POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Master's Degree Thesis

Development of a deep learning-based method for artifact detection and quality controls in digital pathology

Supervisor

Ing. Massimo Salvi

Candidate

Roberta Patti

Co-supervisor

Prof. Filippo Molinari

December 2022

Abstract

In the last years, in an attempt to overcome some of the issues of traditional pathology, whole slide digital scanners have been adopted, enabling the transition of pathology into the digital era. Histological slides can now be digitalized, but the process that goes from the collection of the tissue to the digital image consists of specific sequential steps, typically carried out manually by laboratory technicians, each of which can introduce artifacts such as tissue folds, air bubble, pen marker or dust, that can lower the quality of the histological image. These artifacts may alter the appearance of the tissue, making diagnosis difficult for the pathologist and negatively affecting the performance of automatic algorithms that operate on digital WSI. For these reasons, a quality control mechanism is needed.

Currently, most quality control processes are performed manually, but it is a laborious and subjective task; moreover, there are artifacts that may complicate the quantitative analysis of automated algorithms while not having implications for the pathologist's diagnosis. An automated approach for quality controls can help to overcome these problems.

The aim of this thesis is to develop a fully automated quality control system for histological slides. To achieve this goal, a multi-class approach based on deep learning was used which allows, starting from the pyramidal image, to identify the artifacts present on the entire slide. The dataset, including thousands of whole slide images (WSIs) relating to 9 different organs and 4 different stains, was manually annotated, and for each WSI the images at a magnification of 1.25x and 5x and the corresponding manual masks were extracted. After the data preparation, the tiles extracted from the starting images and masks were used to train the neural networks for the segmentation of tissues and artifacts. The architecture used for the neural networks is the DeepLabV3+ model with a ResNeSt as backbone, that is a split-attention network which uses the typical attention mechanism applied on channel to capture cross-channel feature correlations, while preserving a multi-path architecture to learn independent features. For the test phase, a sliding window approach has been used, keeping only the center of each prediction. The results show that this system has a high generalizability. For this task, DSC values higher than 90% were achieved, overcoming the performances of other methods currently present in literature.

The approach developed proves to be useful for two main aspects: on one side, introducing an automated quality control system into the daily workflow of pathologists allows to quickly identify poor-quality slides that needs to be reproduced or rescanned, thus speeding up the workflow and avoiding delays in the formulation of diagnoses; on the other side, it can help the development of new automatic algorithms, identifying the regions that should to be avoided during the training and the testing of the algorithm, so as to have more robust and faster algorithms and avoid results distorted by the presence of artifacts.

In the future, this model can be further improved by increasing the number of poorly represented organ-color combinations, reducing training times and developing a more structured quality score for histological slides.

Contents

1 Introduction **1**

1.1 Digital pathology 1

1.2 Histological stains 4

1.2.1 Routine stain 4

1.2.2 Special stains and immunohistochemical stains 5

1.3 Artifacts in histology 10

1.4 Quality controls in digital pathology 17

1.4.1 State of the art 17

2 Materials and Method **23**

2.1 Dataset 23

2.2 Pipeline 25

2.3 Manual mask generation 26

2.4 CNNs for tissue and artifact detection 29

2.4.1 Data preparation 29

2.4.2 Training 34

2.4.3 Testing 46

2.5 Validation metrics 49

3 Results **51**

3.1 Tissue network 51

3.2 Artifact network 56

4 Conclusions and further developments **71**

Bibliography **73**

Chapter 1

Introduction

1.1 Digital pathology

The term *digital pathology*, which initially refers to the process of digitizing whole slide images (WSIs) using advanced slide scanning technology, is now a generic term that includes approaches based on artificial intelligence for detection, segmentation, diagnosis, and analysis of digitized images [1].

In an effort to address some of the issues with traditional pathology techniques, complex, new imaging systems and whole slide image (WSI) scanners have been developed and adopted, enabling the transition of pathology into the digital era. Within minutes, WSI scanners capture multiple images of entire tissue sections on the slide, which are digitally stitched together to generate a very high-resolution image that replicates the entire histopathologic glass slide and can be reviewed by a pathologist on a computer monitor [2,3].

To support a streamlined method for loading images, whole slide images are stored at multiple resolutions. By precomputing lower-resolution representations of the whole slide image, this multi-resolution representation, also known as an image pyramid, provides more efficient data flow. Each level of the pyramid is down-sampled from the previous level (Figure 1.1) [4,5].

