Machine Learning in Applications
01-FP4 Project Report

Belloni Sofia (303393), Civiero Vittoria (302839),
De Marco Alessandro (317626), Panuccio Ilaria (302677)
Politecnico di Torino

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Machine Learning in Applications 01-FP4 Project Report

***Abstract*—Identifying glomeruli, which involves detection and classification into normal and sclerosed glomeruli, is an essential step in various nephropathology studies such as those involving DM1 (type 1 Diabetes Mellitus) and DM2 (type 2 Diabetes Mellitus). However, manually counting glomeruli can be tedious and a time-consuming task. That's why it's necessary to use image processing tools that can accurately detect and classify glomeruli. In this paper they have been identified through two semantic segmentation architectures: SegNet and U-Net.**
On the other hand, since there wasn't available ground truth for glomeruli classification, an unsupervised method was required to determine distinct levels of glomeruli necrotization. In particular, to reach this goal a K-means method was performed after an initial feature extraction and dimensionality reduction phase. These methods were tested on a dataset composed of 9 WSIs belonging to human kidney sections. The best approach of semantic segmentation was SegNet which returned slightly better results than U-Net. Regarding clustering, K-means did not provide a comfortable classification, probably due to the complex nature of our initial images. Therefore, in presence of labels, it might be better to approach this problem with a supervised technique rather than an unsupervised method.

I. Introduction

Glomeruli are clusters of capillaries located in the kidneys. They represent an important structure for renal system, indeed their role is to carry out the first filtering step, contributing to the balance of body fluids and the removal of waste.
For several reasons, including type 1 and 2 diabetes, glomeruli could be injured and scarred, losing their functionality and entering in a condition called "glomerulosclerosis" [1].
In everyday practice, kidney biopsy could quantify the number of glomeruli in the histological sample. The importance of indicating the degree of sclerosis and quantifying sick glomeruli is reflected in the identification of the patient's pathology and in making a diagnosis. In particular, if a high number of sclerosed glomeruli in an advanced state is detected from the sample, the patient will have to undergo specific treatments, as he is suffering from chronic kidney disease. The glomeruli count is very tedious and time consuming: there is a need for methods capable of correctly processing images of this type, and therefore of accurately identifying and classifying the glomeruli.
The aim of this study is to identify the glomeruli and classify them according to their sclerosis status. For the identification phase, two semantic segmentation networks were used and compared, which are SegNet and U-Net. As regards the classification phase, not having available a complete dataset of labels indicating which glomeruli are sclerosed, a feature extraction phase was carried out, comparing three different methods, such as *autoencoder*, *VGG19* and *ResNet50*. Thanks to the extracted features and the dimensionality reduction through *t-SNE* and *Isomap*, it was possible to apply *K-means*. The available dataset presented several challenges, as it was composed of large WSIs (Whole Slide Images), therefore difficult to process on common machines. Due to the computing power and memory required to implement machine learning methods mentioned above, it was necessary to connect with the Legion platform, offered by the Politecnico di Torino.

II. Background

Artificial intelligence offers the necessary tools to address the problem highlighted in the introduction.
Identifying objects within images can be achieved using models that perform semantic segmentation, which provides a pixel-by-pixel mapping, assigning each pixel to its respective class. This segmentation is largely attributed to the capabilities of convolutional neural networks (CNNs), with prominent examples including AlexNet, ResNet, U-Net and SegNet [2].
The challenge of identifying glomeruli using machine learning techniques is well-documented in literature, as demonstrated by studies such as those presented in [3]. In the referenced study, two primary approaches to the problem are investigated:

- Three classes semantic segmentation. That is a comparison of SegNet and U-Net methodologies for segmentation between non-glomerular structures, normal glomeruli and complete sclerosed glomeruli;
- Two classes semantic segmentation follows by classification. Based on that, a classification CNN is proposed where glomerular regions obtained by a semantic segmentation are divided into normal and complete sclerosed glomeruli.

Expanding upon the foundational methods, our research delves into the unsupervised classification of glomeruli. Given the complex, high-dimensional nature of image data, directly applying clustering techniques can introduce significant errors. In this multidimensional space, images that appear similar in a low-dimensional projection might actually be very distant from each other, leading to potential misclassifications. To address this challenge, feature extraction emerges as an essential step. By transforming the high-dimensional image data into a reduced feature space, we can more accurately capture the inherent structures and patterns in the data, facilitating more effective clustering.
CNNs have been particularly instrumental in this aspect. Models like VGG-19 and ResNet are adept at capturing hierarchical patterns and intricate details from images, making them suitable for this feature extraction task [4].
Yet, there's another promising avenue in the form of autoencoders [5]. An autoencoder is a type of neural network designed to encode data into a compressed representation and then decode it back to its original form. The compressed

