h:

$$m(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} |x_i - y_i|, \tag{11}$$

$$C(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} x_i y_i - m_{-h} m_{+h}, \tag{12}$$

where m_{-h} is the average of the values of the vectors origin point:

$$m_{-h} = \frac{1}{N(h)} \sum_{i=1}^{N(h)} x_i, \tag{13}$$

and m_{+h} is the average of the values of vectors end point:

$$m_{+h} = \frac{1}{N(h)} \sum_{i=1}^{N(h)} y_i, \tag{14}$$

$$\rho(h) = \frac{C(h)}{\sigma_{-h}\sigma_{+h}}, \tag{15}$$

where σ_{-h} and σ_{+h} are the standard deviations of the values of the origins and the extremities of the vectors, respectively.

As aforementioned, vectors (h) are obtained from a combination of four directions and 11 distances, of which six are fixed and five are proportional to the input image dimensions. Thus, by combining the 44 distance vectors (h) with the four geostatistical functions and applying these to the R, G, and B image channels, the resulting vector contains 528 characteristics.

Deep features

CNNs are commonly applied in the field of machine learning for many tasks as to extract the typical system response profiles of a complex system and represent them as visual outputs [25], measuring the degree of unpredictability in dynamical systems with memory [26]. A significant advantage of these techniques is their ability to automatically detect essential features, since their deep architectures allow to extract a set of characteristics at multiple levels of abstraction. CNNs have been used in the development of diagnosis tools and have outperformed conventional methods of extracting features with better accuracy rates [27].

The common architecture of a CNN includes two sections. The first one is formed of a sequence of convolution operations followed by pooling operations, and the second section is composed of fully connected layers. Figure 3 illustrates a generic CNN architecture.

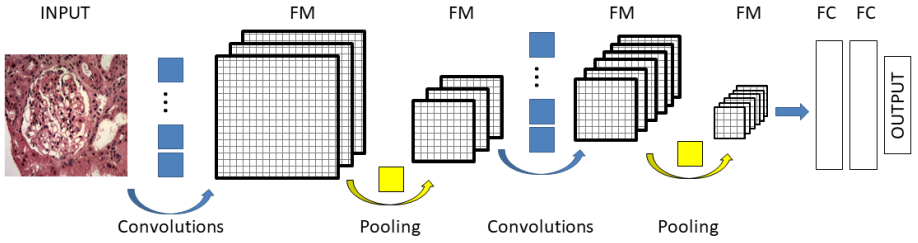

Figure 3: Generic CNN architecture: Blue and yellow blocks represent the convolutional and pooling filters applied to the input image, respectively (FM = Feature maps, FC = Fully connected layer).

The convolution layers apply filters that extract feature maps. When passing through the pooling layers, the map's dimensions are reduced, keeping only those of greater magnitude. After the sequence of convolutions and pooling, the generated feature map forms the input for the fully connected layers. In this step, the architecture and operation mechanisms are similar to those of a traditional neural network, and the last layer generates the output data; i.e., the classification result.

Usually, the training of a CNN is a task with high computational cost, and requires a large amount of data to achieve satisfactory results in terms of power

of generalization [27]. To bypass the training stage, here, it is used transfer learning (TL), which allows the domains, tasks and distributions used in training and testing to be different; i.e., the goal of TL is to reuse the knowledge learned in one field, and apply it to another correlate field [27].

In [19], the authors present one way to apply TL to CNNs used for feature extraction. Using a previously trained network on an extensive dataset, by keeping their weights, a specific image is then supplied to this network. The output data from some internal layers are used as a features vector called deep features.

We apply TL in feature extraction by taking the output vectors of the penult fully connected layer, i.e., the layer before the classification layer, of four CNNs: VGG-16, VGG-19, Xception and ResNet50. All these CNNs were pre-trained on the ImageNet dataset [28], which contains more than 1.2 million images and 1,000 classes.

Table 1 presents a summary of all the individual feature vectors extracted and evaluated in this study. In addition to these six individual vectors, we produced hybrid vectors by concatenating all their combinations, making sets with two, three, four, five and all six vectors, which led to $2^6 - 1 = 63$ vectors in total.

Table 1: Summary of the features evaluated in this study.