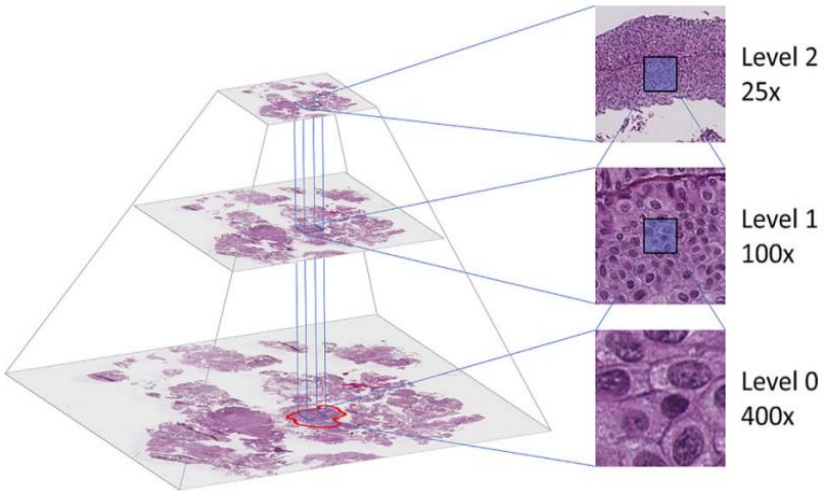


Figure 1.1. Pyramidal histological image with three levels [5].

There are many practical advantages to using these digital pathology image systems, that include the organization and storage of large amounts of data in a centralized location, the simple sharing of image data to enable cross-specialty worldwide remote communication, integration of digital software in the workflow to help streamline procedures and increase efficiency, a shorter testing turnaround time, and the creation of precise and highly reproducible tissue-derived readouts reducing inter-pathologist variability. Digital pathology also has the potential to help improve clinical workflows, reducing the need for storing glass slide on site and reducing the risk of physical slides getting broken or lost.

Digital image analysis is used in pathology to quantitatively assess histological features, morphological patterns, and biologically relevant regions of interest. It can quickly and accurately identify and quantify particular cell types. Performing similar tasks manually can take a lot of time and can be prone to human error. Additionally, data from tissue slides that may not be available during manual evaluation via routine microscopy can be collected using quantitative image analysis tools.

Despite the benefits of applying digital pathology in a clinical context, challenges remain. Algorithm development is a time-consuming endeavor, the success of AI-based approaches relies on the quality and quantity of data used to train the algorithm and data produced from these algorithms can be difficult to extract and interpret. Moreover, limited access to large, well-annotated dataset may limit clinical utility.

However, the efficiency gains, such as faster results and higher throughput, are key motivators for pathologists to adopt digital pathology. As a result, WSI devices have become important tools that may help pathology research and routine diagnostic work [3].

In order to obtain digitalized histological images, specific sequential steps must be followed to ensure an optimal slide preparation (Figure 1.2). These steps are typically carried out manually by laboratory technicians [6]:

a) Collection and fixation: the sample is treated with an appropriate chemical fixative depending on the tissue and the analysis to be performed in order to prevent physical and chemical alterations, preserve the tissue from bacteria or cellular enzymes, and maintain an optimal contrast between histological structures.
b) Dehydration and clearing: the specimen is dehydrated by dipping it into increasing concentrations of alcohol and cleared through an organic solvent, with the aim of removing the fixative agent and water from the tissue, but if not done correctly it can introduce an artifact as water droplets can form in the sample. The consequence is that the opacity of the

affected regions can cause the loss of microscopic details, and the appearance of the tissue may be different from that expected on the basis of the staining used.

c) Paraffin embedding: in this step heated paraffin is used to embed the specimen with the aim of evaporating the solvent used in the dehydration and clearing phase and fill in any remaining spaces in the histological sample. The result of this step is the tissue sample embedded in a paraffin block.

d) Microtomy: a microtome is used to gradually dissect the block obtained in the previous step, after briefly exposing it to low temperatures to harden it. It's crucial to cut uniformly the section of tissue in order to preserve the sample's appearance. A thickness of 5 μm is ideal as it allows to see both the architectures of the tissue, not visible with smaller thicknesses, and the morphology of the cells, not visible with larger thicknesses.

