POLITECNICO DI TORINO

MASTER's Degree in Biomedical Engineering

MASTER's Degree Thesis

Banff score: from clinical practice to computer-aided tools

Supervisors

Prof. SANTA DI CATALDO

Prof. FRANCESCO PONZIO

Dr. XAVIER DESCOMBES

Candidate

SONIA CIUFFREDA

DECEMBER 2023

Summary

The Banff classification is a comprehensive system that strives to standardize the pathological evaluation of kidney transplant. This process is particularly time-consuming and energy-intensive for pathologists, as each sample from the biopsies must be analyzed in detail. The aim of this project is to automate this process. In this work we first focus on the detection of tubules, vessels and nuclei. A color invariant classification of these structure was then performed.
A new database, composed by nineteen WSI images collected by the University Hospital Center (CHU) of Nice, was used. From the original data base, we selected and labeled a total of fifty-six 1024 x 1024 patches, nine per patient. Using the selected data we were able to obtain a train set, a validation set and a test set sufficiently representative of the entire population under analysis. Two new pipelines for the segmentation of blood vessel and tubule lumens, and nuclei are proposed. To classify the structures, supervised machine learning-based methods were utilized, implementing three classifiers: K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM). The novelty of our work lies in the method used to perform feature selection and feature extraction. To replicate the visual classification process carried out by a human, although using conventional and automated techniques, we worked in accordance with two expert pathologist to extract the optimal set of features. We obtain an accuracy of 91.03% and a balanced accuracy of 91.08% on the test set using the SVM classifier. Furthermore, to test the robustness of the method we applied the best performing model on a smaller data set representing the worst case. These images were selected by a skilled pathologist. To validate the no color dependency we also applied the best model on some TRI-stained images. In both cases we get acceptable performance. Although embryonic, this thesis work lays the foundations for the automation of the Banff classification providing an essential tool for increasing the number of diagnoses.

Table of Contents

List of Tables

List of Figures

Acronyms

ABMR
antibody-mediated chronic rejection

AMR
antibody-mediated rejection

AI
artificial intelligence

CHU
centre hospitalier universitaire

DL
deep learning

EM
electron microscopic

GAN
generative adversarial networks

GBM
glomerular basement membrane

KNN
k-nearest-neighbors

LCAP
laboratoire central d'anatomie pathologique

ML

machine learning

MLP

multi-layer perceptron

MV

majority voting

MVI

microvascular inflammation

NLP

natural language processing

PAS

periodic acid–schiff

PTC

peritubular capillaries

RF

random forest

SA

segment anything

SAM

segment anything model

SVM

support vector machine

TCMR

T-cell-mediated acute rejection

TMA

thrombotic micro-angiopathy

TRI

trichrome

WSI

whole slide images

Chapter 1

Introduction

1.1 The Banff classification

Nowadays, the amount of end-stage renal disease is constantly increasing. Renal transplantation represents an effective therapeutic route for people suffering from this particular disease [1]. However, a significant challenge associated with this medical intervention lies in the manifestation of allograft dysfunction. Various factors, including rejection, infection and drug toxicity, can contribute to the occurrence of allograft dysfunction. Statistically, this complication occurs in 50%-60% of cases following kidney transplantation [2].
The Banff classification aims to provide a guide for therapeutic approaches and to outline an objective endpoint for clinical trials on allograft rejection. The first Banff meeting took place in the modest town of Banff, Alberta, Canada, in August 1991.With the participation of 12 nephropathologists and transplant clinicians.The objectives were to guide therapy and to establish an objective endpoint for clinical trials [3]. After this first meeting, these meetings became biannual, leading to a continuous refinement of the standards used for classification [1].
The Banff process systematically identified and delineated lesions within the allograft parenchyma through a comprehensive and semi-quantitative approach. More specifically is possible to identify four steps: definitions of various components or lesions; diagnostic lesions; Semi-quantitative scoring of the lesions; Additional diagnostic features and categories [4]. It is essential to adhere to the criteria established during the Banff conferences in 1991 and 1997. Currently, it is a minimum of 10 glomeruli with at least two arteries, which fulfil the criteria for numerical coding, whereas, the presence of at least seven glomeruli with one artery is considered marginal [5]. It is also necessary to have at least seven slides of the same biopsy, of which three stained with haematoxylin and eosin (H&E), three stained with periodic acid-Schiff (PAS) and one stained with trichrome (TRI) [4].

Definition of various component of lesion

The pathological manifestations of rejection occur in four components of the kidney: glomerules, tubules, interstitium and vessels. Independently or jointly. The histological interpretation of allografts must also take into account factors related to the recipient and the donor. For instance, advanced glomerulosclerosis and tubulo-interstitial fibrosis may be present in the kidneys of elderly donors, and those of brain-dead donors may show ischaemic changes. Occasionally, donor-transmitted diseases may be evident in time-lapse or early biopsies. In addition, drug toxicity and infections may occur at any time after transplantation [4].

Diagnoses of the lesions

Depending on the component involved, different types of lesion may occur.Below is a list of the pathologies involved in the Banff classification, with the respective index used for representation in brackets.

Speaking about the **glomeruli**:

- **Glomerulitis (g)**: index of micro-vascular inflammation (MVI), and indicator of antibody activity and interaction with the tissue, particularly in cases of antibody-mediated rejection (AMR). The glomerulitis accouses when there is a complete or partial occlusion of one or more glomerular capillary by leukocyte infiltration and endothelial cell enlargement [6]. It is essential to note that glomerulitis can also occur in the context of recurrent or de novo glomerulonephritis. However, it is possible to exclude these two eventualities by application of immunotoxins and electron microscopic (EM) examination [7].

- **Mesangial matrix increase (mm)**: indicator of moderate mesangial matrix expansion assesses the proportion of glomeruli that exhibit moderate mesangial matrix expansion compared to all non-sclerosed glomeruli. This occurs due to a matrix expansion in the mesangial inter-space greater than the equivalent width of 2 mesangial cells on average in at least 2 glomerular lobules [5].At present, this Banff lesion score is not used to determine a diagnostic category and remains purely descriptive in nature.[7]

- **Transplant glomerulopathy (cg)**: lesion indicative of antibody-mediated chronic rejection (ABMR). It is characterised by doubling or multi-layering of the glomerular basement membrane (GBM). Similar features can be found in other conditions such as thrombotic micro-angiopathy (TMA) and membrano-proliferative glomerulonephritis associated with hepatitis C viral infection.

However, it is possible to distinguish cases of transplant glomerulopathy by the presence of IgM and C3 immunoglobulins [4].

With regard to **tubules**:

- **Tubulitis (t)**: index that assesses the extent of inflammation within the cortical tubule epithelium. Tubulitis is defined as the presence of mononuclear cells in the basolateral aspect of the renal tubule epithelium. When tubules are dissected longitudinally, the score is calculated by evaluating the average number of epithelial cells per tubular cross-section [7].

- **Tubular atrophy (ct)**: it assesses the degree of cortical tubule atrophy, a phenomenon closely linked to interstitial fibrosis. This is determined by the presence of tubules with a thickened basement membrane or a reduction in tubular diameter of more than 50% [7].

Of great relevance is also the **interstitium**:

- **Interstitial Inflammation (i)**: index of interstitial inflammation, assesses its extent in unhealed areas of the cortex, acting as a marker indicative of T-cell-mediated acute rejection (TCMR). To assess the i-index it is necessary to exclude fibrotic regions, the immediately sub-capsular cortex and the adventitia around the large veins and lymphatics. If more than 5%-10% eosinophils, neutrophils or plasma cells are present, an asterisk is added to this index [7][5].

- **Interstitial Fibrosis (ci)**: indicator of the degree of cortical fibrosis.This pathology only affects the cortex composed of fibrous tissue [7].

- **Total Inflammation (ti)**: it assesses the overall extent of cortical inflammation. In cases where at least mild interstitial fibrosis and tubular atrophy (IFTA) is present, this score is considered a better predictor than the Banff lesion score i for adverse graft outcomes [7][8].

- **Inflammation in Area of IFTA (i-IFTA)**: it estimates the extent of inflammation in the scarred cortex. It is used for the diagnosis of TCMR grade IA or IB in combination with Banff lesion score ti [7][9].

Finally there are the **vessels**:

- **Intimal Arteritis(v)**: it rates the existence and degree of inflammation within the arterial intima. Intima arteritis is specifically defined by the presence of inflammatory cells, predominantly lymphocytes and monocytes, in the sub-endothelial space of one or more arteries [6].This characteristic is observed in both TCMR and AMR [7].

- **Peritubular Capillaritis (ptc)**: it assesses the degree of inflammation of the peritubular capillaries (PTC). PTC includes microvascular inflammation as a feature of antibody-mediated active rejection or chronic active AMR. It may also be observed in cases of TCMR or borderline rejection. The score is determined by observing the severity of the most affected peritubular capillary [7].
- **Vascular Fibrous Intimal Thickening (cv)**: measures the degree of thickening of the intima of the most affected artery and not the average of all arteries [5]. It is important to note that this score does not distinguish between mild arterial intimal fibrosis and leukocyte-containing fibrosis, although the presence of the latter is more indicative of chronic rejection. Chronic rejection includes AMR and TCMR [7][9].
- **Arteriolar Hyalinosis (ah)**:it assesses the extent of arteriolar hyalinosis. Arteriolar hyalinosis is defined as a PAS-positive hyaline arteriolar thickening. An asterisk is added to the index in the case of arteriolitis [5]. It is important to notice that it is not currently used to determine a diagnostic category and remains purely descriptive [7].
- **Hyaline Arteriolar Thickening (aah)**: Alternative index to quantify arteriolar hyalinosis introduced by the poor reproducibility of Banff score index ah [10]. It focuses on circumferential or non circumferential hyalinosis by considering the number of arterioles involved. Similar to the Banff lesion score ah, aah is not currently used to establish a diagnostic category and remains purely descriptive [7].
- **C4d**: Assesses the extent of staining for C4d. C4d is a score determined by the percentage of peritubular capillaries and vas recta that exhibit a linear and circumferential staining pattern. The evaluation is performed by immunofluorescence (IF) on frozen sections of fresh tissue or immunohistochemistry (IHC) on formalin-fixed, paraffin-embedded tissue [7].

Semi quantitative scoring of the lesion

The Banff scoring system has three grades: mild (1), moderate (2), and severe (3). Table 1.1 reports the percentage of component (between glomeruli, tubules, vessels and interstitium) involved in the pathology. There are some exception like for Transplant glomerulopathy (cg) in which is evaluated the amount of double contours of the GBM in peripheral capillaries, using light microscopy (LM) or EM [7]. For the Tubulitis (t) is considered the number of mono-nuclear cells in Foci across section. Is possible to define a t3 condition also for the presence of more than two areas of tubular basement membrane destruction accompanied by i2/i3

inflammation and t2 elsewhere [7]. For Intimal Arteritis (v) is evaluated the severity of the pathology in at least one arteria cross section [7]. For Peritubular Capillaritis (ptc) the number of leukocytes in most severely involved PTC is considered [7].For the Vascular Fibrous Intimal Thickening (cv) is considered the percentage of fibrointimal thickening in the luminal area [7]. For Arteriolar Hyalinosis (ah) the severity of the PAS-positive hyaline thickening in at least 1 arteriole is evaluated [7]. Considering Hyaline Arteriolar Thickening (aah), is adopted the following grading: aah0 = No typical lesions of calcineurin inhibitor-related arteriolopathy; aah1 = replacement of degenerated smooth muscle cells by hyaline deposits in only 1 arteriole, without circumferential involvement; aah2 = Replacement of degenerated smooth muscle cells by hyaline deposits in more than 1 arteriole, without circumferential involvement; and aah3 = replacement of degenerated smooth muscle cells by hyaline deposits with circumferential involvement, independent of the number of arterioles involved [7]. Finally for C4d is taked the percentage of PTC and medullary vasa recta [7][9][11].

Pathology	**intex**	**grades**
Glomerulitis	g	g0= no glomerulitis g1 = less than 25% g2 = between 26% and 75% g3 = more than 75%
Mesangial matrix increase (for non sclerotic glomeruli)	mm	mm0 = No more than mild mesangial mm1 = less than 25% mm2 = between 26% and 50% g3 = more than 50%
Transplant glomerulopathy	cg	cg0 = no GBM observed cg1a = incomplete in at least 3 capillaries cg1b = between 1% and 25% cg2 = between 26% and 50% cg3 = more than 55%
Tubulitis	t	t0 = No mononuclear cells t1 = 1-4 t2 = 5-10 t3 = more than 10%
Interstitial Inflammation (in not scarred cortical parenchyma)	i	i0 = less than 10% i1 = between 10% and 25% i2 = between 26% and 55% i3 = more than 50%
Interstitial Fibrosis (in cortical area)	ci	ci0 = less than 5% ci1 = between 6% and 25% ci2 = between 26% and 50%

		ci3 = more than 50%
i-IFTA (in scarred cortical parenchyma)	i-IFTA	i-IFTA0 = less than 10% i-IFTA1 = between 10 % and 25% i-IFTA2 = between 26% and 50% i-IFTA3 = more than 50%
Total Inflammation (in cortical parenchyma)	ti	ti0 = no or trivial inflammation ti1 = between 10% and 25% ti2 = between 26% and 50% ti3 = more than 50%
Intimal Arteritisa in cortical parenchyma	v	v0 = no arteritis v1 = mild v2 = Severe v3 = Trans-mural arteritis and/or arterial fibrinoid change
Peritubular Capillaritis	ptc	ptc0 = less than 3 ptc1 = between 3 and 4 ptc2 = between 5 and 10 ptc3 = more than 10
Vascular Fibrous Intimal Thickening	cv	cv0 = no chronic vascular changes cv1 = less than 25% cv2 = between 26% and 50% cv3 = more than 50%
Arteriolar Hyalinosis	ah	ah0 = no (PAS)-positive ah1 = Mild ah2 = Moderate ah3 = Severe
C4d	C4d	C4d0 = 0% C4d1 = between 0% and 10% C4d2 = between 10% and 50% C4d3 = more than 50%

Table 1.1: Banff's semi quantitative scoring of the lesion [7][9][11].

Additional diagnostic features and diagnostic categories

The most recent Banff classification has six categories.Rejection categories are Category 2, Category 3, and Category 4 [4][7]:

1. **Normal biopsy or non-specific changes**: to fall into this category, all other categories must be excluded [7].Cases in which the graft has no inflammatory cells or features of acute tubular injury (ATI) and acute tubular necrosis fall into this category [1].

2. **Antibody-mediated changes**: The diagnosis of active and chronic antibody-mediated rejection (ABMR). Morphological, immunohistological and serological tests are reviewed to prove the veracity of this category [4]. While the morphological features, indices of active and chronic lesions, are unique. Active lesions can coexist in chronic active ABMR.In contrast, immunohistological and serological indicators are the same for active and chronic ABMR [7].It is possible to divide this class into two sub-classes referring to acute and chronic TCMR cases. Specifications regarding the indices are given in Table 1.2.

	morphologic	**immunologic**	serologic
Acute ABMR	g≥0 or ptc≥0 v≥0 or acute TMA or acute tubular injury	lienar C4d in PTCs g+ptc ≥2	DSAs
Cronic Active ABMR	cg if no TMA≥0 or arterial intimal fibrosis	lienar C4d in PTCs g+ptc ≥2	DSAs
Chronic ABMR	cg ≥0 if no TMA severe PTC	priopr diagnosis of acute or chronic active ABMR	DSAs

Table 1.2: Banff scores for antibody-mediated changes [4].

3. **Suspicious for acute T cell-mediated rejection**: This category specifically concerns the tubulo-interstitial type of rejection. The true clinical significance of this lesion remains a matter of debate and is highly dependent on factors such as the organ donor, time since transplantation, and the indication and timing of kidney allograft biopsies [1][12][13]. One always falls into this category if one has mild (i1) to severe (i3) interstitial disease and mild tubulitis (t1), or vice versa, and in the presence of the absence of arteritis (v0) [1].

4. **T cell-mediated rejection** : the diagnosis of acute and chronic T-cell-mediated rejection is mainly based on the presence of active inflammation in the tubules, interstitium, and non-atrophic vessels [4].Again, a distinction can be made between chronic and acute cases. The sub-classes and their indices are given in Table 1.3.

Acute TCMR	Banff score
Type 1A	t2i2 or t2i3
Type 1B	t3i2 or t3i3
Type 2A	v1
Type 2B	v2
Type 3	v3

Chronic active TCMR	Banff score
Grade 1A	t2,t1$\geq$2 and i-IFTA$\geq$2
Grade 1B	t3,t1$\geq$2 and i-IFTA$\geq$2
Grade 2	cv1,cv2 or cv3

Table 1.3: Banff scores chronic and acute TCMR [4][7].

5. **Interstitial fibrosis and tubular atrophy**: The IFTA is caused by two conditions: the interstitial fibrosis and the tubular atrophy.Interstitial fibrosis is characterized by excessive accumulation of connective tissue in the interstitium, while tubular atrophy manifests as a reduction in size and functions of the renal tubules.These conditions are often associated with persistent inflammatory processes or repeated renal damage over time. Table 1.4 shows the relative Banff scores.

IFTA severity	Banff score
mild	ci1 or ct1
moderate	ci2 or ct2
severe	ci3 or ct3

Table 1.4: Banff scores for IFTA [7].

6. **Non-rejection conditions/diseases**: often, transplant non-rejection is associated with different types of pathology.Therefore, the necessity of introducing this new class into the Banff classification arose. Was so introduce this sixth Banff category as a separate sheet [3]. It is indeed essential to fully understand the conditions of non rejection for precise diagnostic delineation and tailored therapeutic interventions, especially in the context of transplant dysfunction in the post-transplant environment [4].

1.2 The AI role

In the clinical context, the analysis of the Banff classification is characterised by considerable complexity. A significant workload is required to obtain essential indices for this classification.Moreover, the presence of numerous qualitative factors often introduces considerable variability both within and between practitioners.
To reduce computing costs and improve the accuracy and the objectivity of classifications, Artificial Intelligence (AI)-based tools are a great instrument.The use of such approaches could not only reduce the time needed to obtain a diagnosis, but also improve its consistency and reliability.
Several studies have attempted to replicate some of the steps of Banff classification through AI-based approaches.