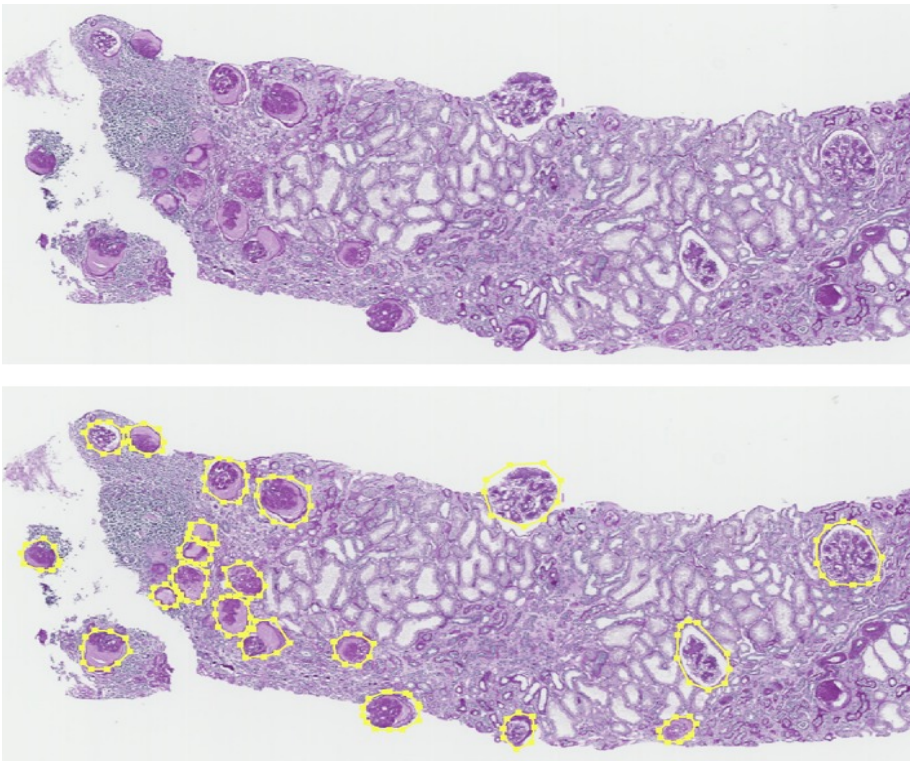

Fig. 1: The first image is the original WSI while the second is the corresponding labelled one

representation, often referred to as the latent space, captures the essential features of the data. By using the encoded features from this latent space, we can achieve a more compact and meaningful representation of the images, which can be advantageous for subsequent clustering tasks.

III. Materials and methods

A. Materials and data preprocessing

The digital tissue images used in this study are WSIs (Whole Slide Images) of renal biopsies. The dataset is composed of 9 different WSIs of human kidney tissue provided with annotations which show glomeruli (Figure 1).

Each histological image was partitioned into smaller patches with a size of 2000×2000. To increase the chance of obtaining a complete glomerulus patch, an overlap technique was performed. In particular, each patch has been moved 1000 pixels along the horizontal, vertical and diagonal axis, taking care not to exceed the image size.

Since the original WSIs were composed of large white areas without any evidence of tissue, we filtered the obtained patches deleting those with a white pixels threshold bigger than 85%. Afterwards, a binary mask was computed for each patch according to the initial annotations to highlight glomeruli contours. At the end all patches were resized into 512x512 pixels to reduce memory and computational effort. After these procedures the number of images was 7271, whose 3154 containing glomeruli.

The following phase consists of class balancing. It was performed reducing the majority class, i.e. the one without glomeruli, by half of its amount obtaining a total of 5213 images. In addition, the number of samples containing glomeruli has been doubled through a specific data augmentation, including vertical and horizontal rotation and contrast mutation. At the end of this balancing process the final sample counted 8367 images.

The dataset was splitted into training, validation and test set with a percentage of 70%, 20% and 10%, respectively.

CNN architectures generally rely on large datasets of images to achieve meaningful results [6]. Therefore, a data augmentation process was conducted to increase the number of samples. Color normalization is a commonly employed data augmentation technique in digital pathology. Despite the use of the same staining marker in immunohistochemical processes, variations in color may arise within the tissue. Color normalization methods overcome this issue by applying a colour transfer between images. In particular, a random data augmentation was applied on our training set of 5856 images. It includes some affine transformations as flips, rotations, crop and local distortion and some color transformations like dropout, brigthness, greyscale and gammacontrast.

Finally, the resulting dataset consists of 23424 images.

B. Segmentation

In order to identify glomeruli, a segmentation was performed using two powerful semantic segmentation methodologies, SegNet and U-Net. These techniques have been compared and their results are shown in Section IV.

Let's see more in details their structural differences and similarities.

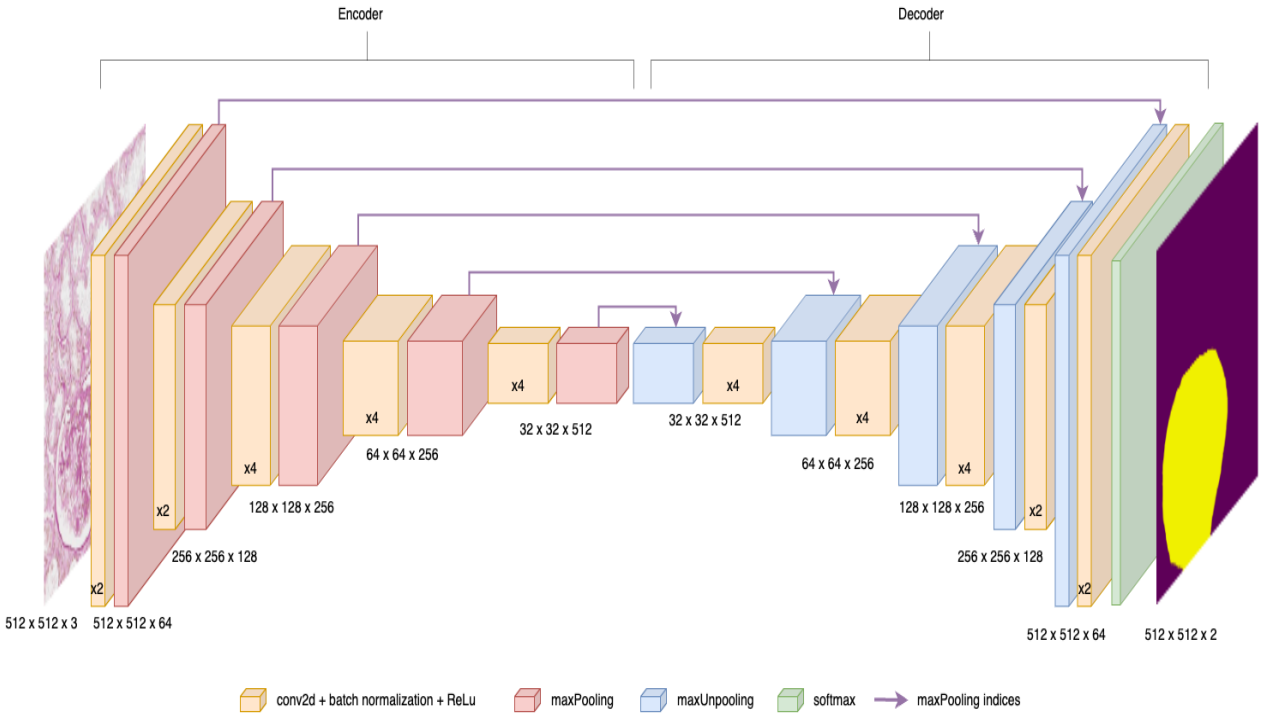

Fig. 2: SegNet architecture

SegNet is an encoder-decoder architecture (Figure 2), followed by a final pixelwise classification layer. At the encoder, convolutions and max pooling are performed, in particular there are:

- Input layer;
- Encoder with 5 convolutional blocks which correspond to VGG19 convolutional blocks (described in section III). Each of them performs convolution with a filter bank to produce a set of feature maps. These are then batch normalized and ReLU is applied;
- Each convolutional block is followed by a maxPooling layer which stores the maximum pooling indices to aid in detailed reconstruction during the decoding phase.