e) Staining: at this stage most cells and other tissue structures of the specimen are transparent and show relatively low contrast when scanned through any microscope. Therefore, specific dyes are used to emphasize the cellular components on the sample. Aspects that affect the final appearance of the tissue sample are the solution's pH and concentration, but also the time for which the procedure is carried out. The exposure time to the stain determines the intensity of staining.

f) Mounting: a transparent coverslip is used to enclose the histological sample in order to prevent the alteration of the tissue caused by external agents. This may result in artifacts such as air bubbles, dust or the presence of microorganisms which contaminate the specimen.

g) Digitalization: this step involves the use of scanners to digitize the histological slide; this creates a high-resolution digital image from the glass slide specimens. Since there are several scanning platforms and acquisition technologies, the same sample may appear in different ways when scanned by different scanners. Moreover, if the histological slide is not lined up with the scanner's focal plane, image blurring may happen.

It is evident that during this manual process different artifacts can be generated, and this can lower the quality of the histological image.

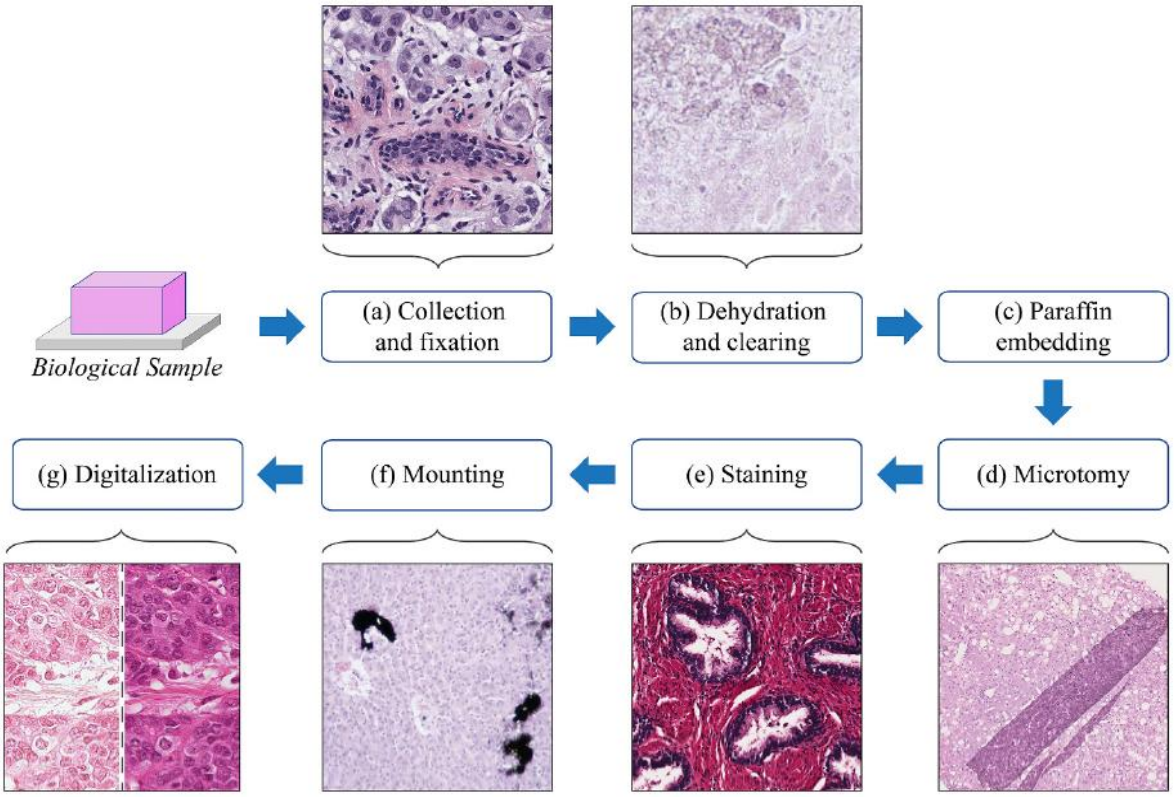


Figure 1.2. Steps followed to obtain the digitalized histological slide starting from the biological sample [6].

1.2 Histological stains

A tissue sample observed under the light microscope appears almost colorless. In order to allow to distinguish between different cells and tissue, staining is used to create differential coloration. A huge range of stains is used in histology, from dyes and metal to labeled antibodies.

1.2.1 Routine stain

The routine stain is haematoxylin and eosin (H&E), a dichromic staining. This is the most common histologic stain, used to differentiate tissue structures and in the diagnoses of various pathologies, and can be used on all organs.