One of the most complete work was done by [14]. A convolutional neural network (CNN) for the histological analysis of renal tissue stained with PAS was developed and validated.A CNN is a specific type of deep learning neural network that is particularly suited for image analysis and computer vision [14][15][16]. The CNN was initially designed to identify five classes (sclerotic glomeruli, proximal tubules, distal tubules, atrophied tubules and interstitium) in biopsies of healthy and pathological kidney tissue [14]. Subsequently, a comparison between the CNN quantification and the elements of the Banff grading system, manually assessed by several renal pathologists, was performed [14].
A U-Net architecture was used for segmentation [17]. The network was trained for 100 epochs, with 300 iterations per epoch and batches of 6 patches. Spatial augmentation (rotation, flipping, elastic deformation, zoom) and colour augmentation (brightness, contrast, saturation, hue shift, Gaussian noise, Gaussian blur) techniques were applied [14][18].Adam was used as the learning rate optimisation algorithm and categorical cross entropy as the loss function [14][19]. Specifically, five U-Net were used. The probability per pixel for all five networks was averaged and only the class with the highest probability obtained was considered [14].Finally, post-processing steps were implemented .Image 1.1 represents the obtained confusion matrix. Subsequently, quantitative and morphometric data were extracted for each class. The sum of the objects labelled as glomeruli and sclerotic glomeruli constituted the glomeruli count. To evaluate the percentage of interstitium area, the number of pixels labelled as interstitium was divided by the total number of segmented pixels [14].Finally, the percentage of atrophied tubules was determined by dividing the number of objects labelled as atrophied tubules by the sum of objects labelled as one of the four tubule classes. For the glomerular count, the Intraclass Correlation Coefficients (ICC) between the pathologists and the CNN was calculated [14]. The ci, ti, ct and IFTA classification scores assigned by expert pathologists were then compared with the results obtained via CNN. To do so,

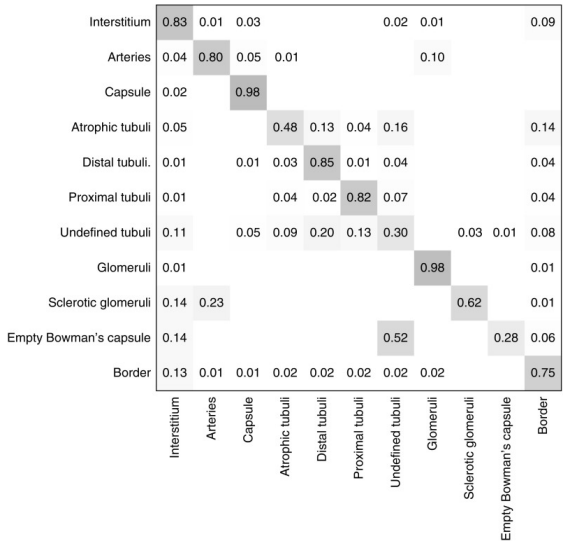


Figure 1.1: Confusion matrix form the segmentation performed by [14]

Spearman's correlation coefficient and coefficient of determination (R2) were calculated [14]. Tables 1.5 and 1.6 show respectively the ICCs for glomerular counting by three pathologists (P1–P3) and the CNN, and the Spearman correlation coefficients for quantification by the CNN and the average visual scores of multiple pathologists for relevant Banff components [14].

pathologist	**ICCs**
P1	0.94
P2	0.96
P3	0.93

Table 1.5: ICCs for glomeruli [14]

	intertubular area	**ci**	**ti**	**IFTA**	**ct**
interstitium	0.81	0.55	0.71	0.33	/
atrophied tubular area	/	0.62	/	0.58	0.58

Table 1.6: Spearman correlation coefficients. The '/' mean that the parameter were not evaluated [14]

[15] proposed a different approach. CNNs for automatic classification of kidney allograft biopsies were developed.Specifically, two sequential CNNs were trained to distinguish between normal (Banff category 1) and disease (all other Banff categories) [15]. The first CNN makes this initial distinction while the second aims

to distinguish rejection (Banff categories 2-4) from other diseases, including Banff category 5. Several CNN architectures were tested, including ResNet18, ResNet50, ResNet101, ShuffleNet and Inceptionv3 [15].The Inceptionv3 architecture was chosen.Figure 1.2 reports the AUROCs coming from the application of the two CNNs on an external data set [15]. The first serial CNN achieved AUROCs of $0 \cdot 83$ for the normal class and $0 \cdot 83$ for the disease class.The second serial CNN generalised less well with AUROCs of $0 \cdot 61$ for the other diseases class and $0 \cdot 61$ for the rejection class [15].

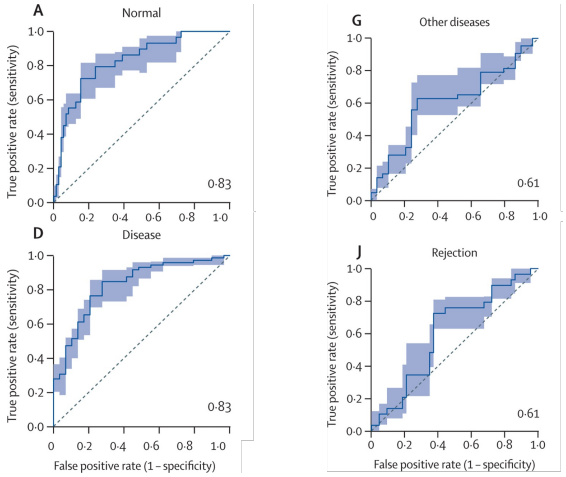


Figure 1.2: Performances obtained by [15].

Instead of proposing an automated method to perform the Banff classification, [20] attempted to create a decision aid system. The aim was to distinguish normal and abnormal renal tubules by developing a U-Net- based segmentation model.Initially, the segmentation of two classes was implemented: glomeruli, normal tubules, abnormal tubules, arteries and interstitium. The ability of the model to distinguish between normal and abnormal tubules was tested.To evaluate the performance of this first stage, Dice coefficients (DC) were calculated [20].The highest DCs were obtained for the interstitium and glomeruli, indicating a strong segmentation performance. Normal and abnormal tubules showed intermediate DCs, while arteries showed lower DCs [20].
In the next step, the number of classes was extended to eight (glomeruli, proximal tubules, distal tubules, atrophied tubules, tubulitis, degenerated tubules, arteries and interstitium) with the aim of assessing the ability to detect various types of abnormal tubules [20]. The evaluation of segmentation performance was again conducted using Dice coefficients (DCs). The highest DCs were observed for the interstitium and glomeruli, as well as for the five classes involved in semantic segmentation [20]. Proximal tubules, distal tubules, atrophied tubules and degenerated

tubules had intermediate DCs, while arteries and tubulites had lower DCs [20].
The final step is to evaluate the actual usefulness of the obtained segmentations. To do this, the concordance ratio between two nephrologists experienced in renal pathology was examined, with and without the aid of U-Net segmented images [20]. The first evaluation was conducted without the aid of automatic segmentations (U-Net- group), while the second evaluation included their use (U-Net+ group) [20]. ICCs values were calculated to assess the agreement ratio for continuous variables, and Cohen's κ for categorical variables [20].The values of these two coefficients are shown in the Table 1.7.

	U-Net- k	group ICC	U-Net+ k	group ICC
glomerular count	-	0,97	-	0.95
t score	0,92	-	0,90	-
ct score	0,+1	-	0,95	-
ci score	0,91	-	0,82	-
% tubulitis	-	0,14	-	0,52
% tubular atrophy	-	0,28	-	0,76
% degenerative tubules	-	0,18	-0,17	
%interstitial space	-	0,59	-	0,81

Table 1.7: Agreement ratios between renal pathologists with and without U-Net-segmented images [20]

A completely different approach was adopted by [21].A computerised decision support system was developed to translate all Banff classification rules and potential diagnostic scenarios into a computer algorithm. This algorithm is capable of automatically diagnosing kidney allograft biopsies [22][21]. It is important to notice that the developed system is not an AI system, but a sophisticated 'if-then' algorithm [22]. The algorithm was subsequently tested by reclassifying both adult and paediatric kidney biopsies.In the adult kidney transplant population, the Banff automation system demonstrated a significant impact by reclassifying 29.75%of antibody-mediated rejection cases and 54.29% of T-cell-mediated rejection cases into alternative diagnostic categories. In contrast, 7.32% of biopsies initially diagnosed as non-rejection by pathologists were reclassified as rejection cases by the Banff automation system [22][21]. In addition, 7.30% of adults initially diagnosed as non-rejection were reclassified into various types of rejection diagnoses using the Banff automation system [22][21].
In the pediatric population, the reclassification rates into other diagnostic categories were 30.77% for antibody-mediated rejection and 30.77% for T-cell-mediated rejection [22][21].

1.3 Thesis scope

The fundamentals to fully understand Banff classification from a medical point of view were presented.Analysing the clinical problem, the magnitude and complexity of developing an algorithm capable of performing the Banff classification automatically becomes clear. This type of classification presents significant intra- and inter-operator variability. At the diagnosis stage, it is not enough to identify the various renal tissues and related pathology, but it is also crucial to determine the degree of these pathology and to establish the relationships between them.
Examination of images obtained from biopsies reveals a wide range of renal tissues and components showing abnormal behaviour. While a human operator is able to select only the relevant elements for analysis, excluding what could be considered as non-relevant abnormalities, replicating this process with a non-human operator is complex.
A further obstacle is the limited availability of studies focusing on the Banff classification. Most of the proposed methods focus on a single pathology included in the classification, facilitating the work of pathologists but not fully implementing the classification itself.Furthermore, many approaches make use of supervised Deep Learning (DL) systems. The absence of a ground truth for all collected data prevents such designs from being exploited to improve performance. In order to address these challenges, the implementation of an algorithm that emulates the decision-making process of a human operator is proposed. This requires the identification and discrimination of all components present in a renal histopathology image, followed by the association of the relevant pathologies through analytical rules defined in collaboration with experienced pathologists.

The renal histopathological image shows several elements that can be grouped into five main categories: tubules, blood vessels, epithelial cells (nuclei), glomeruli and interstitial tissue.The aim is to recognise four of these elements and deduce the fifth by exclusion. Given the availability of a method for identifying and classifying glomeruli, the focus has been on the segmentation and distinction of tubules, blood vessels and nuclei. It is then possible to identify interstitial tissue by exclusion.Specifically, the aim of this project is to provide a tool for the identification of tubules, blood vessels and epithelial cells.

Chapter 2

Materials and methods

2.1 The data sets

To compute the BANFF classification, histopathological images are required. These images are used to analyze pathological tissue alterations with histological cuts. Specifically, within the scope of the BANFF classification, we consider biopsies of the kidney.
A total of 18 Whole Slide Images (WSIs) were utilized, provided by the Central Laboratory of Pathological Anatomy (LCAP) at the University Hospital Center (CHU) in Nice. WSIs are multi-scale high-resolution digital images obtained from slides that can be examined through optical microscopes.Individual sections were obtained through biopsies involving 9 individuals affected by various pathologies, including: humoral rejection, interstitial fibrosis and grade 1 tubular atrophy with vascular lesions, arteriolar lesions, isolated fibrotic and vascular intimal lesions, interstitial fibrosis and chronic active redox, tubulo-interstitial lesions, chronic lesions with vascular lesions, rare lesions of acute tubular necrosis.Pathological tissue was sampled using the microtome Leica RM2245, providing sections of 3-5 μm.Subsequently, each sample was processed through specific staining techniques, such as Periodic Acid-Schiff (PAS) or trichrome staining (TRI). Finally, each slide was scanned and digitized using the Leica Aperio AT2 scanner.

2.1.1 Data collection

To collect the data, I personally visited the CHU, where I had the opportunity to know the reality of LCAP and meet all the pathology personnel working in the laboratory. Given my lack of prior knowledge in the field of histopathology, my initial priority was to acquire the basic knowledge for the correct analysis of samples from both a medical and technical point of view.
For this purpose, the laboratory is equipped with a multi-head microscope, an optical microscope with a tubular structure that connects to multiple body tubes.This device is typically used in an educational contexts, enabling several individuals to simultaneously observe the same slide. One user operates the microscope, moving within the section, while all others observe. A brief introduction to the use of this device was provided.

Subsequently, a detailed analysis of various pathological sections was conducted. After a brief explanation of the methods used to obtain the samples, the main features of five structures were examined: blood vessels, tubules, glomeruli, nuclei, and interstitial tissue. This step was crucial to conduct an initial visual classification and facilitate the identification of the key features necessary for the future implementation of the classifier. Following this, the focus shifted to understanding how these features may vary to recognise a specific pathology and its degree of advancement.
At this point, the samples for the actual analysis were selected.The laboratory holds a database containing all the slides analyzed, with their respective medical reports. To identify the optimal subjects, each report was meticulously examined to create a database that would be as representative as possible of the average population of samples reaching the laboratory. From this research, 19 samples were chosen, 9 subjected stained with PAS staining and 9 with TRI ones.

To fully understand why two separate staining methods are used, it is essential to have a clear understanding of the glass preparation process. Starting with the tissue sample obtained by biopsy, multiple sections are extracted at various depths. In particular, the laboratory uses the microtome Leica RM2245, which is capable of producing sections with a thickness of 3-5mum. These sections allow a three-dimensional view of the organisation of the components within the biopsy sample.
Subsequently, depending on the specific attributes to be highlighted, each section is subjected to a staining process involving an interaction between the sample and various substances that modify tissue coloration. In particular, the PAS and TRI staining methods are introduced. PAS staining highlights the presence of glycogen and gives a magenta hue to the tissue, facilitating the identification of the basement

membrane surrounding the renal tubules and allowing better differentiation from the vessels. TRI staining, on the other hand, uses three dyes to turn collagen-rich tissues blue and erythrocytes and muscle tissue red. This second staining method therefore allows fibrotic tissue to be clearly observed.

Finally, the selected slides were examined using the microscope under the guidance of an experienced pathologist to validate the quality of the samples and exclude the presence of artefacts. Often, during staining and slide preparation, may occur areas with excessive dye concentration or other artefacts resulting from overexposure of the specimen. Other times, the slide may contain a minimum of tissue adequate for medical analysis, but inadequate for the extraction of an adequate set of images. Once the selection was validated, the individual slides were scanned with the Leica Aperio AT2, which can provide high-resolution digital representations.These images are referred to as Whole Slide Images (WSI) and have an average size of 129368 x 59379 pixels in .SVS format.Image 2.1 represents the used microtome,a multi-head microscope and the used scanner.

(a) **(b)** **(c)**

Figure 2.1: Microtome (a), multi-head microscope (b), and scanner (c) used by the laboratory.

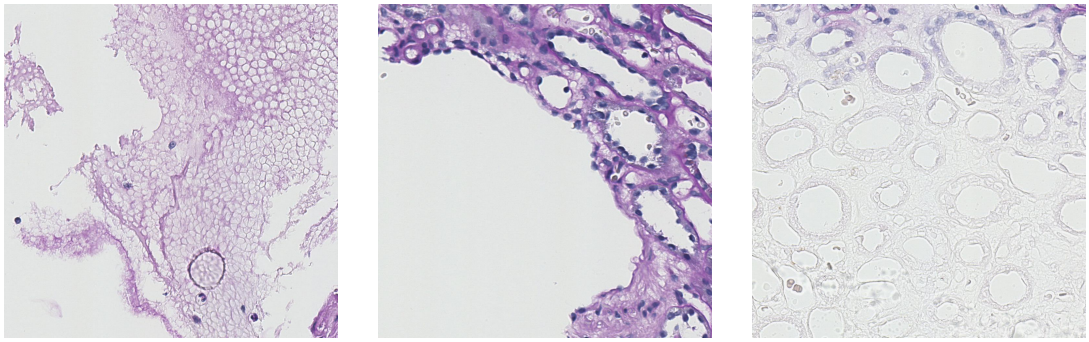
2.1.2 Data preparation

The available database consists of 19 whole slide images (WSI) in .SVS format.This format is commonly used for medical and histological images. In addition to providing high resolution, this file extension allows the data to be managed and analysed in a similar way to a microscopy tool. The .SVS data are defined as multilevel, i.e.they allow the image to be viewed at various resolutions or zoom levels. Using specific visualisation tools, it is possible to view the image at different zoom levels without loss of detail or sharpness.

Furthermore, .SVS files can include metadata such as the location of a particular portion of the image, the type of scanner used or even patient data. Therefore, they are an essential tool in medicine for data storage and diagnostic purposes. In particular, all images were processed using the highest resolution level, with an average size of 129368 x 59379 pixels. Napari was chosen for visualisation. Each image faithfully represents the entire composition of the tissue sample, showing different sections of the same biopsy sample taken a few mum apart, separated by a white background.
Due to the high number of pixels and the presence of a white background, it was necessary to divide the image into patches and remove portions that represent exclusively or predominantly the background. The selection criteria refer to three parameters:

- **m**: Maximum value between the standard deviations evaluated for each channel of the RGB image. This parameter allows the removal of patches with uniformly the same color throughout the entire image (fig.2.2a).
- **k**: Fraction of pixels in the image with average values for each channel above a certain threshold. In particular, the threshold is set at 220. This makes it possible to remove areas with a certain percentage of very light pixels and thus a certain percentage of background.(fig.2.2b).
- **z**: The average of the minimum values for each channel in the RGB image. Z allows the control of the image's brightness (fig.2.2c).

Subsequently, limit values were set for each of the three parameters, specifically m < 4.5, k > 0.6, and z > 240. If at least one of these conditions is respected, the image is not selected. Initially, square patches of the typical medical imaging size, i.e. 512 x 512 pixels, were extracted.However, considering the project's goal and the visual analysis of the individual patches, it was decided to double the size to obtain patches of 1024 pixels per side.This final size represents a good compromise, as it does not result excessive computational expensive, and at the same time preserves most of the structures in the image. Figure 2.3 represents the reconstruction of a



(a) (b) (c)

Figure 2.2: Example of not selected patches during the first data selection step

1024 x 1024 patch using four 512 x 512 patches and the corresponding 1024 x 1024 image. Some biological structures are divided across the smaller patches which may adversely affect their identification and classification. Implementing an alternative patch splitting strategy may be a more effective way to maintain the integrity of these structures in the image.