Type	**Method**	**Features**	**Number of features**
Texture	Geostatistic	Semivariogram, semimadogram, covariogram and correlogram	528
	Haralick	Contrast, dissimilarity, homogeneity, angular second moment and correlation	660
Deep features	VGG-16	Features extracted from the penult fully connected layer	4,096
	VGG-19		4,096
	Xception		2,048
	ResNet50		2,048
TOTAL			**13,476**

3.1.3. Feature selection

We performed a feature selection process for each of the 63 feature vectors obtained in the previous step, sorting the features in the vector in descending order of relevance. The F statistic of variance analysis (ANOVA-F) and mutual information (MI) were used to calculate the relevance of each feature.

ANOVA-F is based on measures of dispersion among the elements belonging to a group and on the dispersion among the means of each group. Given an attribute x of the feature vector, when grouping into two classes (GS or MCD), the more distant the mean value of x for these groups and the less dispersed the values of x within each group, the higher the value of the ANOVA-F, and the higher the relevance of this attribute within the feature set.

The MI between two variables is the amount of information that one variable has from the other [29]. MI measures the extent to which knowledge of one of these variables reduces uncertainty in the other. Thus, for two attributes, x_1 and x_2, when comparing the MI of each of them with variable y (i.e., the GS or MCD class), the one with the highest MI value will be considered more relevant.

We performed tests to define the final feature vector's dimensionality, searching for the best results with the smallest vector size. We used the ranked features and performed an incremental attribute selection approach, starting from the highest relevant feature and adding the others to complete all features.

3.1.4. Classification

The classification step is responsible for receiving the selected attributes and outputting the final prediction, here: GS or MCD. In this study, we evaluated two supervised classifier algorithms: Random Forest and Support Vector Machine.

The initial parameters used in the SVM were a penalty of 1.0, and a radial basis function (RBF) kernel with a gamma coefficient of 1/number of features. For RF, 100 trees and no in-depth limit growth were adopted. After analyzing the classification results, we chose the five best results and searched for the best set of classifier hyperparameters.

Here, the stratified k-fold cross-validation technique ($k = 5$), which consists of randomly distributing the dataset instances into k mutually exclusive subsets (folds) of approximately equal size, and in the same proportion observed in the original dataset, was employed. The classifier is then trained and tested k times, and in each round, a different subset is used for testing, and the remaining k–1 subsets are used for training. This mechanism ensures that each dataset element is used to evaluate the classifier and train it, and the classification ability for both classes is assessed in all folds. A confusion matrix was computed for each fold, and the arithmetic average of the five values achieved from each studied classifier was taken into account.

3.2. Evaluation metrics

The confusion matrix confronts the classifier predicted results and the actual results for the same set of tests. Here, there are two classes, GS and MCD, and the problem is, therefore, a binary classification problem. Thus, there are four values in this matrix: the true positive (TP), which indicates the number of images correctly classified as MCD; the true negative (TN), corresponding to the number of correct GS classifications; the false positive (FP), representing the number of images classified as MCD, but which are GS; and finally, the false

negative (FN), which refers to the number of images erroneously classified as GS.

The accuracy assesses the overall rate of correct classification for both classes together, and is obtained as the ratio between the number of correctly classified images and the total number of images.

Here, Cohen's kappa [30] is used as the primary evaluation metric because it is more challenging than accuracy, since kappa considers the probability distribution of the expected classes. Also, kappa gives a value that represents the degree of agreement between nominal classifications performed by two evaluators. In the present case, those predicted by the classifier and those annotated by a pathologist:

$$kappa = \frac{accuracy - P_e}{1 - P_e}, \tag{16}$$

where P_e represents the expected probability of the evaluators agreeing on the classification; in other words, it means the overall random agreement probability.

The maximum value of kappa is 100%, which indicates a perfect agreement among the evaluators. The labels shown in Table 2, proposed by [31], are usually used to maintain consistent vocabulary when describing the relative strength of agreement associated with the kappa metric, and therefore, they were also adopted in the current study.

Table 2: Labels assigned to the corresponding ranges of kappa.

kappa (%)	**Strength of Agreement**
<0	Poor
0 ⊢ 20	Slight
20 ⊢ 40	Fair
40 ⊢ 60	Moderate
60 ⊢ 80	Substantial
80 ⊢ 100	Almost Perfect

In this study, two additional metrics were computed to give a further evaluation: *precision*, which can be understood as the ability of the classifier to avoid labeling a negative example as positive, and *recall*, which can be interpreted as the ability of the classifier to identify all positive instances.