Haematoxylin is a naturally occurring dye derived from the tree *Haematoxylum campechianum*. Hematein, the real dye employed in an H&E stain, is created by oxidizing hematoxylin. This dye cannot bind to the anionic components of the tissue as it is also negatively charged, consequently a mordant is required that introduces positive charges in order to generate the staining. Actually, the staining is caused by the mordant in conjunction with the hematein. On the basis of the mordant used, different types of hematoxylin are distinguished; the most

commonly used mordants are salts of aluminum, iron and tungsten. This substance is the *hemalum* and when applied to a tissue section stains nuclei blue. The histological sample is then counterstained using a solution of eosin (either alcohol or water), an acidic dye which attaches to the cationic (positively charged) components of the tissue staining proteins and cytoplasm in various shades of pink (Figure 1.3). Eosin have different types, but Eosin Y is commonly used [7,8].

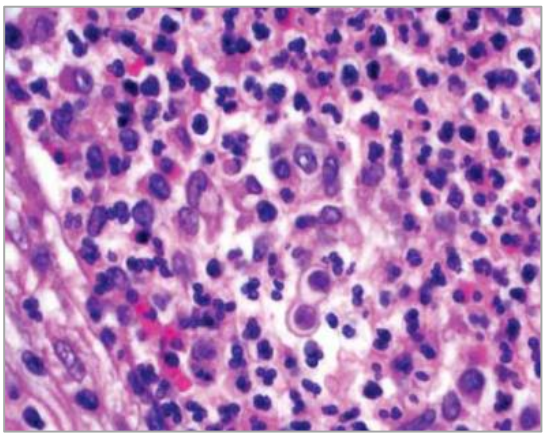


Figure 1.3. H&E-stained section of skin [9].

The two main benefits of H&E staining are the recognizability of the staining pattern and the persistence of the preparations, while among the shortcomings there are the inability to distinguish many cytoplasmic structures and the ambiguity surrounding the nuclear stain's mechanism. For example, with this stain is not possible to detect mitochondria, the Golgi apparatus, fine collagen fibers and the majority of neuronal and glial cytoplasmic elements in the nervous tissue [9].

1.2.2 Special stains and immunohistochemical stains

Special stains, so called because they are not routine, are used to identify and demonstrate particular structures and tissues which are not visualized by H&E stains. The areas of application are research and diagnostic.
Commonly used special stains are the trichrome stains and the periodic acid-Schiff (PAS) stain, but there is a variety of other special stains commonly used for specific purposes.

Trichrome stains

A group of commonly used special stains is the one of the trichrome stain. These stains use a mixture of three colors to differentiate nuclei, muscle fibers and collagen. The trichrome stain's

main function is to show collagen and muscle in healthy tissue or to distinguish collagen and muscle in tumors. It can also be used in the presence of some renal disease or liver cirrhosis as it allows to detect fibrosis, or to differentiate tumors from fibroblast and muscle cells [9].

Widely utilized techniques are [9–12]:

- Masson's trichrome stain: it sequentially uses three different dyes that are iron hematoxylin to stain the nuclei black, Biebrich scarlet to stain cytoplasm red, and aniline light green or aniline blue to stain collagen green or blue depending on the type of aniline used. In order to balance the solution used for staining, phosphomolybdic or phosphotungstic acid is added to the anionic dyes (Figure 1.4 – a).

  Masson's trichrome allows to identify and distinguish between different components of soft tissue and connective tissue, but also to highlight sclerotic lesions, scarring, perivascular fibrosis or collagen's subtler features in the dermis.
- Mallory's trichrome stain: it consists of a combination of aniline blue, orange G and acid fuchsin which stain fibrils of collagen in blue, erythrocytes in orange, and chromatin, nuclei, basophilic cytoplasm and muscle cell cytoplasm in red color (Figure 1.4 – b). This method is especially suitable for studying connective tissue.

  A variation of this method is the Heidenhain's Azan trichrome stain in which acid fuchsin is replaced by azocarmine G.
- Gomori's trichrome stain: this is a one-step technique which combine all dyes and reagents in a single solution. In a phosphotungstic acid solution with glacial acetic acid added, a plasma dye (2R chromotropic) and connective tissue fiber dye (fast green FCF, light green, or aniline blue) are combined. Muscles and cytoplasm exhibit red staining due to phosphotungstic acid. Collagen specifically absorbs tungstate ions, and connective tissue fiber staining is then bound to this complex, staining collagen in green or blue, depending on the counterstain used. Nuclei are stained in black (Figure 1.4 – c).