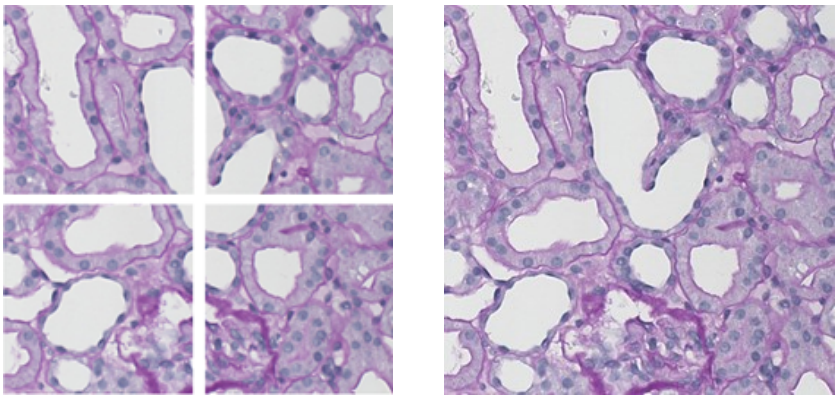


Figure 2.3: Reconstruction of a 1024 x 1024 patch using 512 x 512 patches and the corresponding 1024 x 1024 patch.

2.1.3 Data selection

The morphology of the kidney reminds one of a bean.The outermost layer is known as the renal capsule, adjacent to which is the cortical layer, followed by the medullary layer. In the context of the BANFF classification, the layer involved is the cortical layer, which makes the ability to distinguish and isolate the various layers that make up the kidney crucial.
During a biopsy procedure, the pathologist extracts a portion of kidney tissue transcutaneously using a needle, collecting both cortical and medullary tissue.

Because of the shape of the kidney, the exact sequence of these tissues within the specimen cannot be predicted. Often the section contains a portion of medullary tissue enclosed between two layers of cortical tissue.However, in some cases, there is a clear separation between the cortical and medullary layers.Figure 2.4 shows two different ways to extract kidney's tissue portions trow biopsy. A simple method

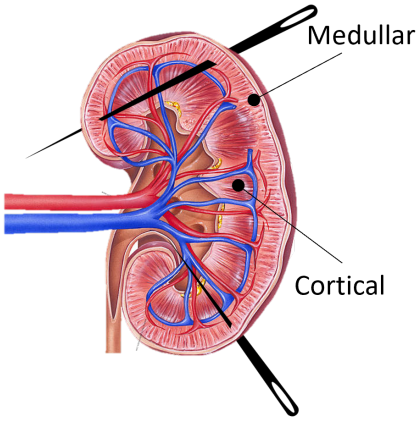


Figure 2.4: Biopsy procedure

to rule out the presence of cortical tissue is the presence of glomeruli, as these structures are found exclusively in the medullary region. Figure 2.5 shows a biopsy specimen in which the succession of cortical tissue and medullary tissue containing glomeruli can be observed. Using this information, samples belonging to cortical tissue could be excluded. A second selection was performed by selection the appropriate staining, particularly for the final proposed approach. For the SAM approach, it was possible to use all available data, since SAM is not affected by the type of staining.On the other hand, for the final approach, both segmentation and classification methods involve the use of thresholds and the extraction of features that are highly dependent on the staining type. Therefore, it was necessary to proceed using a single stain type. The model previously implemented by within the project was trained on PAS-stained images; thus, the decision was made to focus solely on PAS staining for this segmentation, at least initially, abandoning TRI staining.

In addition, during the implementation of the supervised machine learning models, it was necessary to manually annotate the images. As a result, a data set reduction was implemented to obtain a smaller data set that was representative of the population under analysis.

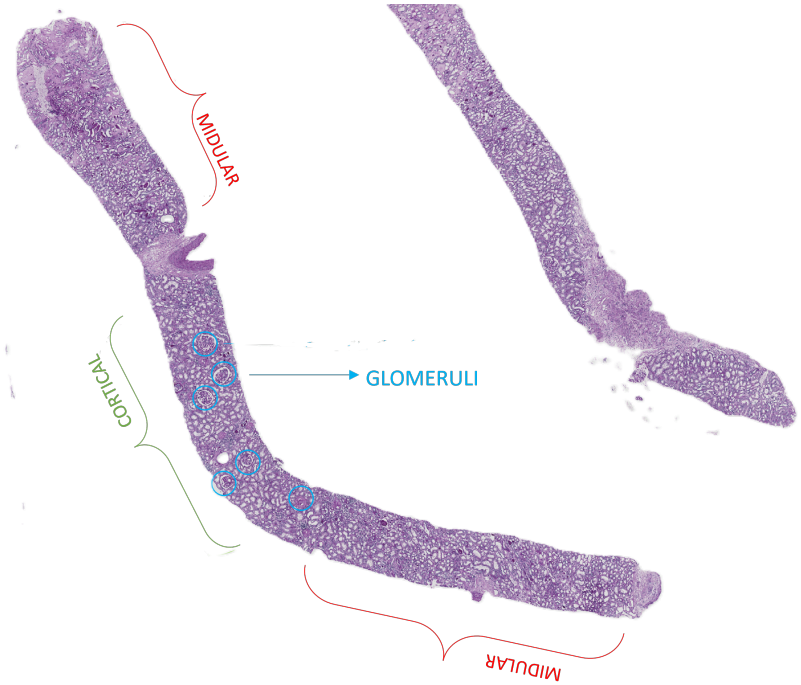


Figure 2.5: Biopsy sample

2.2 SAM approach

A first task was to evaluate the deep learning approach to segment our images using the recent network SAM (Segment Anything Model). SAM is an unsupervised deep learning segmentation model developed and released by Facebook. It allows the segmentation of every object within an image without the need for any additional training [23].

The initial idea was to use this model which provides a multi-scale segmentation of the image. Subsequently, the analysis of these masks would be used to implement an algorithm for the classification of nuclei, blood vessels, tubules and interstitial tissues.

2.2.1 Segment Anything Model

The "Segment Anything Model" is a segmentation model that drives data annotation and enables zero-shot approach, it refers to the model's ability to perform optimally even when faced with data different from those used during the training phase [23]. The "Segment Anything Model" is a segmentation model that facilitates data annotation and enables the zero-shot approach. This approach refers to the model's

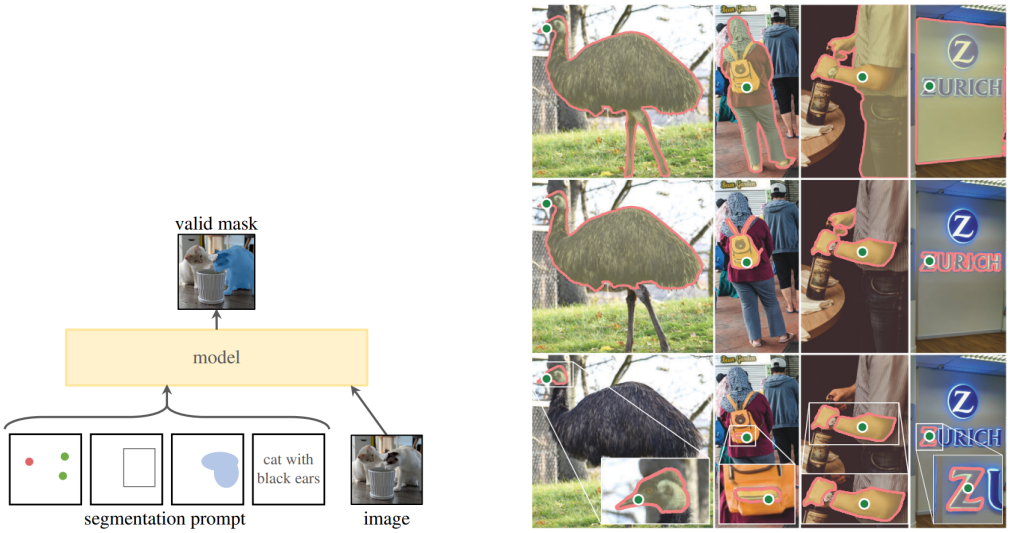
ability to optimally perfrom even in the presence of data different from those used during the training phase [23]. SAM constitutes, together with the promptable segmentation task, one of the main components of the Segmentation Anything (SA) project [23]. Another great strength of this project is the used data base. It is named SA-1B and includes more than 1 billion masks.Comparing the numerosity of this data base with any other data base used for segmentation, it turns out to be 400 times more numerous [24][25][26][27].

Promptable segmentation task

Task definition is inspired by natural language processing (NLP), in which model pre-straining is based on prediction of the next token. Through the given prompts, the model becomes capable of handling a number of downstream tasks [23]. In this context, a prompt can consist of any information that guides the model in segmenting an image (foreground and background points, rough bounding box, written text, etc.)[23]. Figure 2.6a better explain how a promptable segmentation works. The model uses the image and the user prompt to find the mask of the required element of the image.

A fundamental aspect within the context of the SA project is the ability to generate a valid mask even when the provided instructions are ambiguous. In other words, if the prompt refers to more than one object within the image, the model must be able to produce a suitable mask for at least one of these objects. This approach can be seen as a kind of iterative segmentation. However, unlike traditional iterative segmentation, which requires a minimum number of commands to obtain a valid mask, here the goal is to provide a valid mask based on any type of instruction[23]. In the figure 2.6b, it is possible to notice that starting from the same prompt (green point), which might refer to multiple elements within the image, the task aims to provide a mask for, at least, one of these elements.

To improve the model's generalization some novel tasks were incorporated: edge detection [23][28]; super pixelization [23][29]; object proposal generation [23][30]; foreground segmentation [23][31]; semantic segmentation [23][32]; instance segmentation [23][27]; panoptic segmentation [23][33]; and more. This approach might seem to be a form of multi-task classification. Whereas, however, in multi-task classification the model is only able to satisfy tasks on which it has been trained, in this case the model is able to adapt to novel tasks.So, it is a form of task generalization [34].



(a) (b)

Figure 2.6: How the promptable segmentation and the resolution of ambiguity work[23].

The model

The Segment anything (SAM) model consists of three elements: an image encoder that computes an embedding of images, a prompt encoder that embeds prompts, and a lightweight mask decoder that combines the two sources of information and predicts segmentations[23].

- **Image encoder**: a pre-trained MAE [23][35] Vision Transformer (ViT)[23][36] is used. The output of the encoder is an image that is 16 times smaller than the original. Specifically, images with a resolution of 1024x1024 were chosen, which are scaled by the encoder to 64x64. To reduce the number of channels, a convolution layer with a 1x1 kernel and 256 channels is introduced, followed by another convolution layer with a 3x3 kernel and the same number of channels [23][37]. Each of these is followed by a layer for channel normalization [38]. Both the convolutional layer are followed by a normalisation layer [23][38].

- **Prompt encoder**: the type of prompt encoding depends on the type of given prompt. The goal is to transform them into a vector of the same size as the number of channels, which in this case is an embedded vector of 256 elements citekirillov2023segment.

- **Lightweight mask decoder**:the aim of the decoding process is to generate a mask by combining the results of the image decoder and the prompt decoder. Firstly, the most relevant prompts are selected through a self and cross

attention process. Each selected item is individually processed through a multi-layer perceptron (MLP). The MLP's output is then combined with the decoded image through a second cross-attention step. Next, the size of the embedded image is increased by a factor of 4 using two levels of transposed convolutional layers and is again associated with the encoded prompts. The result of this association is processed by a small three-layer MLP, which produces a vector with dimensions corresponding to the number of channels in the enlarged image. Element-wise product between the scaled and encapsulated image and the output of the MLP [23].

To address the ambiguities, SAM was designed to generate multiple masks from a single prompt.The model is also designed to handle the generated masks. Specifically, the loss between the ground truth and each mask is computed, and only the masks that result with the lowest value are involved in the back-propagation process [23][39][40][41]. Finally, a confidence index is evaluated, such as Intersection over Union (IoU), to determine which mask is the most reliable. Figure 2.7 presents an overview of SAM.

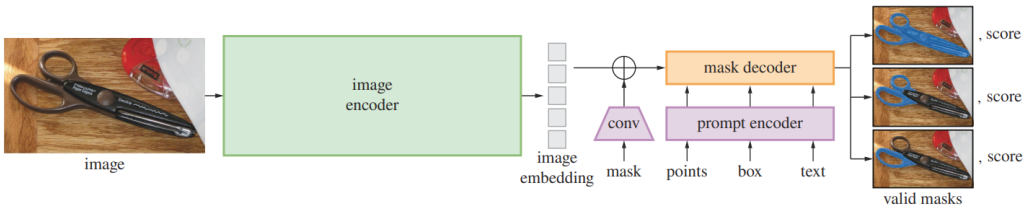


Figure 2.7: Segment Anything Model (SAM) overview[23]

The training

The training process relies on an iterative segmentation approach.In the initial phase, a foreground point or a bounding box containing the target element is randomly chosen with equal probability. Both are extracted from the ground truth mask. In the case of the bounding box, Gaussian noise is added to each coordinate, with a standard deviation equal to 10% of the bounding box's size, up to a maximum of 20 pixels [23][42][43].

During each iteration, a sequence of points corresponding to the classification errors is extracted. These, along with the resulting mask, are used as additional prompts. Subsequently, back propagation of the outputs is implemented, involving only the masks with a sufficiently low loss value. After 8 cycles, the number of classification errors begins to decrease, so 8 was chosen ad the optimal number of iterations[23]. Two additional iterations are added in the case in which no additional points are sampled, one randomly placed among the first 8 iterations and the other at the

end. The aim of these additional iterations is to encourage the model to refine its own mask predictions [23].

Regarding the parameters, the AdamW optimizer is used [44]. The entire process is iterated for a total of 90,000 iterations, with periodic changes in the learning rate. A batch size of 256 images is used, and regularization techniques are applied[23].

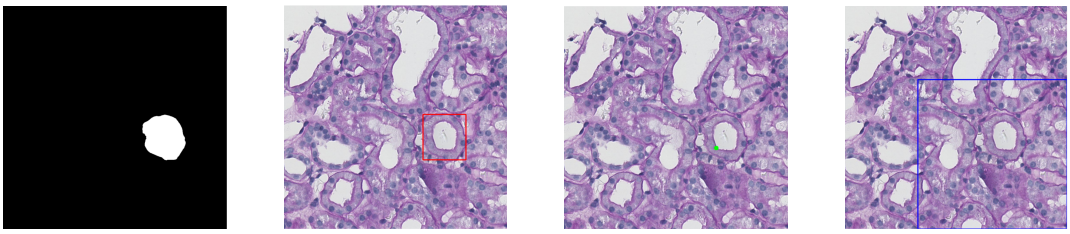
2.2.2 SAM's implementation

The main strength of SAM lies in the dimensionality of the training data set. Typically, to achieve high-quality masks using machine learning or deep learning techniques, it is necessary to collect and manage a huge amount of data. This process can be very time and computational consuming. However, SAM is a pre-trained model on a wide range of data and is open-source, making it applicable to new images with a high probability of obtaining accurate masks for objects of interest. Consequently, the initial phase of SAM implementation involves applying the model provided by the developers to the set of available images.

Moreover, SAM provides various types of information that enhance segmentation. Specifically, the output of SAM consists of a list of dictionaries. Each dictionary is associated with an element identified in the masks and includes the following attributes:

- **segmentation**: array of size (W, H), matching the dimensions of the original image, in this specific case, 1024 x 1024. These are binary images that represent the masks of the elements of interest.
- **area**: integer representing the number of white pixels comprising the mask, which is equivalent to the area of the segmented object.
- **bounding box**: list of four integers that describe the bounding box containing the respective mask. Specifically, it provides the coordinates (x, y) of the top-left corner and the width and height values of the bounding box.
- **Predicted IoU**: value indicating the quality of the predicted mask.
- **point coords**: sampled points from the input image used in predicting the mask.
- **stability score**: additional measure of the quality and stability of the mask.
- **crop box**: square that includes the portion of the image where the mask is. It's provided in the same format as the bounding box.

In figure 2.8 are showed the different outputs provided by SAM. Fig.2.8a presents the mask of the i-th element within the image, fig.2.8b the bounding box that precisely encloses the mask, fig.2.8c the coordinates of the i-th point sampled from the original input image and fig.2.8d the crop box.