3.3. Image datasets

The images used in this study were from two datasets. The first one, named DME, contains 83 RGB images from the Department of Specialized Medicine of the Federal University of Piauí, of which a specialist had classified 42 as GS and 41 as MCD.

The images of the DME dataset images were acquired using a Nikon e220 and a Nikon e200 microscopes adapted with immunofluorescence. Pigmentation was applied to the slides using the following dyes: H&E, Masson's trichrome, PAS and silver methenamine.

The second used dataset, named InetDB, was built by collecting images in public domain available on the internet, and is composed of 21 RGB images, of which a specialist had classified 10 as GS and 11 as MCD.

The original images have different aspect ratios (width/height), and therefore, they were manually adjusted, keeping glomerulus in the centre and leaving the width equal to the height. Only cutouts and padding were used in this operation to not cause distortions in the glomerulus image, not even when it was resized in the pre-processing step.

Figure 4 shows examples of the used images. In these images, visual heterogeneity can be observed between images belonging to the same class and similarities between images in distinct classes, which is always challenging for classification.

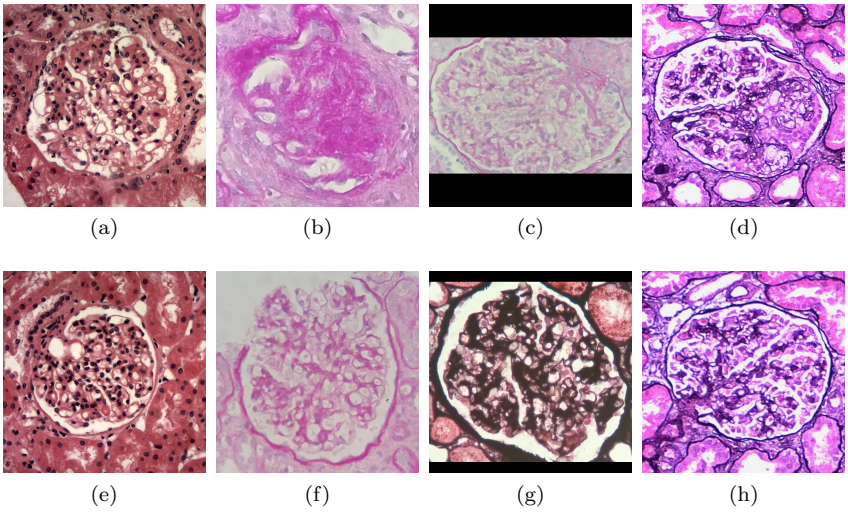

Figure 4: Samples of images from the used dataset: a-d, glomerulosclerosis images; e-h, minimal change disease images; c and g, padded images. (The images on the last column are from the InetDB dataset, the others are from the DME dataset.)

Hence, the two datasets together contain 104 images, where 52 are from GS and 52 from MCD. Despite being balanced, the number of images may not be enough for the proper training of the classifiers. Thus, to obtain a more extensive training set, data augmentationtechniques were applied to the original images.

To preserve the characteristics of the input images and not lose critical information, changes such as shear or zoom were avoided, and three rotation transformations: 90, 180, and 270°, were applied to each original image. These transformations represent situations that can naturally happen, e.g., a slide is rotated under the microscope at the time of analysis. As a result, the total number of samples, i.e., original + transformed images, used in this studied were increased from 104 to 416. Thus, on each of the five rounds of cross-validation, the training was performed with an average of 332 images, while the test was

applied with an average of 21 images.

4. Results and discussion

In this study, 252 scenarios were evaluated: 63 feature vectors combined with two feature selection methods and two classifiers. An incremental feature selection approach was also applied by computing the importance of each feature. Then, the highest mean *kappa* was taken into account in order to select the best scenario.

In the sections below, the following abbreviations: v16 = VGG-16, v19 = VGG-19, xce = Xception, rsnet = ResNet50, hrlk = Haralick's feature and geo = Geostatistics, are used to refer the feature vectors in order to improve the comprehension and layout.