At the decoder, upsampling and convolutions are performed, in particular:

- Each encoder layer has a corresponding decoder layer, hence the decoder network is composed by 5 convolutional blocks;
- Each convolutional block is preceded by a maxUnpooling layer. During upsampling, the max pooling indices at the corresponding encoder layer are recalled to upsample;
- Finally, a K-class softmax classifier is used to predict the class for each pixel. The predicted segmentation corresponds to the class with maximum probability at each pixel.

U-Net is a unique architecture designed for image segmentation tasks. It was built integrating specific layers of MobileNetV2, a CNN pre-trained on ImageNet dataset.

The process consists of reducing image dimensions while extracting high-level features, scaling it to its original size combining these details with the original look of the image, and labelling each pixel. This structure provides a balance between capturing hierarchical features and maintaining spatial resolution, essential for accurate segmentation.

U-Net (Figure 3) is described as follows:

- Input layer;
- Encoder with 5 convolutional blocks consisting of convolutional layers and ReLU layers. These blocks capture the hierarchical features of the input image;
- Decoder made by 5 convolutional blocks with up-convolutional layers, upReLU layers followed by depth concatenation layers. The decoding path progressively increases the spatial dimensions of the feature maps;
- 4 skip-connection: after each upsampling step, features from the corresponding block of the encoder are concatenated. This skip-connection approach helps in retaining the spatial information which might get lost during aggressive downsampling;
- 1x1 convolution layer to upscale the 64 feature map extracted to the original image's spatial dimensions, ensuring that each pixel of the image has an associated class;
- Softmax layer which calculates pixel-wise classification scores;
- Output layer with the predicted pixel map composed by two channels. Each channel corresponds to the probability of a pixel belonging to one of the two classes.

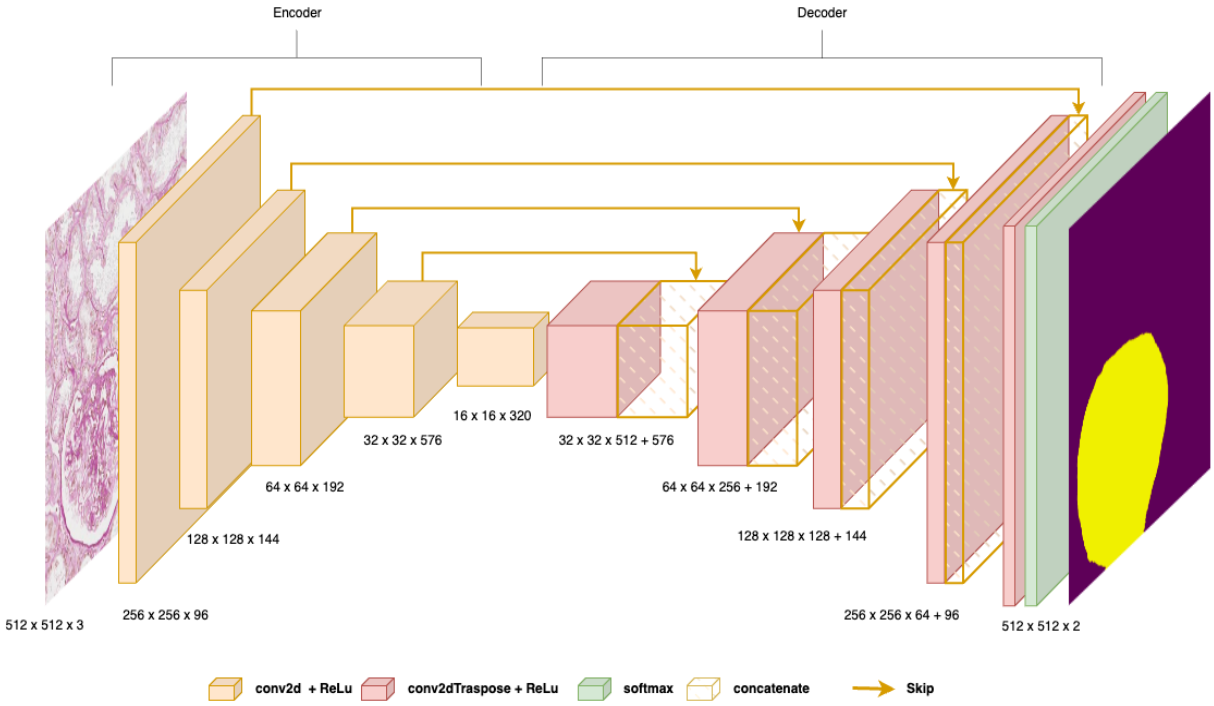

Fig. 3: U-Net architecture.

C. *Features extraction*

In order to perform cluster methods to classify glomeruli, we applied, in parallel, some feature extraction techniques, specifically: autoencoder, VGG19 and ResNet50.

Autoencoder is a pivotal neural network model. The fundamental architecture comprises an encoder, a bottleneck containing the latent space, and a decoder (Figure 4). Its primary objective is to reconstruct input data in an unsupervised manner.

Autoencoders have demonstrated proficiency in image processing, data generation, and anomaly detection. Their structure is particularly conducive for unsupervised feature extraction. Unlike other feature extraction models such as VGG-19 or ResNet, autoencoders leverage the latent space for feature representation.