(a) **(b)** **(c)** **(d)**

Figure 2.8: SAM output

During the model's implementation, it is also possible to define certain parameters. Specifically:

- **points_per_side**: Number of points to sample on each side of the image. The total number of sampled points will be equal to this value raised to the power of two. Providing the coordinates of points directly allows for targeted segmentation. The greater the number of sampled points, the more masks are obtained, but this also increases the computational time.**Default value = 32**

- **point_per_batch**: The number of processed points per batch in which the input image is divided. Increasing this value reduces computational time but raises memory requirements. **Default value = 64**

- **pred_iou_thresh**: Threshold related to the Intersection over Union (IoU) index evaluated between the masks obtained from the same sampled point and the ground truth. Increasing this threshold results in a higher number of correct masks. **Default value = 0.88**

- **stability_score_threshold**: Threshold applied to select masks based on stability score.A mask is considered stable if its characteristics do not vary over time. Lowering this parameter makes the algorithm more permissive, potentially resulting in an increase in false positives. **Default value = 0.95**

- **stability_score_offset**:Value added to the threshold related to the stability score.This allows even less stable masks, those with a stability score not greater than or equal to the threshold, to be selected. It further increases the permissiveness of the model. **Default value = 1**

- **box_nms_thresh**: Threshold used during non-maximum suppression (NMS), specifically for the suppression of bounding boxes. In the event that the same mask is associated with two bounding boxes, the overlap index is calculated. If this index is lower than the specified threshold, the bounding box with the lower overlap index related to the mask is eliminated.A higher threshold reduces the algorithm's selectivity. **Default value = 0.7**

- **crop_n_layers**: Levels of convolution to apply to the feature map.Increasing this value enhances the level of extracted features, resulting in improved segmentation. However, excessively high values may lead to excessively long computational times. **Default value = 0**

- **crop_nms_thres**: Threshold related to the NMS of the extracted features.It is directly proportional to permissiveness. **Default value = 0.7**

- **crop_overlap_ratio**: Overlap fraction between the crops. This is an initial value that decreases during implementation. **Default value = 512/1500**

- **crop_n_points_downscale_factor**: Factor used to scale down the number of sampled points with respect to the number of crop layers and the current layer index. Increasing this value enhances the reduction effect due to scaling. **Default value = 1**

- **point_grid**: Allows specifying the points to sample in case targeted segmentation is desired. **Default value = None**

- **min_mask_region_area**: Threshold applied to the area of the masks.If the area of the i-th mask is lower than this threshold, the segmentation is not considered. **Default value = 0**

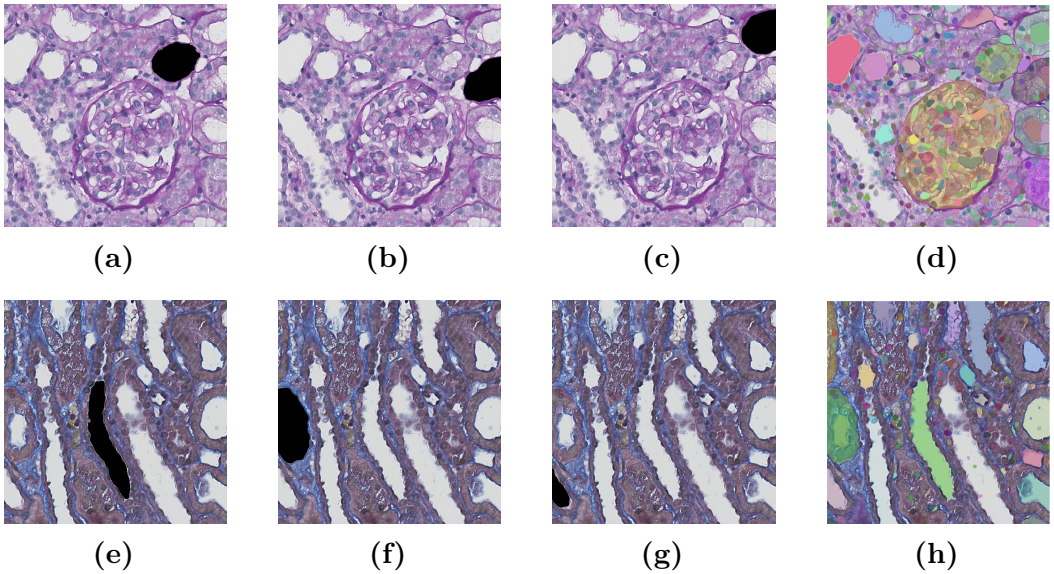
- **output_mode**: Allows specifying the type of output mask. **Default value = 'binary_mask'**

From empirical tests, it was chosen to modify some parameters from their default values.Specifically, points_per_side= **64**, pred_iou_thresh= **0.9**, stability_score_thresh= **0.96**, crop_n_layers= **1** and crop_n_point_downscale_factor= **2**.

In the implementation, each list of dictionaries was saved as a NumPy (.npy) object to allow all the above information to be used even after the model was applied. For each patch, a directory was created to store all the masks obtained. These masks are displayed directly on the original patch, instead of being saved as binary images; they are represented as black objects on an RGB image.
In this initial phase both types of staining were used.Due to the absence of a GPU,

the computational time required for processing a single patch is considerably high. Consequently, it was chosen to apply the model to all patches of only two WSI images, one with PAS staining and one with TRI staining.

To facilitate visualization, all masks pertaining to a specific patch were represented within the same image. In particular, in order to improving understanding of the information displayed,each mask was assigned a random color. In figure 2.9 are presented both the single mask colored in black on the original image and the representation of the all the generated masks.



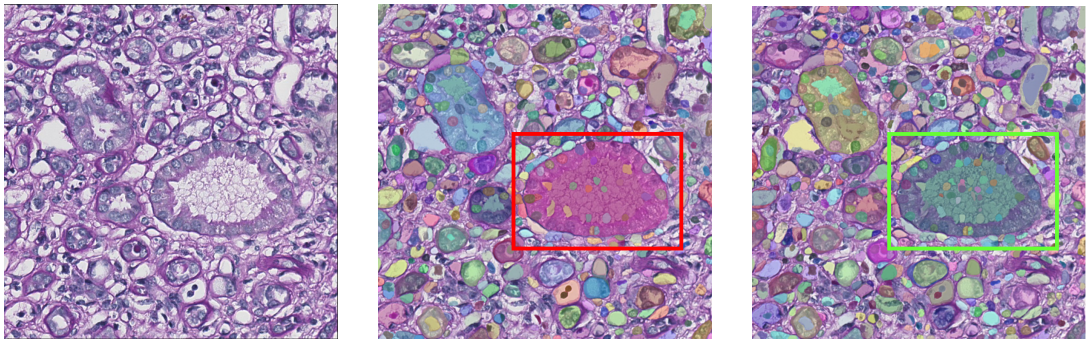
(a) **(b)** **(c)** **(d)**

(e) **(f)** **(g)** **(h)**

Figure 2.9: Images (a),(b) e (c) represent the single generated mask on PAS-training image. The same is for images (e),(f) e (g) which refers to TRI staining. Image (d) e (h) show the representation of all the mask given as output.

As can be seen in figure 2.9d and 2.9e ,this initial segmentation is not optimal. Therefore, it was decided to modify the selected values to identify a new combination that could offer better performance. The choice was made considering computation time and the successive application of masks. A model capable of segmenting not only larger elements, such as tubules or glomeruli, but also smaller ones, such as lumens or single nuclei, is needed. Although having only the tubule mask rather than the lumen mask is useless for classification. The segmentation of each individual element that constitutes the tubule had to be used to establish a kind of hierarchy among the masks. In fact, by using the characteristics of the "lower level" masks, larger ones can be classified.
The following indicators were considered to select the best combination of model parameters: the average number of masks extracted per patch, the average number of masks representing the innermost elements but not included in masks related

to larger elements (as in the case where you only have the lumen mask of a glomerulus without having the glomerulus mask), and the average number of masks representing the larger elements but lacking smaller masks within them (as in the case where you only have the glomeruli mask without having the masks of the numerous lumens and nuclei within it).
After numerous trials, the best combination of values was found. It differs from the previously described model in the following parameters: stability_score_thresh= **0.95**, stability_score_offset= **1.2**, box_nms_thresh= **0.8**, crop_n_layer= **2**, crop_nms_thresh= **0.8** and min_mask_region_area= **50**.
In figure 2.10b can be seen that the masks enclosed within the red rectangle cannot be used for classification as they only represent the tubule and some nuclei within its crown. On the other hand, the figure 2.10c represents the output of the system using the last set of parameters. In this case, the green rectangle contains not only the tubule mask but also the lumen mask. It is also evident that the number of identified masks is significantly higher.



(a) original patch **(b)** first parameters set **(c)** second parameters set

Figure 2.10: SAM parameters fine-tuning results

2.2.3 SAM's application

The primary purpose of this project is the classification of tubules and blood vessels in kidney histopathological images. SAM is capable of generating masks for most of the elements in the image but does not have the ability to perform the classification. Therefore, it was necessary to develop an algorithm capable of using these masks to conduct the desired classification.
The available images do not come from any existing database, so ground truth is not available. Furthermore, as the database is very large, an unsupervised classification approach was chosen before manually annotating each image.
Initially, the idea was to identify patterns that would allow the masks of tubules to be distinguished from those of blood vessels. Subsequently, these patterns would be used to define classification rules to be automatically applied. In essence,

an attempt was made to develop an algorithm that would emulate the visual recognition ability of a pathologist.
For this purpose, the knowledge learned during the visit to the histopathology laboratory was very useful. In general, there are a few key features that allow tubules to be immediately distinguished from blood vessels:

- **lumen size**: generally, lumens related to blood vessels are much smaller than those related to tubules.
- **crown around lumen**: the blood vessels's wall is usually very thin, so there is no real separation between the lumen and the interstitial tissue.On the other hand, as far as tubules are concerned, it is always possible to identify a sort of crown surrounding the lumen and separating it from the interstitial tissue; this crown is very evident in PAS staining but is also present and clearly visible in TRI staining.
- **presence and size of nuclei**: when faced with a dubious situation, it is necessary to observe the conformation and the amount of nuclei around the lumen. With regard to blood vessels, these are generally present in small quantities and are adjacent to the lumen. In tubules it is almost always possible to observe a large quantity of nuclei within the crown, these nuclei never touch the lumen but always remain confined within the crown.Furthermore, whereas the nuclei relative to the tubules have an almost rounded shape, those relative to the blood vessels are usually flatter with an elliptic shape.

We aim to classify the masks relating to tubules and vessels form the masks of lumens, crown, and nuclei.To do this, the first idea was to create a tree structure in which each node represents a mask.The mask associated to the i-th node is connected to the j-th node only if it overlaps the mask associated to it. Therefore, in such tree structure, the root is the starting image. Then connected to this first node are the larger SAM-generated masks that will form the second level of nodes.Further nodes and sub-levels are obtained by the masks included in the first level masks and so on.
In this way, by identifying the level related to the lumen masks, it is possible to perform a classification by analyzing the masks of the levels below.

Tree structure creation

The first step in the implementation involves the creation of the tree structure. To do this, it was first necessary to name and sort the masks according to their size, and specifically in descending order. Therefore, for each patch, a folder was created in which all the N masks produced by the model were saved, with the mask named 'M_0' being the largest and the mask named 'M_N-1' the smallest.

Starting from the smallest mask, the intersection between the i-th mask and all other masks related to the image was evaluated. In the case in which the result of all the intersections is zero, the i-th mask is directly linked to the root node. In the case in which the i-th mask is subtended by other larger masks, it is assigned to the mask that originated the maximum intersection.
This type of algorithm made it possible to create a tree structure for each patch, specifically the library *networkx* was used to visualize and manage this structure. Figure 2.11 represent the output of the algorithm. It is possible to see the second-level masks directly connected to the main node (fig.2.11a) and, to simplify visualization, only some of the masks connected to the mask M_2 (fig.2.11c).

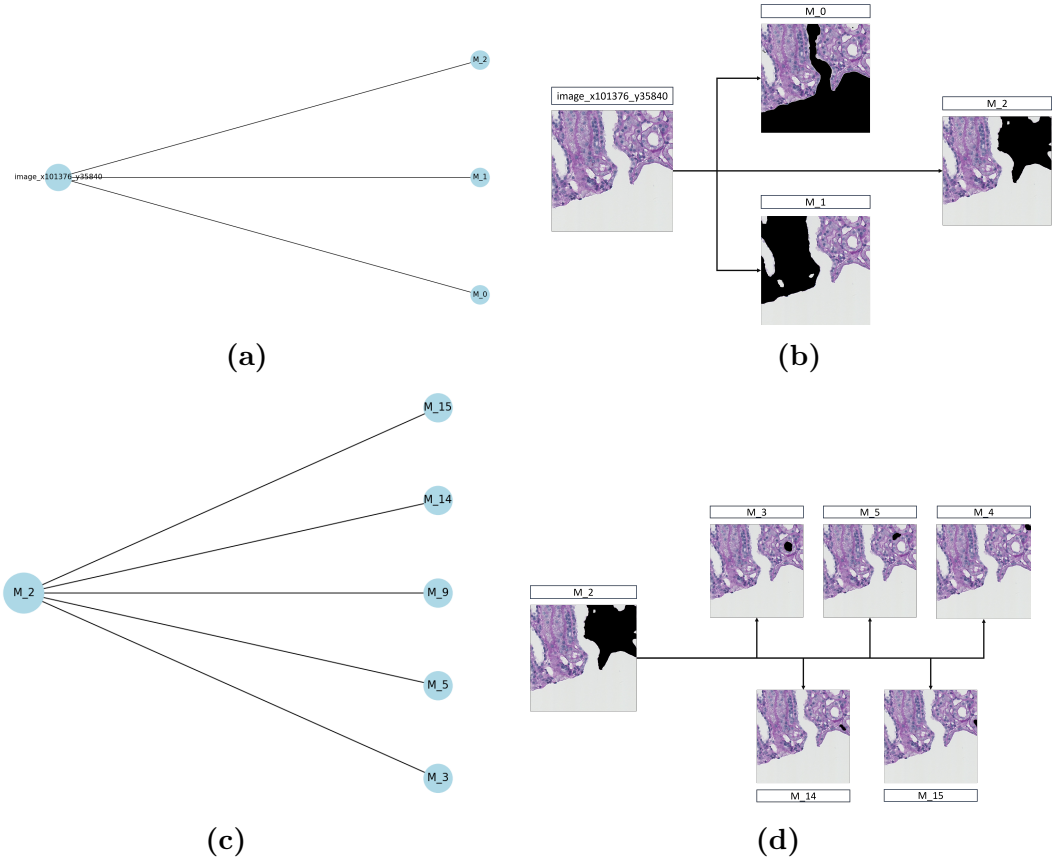


Figure 2.11: Images (a) and (c) represent the output of the algorithm used to generate the pyramid structure. Images (b) and (d) show the relative images

Masks analysis

As mentioned above, to classify tubules and vessels it is necessary to analyze the characteristics of three elements: the lumens, the nuclei, and the crown around the nuclei. First, it is necessary to distinguish the different types of masks. The identification of lumens and nuclei is based on size and color.
For the identification of lumens, background masks, which are easily mistaken for lumens, and masks of extremely small size were excluded from the analysis. Two thresholds were defined.The chosen threshold define the maximum and minimum size of a mask to be considered a lumen. Specifically, all masks with a size between 21.9% and 0.12% of the total patch size were considered as potential lumen masks. Considering the color, a third threshold was applied at an α value.α is defined as the ratio between the average of the pixels relative to the grey scale image and belonging to the i-th mask and the maximum value assumed by the pixels in the entire image.

$$\alpha = \frac{\text{mean}(image\_gray[mask == 1])}{\max(image\_gray)}$$

For images with PAS staining, only masks with α greater than 0.75 have been considered as luminescent masks, while for TRI staining α is set greater than 0.63. As far as nuclei are concerned, only masks with a size smaller than 0.12% of the total size and with α lower than 137 were taken into consideration.
After identifying lumens and nuclei, it was necessary to define a strategy to obtain the masks of the crowns around lumens. SAM does not always provide masks for all the elements in the image, so that three different cases can are possible:

- **complete masks**: for each element is given the mask relating to the lumen, the mask relating to the crown, the masks of all the nuclei and finally the mask relating to the total element given by the union of all the other masks mentioned.

- **partially complete masks**: for each element, is possible to have the lumen mask, the global mask of the element and the mask of some nuclei. In this case, it is possible to evaluate the crown mask by subtracting the lumen mask from the total mask.With regard to the nuclei, it is necessary to assess their numerosity to see whether they are sufficient to carry out the classification.

- **single masks**: for each element there is only one mask relating either to the lumen or to the total element. In this case it is impossible to perform the classification as there is no information on either the crown or the nuclei.

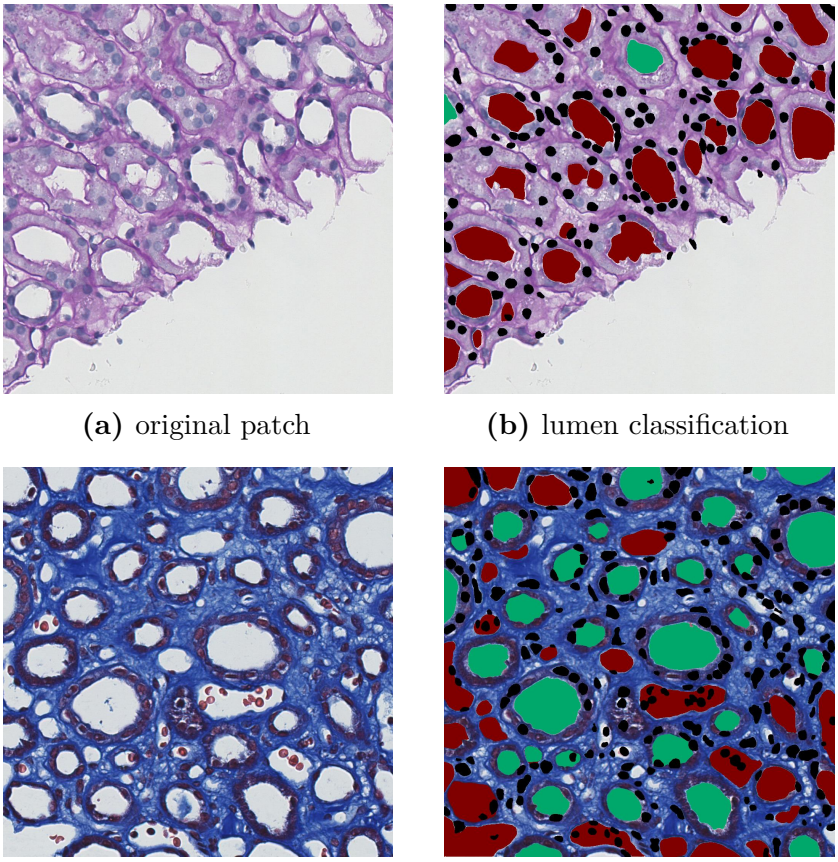
The next step involves the described tree structure. For each i-th mask that respects the conditions to be considered as lumen, the j-th mask to which it is directly

linked in the tree structure is analyzed. Depending on the features of the j-th mask, it was possible to distinguish two types of lumens.

- **useful lumen**: the lumen is directly connected to the mask of an element to be classified. The j-th mask is therefore neither a lumen or part of the background, and the lumen in question is not directly connected to the root of the tree structure. This is again the case for a complete or partially complete mask for which the crown mask can be derived.
- **isolated lumen**: the lumen is directly connected to the root of the tree or the j-th mask is itself a lumen or part of the background. In this case, it is impossible to deduce information about the crown surrounding the lumen itself.

Figure 2.12 shows the classification of lumens. In green are represented the lumens contained within other masks, and in red the isolated lumens. The nuclei masks are also shown in black. Analyzing only the TRI-colored image, it might appear that the isolated lumens mostly belong to the blood vessels. If, however, we consider PAS images, it is clear that it is impossible to classify many of the tubules as there is no information about the crown.

Regarding masks that do not refer to lumens, a further distinction has also been made. As mentioned above, in addition to the crown and the lumens, it is also important to analyze the number and the position of the nuclei within the crown or adjacent to the lumens. It may happen that some lumens are not detected by SAM. For these reasons we refer to "full" masks and "empty" masks. The former are masks that contain other smaller masks, such as those of the lumen or nuclei; in the tree structure they therefore represent a node from which other branches branch off. Empty masks, on the other hand, are useless for classification purposes as it is impossible to extract further information from them; they therefore constitute a leaf node.