Table 3 presents the best five classification results obtained using the RF and SVM classifiers, and the average processing time, in seconds, for each evaluated method. The highest 35 *kappa* values were achieved using RF and MI. Random Forest is a classifier formed by a committee of decision trees where each tree considers subsets of different features. Thus, it is more robust to overfitting and here performed better than the tested SVM classifier.

Table 3: Best five kappa achieved using each classifier under comparison and the respective average processing in seconds.

I	Vector	#F	Accuracy	Kappa	Precision	Recall	Time
			RF				
1	v16+v19+rsnet+ hrlk+geo	25	**88.4 5.1**	**76.7 10.2**	89.0 7.0	88.4 9.5	26.8
2	v16+xce+rsnet+ hrlk+geo	29	**88.4 5.1**	**76.7 10.2**	**91.0 8.3**	86.5 11.3	32.2
3	v16+v19+xce+ rsnet+hrlk+geo	46	88.3 6.9	76.5 13.9	89.5 10.4	88.4 9.5	37.3
4	v16+v19+xce+ hrlk+geo	**22**	87.6 4.5	75.3 9.0	84.7 4.6	**92.5 10.6**	24.9
5	v16+v19+ xce+geo	31	87.5 6.1	75.1 12.2	88.2 3.5	86.7 12.2	**23.9**
			SVM				
36	v19+rsnet	87	**84.5 9.3**	**68.9 18.6**	**83.8 10.4**	86.5 9.4	13.6
39	v19+xce+rsnet	261	83.7 7.1	67.5 14.2	82.5 7.7	**86.7 9.4**	22.3
43	v16+v19+rsnet	106	83.5 7.9	67.1 15.7	82.5 10.0	**86.7 9.4**	18.1
51	v16+v19+ xce+rsnet	**49**	82.6 5.8	65.3 11.5	82.2 7.7	84.7 7.2	26.8
52	v16+v19	122	82.6 7.5	65.3 15.0	81.7 10.7	**86.7 13.0**	**7.9**

- I: Ranking index in a global comparison (252 scenarios); #F: number of features
- For each classifier, the best results found are in bold.

From the data in Table 3, one can realize that the scenario that required the longest time to run took, on average, 37.3s. The tests were carried out on a computer with an Intel Core i5-1135G7 CPU and 8GB of RAM memory. There was no substantial difference between the execution time of the classifiers. The extraction of geostatistical information was the most time-consuming

operation. However, according to a consulted pathologist, the visual examination of a glomerulus can take about three minutes. Thus, the response time of the proposed solution can be considered applicable in real environment. It is essential to point out that the use of a GPU-appropriate implementation would decrease the reuired processing time.

The results obtained with each of the descriptors individually were also analyzed. Among the deep features, the best result was found using the exception vector ($kappa$ = 71.1%), classified with RF. Among the texture descriptors, geostatistics were outstanding ($kappa$ = 57.8%) when classified with RF. These results indicate that the use of a hybrid descriptor is more efficient to characterize the images under study.

We also used a Grid search for the best hyperparameters for the best five results in order to improve the classification results reached using the RF classifier. A Grid search is a process that searches exhaustively through a manually specified subset of the hyperparameter space of the targeted algorithm. A total of 211,200 hyperparameter settings concerning the RF classifier were evaluated. The classification task was performed for each of them, keeping the same 5-folds and using the same previously selected feature set. Table 4 presents the results reached after the RF tuning, and Table 5 the parameters found from the tuning process of the RF classifier that gave the best classification results.

Table 4: Classification results after the RF hyperparameter tuning for the best five achieved kappa.

Vector	**#F**	**Accuracy**	**Kappa**	**Precision**	**Recall**
v16+v19+rsnet+ hrlk+geo	25	89.4 5.8	78.7 11.5	91.0 8.3	88.5 10.8
v16+xce+rsnet +hrlk+geo	29	**90.3 3.3**	**80.5 6.6**	**91.4 7.8**	90.4 8.6
v16+v19+xce+ rsnet+hrlk+geo	46	88.4 5.1	76.7 10.2	↓89.0 7.0	88.4 9.5
v16+v19+xce+ hrlk+geo	**22**	87.7 7.7	75.5 15.4	85.3 4.8	**↓90.7 14.1**
v16+v19+ xce+geo	31	88.5 6.0	77.1 12.0	88.5 2.9	88.7 13.3

- #F: number of features.
- Down arrow means metric worsening; Best found results are in bold.