In this study, we employed a convolutional autoencoder that harnesses convolutional layers both to derive the latent space and to produce the output. The encoder's structure is delineated as follows:

- A convolutional layer with 128 kernels of size 3x3, succeeded by a ReLU activation layer and max pooling;
- A convolutional layer with 256 kernels of size 3x3, followed by a ReLU activation layer and max pooling;
- A convolutional layer with 512 kernels of size 3x3, followed by a ReLU activation layer and max pooling;
- A flatten layer, which reshapes the multidimensional output from the previous convolutional layers into a one-dimensional vector, facilitating the transition to fully connected layers;
- A dense layer comprising 512 neurons, the output of which represents the latent space.

The decoder is structured as:

- A dense layer with 320,000 neurons, which serves to expand the latent representation. This is then reshaped to a format of (25x25x512), matching the dimensions prior to the flatten operation, and thereby enabling subsequent convolutional processing in the decoder;
- A deconvolutional layer with 512 kernels of size 5x5, succeeded by a ReLU activation layer;
- A deconvolutional layer with 256 kernels of size 3x3, followed by a ReLU activation layer;
- A deconvolutional layer with 128 kernels of size 2x2, succeeded by a ReLU activation layer;
- A deconvolutional layer with 3 kernels of size 3x3, followed by a sigmoid activation layer. The output of this layer is the reconstructed image.

The autoencoder was trained on 200x200 images. Leveraging prior segmentation, images were processed to solely represent glomeruli. This resulted in cropped images depicting only glomeruli. This preprocessing was undertaken to ensure the autoencoder's focus remained solely on the glomerulus surface. Upon training, the desired features are extracted from the encoder's output.

VGG19 is a deep convolutional neural network with nineteen layers, capable of extracting meaningful features from images, such as lines, curves, textures and shapes. Thanks to its ability, this model is exploited in various applications, including object detection and image classification. VGG19 stands out from

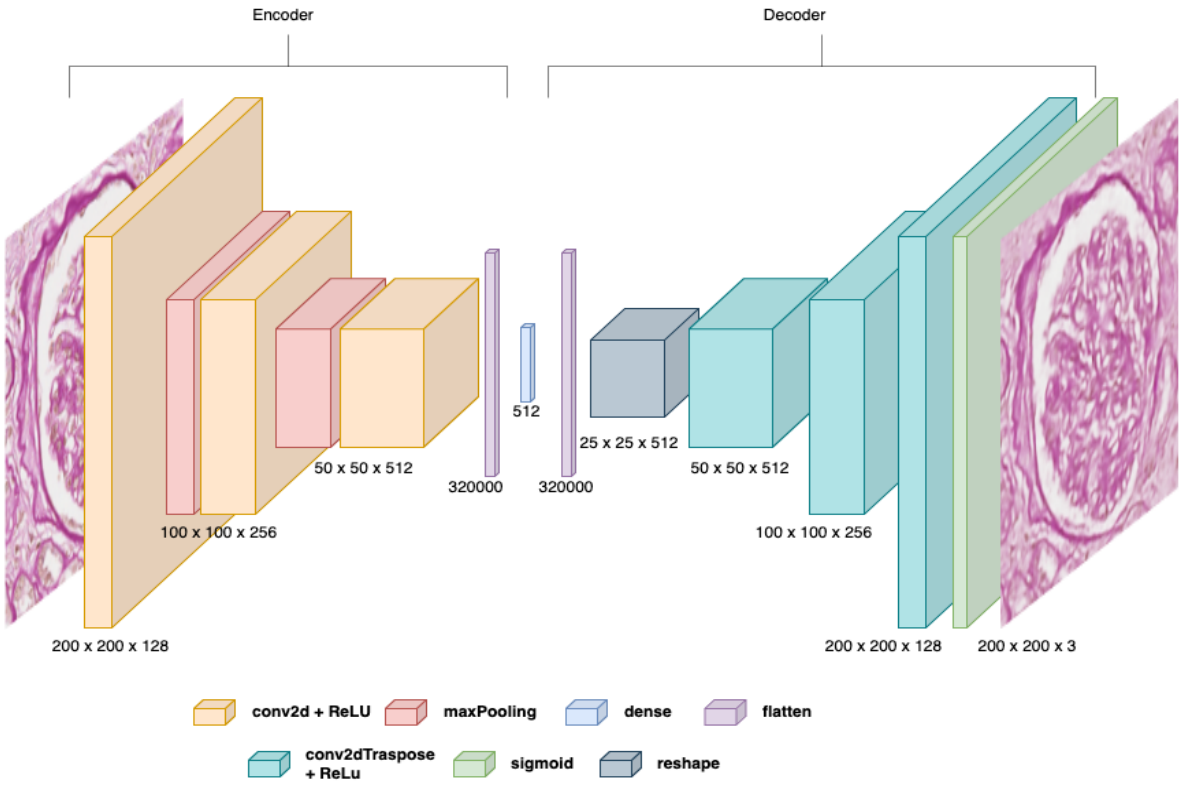

Fig. 4: Autoencoder architecture

other CNNs for its simplicity and interpretability given by its uniform architecture, where all convolutional layers have 3x3 kernels with stride 1 and padding 1, followed by 2x2 pooling layers with stride 2. Therefore, the convolutional layers are repeated in a similar way, but with a progressive increase in the number of kernels, which makes the learning of the feature complexity increasing gradually in the images. This can be advantageous for the feature extraction phase.
Specifically, it is composed by:

- Input layer;
- 16 convolutional layers, which can be divided into five convolutional blocks:
  - Block 1: 2 convolutional layers wit 64 kernels each of size 3x3 and a final MaxPooling with dimension 2x2 and stride 2;
  - Block 2: 2 convolutional layers with 128 kernels each of size 3x3 and a final MaxPooling with dimension 2x2 and stride 2;
  - Block 3: 4 convolutional layers with 256 kernels each of size 3x3 and a final MaxPooling with dimension 2x2 and stride 2;
  - Block 4: 4 convolutional layers with 512 kernels each of size 3x3 and a final MaxPooling with dimension 2x2 and stride 2;
  - Block 5: 4 convolutional layers with 512 kernels each of size 3x3 and a final MaxPooling with dimension 2x2 and stride 2;
- 3 fully connected layers with 4096, 4096 and 1000 neurons respectively. The final number of neurons refers to the classes of the training dataset, specifically ImageNet, described later. To introduce non linearity into the network, the ReLU activation function is applied between each fully connected layer. Finally, for the classification of the processed image, a Softmax function was performed.