(a) original patch **(b)** lumen classification

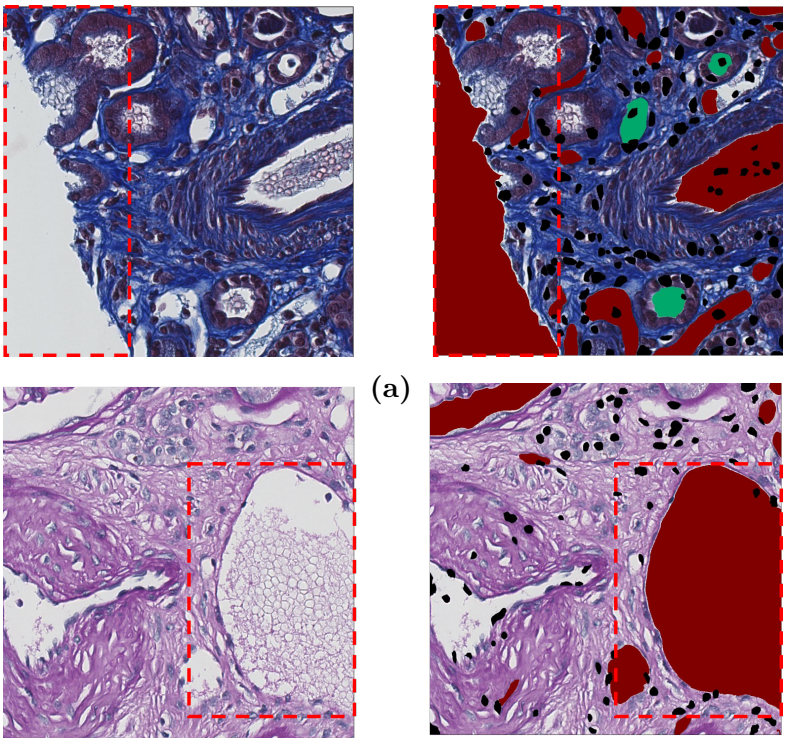
(c) original patch **(d)** lumen classification

Figure 2.12: Classification of lumens using SAM

2.2.4 Evaluation of SAM

Although SAM is a particularly innovative model and potentially applicable to any type of context, it was decided to abandon this approach. In addition to the problems previously exposed concerning the presence of isolated lumens that cannot be used for classification, other critical aspects have emerged. One of the main limit concerns the threshold used to exclude background masks. Often the background occupies a portion of the image smaller than 20%. At the same time some large elements with characteristics such as to be classified as lumens can be found. Consequently, it is not possible to change the threshold values as this would lead to loose the masks of such elements. This leads to many false positives in the detection of lumens from the background. In figure 2.13 are represented in red the lumens classified as isolated, in green the useful lumens and in black the nuclei. In fig.2.13a, it is possible to see how the portion of the image enclosed in the red rectangle is part of the background and is classified as a lumen due to its

size. On the other hand, in fig.2.13b the red rectangle highlights an element useful for classification that covers more than 20 % of the image.
Looking at the images it is also clear that the detection of nuclei and lumens is not optimal. Since there is no ground truth available, it is not possible to determine the percentage of missed elements. Nevertheless, form the visual analysis, it is clear that the number of false negatives, i.e. pixels classified as background but which actually belong to an element of interest, is very high. This second problem mostly does not involve nuclei, which are extremely important not only for the classification of tubules and blood vessels but also for the detection of a given pathology.



(a)

(b)

Figure 2.13: Representation of some SAM's criticism

2.3 Proposed solution

Due to the reasons outlined in the preceding section, it was decided to propose a new approach based on features detection followed by a machine learning step. The first step involves the segmentation of nuclei and lumens. Subsequently, the obtained masks are used to compute some geometric parameters that are used for lumens classification. About this second step, two different strategies were tested. The first involves the classification of the nuclei and its subsequent use to to obtain the lumen classes. The second involves directly classifying the lumen masks using both the features of the masks of the lumens' and of nuclei.

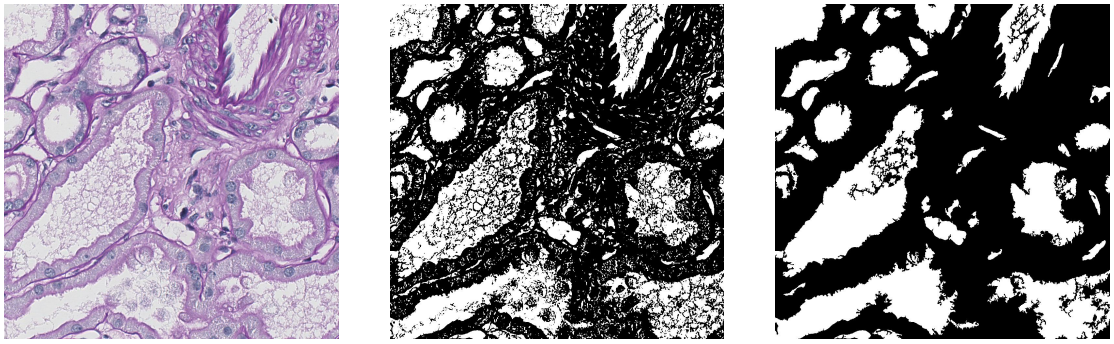
2.3.1 Segmentation

The final goal of segmentation is to isolate a object or a class of objects within an image. To do this, it is necessary to obtain a mask. A mask is a binary image with the same dimensions as the source image,where each pixel can only assume one of two values. Specifically, pixels belonging to the object of interest are assigned a value of 1, while those belonging to the background are assigned a value of zero. Segmentation thus produces an image of the same size as the source image in which the objects of interest are represented as white objects on a black background.
There are numerous segmentation techniques, some involving artificial intelligence, while others directly analyze image at the pixel level.
In the context of this project, the elements of interest are the lumens of the tubules and blood vessels, and the nuclei. It was therefore necessary to divide the segmentation problem into two sub-problems: the segmentation of the lumens and the segmentation of the nuclei.

Lumen segmentation

Before carrying out the actual segmentation, pre-processing strategies were adopted to improve the segmentation results. We first increase the image contrast by extending the image histogram. The lumen segmentation is then obtained by thresholding. This method involves applying a threshold to all pixels in the image. Pixels with a value greater than the threshold will be considered as belonging to the object, assuming a value of 1 in the final mask. This is a very simple method, but at the same time very effective, especially when used to segment highly contrasted which is the case of lumens. The Otzu method was chosen to obtain the threshold. In Othu's method the optimal threshold is chosen from the image histogram. It is an iterative method in which the histogram is divided into two classes using a variable threshold, at each iteration the distance between the two classes is evaluated and the threshold is chosen to minimize the intra-calss variance.

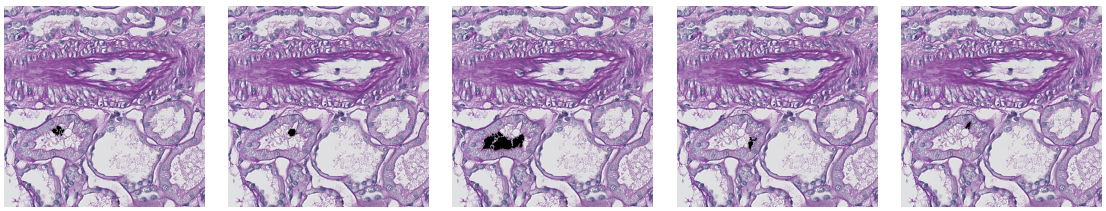
To further improve the quality of the segmentation, post-processing was also carried out. First of all, all masks exceeding the threshold chosen to exclude the background (21.9% of the image size) and very small masks (0.04% of the image size) were eliminated.Finally, to increase masks uniformity and to eliminate some dark pixels within the lumens, the morphological operator of closure was used. Closure consists of the sequential application of a dilation and an erosion. Specifically, the *remove_small_holes* method of the *sklearn* library was applied with an area of 120 pixels.Figure 2.14 shows the effects of this post processing. After labelling



(a) **(b)** **(c)**

Figure 2.14: Lumen segmentation. Image (a) represents the original patch, image (b) the mask obtained by applying global thresholding and image (c) represents the final mask after post-processing.

the segmented image we observed that in some cases masks with different values coexisted within the same lumen which reflects a segmentation error that divides a lumen into several masks. This is caused by the presence of pixels within the lumen with similar values to those of the surrounding tissue. These pixels are recognized as background and lead to the fragmentation of the lumen mask. Figure 2.15 shows how a single lumen is segmented in five different masks.
To solve this problem two tools, the color deconvolution and the Bresenham algorithm, were used. Specifically, it was necessary to find a method to provide the

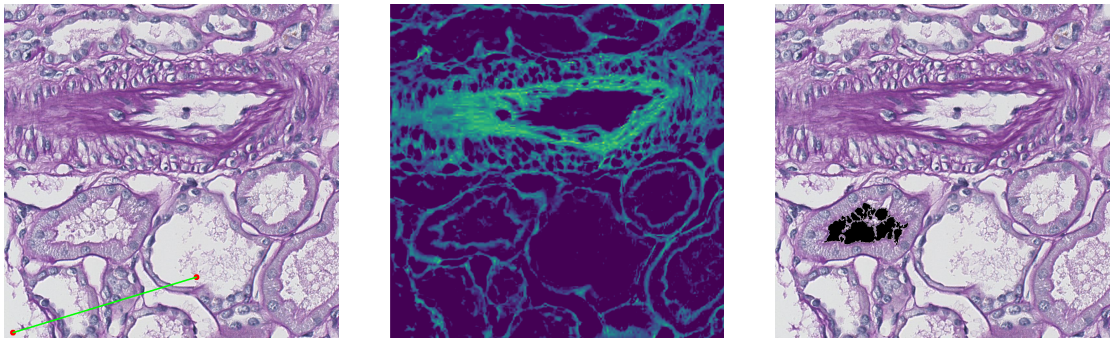


(a) **(b)** **(c)** **(d)** **(e)**

Figure 2.15: First problems in lumen segmentation. Images (a),(b),(c),(d) and (e) represent five distinct masks belonging to the same lumen.

same value to masks belonging to the same lumen. On image resulting from the difference between the red and green channels, each lumen is well separated from the others by a layer of pixels with a much higher value than those within the lumen itself. This behavior also occurs in cases in which also dark pixel are present inside the lumen. It was therefore decided to apply Bresenham's algorithm to this image obtained by implementing the color deconvolution.
Bresenham's algorithm provides the coordinates of the pixels joining two points by straight line.Applying it to the centroids of each mask, it is then possible to obtain the value of the pixels that separate them. Consequently, it is very easy to trace the masks contained within the same lumen, since the value of all the pixels separating the corresponding centroids will be low.

In practice, all masks whose centroids are separated by pixels with a value lower than twice the average value of the image created by color deconvolution were considered to belong to the same lumen. To further emphasize the contours of the lumens, the contrast of this image was increased.
Figure 2.16 shows the application of the mask aggregation algorithm. Specifically, fig.2.16a represents the Bresenham path (green line) between the centroids of two masks (red points). Fig.2.14b shows the image resulting form the difference between the the red and the green channel of the RGB image. Fig.2.14c shows the result of the union of the masks belonging to the same lumen exposed in 2.15.



(a) Bresenham path **(b)** R channel - B channel **(c)** joined masks

Figure 2.16: Application of the mask aggregation algorithm.

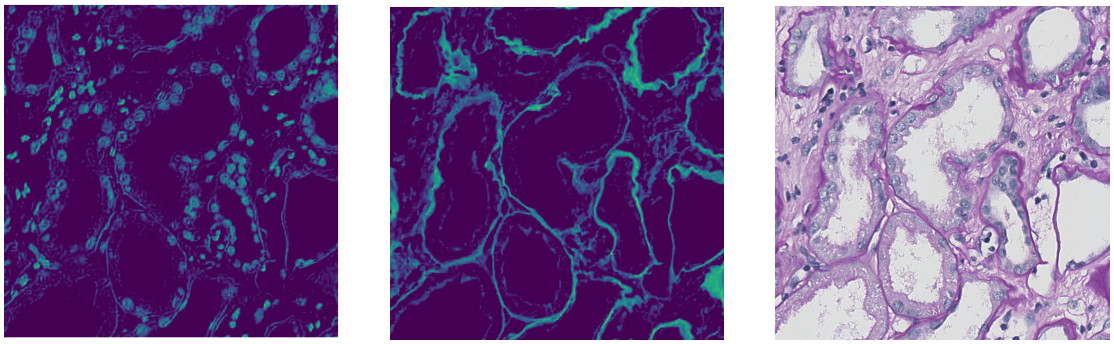
Nuclei segmentation

While the first identified strategy was successful for lumens, for nuclei two different methods were tested. In the first method, an attempt, to replicate the pipeline used for obtaining lumen masks, was made. Starting with pre-processing strategies, global thresholding was once again applied. With the aim of improving the quality of the masks, post-processing was performed. In the second case, however, a

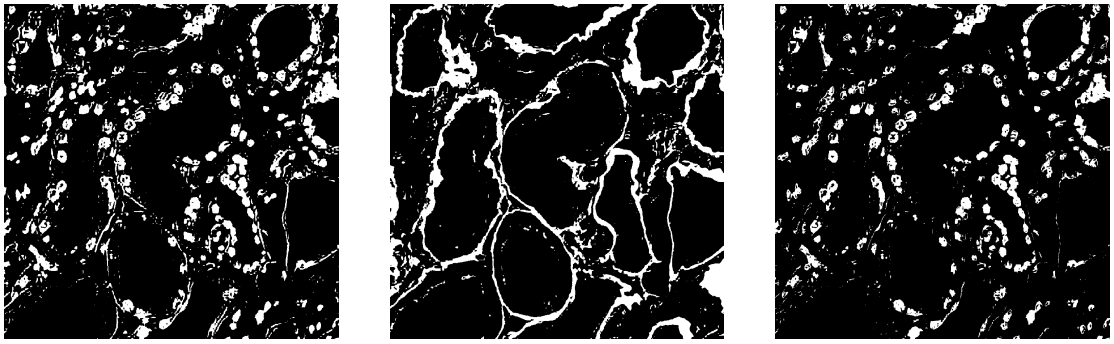
completely new method was used.
While the segmentation of lumens focused on the gray scale image, for nuclei, segmentation was performed on an image derived from color deconvolution. In the image resulting from the difference between the red and blue channels, nuclei and the membrane surrounding tubules and blood vessels are particularly accentuated. In this case too, contrast augmentation was permed, using true luminance.
Subsequently, global thresholding was applied to this image using Otsu's method. The obtained segmentation includes both nuclei and cell membrane masks. This segmentation error can be explained by the features of the segmented image. In fact in the image nuclei and the walls of the tubules have approximately the same color. To solve this problem and obtain only the masks of the nuclei, a number of post-processing strategies were applied.

The initial step involved setting all pixels identified as belonging to a lumen to zero. It was then found that the image obtained by subtracting the green channel from the blue channel only contained the membrane of blood vessels and tubules. It was therefore decided to implement a second segmentation of these elements to isolate the nuclei in the first mask. The contrast of the image was increased and global thresholding was again applied using the Otzu's method. Once the two masks were obtained, all the segmented elements in the second mask were set to zero.
To further improve the quality of the segmentation, only masks larger than 0.01% of the total image size were selected. Figure 2.17 shows the images that came out from the color deconvolution and the respective masks.

Reviewing at the final result, it can be seen that there is a considerable number of missing elements. Moreover, the shape of the recognized nuclei also does not reflect their real profile. Considering the importance that the number of nuclei and their shape have in the classification, it was decided to pursue an alternative approach. This second approach is inspired from the marked point process modeling which consists in fitting a collection of geometrical objects onto the image based on a contrast term and a non-overlap constraint. In our case we define a shape dictionary composed of ellipses of different size and orientation. We then associate to each shape a kernel defined by a positive value within the object and a negative value on the object contour. We compute the convolution of the image with each kernel and consider as candidates the objects maximizing the convolutions on each pixel if greater than a given threshold. We sort the candidates with respect to the convolution values and select the best ones without overlap.
Figure 2.18 shows the difference between the nuclei mask obtained with the first approach (fig.2.18b) and the second one (fig.2.18c).

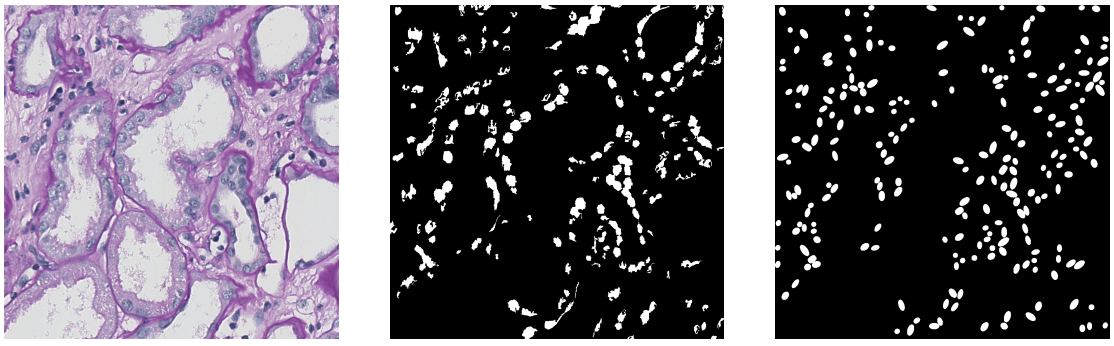


(a) R channle - B channel **(b)** G channle - B channel **(c)** original patch



(d) **(e)** **(f)**

Figure 2.17: First nuclei segmentation strategy. Images (d) and (e) respectively represent the result of the segmentation of image (a) and (b). Image (f) represents the result of eliminating the contours from image (d) using image (e)



(a) **(b)** **(c)**

Figure 2.18: Final nuclei segmentation strategy. Image (b) represents the mask resulting from the first pipeline after all post-processing actions. Image (c) instead represents the mask obtained using the second type of algorithm.