Finally, after RF hyperparameters tuning, the best feature vector was: v16+xce+rsnet+hrlk+geo. From the 9,380 features of this vector, 29 were selected based on MI. Then, it was classified using the RF classifier. Figure 5 shows the number of attributes chosen from each descriptor, and their distributions along with the positions of the vector sorted by the MI algorithm.

By analyzing Figure 5, one can notice that of the 1,188 texture characteristics, 23 were selected; while of the 8,192 deep features, only six were selected. Notably, the texture had greater representativeness in the set of more relevant

Table 5: Values of the parameters found through tuning of the RF classifier for the best 29 MI features obtained with v16+xce+rsnet+hrlk+geo.

Parameter description	**Range search**	**Value found**
Number of trees	40 values from [5 to 200]	80
Minimum number of samples required to split an internal node	[2 to 6]	6
Minimum number of samples required to be at a leaf node	[1 to 4]	1
Grow limit in the depth way	Unlimited and 10 values from [10 to 100]	Unlimited
Split quality measure function	Gini impurity and Information gain	Gini impurity
Amount of features to consider when looking for the best split	$\sqrt{N}$, $log_2(N)$ e N, where N is number of features.	$\sqrt{N}$
Value which grow trees in best-first fashion.	Unlimited, 15, 30 and 60	Unlimited
Total searched combinations	**211,200**	

attributes; this simulates the pathologists' practice when observing these visual characteristics in order to differentiate GS from MCD.

As for deep features, we believe that, although the used CNNs were pre-trained on a varied and dense image dataset, their ability to extract relevant characteristics is low for the specific type of image under study. It is possible that other TL techniques, such as, for example, fine tunning, or the complete training of the same architectures, could change this scenario. However, this task requires a dataset with a more significant number of images.

Although deep features represent about 21% of the definitive vector, they are not expendable. The metrics achieved using the vector composed only by the two texture descriptors (hrlk + geo), and *kappa* reached 61.8%, which is a 19.4% lower result relatively to the vector composed with the included deep features, and that before even adjusting the hyperparameters of the RF.

When using the classifier with the hyperparameters found in the search, in addition to good metrics,a perfect balance in the classification results was reached: The correct answers (TP and TN) obtained equal values, 47 images in each class. Similarly, the classification errors (FP and FN) also occurred in a balanced way, with 5 images in each class. This points out that the proposed solution is not biased when classifying instances. Figure 6 depicts some results obtained using the proposed solution.

The proposed computational solution was evaluated on two image datasets. This is a positive aspect, as it brings more robustness to it, since images from different sources reduce the possibility of carrying patterns of color, brightness, etc., that could generate bias. However, there are two limitations to consider.

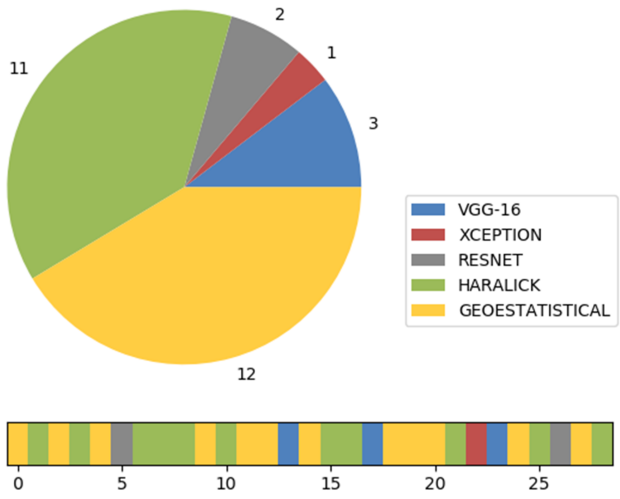

Figure 5: Origin of the 29 attributes with best mutual information for the v16+xce+rsnet+hrlk+geo vector. (The bar at the bottom depicts the source of the characteristics at each position.)

The former is the number of examples; for the solution be more reliable, it is interesting to test it using more images. And the second one is due to the fact the used images contain the centralized glomerulus, and not the entire biopsy slide. So, in practice, a previous step to segment these structures will be necessary. However, there are works, such as [5], [6], [7] and [8], that proposed computational methods that can be used to automatically perform this task.