ResNet50 is a deep convolutional neural network capable of classifying, detecting and localizing objects in even very complex images. Resnet introduces the concept of "residual blocks", which solve the problem of the "disappearance of the gradient". This problem occurs often in deep neural networks and leads to an incorrect update of the weights in the first layers. In a residual block, the input *"x"* follows a path called "*skip connection*" and it is added directly to the network's output, allowing the model to learn the differences and simplify training of deep neural networks.
In this specific case, ResNet contains 50 layers, with the following structure:

- Input layer;
- 1 convolutional layer with a kernel size of 7x7 and 64 different kernels each with stride 2;
- MaxPooling with stride 2;
- 9 convolutional layers, which are of 3 types, repeated 3 times. Specifically, there are convolutions with: 64 kernels of size 1x1; 64 kernels of size 3x3; 256 kernels of size 1x1;

- 12 convolutional layers, which are of 3 types, repeated 4 times. Specifically, there are convolutions with: 128 kernels of size 1x1; 128 kernels of size 3x3; 512 kernels of size 1x1;
- 18 convolutional layers, which are of 3 types, repeated 6 times. Specifically, there are convolutions with: 256 kernels of size 1x1; 256 kernels of size 3x3; 1024 kernels of size 1x1;
- 9 convolutional layers, which are of 3 types, repeated 3 times. Specifically, there are convolutions with: 512 kernels of size 1x1; 512 kernels of size 3x3; 2048 kernels of size 1x1;
- Global Average Pooling layer, which reduces the size of the extracted features to a fixed size;
- 1 final fully connected layer with 1000 neurons. Finally, for the classification phase, the Softmax function was performed.

For the feature extraction, have been used the just decribed VGG19 and ResNet50 models of the open source library *Keras*, both pre-trained with "ImageNet" dataset, which consists of about 14 million images with different subjects. The last three fully connected layers of the two neural networks are not included, achieving the purpose of feature extraction rather than image classification.
With the aim of better adapting these two pre-trained models to the dataset of this project, the fine tuning has been performed. In this phase, the models were modified, adding two fully connected layers, each with 4096 neurons and a ReLU activation function and a third fully connected layer with 2 neurons (representing the classes glomerular/ non glomerular) and a Softmax function for the final classification. Finally, the modified models are trained on the same training set of semantic segmentation made of 512x512 images, which are labeled based on the presence or absence of glomeruli (1 if at least one glomerulus is present, 0 otherwise). The adapted features were extrapolated by excluding the last three fully connected layers from the resulting tuned models.

D. Clustering

After the feature extraction process from glomeruli-containing images, the subsequent step involves clustering these features with the objective of classifying them based on their respective sclerosis statuses.
Dimensionality reduction plays an important role in enhancing the effectiveness and efficiency of clustering methods, facilitating a more precise identification and comprehension of data patterns. Nevertheless, it is crucial to execute this reduction accurately to avoid the loss of information during the process.

Nonlinear dimensionality reduction techniques were applied to deal with the analysis and visualization of high-dimensional data. It is also known as manifold learning and it aims to discover the underlying structure, or manifold, of the data and represent it in a lower-dimensional space while preserving important relationships and patterns. The main idea behind manifold learning is to learn a mapping from the high-dimensional input space to a lower-dimensional representation that captures the essential structure of the data. In particular, the isomap and t-sne techniques were used in our project.

Isomap algorithm, short for Isometric Mapping, aims to find a low-dimensional representation of data that preserves the geodesic distances between points in the high-dimensional dataset. Isomap constructs a nearest neighbour graph among data points and calculates geodesic distances between graph points. It then performs a dimensionality reduction technique like Multi-Dimensional Scaling (MDS) on these geodesic distances to obtain a low-dimensional representation of the data.

T-SNE, short for t-Distributed Stochastic Neighbor Embedding, measures similarity between high-dimensional points using a probability distribution based on a t-distribution. It then creates a low-dimensional representation where it minimizes the Kullback-Leibler divergence between the probability distributions of point pairs in the two-dimensional spaces. This process results in an arrangement of points in a low-dimensional space where similar points are close to each other.

Next, we clustered the reduced-dimensional features. **K-means** is a clustering algorithm that partitions a set of data points into homogeneous groups (clusters) based on their similarity. It requires specifying the number of clusters "K" beforehand, which can be a critical point and can influence the results. In fact, once the value of K is fixed, it assigns K centroids (representative points) randomly and iterates until convergence. During each iteration, it assigns each point to the cluster whose centroid is closest and then updates the centroids by calculating the mean of the points assigned to each cluster. The algorithm continues to iterate until the centroids stop changing significantly or a convergence criterion is met.
Silhuette score, Calinski-Harabasz Score and Davies-Bouldin Score were used as evaluation metrics both to identify the optimal number of clusters and to measure the 'goodness' of the clustering structure. They provide different perspectives on the effectiveness of K-means and help determine how well the data points are grouped into clusters.

Silhouette Score measures the quality of clustering by evaluating how well-separated the clusters are and how similar the data points within each cluster are to each other. The overall Silhouette Score for the entire dataset is the average of the Silhouette Scores for all data points and it ranges from -1 to 1. A higher Silhouette Score indicates that the clusters are well-separated, and data points within each cluster are similar to each other.

Calinski-Harabasz (CH) Score is a measure of cluster separation or compactness. It calculates the ratio of between-cluster variance to within-cluster variance, so an higher CH score indicates better-defined, well-separated clusters.

Metrics	Parameters	Equations
Silhouette Score	a is the mean intra-cluster distance; b is the mean nearest-cluster distance	$\frac{b-a}{max(a,b)}$
Calinski-Harabasz Score	B is the intra-cluster variance; W is the inter-cluster variance; N is the total number of data; k is the number of clusters.	$\frac{B}{W} \cdot \frac{N-k}{k-1}$
Davies-Bouldin Score	k is the number of clusters; S_i is the inter-cluster variance; $d(c_i, c_j)$ is the distance between the centers of cluster i and j.	$\frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i+S_j}{d(c_i,c_j)}\right)$

TABLE I: Performance metrics for clustering

Davies-Bouldin (DB) Score measures the average similarity between each cluster and its most similar cluster, relative to the size of the clusters. It helps to evaluate how distinct and well-separated the clusters are. A lower DB score indicates better clustering, where lower values imply that clusters are well-separated and distinct.