2.3.2 Classification

To classify lumens we first propose a two steps approach, where the nuclei are subject first classify between tubule and vessel, and the lumen masks are classified depending on the neighboring nuclei. As mentioned above, to distinguish a tubule from a blood vessel, it is essential to analyze the number, position and shape of the nuclei surrounding the lumen. The approach consists in training a ML models to classify the nuclei.Each nuclei is then assigned to the closest lumen. Each lumen finally takes the class from a majority vote among its associated nuclei.

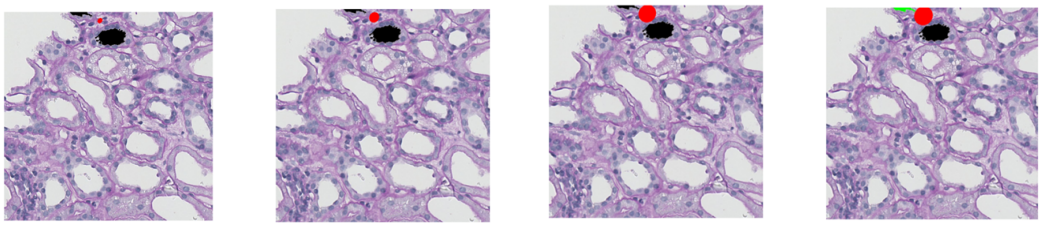
Nuclei classification

The nuclei surrounding the tubules are always inserted within a crown surrounding the lumen. They are mostly round and never come into close contact with the lumen.On the other hand, in the case of blood vessels, the nuclei are adjacent to the lumen and tend to have a more elongated shape.Another distinguishing parameter is the number of nuclei. Tubules tend to be surrounded by more more nuclei than blood vessels. Exploiting these features chosen in agreement with two experienced pathologists, several classifiers were trained to identify each nuclei as belonging to a tubule or a blood vessel.

The first step in the implementation of this strategy is to develop an algorithm capable of associating each lumen with its respective contour nuclei. The simplest way to associate a nucleus with a lumen is to assess whether that nucleus is located in the neighbour of the lumen itself. To do this, it was sufficient to create a mask for the hypothetical crown surrounding each lumen. To obtain the crown we apply the morphological dilation operator using a square kernel with dimensions of 90 pixels per side to each lumen's mask. Subsequently, the mask of the lumen itself is subtracted from the dilated mask. To understand which lumen the i-th nuclei was connected to, is necessary to analyse the overlap between the nuclei mask and the mask of each obtained crown. Evaluating the pixels that these two segmentations have in common, three scenarios can arise:

- **no overlap**: if the i-th nucleus is not inserted in any of the obtained crowns, it is classified as a nucleus belonging to the interstitial tissue or within a glomerulus. Such nuclei are excluded from the analysis as they have atypical features that could load to misclassifications.
- **Single overlap**: if the product of the nuclei mask and the mask of each crown gives a non-zero result for a single crown, the nuclei is uniquely assigned to the lumen that gave rise to the crown.
- **multiple overlap**: having chosen a large kernel some nuclei overlaps with

several crowns. In this case the nuclei are assigned to the nearest mask. Considering the number of possible lumens and the size of the images, it was necessary to optimize the distance evaluation. We therefore compute dilatation of the mask of each nuclei iteratively. However, in this case, a circular kernel is used, the radius increases by one unit at each iteration. When the dilated nuclei mask comes into contact with the mask of one of a lumens being analysed, it is assigned to this lumen. Figure 2.19 shows the expansion process of a nucleus inserted into the crowns of the two masks shown in black. The moment when the nuclei is assigned to a mask (colored in green) is represented in 2.19d. However, it may happen that the nuclei expansion intersects with more lumen masks at the same time. In this case the assignment is made by evaluating the value of the pixels separating the centroids of the nuclei and the lumen. Specifically, the pixels are selected from the Bresenham path. The image used to extract their values is the difference between the red channel and the blue channel. Each lumen is surrounded by a magenta-colored membrane, this membrane is particularly highlighted by the image obtained by the color deconvolution. As soon as the centroid of the nucleus and the centroid of the lumen are separated by pixels belonging to this membrane, the nucleus cannot be assigned to the lumen. It is therefore sufficient to identify the Bresenham path that gives rise to the lowest maximum value in order to have an unique association between nucleus and lumen.



(a) **(b)** **(c)** **(d)**

Figure 2.19: Nucleus expansion

Once the nuclei and the lumens to which they belong have been identified, it is necessary to proceed with feature extraction and selection. To select features, we followed the practice of two experienced pathologists. Specifically, were extracted 18 features for each selected nucleus:

- **lumen size**: single feature concerning the size of the lumen to which the i-th nucleus is connected. Usually, lumens referring to tubules are larger than those referring to blood vessels.
- **nuclei size**: single feature concerning the size of the i-th nuclei. Here again, a larger size is found for nuclei circumscribing tubules.

- **nucleus color**: the mean, median, and standard deviation of the pixel values belonging to the i-th nuclei mask were extracted using the red channel, blue channel, green channel, and gray scale image as the sources for extracting values. For each image, three features were then extracted, resulting in a total of twelve features related to the color characteristics of the nuclei. It was noted that the nuclei of blood vessels tend to assume blue tones while those relating to tubules of magenta.

- **lumen shape**: single feature regarding the eccentricity of the lumen to which the nuclei is attached . The eccentricity is calculated as the ratio of the focal half-distance c to the semi-major axis of the ellipse circumscribing the lumen mask.

$$e = \frac{c}{a} \tag{2.1}$$

$$c = \sqrt{a^2 - b^2} \tag{2.2}$$

$$a = \text{major semi-axis}$$

$$b = \text{minor semi-axis}$$

  It is therefore a value between 0 and 1 that indicates how far the mask deviates from the circular shape. Specifically, it takes on an increasingly helical shape as e increases.

- **nuclei shape**: eccentricity is also extracted for the nucleus mask using the same formula and criteria as before.This feature is chosen because blood vessels are usually very elongated, whereas it is not uncommon to find tubules with almost circular nuclei.

- **number of nuclei**: for each i-th nucleus, the number of nuclei assigned to the same lumen is extracted. As already mentioned, tubules have a much higher numerosity than blood vessels.

- **convex Hull of the lumen**: the convex hull represents the smallest convex shape capable of enclosing a set of data points. It is therefore the simplest curve capable of surrounding the contour of a specific mask. By evaluating the convex hull on the mask of a lumen, information regarding the regularity of the mask edges are obtained. Specifically, the difference between the envelope and the mask was extracted. If the mask has regular and slightly jagged contours, this difference will be very low. While if the lumen has protuberances or recesses, the calculated value will increase. This feature was selected because the lumens referred to tubules are more irregular than those of blood vessels.

Once the features had been selected, they were collected in a feature matrix with all the selected nuclei as rows and the corresponding features as columns. regarding the train set, class balancing was performed in order to train the classifier to equally recognize both elements.Therefore, the redundant elements of the most represented class were eliminated.
Finally, to increase the robustness of the models, a random shuffle of the matrix rows is implemented. Both the train set and the test set were then normalized using min-max scaling.

To implement supervised classification methods, it is necessary to extract the ground truth. To do this, the masks of all nuclei belonging to the image were manually labeled. This process was carried out under the supervision of an experienced pathologist who carried out a second validation after the labeling.

Once the ground truth and feature matrix had been obtained, it was possible to proceed with the training of four classifiers:

- **K-Nearest Neighbors algorithm (KNN)**: it is a non-parametric classification method capable of classifying the elements of the data set on the basis of the elements that are part of its neighborhood. Specifically, a similarity measure is defined to determine the distance between each element of the data set and all the other elements[45]. All k most similar elements are considered part of the neighborhood.After that, the class corresponds to the class most represented in the neighborhood. In this first step, the default values of the function *neighbors.KNeighborsClassifier* from the *sklearn* library were used, whereby the Euclidean distance was used as the similarity measure and fixed k equal to 5.

- **Support Vector Machine (SVM)**: this method of classification involves remapping the elements involved in the classification in a space of higher dimensions using a special function named kernel. In this pace an hyperplane that is able to separate the points into two classes is identified.This hyperplane is estimated to maximize the margins.The margines are the distance between the hyperplane and the elements of the classes closest to the hyperplane itself. To implement this classifier, the *svm.SVC* function of the *sklearn* library was used, again using default parameters. Specifically, we have a kernel of type *rbf*,the Gaussian kernel.A value of C equal to 1 where C is the parameter that governs the bias-variance trade-off i.e.the relationship between the complexity of the model, the accuracy of the prediction and the quality of the predictions made on data never seen by the classifier.On the other hand, the term *coef0* is set equal to 0; it is an additive term inserted into the function describing the kernel. Finally, as regards the tolerance evaluated by the stop criterion,

the term *tol* is set equal to 0.001.

- **Random Forest (RF)**: It is a set of decision trees. A decision tree is a classifier based on a tree structure in which each branch is associated with a feature and a threshold, each node represents a partition of the data set and the leaf nodes, the final nodes of this structure, give an indication of the class to which the relevant partition belongs. Classification by means of decision trees follows a top-down approach in which, starting from the entire data base represented by the first node, also called the root-node, smaller and smaller partitions are created according to the value assumed by the features extracted until arriving at the roots of the tree. Random Forests are more complex structures that use the classification performed by several decision trees and then apply majority voting between the output of each tree to assign it the final class. Again, a function of the *sklearn* library was used, specifically *RandomForestClassifier*, which sees several default decision trees equal to 100.

- **Multi-layer Perceptron (MLP)**: it is a particular type of neural network in which the hidden neurons first process the data received from the input neurons by means of a linear weighted sum and then apply a non-lien activation function. Back-propagation was also implemented using the *sklearn MLPClassifier*. Again, the default parameters listed in table 2.1 were selected.

parameters	**value**
number hidden layers	100
activation function	relu
optimizer	Adam
batch size	min[200,n_sample]
learngin rate	costant
initial learning rate value	0.001
number of max iteration	200

Table 2.1: default values for *MLPClassifier* function

The four trained models were then applied to the test set.To evaluate their performance, accuracy and balance accuracy were calculated according to the eq.2.3 and eq.2.4.

$$accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2.3}$$

$$balance\ accuracy = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right) \tag{2.4}$$

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

The last step in this first method involves the lumen classification. To do this, the majority vote is applied, using the classes of the nuclei around each lumen. Accuracy and balance accuracy were evaluated.
Considering the performance obtained and the computational cost required to extract features from each nucleus, it was decided to abandon this method and proceed with the classification of the lumens without going through the classification of the nuclei.

Lumen classification

In this second approach we directly classify the lumen. It is visually difficult to distinguish the tubules lumen from the blood vessels one is no real color differences, whereas considering the shape and the size there are too many outliers to view these features as discriminating the two classes. The area around the lumen and the nuclei surrounding it seem more appropriate to discriminate both lumen types.

Before proceeding with feature extraction and selection, the algorithm for associating nuclei with lumens was modified. Similar to the previous algorithm, an initial association is established using the circular crown around the lumens. The association remain unambiguous and immediate if the nuclei is inserted in no or only one crown. We revised the case when the nuclei is subtended by several crowns.It was observed that, although a circular kernel is used, dilatation tends to alter the original shape of the nuclei. Moreover, tubules often have a particularly thick crown that might border on the lumens of some blood vessels.Consequently, even if the nuclei is within the crown of the corresponding lumen, it might be closer to the lumen of another element.
For all these reasons, the allocation criterion for nuclei inserted in several crowns has been modified. Specifically, the iterative expansion process is eliminated and only the evaluation of the pixels value belonging to the Bresenham path and separating the centers of mass of the i-th nuclei and the possible lumens is taken into account.

The extraction and the initial selection of features was then proceeded. Again, the selection of features was carried out under the advice of two experienced pathologists. This allowed an initial and substantial reduction in the number of features extracted.For each lumen of each image in the data set, 46 features were extracted:

- **lumen size**: number of pixels in the lumen mask.

- **lumen eccentricity**: eccentricity of the ellipse circumscribing the lumen mask eq.2.1.
- **lumen width and height**: although this information is already partly contained in the eccentricity value, it was decided extract the width and height of the lumen mask. To extract these two features, were not used the dimension of the bounding box surrounding the lumen mask, but the axes of the ellipse in which the mask is inscribed.
- **Convex Hull**: for the same reasons as above, the difference between the Hull convex envelope and the lumen mask, all normalised by the size of the lumen, is also evaluated.
- **number of nuclei**: the number of nuclei around the lumen.
- **crown color**: the mean, median and standard deviation of the pixel value belonging to the mask of the crown around the lumen are extracted with reference to four different images (red channel, green channel, blue channel and grey scale image). Specifically, the crown of the nuclei is considered. To extract the crown, the same procedure as for the assignment of the nuclei is implemented but using a smaller kernel. Specifically a square kernel of 40 pixels per side was chosen.So, three feature per image are extracted for a total of 12 features referring to the color of the crown. It was decided to take these features for the presence of the crown only around the tubules.It is clearly that it is characterized by a darker shade of magenta than the color of the interstitial tissue. Therefore, as far as the blood vessels are concerned, without such a crown, the mask detected will segment part of the interstitial tissue and will therefore be characterized by a generally lighter color.
- **nuclei size**: number of pixels that make up the masks of the nuclei outlining the lumen. As more than one nucleus is to be considered at a time, the mean and standard deviation of the sizes of all nuclei assigned to the considered lumen is evaluated. This results in a total of two features referring to the size of the nuclei.
- **nuclei shape**: the eccentricity of the ellipse circumscribing the nuclei mask is used. Again, the mean and standard deviation of the eccentricities of the nuclei contour the i-th lumen are evaluated for a total of two features referring to the shape of the nuclei.
- **nuclei color**: for each nucleus around the lumen, the mean, median and standard deviation of the pixel value belonging to the nuclei mask and four images (red channel, blue channel, green channel and grey scale image) are extracted. A total of 12 features are then extracted for each nuclei surrounding

the lumen. Again, what is inserted into the feature matrix is the mean and standard deviation of the values that each feature takes considering all the surrounding nuclei.This results in a total of 24 features referring to the color of the nuclei. These features have been extracted because the nuclei referring to blood vessels tend to be more saturated, while the color of the nuclei belonging to the tubules is often mixed with that of the crown.

The feature matrix obtained from the feature extraction process presents 46 columns and a number of rows equal to the number of lumens present in the entire data set. Again, to train the classifiers to equally recognize each class, a balancing of the train set was implemented by eliminating the additional elements of the most represented class. To reduce bias errors, a shuffle was performed between the rows of the feature matrices. Regarding normalisation, min-max scaling was again implemented on all the data sets.Since these are still supervised learning methods, lumen mask labelling was implemented under the guidance of an experienced pathologist.

To further improve performance, it was decided to carry out a validation of the input parameters of each classifier. The validation process involves setting a number of parameters and a set of values they can take. The generic classifier is trained using all possible combinations and then each trained model is validated on a set of new images contained in the validation set. For each validation, the balance accuracy on the validation set is evaluated. Only the set of values capable of maximizing it, is selected.
It is then necessary to extract a validation set; feature extraction and normalization of the feature matrix is also implemented for this new data set.
It is decided to implement only three of the four classifiers used for the classification of the nuclei, specifically: a K-Nearest-Neighbors (KNN), a Support Vector Machine (SVM) and a Random Forest (RF). Depending on the type of classifier, the number of parameters involved in the validation changes. For the KNN, only the value of k was chosen to vary. For the RF, the only parameter chosen was the number of trees t. For the SVM, on the other hand, several values were made variables, specifically the kernel, the degree of the polynomial kernel function (only in the case where the chosen kernel is of type *poly*), the value of C, *coef0* and *tol*. The values chosen for each parameter are shown in table 2.2.

classifier	parameter	chosen values
KNN	k	[3,5,7,15,25,35,45,55,65,75,85,95,105]
RF	t	[20,70,50,100,150,200]
SVM	kernel	['poly','rbf','sigmoid']
	C	[0.01,0.1,1]
	degree	[2,3,4,5]
	coef0	[0.5,1,5,10,20,30]
	tol	[1e-8,1e-6,1e-5,1e-4]

Table 2.2: values assumed by the parameters of the classifiers during the validation phase

The subsequent testing phase therefore involves only the pre-trained models chosen during validation. For each classifier, the *accuracy* (eq.2.3), the *balance accuracy* (eq.2.4) and the values of *precision* (eq.2.5) and *recall* (eq.2.6) were extracted for each of the classes.

$$precision = \frac{\text{TP}}{\text{TP+FP}} \tag{2.5}$$

$$recall = \frac{\text{TP}}{\text{TP+FN}} \tag{2.6}$$

In the first instance, the classifiers were trained and tested to discriminate tubules from blood vessels using all 46 extracted features. Subsequently analyzing the performance and considering the large number of available features, a manual feature selection was implemented. At each iteration, each model is trained using a single feature. The performance on the train set referring to the feature is then evaluated using balance accuracy. In this way any deterioration or improvement in classification can be recorded. Finally, only those features that achieve a higher balance accuracy value than the previously selected feature are selected. This method allows to isolating the redundant features, which therefore do not entail any type of variation in performance, and the features that instead worsen the quality of the classification. So, for each classifier, a different set of features is extracted and the balance accuracy trend is graphed.
Finally, we consider features that do not dependent on the pixel value of the image to obtain a color invariant classification and thus robust to variations in the tissue staining. To achieve this, all features related to the color of the crown and nuclei were eliminated. To validate each decision, various graphical methods of visualizing the features were implemented, such as box plot, histogram of the frequency of each feature within the train set, and heat map. These tools make it possible to distinguish the truly relevant features from those that are superfluous or redundant.

Once the best set of features and parameter values for each classifier had been chosen, an experienced pathologist was assigned to choose and label 7 patches representing the worst case. In this way, it was possible to test the algorithm robustness. As with the images of the test set, feature selection and normalisation of the feature map were implemented.
Finally, it was decided to exploit the non-color dependency by applying the models trained with the feature set regarding only shape and size, to 6 TRI-stained patches and their corresponding PAS-stained images generated by a Generative Adversarial Networks (GAN). More specifically a Pythorc implementation of a CycleGAN was used. The CycleGAN was trained to generate PAS-stained images from TRI-stained images. All patches extracted from the 18 available WSI images were used in the training phase. For the parameters tuning the default parameters were mostly used. It was therefore possible to test the true non-color dependence and to demonstrate a potential application of the developed system.