  Gomori's trichrome stain is the most used for differentiating histological alterations that take place in neuromuscular disorders.
- Masson-Goldner trichrome stain: the basic composition is Weigert's hematoxylin for nuclei, a mixture of acid dyes (Fuschine Biebrich acid scarlet) for cytoplasm and brilliant green for collagen. As a result nuclei are stained in black, muscle fibers, keratin and cytoplasm are stained in bright red, collagen and mucus are stained in green and erythrocytes are stained in red-orange (Figure 1.4 – d).

Masson-Goldner trichrome stain is used for the visualization of muscles, collagen fibers, connective tissues, gametes, nuclei, neurofibrils, neuroglia, intracellular keratin fibers, as well as the negative visualization of the Golgi apparatus. Is additionally employed to visualize the increase in collagen accumulation associated with functional tissues taken for scar tissue (diagnosis of liver sclerosis).

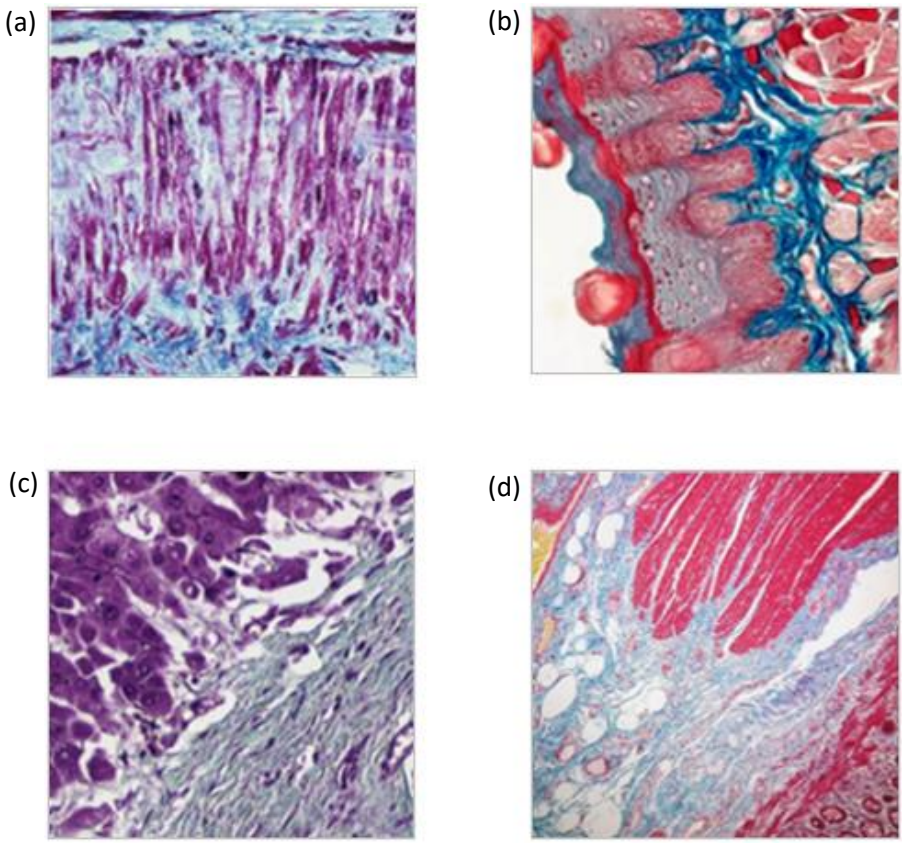


Figure 1.4. Example of tissues stained with (a) Masson's trichrome stain, (b) Mallory's trichrome stain, (c) Gomori's trichrome stain, (d) Masson-Goldner trichrome stain [9,36,37].

Periodic acid–Schiff (PAS) stain

Periodic Acid-Schiff (PAS) is a special stain frequently employed in the dermatopathology laboratory. The process of staining consists of oxidation of sugars with periodic acid, which exposes aldehyde groups that react with the chromophores in Schiff's reagent, created by combining basic fuschin with sulfuric acid, producing a vivid pink color. This stain allows to identify glycogen, neutral mucosubstances and basement membranes which are stained in bright pink, while nuclei are stained in blue (Figure 1.5).