Table II reports the mathematical formula of each score and the related values obtained in our project.

IV. Results and discussion

A. Performance of semantic segmentation

SegNet and U-Net were applied to categorize kidney tissue in two classes: glomeruli and non-glomeruli.
Image visualization by itself is not sufficient to evaluate the efficiency of the CNNs. Therefore, some metrics are necessary to guarantee information about the correctness of glomeruli detection. In particular, Accuracy, Loss and MeanIoU have been considered in this paper and final results are reported in Table II.

Accuracy represents the proportion of correct predictions. Mathematically, it's computed as the ratio of the number of correct predictions to the total number of predictions. A higher accuracy indicates better results, with a maximum value of 1 indicating that all predictions are correct.

Mean Intersection over Union (meanIoU) is a metric used to assess the performance of semantic segmentation models which computes the amount of overlapping between the predicted segmentations and the ground truth across all classes. It consists of the average of the IoU values for each class, that is the ratio of the intersection (correctly predicted pixels) to the union (total pixels belonging to that class in either the prediction or ground truth) of the segmented areas. This is its corresponding formula, with TP, FP, FN indicating True Positive, False Positive and False Negative:

$$\frac{TP}{FP+FN+TP}$$

It ranges from 0 to 1, with higher values indicating better performances and in many contexts a meanIoU above 0.5 is considered good.

Categorical Cross Entropy is a loss function used in multi-class classification problems, where an instance can belong to only one among several possible classes. It measures the distance between two probability distributions: the "true" distribution (expressed as a one-hot vector representing the actual category) and the distribution predicted by the model (the probabilities for each class, as produced by, for instance, a softmax layer in a neural network). Mathematically, for a single instance, the Categorical Cross Entropy L between the ground truth y and the prediction p is given by:

$$L = -\sum_{i=1}^{C} y_i log(p_i)$$

- C is the total number of classes.
- y_i is 1 if the class is the ground truth and 0 otherwise.
- p_i is the predicted probability that the instance belongs to class i.

In table II, accuracy shows that SegNet and U-Net obtained valuable results, approximately 99% and 92%. Since these values represent the percentage of correct predictions, then almost the total of glomeruli has been well identified.
Loss, equal to 0.00869 and 0.44465 respectively, is close to 0 whereas MeanIoU, 0.98381 and 0.84870, is close to 1. According to these results, both methodologies produce predictions really similar to the ground truth but SegNet seems to perform slightly better than U-Net and visual results confirm that.

Non-Glomerulus / Glomerulus		
Metrics	SegNet	U-Net
Accuracy	0.99183	0.91815
Loss	0.00869	0.44465
MeanIoU	0.98381	0.84870

TABLE II: Metrics computed for glomeruli segmentation

Indeed, Figure 5 shows labels of glomeruli after SegNet and U-Net segmentation. Labels have been correctly identified but U-Net is generally less precise in highlighting the areas of interests.

B. Performance of clustering

Starting from the results of SegNet, we extracted features from images containing glomeruli. For feature extraction, all three approaches described in the section III were utilized: autoencoder, VGG-19, and ResNet-50. Subsequently, the K-means clustering algorithm was applied:

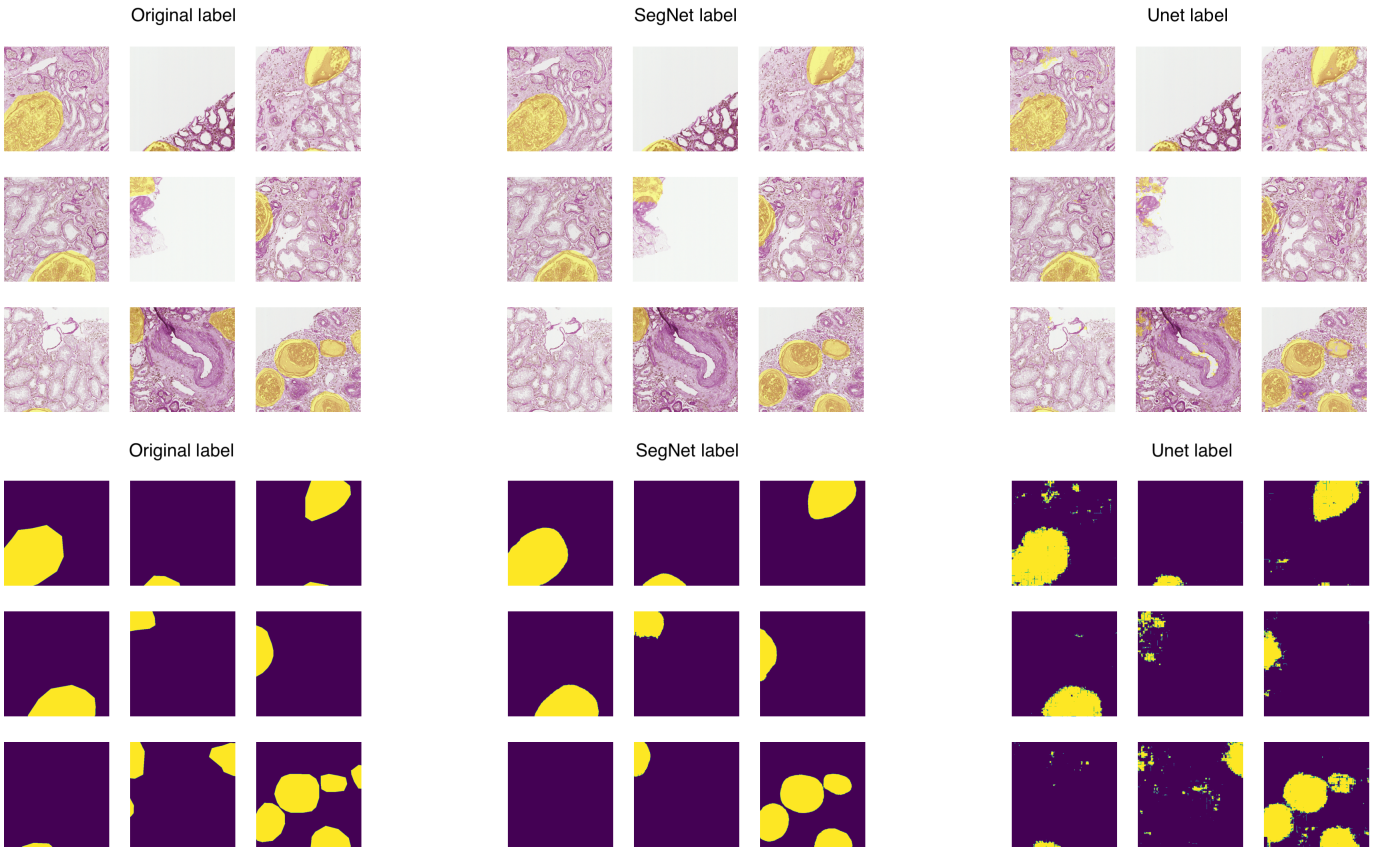

Fig. 5: Comparison of results between ground truth, SegNet and U-Net segmentation

- Directly to the previously obtained features;
- To the features reduced through Isomap;
- To the features reduced through T-SNE.

The optimal value of k was identified using the grid search approach. The results obtained from K-means, evaluated through the techniques described in the previous section (III), are presented in Tables III,IV and V, along with the corresponding optimal number of clusters.

Metrics	Autoencoder	Autoencoder Isomap	Autoencoder t-SNE
Silhouette Score	0.01955 $k=5$	0.03587 $k=5$	0.35767 $k=10$
Calinski-Harabasz Score	29.36729 $k=5$	27.80017 $k=5$	21.61575 $k=5$
Davies-Bouldin Score	3.64542 $k=10$	3.66478 $k=10$	4.55856 $k=7$

TABLE III: Results of K-means applied to features extracted with autoencoder

Metrics	ResNet50	ResNet50 Isomap	ResNet50 t-SNE
Silhouette Score	0.06706 $k=6$	0.06503 $k=7$	0.39665 $k=7$
Calinski-Harabasz Score	53.22224 $k=5$	52.37538 $k=5$	44.09793 $k=5$
Davies-Bouldin Score	2.85266 $k=10$	2.82006 $k=10$	3.12705 $k=10$

TABLE IV: Results of K-means applied to features extracted with ResNet50

Metrics	VGG-19	VGG-19 Isomap	VGG-19 t-SNE
Silhouette Score	0.07741 $k=6$	0.07617 $k=7$	0.39738 $k=5$
Calinski-Harabasz Score	66.21826 $k=5$	64.66937 $k=5$	59.64050 $k=5$
Davies-Bouldin Score	2.60363 $k=6$	2.60726 $k=5$	2.95712 $k=5$

TABLE V: Results of K-means applied to features extracted with VGG-19

In addition to the metrics, we also report in Figures 6, 7 and 8 some of the images belonging to the different clusters obtained for each of the 3 approaches.

Looking at the results obtained from different metrics and taking into account the context of glomeruli classification based on their disease severity, the unsupervised approach utilizing K-means clustering did not produce promising results. The clustering technique failed to effectively

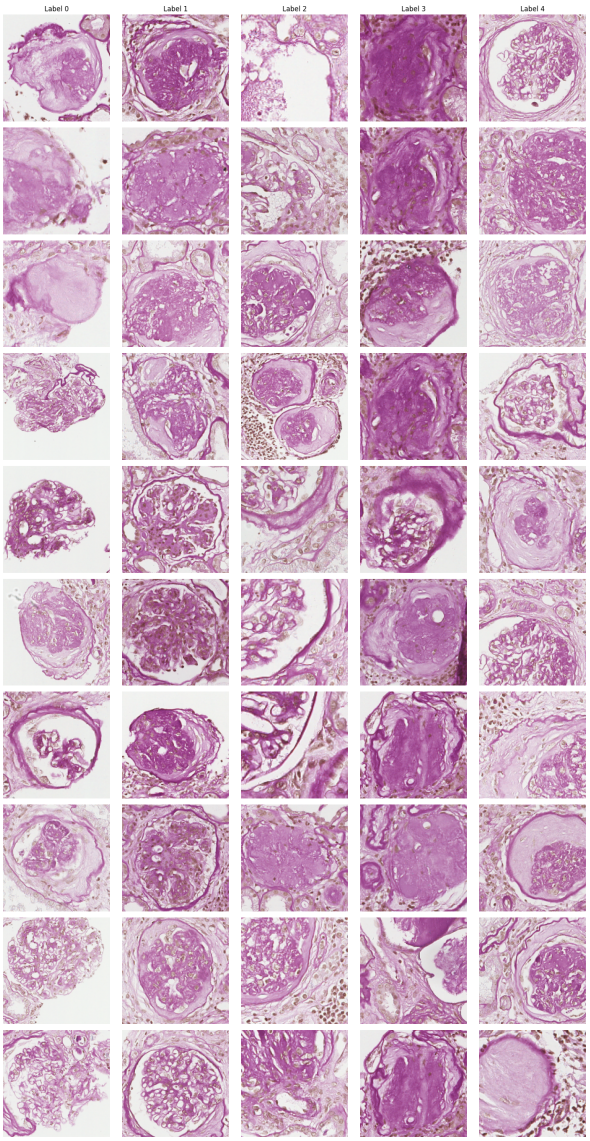

Fig. 6: Images from K-means (silhouette score) applied to autoencoder features reduced throw isomap

discriminate between glomerulosclerosis levels, indicating its limitations in this specific application. Investigating the limitations of the unsupervised approach further, it becomes apparent that the K-means clustering method, which relies only on the intrinsic structure of the data without external labels, faced significant challenges in this specific task of glomeruli classification.

Glomeruli images in the dataset exhibit high variability in terms of shape, size, and texture, which can lead to overlapping clusters. K-means assumes that clusters are spherical and equally sized. As a result, it struggled to create distinct and meaningful clusters in the feature space.