Chapter 3

Results

3.1 The data sets

During the visit to the Laboratoire Central d'Anatomie Pathologique (LCAP), **19 wsi** images were selected, 9 with PAS staining and 9 with TRI staining.
In the subsequent division of the WSI images into 1024x1024 patches and a first data selection, aimed to eliminate patches representing only the background and those with compromised quality due to artifacts, a total 15876 patches were extracted.
A second data selection was then performed to label the images. Two data sets are extracted. The first data set is used to implement the first classification method for distinguishing lumens from the classification of nuclei. This data set comprises **19 patches**, divided into a train set and a test set. Approximately 80% of the data were assigned to the train set while the remaining 20% to the test set. This results in a train set of 12 patches and a test set of 4. On the other hand, regarding the data set used for the second classification method, to improve the classification quality, its size was increased. Six patches per patient were extracted, for a total of **54** (1024 x 1024) patches. Having to implement a training phase, a validation phase and a test phase, it was necessary to divide the data set into three groups. To eliminate any kind of bias each data set contains images from different patients, specifically the train set contains images from 5 patients while the validation and test set contains images from 2 other patients. Thus, the train set consists of 30 patches while the validation and test set consists of 12. The patients were randomly assigned to the three data sets.

3.2 SAM approach

Considering the computational cost, in term of time, required to process a single image, the Segment Anything Model was applied to a reduced set of images. Specifically, only two WSI images patches and to four patches chosen for all other available images were used. This results in a total of **859 segmentations**.
Since the ground truth of each processed patch was not available,indicators were extracted to determine the quality of the segmentation. Specifically, the following were considered:

- **output**: number of masks obtain from the SAM implementation.
- **empty masks**: masks representing only larger elements without smaller masks within.
- **no empty masks**: masks representing larger elements and also smaller masks within them.
- **isolated lumens**: lumen masks not included in larger masks.
- **non isolated lumens**: lumen masks included in larger masks representing tubule, glomerulus, or blood vessel.

For each of these parameters, the maximum, minimum, and average values were evaluated. An average percentage value, calculated across all processed patches, was then added. This value represents the percentage that each type of mask occupies of the total number of extracted masks.
Table 3.1 shows the results referring to the first version of the model in which only some parameters were changed with respect to the default values. Specifically, points_per_side= **64**, pred_iou_thresh= **0.9**, stability_score_thresh= **0.96**, crop_n_layers= **1** and crop_n_point_downscale_factor= **2**.

MASK TIPE	N°MAX	%	N°MIN	%	N°AVERAGE	%
output	597	/	117	/	277	/
empty masks	21	46	0	0	5	11
no empty masks	48	72	1	11	17	33
isolated lumes	64	83	1	2	17	35
non isolated lumes	36	52	0	0	7	13

Table 3.1: Performance evaluation of the first implemented SAM

During parameter tuning, the same performance indicators were extracted. Considering the computation time required, the best combination of values differs from the model previously described for the values of stability_score_thresh= **0.95**, stability_score_offset= **1.2**, box_nms_thresh= **0.8**, crop_n_layer= **2**, crop_nms_thresh= **0.8** and min_mask_region_area= **50**. In table 3.2 the performances referred to this second SAM version are showed.

MASK TIPE	N°MAX	%	N°MIN	%	N°AVERAGE	%
output	839	/	167	/	411	/
empty masks	15	35	0	0	3	6.1
no empty masks	66	77	3	13	24	40
isolated lumes	67	81	0	0	19	33
non isolated lumes	36	48	0	0	8	13

Table 3.2: Performance evaluation of the second implemented SAM

Analyzing the table, specifically focusing on the average values and on the percentages, a notable improvement in the number of masks is evident. In fact, the number of masks obtained almost doubles. Regarding the number of isolated lumens, despite the increased number of masks, a percentage value decreases can be appreciate. On the other hand, there is no appreciable improvement from the point of view of non-isolated lumens, and even with the second model, images for which this value is zero are still present. However the performances are not sufficient to conduct lumen segmentation and classification.

3.3 Proposed solution

3.3.1 Segmentation

Exactly as in obtaining the performance regarding segmentation by SAM, ground truth is not available. Therefore, it is not possible to quantitatively evaluate the performance of the two segmentation algorithms used for lumen and nuclei segmentation. All the choices made were therefore based on visual analysis of the masks obtained, trying to define a pipeline that could be optimal for most of the images in the used data set.

3.3.2 Classification

Nuclei classification

For this first approach, two data sets are available: a train set consisting of a total of 4247 nuclei and 555 lumens and a test set with 1245 nuclei and 167 lumens. Only nuclei connected to a tubule or blood vessel are considered during the feature extraction process. Consequentially, the dimension of both data sets is reduced to **3769 nuclei** for the train set and **1095 nuclei** for the **test set**. The number of masks useful for classification is further reduced for the train set as class balancing is implemented. Specifically, the final **train set** includes **3242 nuclei**. Subsequent implementation of the four classifiers exhibited in 2.3.2 yielded the results exhibited in table 3.3. Specifically, for each classifier, the accuracy (eq. 2.3) and balance accuracy (eq. 2.4) values referring to the test set are reported. Finally, the best performance is achieved by the MLP.

classifier	**accuracy**[%]	**balance accuracy.**[%]
KNN	65,2	66,9
SVM	65,8	70,3
RF	66,4	69,5
MLP	66,1	70,8

Table 3.3: Accuracy and balance accuracy values evaluated on the test set for nuclei classification

Using these classifications, it was possible to implement lumen classification by majority voting. This process resulted in the performance shown in table 3.4. Again For each classifier the accuracy, balance accuracy, and percentage of unclassified elements are reported. Lumen to which no nucleus is assigned, or lumen that have in their surroundings a number of nuclei classified as belonging to a tubule equal to those classified as belonging to a blood vessel, are defined as not classifiable. In both these cases it is indeed not possible to assign a class to the lumen by majority voting. All classifiers, except the KNN, experienced an increase in performance.

classifier	**accuracy**[%]	**balance accuracy**[%]	**not classified**[%]
KNN	60,0	59,9	4,0
SVM	77,1	77,5	2,6
RF	74,3	74,3	2,0
MLP	77,1	77,3	0,6

Table 3.4: Accuracy, balance accuracy and missed values evaluated on the test set for lumen classification

It is also noted that majority voting reduces the gap between classifiers and that from a performance point of view the MLP was outperformed by the SVM. Considering, however, the missing values the MLP remains the first choice.
The presence of impossible-to-classify items remains in any case a problem that this type of approach is not able to overcome. Although the percentage of unclassified is low for MLP classification, if we consider applying this method to a larger data set we have a non-negligible number of unclassified masks.Finally, the performances are not satisfactory and it was decided to abandon this approach.

Lumen classification

To improve the performance of the classifiers, the dimensionality of the data set was increased. Specifically in this second approach, a training set consisting of 1555 lumens, a validation set comprising 731 lumens and a test set with 624 lumens are created. Again, during the feature extraction process, the size of each data set was reduced due to the presence of lumens within the glomeruli. Specifically, the feature extraction process resulted in the selection of 1296 masks for the training, **601** for the **validation** and **502** for the **test** set. The train set was finally balanced to make the classifiers able to equally recognize each class. At the end of this process, the number of lumens within the **train** set is further reduced to **1008**.

The first training implemented involves using all the extracted features.
In table 3.5 are reported the parameters selected during the validation phase and the corresponding balance accuracy value evaluated on the validation set. Table 3.6 shows the values of accuracy (eq. 2.3), balance accuracy (eq. 2.4), recall (eq. 2.6) and precision (eq. 2.5) for the three selected models and for classification by majority voting (MV). To improve the analysis of the results, the confidence interval was also evaluated. The value of the confidence interval is calculated by evaluating the standard deviation of the performances between the images in the test set.

classifier	**selected values**	**balance accuracy**[%]
KNN	k=15	76,8
SVM	kernel='poly' C=1 degree=3 coef0=20 tol=1e-08	89,9
RF	t=100	79,2

Table 3.5: Values of the classifiers' parameters during the validation phase

classifier	accuracy[%]	balance accuracy[%]
KNN	73,9±7,0	73,4±7,5
SVM	89,4±5,7	89,4±5,7
RF	78,9±6,0	78,6±5,8
MV	84,1±5,3	83,8±6,5

classifier	tubul precision[%]	tubul recall[%]
KNN	69,2±8,6	89,2±6,4
SVM	89,3±9,1	90,3±7,3
RF	75,9±9,2	86,5±10,7
MV	80,8±10,5	90,7±6,2

classifier	vessel precision[%]	vessel recall[%]
KNN	83,3±11,9	57,6±14,9
SVM	89,6±12,2	88,5±13,5
RF	83,1±16,9	70,8±10,7
MV	88,6±11,2	77,0±14,8

Table 3.6: Performances obtained using all extracted features

The best classification is obtained by the SVM, followed by majority voting. Although the accuracy and balance accuracy values are acceptable there is a very high standard deviation. The number of used features is also too high. For these reasons an iterative feature selection was implemented.
The features selected for each classifier are shown in table 3.7, specifically 21 features are extracted for the KNN and SVM and 24 for the RF. Variables related to color are expressed in the format *statistic parameters_image type_mask type*. Statistical parameters include mean and standard deviation (std). As for the type of image used instead red channel (r), green channel (g), blue channel (b) and gray scale image (gr). Finally, as mask type the crown around the lumens and mask of the nuclei in the inside of the lumen can be found.

Figure 3.1 shows the graphs obtained from the extracted balance accuracy from the application of each feature individually. The three graphs have approximately the same trend. It is also possible to observe the presence of features that are redundant or cause performance decreasing.

KNN	SVM	RF
lumen_size	lumen_size	lumen_size
lumen_width	hight_lumen	lumen_width
convex_hull	number_nuclei	hight_lumen
median_r_crown	median_r_crown	mean_r_crown
std_r_crown	std_r_crown	std_r_crown
std_g_crown	std_g_crown	mean_g_crown
median_b_crown	median_b_crown	std_g_crown
std_b_crown	std_b_crown	mean_gr_crown
std_gr_crown	median_gr_crown	std_gr_crown
mean_mean_r_nuclei	std_gr_crown	std_nuclei_size
mean_median_r_nuclei	std_mean_r_nuclei	std_nuclei_size
std_median_r_nuclei	mean_median_r_nuclei	mean_mean_r_nuclei
mean_std_r_nuclei	mean_std_r_nuclei	std_median_r_nuclei
mean_mean_g_nuclei	std_mean_g_nuclei	mean_std_r_nuclei
mean_median_g_nuclei	mean_median_g_nuclei	mean_mean_g_nuclei
mean_mean_b_nuclei	mean_std_g_nuclei	std_median_g_nuclei
mean_median_b_nuclei	mean_median_b_nuclei	mean_std_g_nuclei
mean_std_b_nuclei	mean_std_b_nuclei	mean_mean_b_nuclei
mean_mean_gr_nuclei	std_mean_gr_nuclei	std_median_b_nuclei
mean_median_gr_nuclei	mean_median_gr_nuclei	mean_std_b_nuclei
mean_std_gr_nuclei	mean_std_gr_nuclei	mean_mean_gr_nuclei
		std_median_gr_nuclei
		mean_std_gr_nuclei

Table 3.7: Feature selected by manual feature selection.

Table 3.8 and 3.9 show, respectively, the values of the parameters that maximize the balance accuracy on the validation set, and the performance evaluated on the test set. From these values, feature selection resulted in a noticeable increase in performance, achieving accuracy and balance accuracy values above 80% in the case of the SVM.

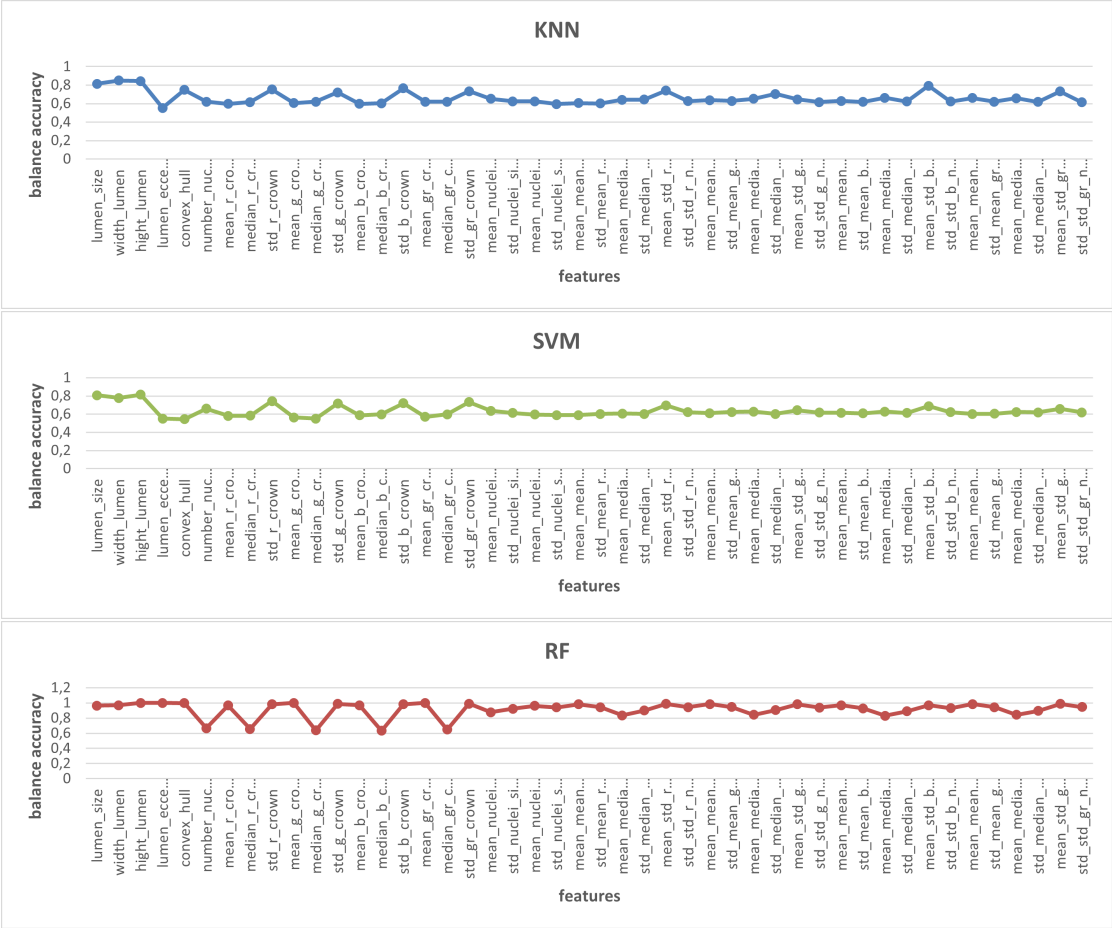


Figure 3.1: Graphs related to the performance obtained on the train set by applying each feature individually.

classifier	selected values	balance accuracy[%]
KNN	k=15	72,0
SVM	kernel='poly' C=0.1 degree=3 coef0=20 tol=1e-08	88,8
RF	t=200	73,7

Table 3.8: Values of the classifiers' parameters during the validation phase

classifier	accuracy[%]	balance accuracy[%]
KNN	74,7±7,4	74,1±8,3
SVM	**88,6±6,8**	**88,5±6,6**
RF	81,5±7,4	81,1±7,1
MV	82,7±6,0	82,3±5,9

classifier	tubul precision[%]	tubul recall[%]
KNN	69,3±8,8	91,5±6,4
SVM	**85,6±10,2**	**93,8±5,4**
RF	75,9±10,4	93,8±6,8
MV	77,4±9,2	93,8±6,1

classifier	vessel precision[%]	vessel recall[%]
KNN	86,3±13,4	56,8±14,6
SVM	**92,7±7,4**	**83,1±11,8**
RF	91,2±10,2	68,3±12,5
MV	91,5±8,8	70,8±11,6

Table 3.9: Performances after the manual feature selection

To make the classification color invariant, all color-related features were finally eliminated. Specifically, 10 features are selected:

- **lumen_size**
- **width_lumen**
- **hight_lumen**
- **convex_hull**: difference between Hull's convex envelope evaluated on the lumen mask and the lumen mask
- **number_nuclei**
- **mean_nuclei_size**: average of the size of the nuclei around the lumen
- **std_nuclei_size**: standard deviation of the size of nuclei around the lumen
- **mean_nuclei_shape**: average of the eccentricity of the nuclei around the lumen
- **std_nuclei_shape**: standard deviation of the eccentricity of the nuclei around the lumen

Tables 3.10 and 3.11 report the values selected for the input parameters of the classifiers during the validation phase, and the performance on the test set.

classifier	selected values	balance accuracy[%]
KNN	k=55	74,4
SVM	kernel='poly' C=0.01 degree=3 coef0=10 tol=1e-08	80,2
RF	t=50	64,2

Table 3.10: Values of the classifiers' parameters during the validation phase

classifier	accuracy[%]	balance accuracy[%]
KNN	82,5±5,22	82,4±6,3
SVM	**91,0±5,3**	**91,1±5,3**
RF	74,3±9,8	73,6±7,0
MV	86,8±5,4	86,7±5,1

classifier	tubul precision[%]	tubul recall[%]
KNN	82,0±9,2	84,6±8,0
SVM	**92,8±8,0**	89,6±6,4
RF	67,5±13,4	**96,9±3,0**
MV	84,1±10,0	92,0±5,0

classifier	vessel precision[%]	vessel recall[%]
KNN	83,0±13,0	80,2±7,0
SVM	89,3±8,5	**92,6±6,0**
RF	**93,8±5,6**	50,2±13,0
MV	90,4±8,3	81,5±7,4

Table 3.11: Performances obtained from color invariant classification

This new feature set not only allows for a system that can classify lumens regardless of the pixel value of the image, but leads to accuracy and balance accuracy values higher than 90% in the case of SVM. RF, on the other hand, produces the best results for recall associated with tubules and precision associated with blood vessels. However, given the substantial gap between precision and recall for both classes, RF was discarded and SVM was chosen as the best classifier.