5. Conclusion and future works

This article proposed a computer solution to differentiate MCD and GS in microscopic images. The most relevant features from VGG-16, Xception, and ResNet50 CNNs, with Haralick and Geostatistical textural descriptors, were evaluated using the Random Forest classifier. This solution gave results that are in near-perfect agreement with the diagnosis made by a pathologist.

From the findings, one can conclude that hybrid deep and texture features can convey the attributes from the image under study more competently than a single descriptor. Therefore, a new hybrid deep and texture feature space was built to differentiate glomerulosclerosis from minimal change disease in glomerulus biopsy images. The used features selection algorithms determined that the texture characteristics are most relevant than those obtained from CNNs by deep features.

As future works, we intend to evaluate methods that use the fine-tuning technique in CNNs. This technique uses a pre-trained CNN on a large image dataset and re-train some layers with a small learning rate for fine adjustment

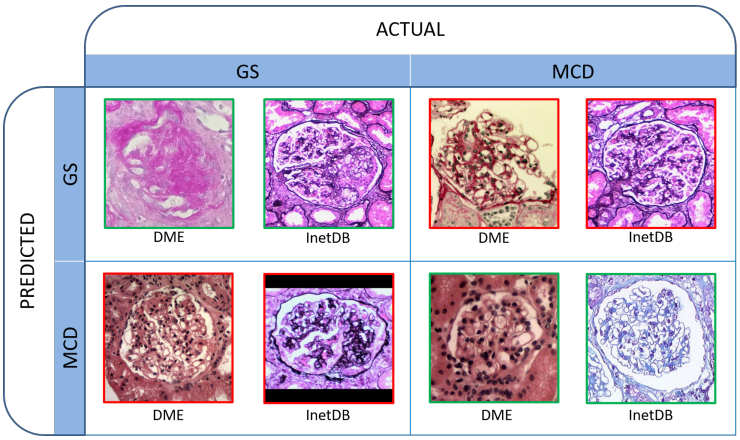

Figure 6: Sample results achieved by the proposed solution.

weights. Another future work concerns using larger image datasets to train and test the proposed solution and validate it in clinical scenarios.

Acknowledgment

This study was financed in part by the Universidade Federal do Piauí, in Brazil. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used in this study.

Conflict of interest

The authors declare that they have no conflict of interest.

[1] D. M. d. N. Costa, L. M. Valente, P. A. d. C. Gouveia, F. W. Sarinho, G. V. Fernandes, M. A. G. d. M. Cavalcante, C. B. L. d. Oliveira, C. d. A. J. d. Vasconcelos, E. S. C. Sarinho, Comparative analysis of primary and secondary glomerulopathies in the northeast of Brazil: data from the Pernambuco registry of glomerulopathies-repeg, Brazilian Journal of Nephrology 39 (1) (2017) 29–35.

[2] Centers for Disease Control and Prevention, Chronic kidney disease in the united states, 2019, gA: US Department of Health and Human Services, Centers for Disease Control and Prevention (2019).

[3] L. R. Moura, M. F. Franco, G. M. Kirsztajn, Minimal change disease and focal segmental glomerulosclerosis in adults: response to steroids and risk of renal failure, Brazilian Journal of Nephrology 37 (4) (2015) 475–480.

[4] J. Yanas, E. Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: Past and present developments, Expert Systems with Applications 138 (2019) 112821.

[5] Y. Zhao, E. F. Black, L. Marini, K. McHenry, N. Kenyon, R. Patil, A. Balla, A. Bartholomew, Automatic glomerulus extraction in whole slide images towards computer aided diagnosis, in: 12th International Conference on e-Science (e-Science), IEEE, 2016, pp. 165–174.

[6] P. Sarder, B. Ginley, J. E. Tomaszewski, Automated renal histopathology: digital extraction and quantification of renal pathology, in: Medical Imaging 2016: Digital Pathology, Vol. 9791, International Society for Optics and Photonics, 2016, p. 97910F.

[7] B. Ginley, J. E. Tomaszewski, P. Sarder, Automatic computational labeling of glomerular textural boundaries, in: Medical Imaging 2017: Digital Pathology, Vol. 10140, International Society for Optics and Photonics, 2017, p. 101400G.