Furthermore, the absence of ground truth labels hindered the evaluation of clustering quality. Unlike supervised methods, where the performance can be quantified using metrics like accuracy or F1-score, unsupervised methods lack a clear benchmark for comparison. This made it challenging to assess

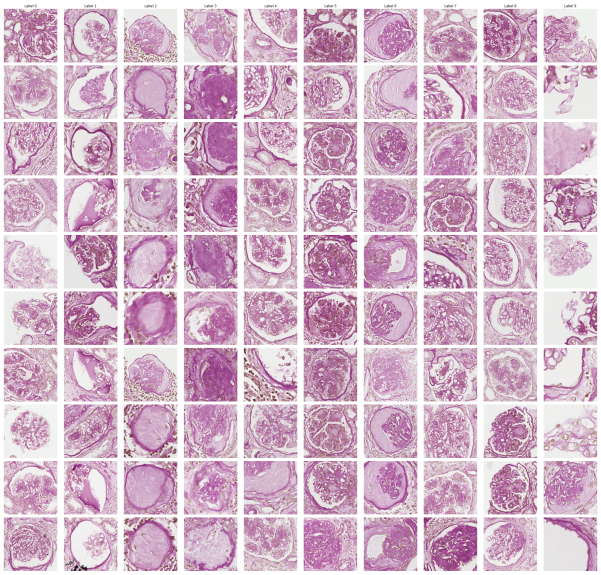

Fig. 7: Images from K-means (DB score) applied to ResNet50 features

the effectiveness of the K-means clustering in an objective manner.

Although the metrics do not offer interesting results, it can be interesting to visually observe the clusters obtained. In fact, in Figures 6, 7 and 8, it is still possible to notice a good coherence between the images belonging to each cluster. From our perspective, Figure 7, in particular, seems to exhibit a more coherent clustering of glomeruli. Given that this is an unsupervised approach, it is not possible to assign specific glomerulosclerosis levels to each cluster due to the absence of ground truth labels. In this context, involving a pathologist's expertise could offer valuable insights for a medically informed evaluation of the results.

V. Conclusions and Future Works

This article has presented a study on the application of supervised models for glomeruli detection and unsupervised models for classification.

For the first part, following previous studies (see [6]), semantic segmentation methods were applied and compared: Segnet and U-net. Although the former provided slightly better results than the latter, we can conclude that both achieved a good segmentation.

The method used for the glomeruli classification distinguishes this study from previous ones: given the absence of labels indicating the degree of the glomerulosclerosis, it was necessary to apply an unsupervised clustering technique, K-means. From the results reported in section IV it can be noted that K-means did not provide comfortable results, despite an initial feature extraction phase being carried out on the glomeruli images. Therefore, the nature of WSIs is maybe too complex and not ideal for unsupervised classification techniques.

In the future, this study could be continued by consulting a pathologist who knows how to evaluate the degree of scle-

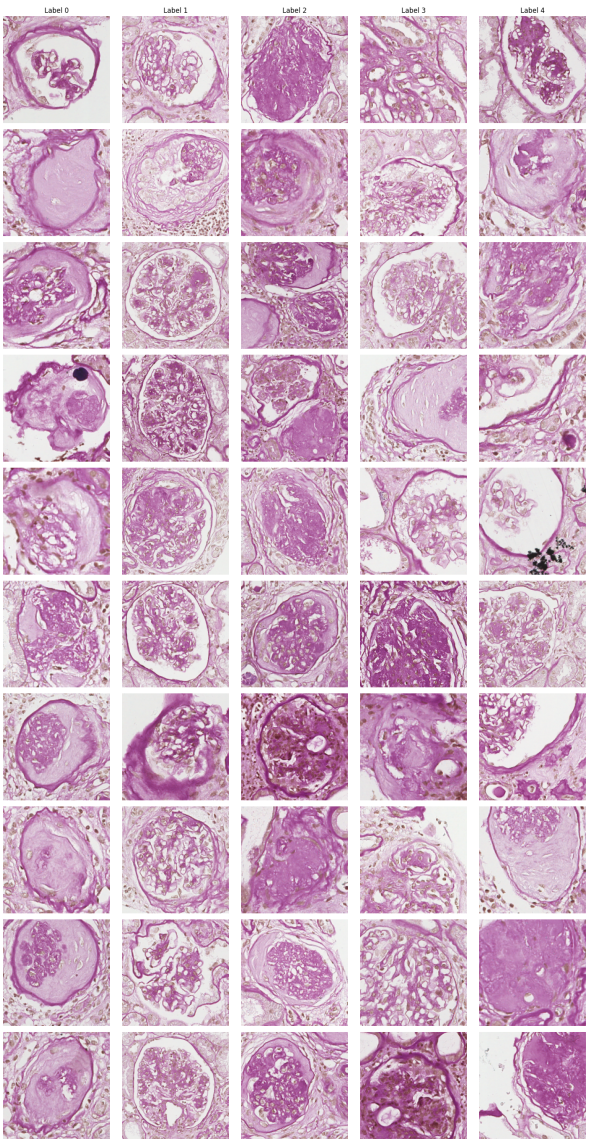

Fig. 8: Images from K-means (CH score) applied to VGG-19 features

rosis of the glomeruli, and consequently applying supervised methods, as reported in several studies (see [1], [6]).

REFERENCES

[1] M. T. R. Radica Z. Alicic and K. R. Tuttle, "Diabetic kidney disease: Challenges, progress, and possibilities," *Clinical Journal of the American Society of Nephrology*, 2017.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[3] G. Bueno, M. M. Fernandez-Carrobles, L. Gonzalez-Lopez, and O. Deniz, "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Computer Methods and Programs in Biomedicine*, vol. 184, p. 105273, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260719311381

[4] M. Shaha and M. Pawar, "Transfer learning for image classification," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 656–660.

[5] A. Chefrour and S. Drissi, "K-cae: Image classification using convolutional autoencoder pre-training and k-means clustering," *Informatica*, vol. 47, no. 7, 2023.

[6] G. B. M. M. F.-C. L. G.-L. O. Deniz, "Glomerulosclerosis identification in whole slide images using semantic segmentation," *Computer Methods and Programs in Biomedicine*, vol. 184, 2020.