To test the robustness of the system, the best models were subsequently applied to a set of images chosen by an experienced pathologist and representing the **worst case**. Specifically, 7 images with a total of **554** nuclei were chosen. Table 3.12 shows the performance obtained by applying the previously trained models to this new set of images.

classifier	**accuracy**[%]	**balance accuracy**[%]
KNN	**69,3±10,1**	71,9±6,6
SVM	66,4±9,9	**73,7±6,6**
RF	46,2±9,0	62,2±6,9
MV	64,4±12,6	71,9±8,5

classifier	**tubul precision**[%]	**tubul recall**[%]
KNN	**45,2±8,0**	77,4±11,5
SVM	43,3±6,6	89,0±7,2
RF	32,4±6,0	**95,9±3,6**
MV	41,7±7,5	87,7±8,3

classifier	**vessel precision**[%]	**vessel recall**[%]
KNN	89,1±4,3	66,4±15,8
SVM	93,7±4,4	**58,3±16,0**
RF	**95,1±9,0**	28,4±14,8
MV	92,7±6,0	56,1±19,8

Table 3.12: Performances obtained by applying best models to worst-case images

As expected, there is a notable deterioration in performance. In fact, the maximum value of balance accuracy achieved is just over 70%.
However, it is important to note that the feature set that achieved the best performance on the test set is still able to provide precision and recall values above 90%.

Finally, to test the non color dependence classification, TRI-stained images were classified using the pre-trained optimal models. These images were also used within a GAN capable of recreating the corresponding PAS-staining images. In this last application of the model, 6 images with TRI-staining and the corresponding images generated by the GAN are used. Considering that the segmentation process has some color-dependent parameters, a different number of lumen useful for segmentation are obtained. Specifically, the data set containing **TRI-staining** images includes **200 lumen** while the data set of **generated images 211**. In tables 3.13 and 3.14 the performances, obtained by applying the pre-trained models,related to the data set of TRI-stained and PAS-stained images, respectively, are shown.

classifier	accuracy[%]	balance accuracy[%]
KNN	**72,5±4,3**	**72,9±5,4**
SVM	69,5±5,2	69,9±5,4
RF	55,0±8,1	55,9±2,5
MV	67,5±5,5	68,0±5,5

classifier	tubul precision[%]	tubul recall[%]
KNN	**65,9±5,9**	90,8±7,5
SVM	63,0±9,1	89,8±9,0
RF	52,1±8,4	**100,0±0,0**
MV	60,9±7,0	93,9±5,7

classifier	vessel precision[%]	vessel recall[%]
KNN	86,2±12,8	**54,9±13,1**
SVM	83,6±11,7	50,0±14,4
RF	**100,0±0**	11,8±5,0
MV	87,8±13,3	42,2±12,2

Table 3.13: Performances obtained using images with TRI type staining

classifier	accuracy[%]	balance accuracy[%]
KNN	75,8±8,8	75,1±9,3
SVM	**84,4±9,6**	**83,6±10,5**
RF	78,2±12,4	77,1±12,8
MV	83,4±9,6	82,4±10,9

classifier	tubul precision[%]	tubul recall[%]
KNN	79,7±12,7	64,3±20,7
SVM	**91,1±8,2**	**73,5±18,7**
RF	87,1±11,3	62,2±23,2
MV	94,4±7,4	68,4±19,6

classifier	vessel precision[%]	vessel recall[%]
KNN	73,5±11,7	85,8±15,7
SVM	**80,3±11,8**	93,8±7,3
RF	73,8±14,2	92,0±6,9
MV	77,9±11,1	**96,5±3,6**

Table 3.14: Performance obtained using PAS-type stained images generated by a GAN from TRI-type stained images.

From the latter tables the trained models are indeed color invariant. Despite a decrease in performance, accuracy and balance accuracy values exceeding 70% are still achievable when applying the models to TRI stained images.
It's also interesting to notice how the use of images generated through GAN demonstrates performance comparable to that of the best model. This kind of phenomena suggest the potential resolution of issues related to the dependency between classification and image color.

Chapter 4

Conclusion

The aim of this project is to identify three biological structures of the kidney: tubules, blood vessels and epithelial cells. The aim is to use this system in combination with a previously developed algorithm for the recognition of glomeruli, in order to select the four protagonists of the Banff analysis: tubules, blood vessels, interstitial tissue and glomeruli. Analising the results, it is clear that the trained models, using the selected feature set, are capable of recognising and classifying tubules and glomeruli in the test set images with high accuracy and precision. In particular, the best performing Machine Learning model is the Support Vector Machine, which achieves a test set a balanced accuracy of 91%, with precision and recall levels of no less than 89% for both classes.The model's ability to classify images independently of colour was also demonstrated, allowing the classifier to be applied not only to images with different PAS staining concentrations, but also to images with other types of staining. One of the main innovations of the proposed method is the use of a Cycle-GAN to convert images with different types of staining to the PAS-type staining, on which the parameters were set and on which the model was trained. Despite the sufficient performance on TRI-type stained samples, the images generated by the GAN show significantly superior results. In this context, SVM is confirmed as the best classifier, with a balanced accuracy of 83.6%, and precision of 91% and 80% for tubules and blood vessels, respectively. Considering that the Banff classification involves the use of three different types of staining (H&M, PAS and TRI), the combined use of Cycle-GAN and the classification system would allow the combined use of the different staining techniques.

In order to replicate the classification process performed by a human operator, features were extracted and selected under the guidance of a team of external pathologists. It was therefore possible to identify the set of variables that best represent the differences between tubules and blood vessels, despite the fact that

these two elements are often very similar to each other. Despite the innovative feature extraction method, some classification errors were found. Through visual analysis, it can be seen that the calssification errors can be attributed to excessive dependence on the size of the mask. Often there are blood vessels with lumens of comparable size to tubules, or vice versa. In many of these cases, the classifier assigns the mask to the wrong class. Segmentation errors were also observed. In many cases, the algorithm designed to aggregate masks associated with the same lumen, although recognised as distinct, fails to joint all masks related to the same lumen. Consequently, the classifier sees the isolated mask as an independent element. Even in terms of worst-case performance, there are some possible improvement. Considering that the Banff classification is applied to images from pathological kidneys, it is not uncommon to find patches with similar characteristics to worst case ones. Finally, as mentioned above, retrieved tissues have elements that even a seasoned pathologist finds challenging to categorize. Furthermore, in the case of patients with advanced IFTA, tissue features significantly change. In these cases, the algorithm implemented may have difficulty to recognise all structures of interest.

To address these challenges with the method devised, various strategies for improvement can be considered. A first solution would be to apply post-processing systems to the individual lumen masks. Upon examining certain images, it becomes evident that lumens belonging to different structures are very close to each other.This makes it very difficult to apply certain morphological operators that aim to eliminate black pixels within the masks. Applying expansion to the entire mask, for instance, distinct but very close lumens would form a single mask. This issue can be effectively circumvented by applying morphological operators to individual masks. Notably, an analysis of this parameter reveals that incorrectly classified masks are characterized by a smaller distance compared to correctly classified ones. Nonetheless, by eliminating the size of the lumens from the set of extracted features, the classification performance considerably decreases. However, one potential improvement could involve applying a threshold to the distance between the individual feature to be classified and the plane utilized by the SVM classifier to differentiate between the two classes. This strategic approach is supported by the analysis of this parameter, which indicates that incorrectly classified masks exhibit a smaller distance compared to correctly classified ones. In the pursuit of improving the segmentation quality, techniques involving AI could be adopted instead of global thresholding.
Additionally, a consideration could be given to combine the results obtained from the segmentation of lumens and epithelial cells to improve SAM performance. One of the main strengths of the SAM model is the possibility of providing prompts to direct the segmentation. For instance, if the coordinates of the centroids of masks obtained through global thresholding were supplied, SAM would be able to perform

a more selective and precise segmentation.

Despite the many areas open to improvement, the work undertaken represents a good starting point for the development of a method capable of conducting the Banff classification independently.

Bibliography

[1] Muhammed Mubarak and et al. «Evolution of human kidney allograft pathology diagnostics through 30 years of the Banff classification process». In: *World Journal of Transplantation* 13.5 (2023), pp. 221–238. DOI: 10.5500/wjt.v13.i5.221 (cit. on pp. 1, 6, 7).

[2] Jin Zhang and et al. «Etiological analysis of graft dysfunction following living kidney transplantation: a report of 366 biopsies». In: *Renal Failure* 40.1 (2018), pp. 219–225. DOI: 10.1080/0886022X.2018.1455592 (cit. on p. 1).

[3] K Solez and et al. «International standardization of criteria for the histologic diagnosis of renal allograft rejection: the Banff working classification of kidney transplant pathology». In: *Kidney International* 44.2 (1993), pp. 411–422. DOI: 10.1038/ki.1993.259 (cit. on pp. 1, 8).

[4] Hyeon Joo Jeong. «Diagnosis of renal transplant rejection: Banff classification and beyond». In: *Kidney Research and Clinical Practice* 39.1 (2020), pp. 17–31. DOI: 10.23876/j.krcp.20.003 (cit. on pp. 1–3, 6–8).

[5] L. C. Racusen and et al. «The Banff 97 working classification of renal allograft pathology». In: *Kidney International* 55.2 (1999), pp. 713–723. DOI: 10.1046/j.1523-1755.1999.00299.x (cit. on pp. 1–4).

[6] M. Haas and et al. «Banff 2013 meeting report: inclusion of c4d-negative antibody-mediated rejection and antibody-associated arterial lesions». In: *American Journal of Transplantation* 14.2 (2014), pp. 272–283. DOI: 10.1111/ajt.12590 (cit. on pp. 2, 3).

[7] Candice Roufosse and et al. «A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology». In: *Transplantation* 102.11 (2018), pp. 1795–1814. DOI: 10.1097/TP.0000000000002366 (cit. on pp. 2–8).

[8] M. Mengel, J. Reeve, S. Bunnag, G. Einecke, G.S. Jhangri, B. Sis, K. Famulski, L. Guembes-Hidalgo, and P.F. Halloran. «Scoring Total Inflammation Is Superior to the Current Banff Inflammation Score in Predicting Outcome and the Degree of Molecular Disturbance in Renal Allografts». In: *American*

Journal of Transplantation 9.8 (2009), pp. 1859–1867. ISSN: 1600-6135. DOI: 10.1111/j.1600-6143.2009.02727.x (cit. on p. 3).

[9] M. Haas and et al. «The Banff 2017 Kidney Meeting Report: Revised diagnostic criteria for chronic active T cell-mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials». In: *American Journal of Transplantation* 18.2 (2018), pp. 293–307. DOI: 10.1111/ajt.14625 (cit. on pp. 3–6).

[10] K. Solez and et al. «Banff 07 classification of renal allograft pathology: updates and future directions». In: *American Journal of Transplantation* 8.4 (2008), pp. 753–760. DOI: 10.1111/j.1600-6143.2008.02159.x (cit. on p. 4).

[11] A. Loupy and et al. «The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology». In: *American Journal of Transplantation* 17.1 (2017), pp. 28–41. DOI: 10.1111/ajt.14107 (cit. on pp. 5, 6).

[12] Brian J. Nankivell. «The meaning of borderline rejection in kidney transplantation». In: *Kidney International* 98.2 (2020), pp. 278–280. DOI: 10.1016/j.kint.2020.04.052 (cit. on p. 7).

[13] Erik Stites and et al. «High levels of dd-cfDNA identify patients with TCMR 1A and borderline allograft rejection at elevated risk of graft injury». In: *American Journal of Transplantation* 20.9 (2020), pp. 2491–2498. DOI: 10.1111/ajt.15822 (cit. on p. 7).

[14] Meyke Hermsen and et al. «Deep Learning-Based Histopathologic Assessment of Kidney Tissue». In: *Journal of the American Society of Nephrology* 30.10 (2019), pp. 1968–1979. DOI: 10.1681/ASN.2019020144 (cit. on pp. 9, 10).

[15] Jesper Kers, Roman D Bülow, Barbara M Klinkhammer, Gerben E Breimer, Francesco Fontana, Adeyemi Adefidipe Abiola, and et al. «Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study». In: *Journal Name* Volume Number.Issue Number (2021), Page Numbers. DOI: 10.1016/S2589-7500(21)00211-9 (cit. on pp. 9–11).

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. «Deep Learning». In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539 (cit. on p. 9).

[17] Anders Brun, Hans Knutsson, Hae-Jeong Park, Martha Elizabeth Shenton, and Carl-Fredrik Westin. «Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004». In: *Lect Notes Comput Sci* (2004) (cit. on p. 9).

[18] D. Tellez and et al. «Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks». In: *IEEE Transactions on Medical Imaging* 37.9 (2018), pp. 2126–2136. DOI: 10.1109/TMI.2018.2820199 (cit. on p. 9).

[19] Diederik P. Kingma and Jimmy Lei Ba. «Adam: A Method for Stochastic Optimization». In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Presented at the Third International Conference on Learning Representations. San Diego, CA, USA, May 2015 (cit. on p. 9).

[20] Satoshi Hara and et al. «Evaluating tubulointerstitial compartments in renal biopsy specimens using a deep learning-based approach for classifying normal and abnormal tubules». In: *PloS One* 17.7 (July 2022), e0271161. DOI: 10.1371/journal.pone.0271161 (cit. on pp. 11, 12).

[21] Daniel Yoo and et al. «An automated histological classification system for precision diagnostics of kidney allografts». In: *Nature Medicine* 29.5 (2023), pp. 1211–1220. DOI: 10.1038/s41591-023-02323-6 (cit. on p. 12).

[22] Michael Mengel and Xian C. Li. «Automation of Banff rules for precision diagnosis». In: *American Journal of Transplantation* (July 2023). DOI: 10.1016/j.ajt.2023.07.005 (cit. on p. 12).

[23] Alexander Kirillov et al. «Segment Anything». In: *Proceedings of the International Conference on Computer Vision (ICCV)*. ICCV 2023, Alexander Kirillov and Eric Mintun are joint first authors. ICCV. 2023 (cit. on pp. 20–24).

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. «Microsoft COCO: Common Objects in Context». In: *European Conference on Computer Vision (ECCV)*. 2014. URL: https://openaccess.thecvf.com/content_eccv_2014/html/Lin_Microsoft_COCO_Common_2014_ECCV_paper.html (cit. on p. 21).

[25] Agrim Gupta, Piotr Dollar, and Ross Girshick. «LVIS: A Dataset for Large Vocabulary Instance Segmentation». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on p. 21).

[26] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. «Semantic Understanding of Scenes Through the ADE20K Dataset». In: *International Journal of Computer Vision* 127.3 (Mar. 2019), pp. 302–321. DOI: 10.1007/s11263-018-1140-0 (cit. on p. 21).

[27] Alina Kuznetsova et al. «The Open Images Dataset V4». In: *International Journal of Computer Vision* 128 (2018), pp. 1956–1981. URL: https://api.semanticscholar.org/CorpusID:53296866 (cit. on p. 21).

[28] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. «Contour Detection and Hierarchical Image Segmentation». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2010) (cit. on p. 21).

[29] Xiaofeng Ren and Jitendra Malik. «Learning a Classification Model for Segmentation». In: *IEEE International Conference on Computer Vision (ICCV)*. 2003 (cit. on p. 21).

[30] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. «What is an Object?» In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010 (cit. on p. 21).

[31] Chris Stauffer and W. Eric L. Grimson. «Adaptive Background Mixture Models for Real-Time Tracking». In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1999 (cit. on p. 21).

[32] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. «TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation». In: *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–15. ISBN: 978-3-540-33833-8 (cit. on p. 21).

[33] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. «Panoptic Segmentation». In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on p. 21).

[34] Bruno C. da Silva, George Dimitri Konidaris, and Andrew G. Barto. «Learning Parameterized Skills». In: *International Conference on Machine Learning*. 2012. URL: https://api.semanticscholar.org/CorpusID:9098305 (cit. on p. 21).

[35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. «Masked Autoencoders Are Scalable Vision Learners». In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 15979–15988. URL: https://api.semanticscholar.org/CorpusID:243985980 (cit. on p. 22).

[36] Alexey Dosovitskiy et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». In: *ArXiv* abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882 (cit. on p. 22).

[37] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. «Exploring Plain Vision Transformer Backbones for Object Detection». In: *ArXiv* abs/2203.16527 (2022). URL: https://api.semanticscholar.org/CorpusID:247793203 (cit. on p. 22).

[38] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. «Layer Normalization». In: *ArXiv* abs/1607.06450 (2016). URL: https://api.semanticscholar.org/CorpusID:8236317 (cit. on p. 22).

[39] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. «Automatic Image Colorization Via Multimodal Predictions». In: *Computer Vision – ECCV 2008*. Ed. by David Forsyth, Philip Torr, and Andrew Zisserman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 126–139. ISBN: 978-3-540-88690-7 (cit. on p. 23).

[40] Abner Guzmán-Rivera, Dhruv Batra, and Pushmeet Kohli. «Multiple Choice Learning: Learning to Produce Multiple Structured Outputs». In: *Neural Information Processing Systems*. 2012. URL: https://api.semanticscholar.org/CorpusID:5730937 (cit. on p. 23).

[41] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. «Interactive Image Segmentation with Latent Diversity». In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 577–585. URL: https://api.semanticscholar.org/CorpusID:3895849 (cit. on p. 23).

[42] Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. «Reviving Iterative Training with Mask Guidance for Interactive Segmentation». In: *2022 IEEE International Conference on Image Processing (ICIP)* (2021), pp. 3141–3145. URL: https://api.semanticscholar.org/CorpusID:231918551 (cit. on p. 23).

[43] Marco Forte, Brian L. Price, Scott D. Cohen, Ning Xu, and Franccois Piti'e. «Getting to 99% Accuracy in Interactive Segmentation». In: *ArXiv* abs/2003.07932 (2020). URL: https://api.semanticscholar.org/CorpusID:212747938 (cit. on p. 23).

[44] Ilya Loshchilov and Frank Hutter. «Decoupled Weight Decay Regularization». In: *International Conference on Learning Representations*. 2017. URL: https://api.semanticscholar.org/CorpusID:53592270 (cit. on p. 24).

[45] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, and Xuelian Deng. «A novel kNN algorithm with data-driven k parameter computation». In: *Pattern Recognition Letters* 109 (2018). Special Issue on Pattern Discovery from Multi-Source Data (PDMSD), pp. 44–54. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2017.09.036. URL: https://www.sciencedirect.com/science/article/pii/S0167865517303562 (cit. on p. 43).