[8] J. M. C. Rehem, W. L. C. dos Santos, A. A. Duarte, L. R. de Oliveira, M. F. Angelo, Automatic glomerulus detection in renal histological images, in: Medical Imaging 2021: Digital Pathology, Vol. 11603, International Society for Optics and Photonics, SPIE, 2021, pp. 115 – 125. doi:10.1117/12.2582201.

[9] G. O. Barros, B. Navarro, A. Duarte, W. L. Dos-Santos, Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys, Scientific reports 7 (2017) 46769.

[10] I. C. d. Araújo, L. Schnitman, A. A. Duarte, W. L. dos Santos, Automated detection of segmental glomerulosclerosis in kidney histopathology, in: XIII Brazilian Congress on Computational Intelligence, 2017.

[11] B. G. Ginley, J. E. Tomaszewski, K.-Y. Jen, A. Fogo, S. Jain, P. Sarder, Computational analysis of the structural progression of human glomeruli in diabetic nephropathy, in: Proceedings of SPIE Medical Imaging, Vol. 10581, 2018, pp. 105810A–1–105810A–6.

[12] S. M. Sheehan, R. Korstanje, Automatic glomerular identification and quantification of histological phenotypes using image analysis and machine learning, American Journal of Physiology-Renal Physiology 315 (6) (2018) F1644–F1651.

[13] T. K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, 1995, pp. 278–282.

[14] J. N. Marsh, M. K. Matlock, S. Kudose, T. Liu, T. S. Stappenbeck, J. P. Gaut, S. J. Swamidass, Deep learning global glomerulosclerosis in transplant kidney frozen sections, IEEE Transactions on Medical Imaging 37 (12) (2018) 2718–2728.

[15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[16] P. Chagas, L. Souza, I. Araújo, N. Aldeman, A. Duarte, M. Angelo, W. L. dos Santos, L. Oliveira, Classification of glomerular hypercellularity using convolutional features and support vector machine, arXiv preprint arXiv:1907.00028 (2019).

[17] J. A. de Sousa, A. C. de Paiva, J. D. S. de Almeida, A. C. Silva, G. B. Junior, M. Gattass, Texture based on geostatistic for glaucoma diagnosis from fundus eye image, Multimedia Tools and Applications 76 (18) (2017) 19173–19190.

[18] E. Isaaks, R. Srivastava, K. (Firm), An introduction to Applied Geostatistics, Oxford University Press, 1989.

[19] L. H. Vogado, R. M. Veras, F. H. Araujo, R. R. Silva, K. R. Aires, Leukemia diagnosis in blood slides using transfer learning in cnns and svm for classification, Engineering Applications of Artificial Intelligence 72 (2018) 415–422.

[20] R. M. Haralick, K. Shanmugam, et al., Textural features for image classification, IEEE Transactions on systems, man, and cybernetics 3 (6) (1973) 610–621.

[21] F. Chollet, Xception: Deep learning with depthwise separable convolutions, arXiv preprint (2017) 1610–02357.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 770–778.

[23] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29 (1) (1996) 51 – 59.

[24] A. C. Silva, P. C. P. Carvalho, M. Gattass, Analysis of spatial variability using geostatistical functions for diagnosis of lung nodule in computerized tomography images, Pattern Analysis and Applications 7 (3) (2004) 227–234.

[25] M. Martínez-García, Y. Zhang, J. Wan, J. McGinty, Visually interpretable profile extraction with an autoencoder for health monitoring of industrial systems, in: 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), 2019, pp. 649–654. doi:10.1109/ICARM.2019.8834281.

[26] M. Martínez-García, Y. Zhang, K. Suzuki, Y.-D. Zhang, Deep recurrent entropy adaptive model for system reliability monitoring, IEEE Transactions on Industrial Informatics 17 (2) (2021) 839–848. doi:10.1109/TII.2020.3007152.

[27] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, IEEE transactions on medical imaging 35 (5) (2016) 1299–1312.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[29] T. M. Cover, J. A. Thomas, Entropy, relative entropy and mutual information, Elements of information theory 2 (1991) 1–55.

[30] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1) (1960) 37–46.

[31] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.