PAS stain allows for the highlighting of hepatocytes since a lot of glycogen is contained in liver. Here, it can be used to detect abnormal form of alpha-1 antitrypsin globules which accumulates in people with genetic mutations. Additionally, it emphasizes storage cells in Niemann-Pick and Gaucher's diseases. At last, the basement membranes surrounding bile ducts, which may be damaged or thickened in biliary disorders, can be shown using PAS stain [9].

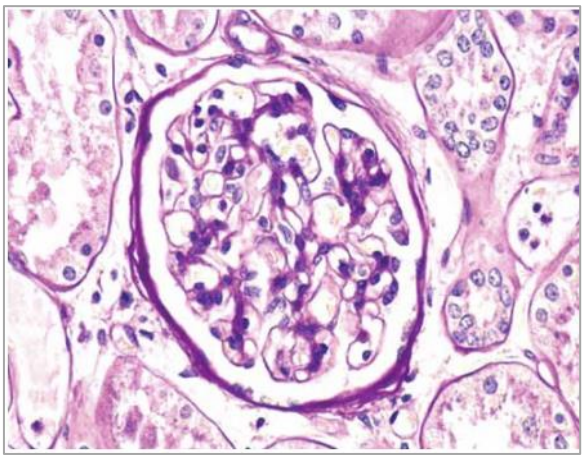


Figure 1.5. PAS-stained section of kidney [9].

Other special stains

There is a variety of other special stains commonly used for specific purposes. Among these, Giemsa stain is used to stain blood smears and bone marrow specimens in order to differentiate between the various hematopoietic elements, to stain a variety of microorganisms including bacteria and several protozoans, to visualize *H. pylori* and for chromosome staining; Wright stain is widely used for performing differential white blood cell counts and evaluate the morphology of blood cells; Gram stain allows to stain bacteria in the bacillary and coccal forms and allows their classification by separating them into Gram-positive and Gram-negative, depending on whether or not they absorb the Gram's stain; myelin stains are used to assess pathological changes white matter, as if present they appear as pale regions in the stained tissue and this could represent foci of demyelination or bundles of degenerating axons that comprise predominantly intact but non-functional axons; silver stains use silver to selectively modify a target's appearance in histological sections, as done in Bielschowsky silver stain that allows to demonstrate neurofibrils and the plaques and tangles that occur in presence of Alzheimer disease, in reticuline stains used to highlight liver architecture, or in the Verhoeff - van Gieson stain used for demonstration and evaluation of the quantity and quality of tissue elastic fibers;

mucicarmine and Alcian Blue stains are specific for mucins; Congo Red stain in a variety of pathologic situations allows the detection of an abnormal protein product that may be present, the amyloid; the Papanicolaou stain allows the detection of cervical cancer in the Pap test [9], [13–15].

Immunohistochemical stains

Immunohistochemistry (IHC) is used in histology to detect the presence of specific protein marker that can help with accurate tumor classification and diagnosis. It also allows pathologists to identify specific proteins that can help predict the behavior of a disease or the response to drugs such chemotherapy. IHC has evolved to complement the H&E and Special Stain techniques which are non-specific.

The *direct method* is a one-step staining method which involves a labeled primary antibody reacting directly with the antigen in tissue sections. Although this method just needs one antibody and is therefore quick and easy, the sensitivity is reduced because there isn't much signal amplification.

Antigens are proteins that are within or on the surface of a cell and are diagnostically useful since their presence or absence may help to fully classify a particular tumor. There are two main types of antibody: polyclonal, which bind to multiple epitopes or the target antigen, and monoclonal, which have an affinity to only one epitope and tend to produce more specific but less intense staining.

In the *indirect method* an unlabeled primary antibody binds to the target antigen in the tissue and a secondary antibody that has been labeled reacts with the primary. Since a single primary antibody can be attached by multiple secondary antibodies to increase the staining intensity, this technique is more sensitive than the direct one.

The detection system builds on enzymes that are joined to a secondary antibody. When several enzymes are linked to the antibody they cause more intense staining since the chromogen can attach a greater number of molecules. Using a microscope, it’s possible to see the chromogen as a colored insoluble precipitate (Figure 1.6). The two chromogens generally employed are DAB, which is brown, and AP, which is red. The first one allows for permanent and intense stains, but in skin sections it can be confused with melanin pigments which have the same color, so here AP is typically used. At the end of the process is generally used a counter-stain to create contrast with the chromogen, for example hematoxylin which produces a blue background [18,19].

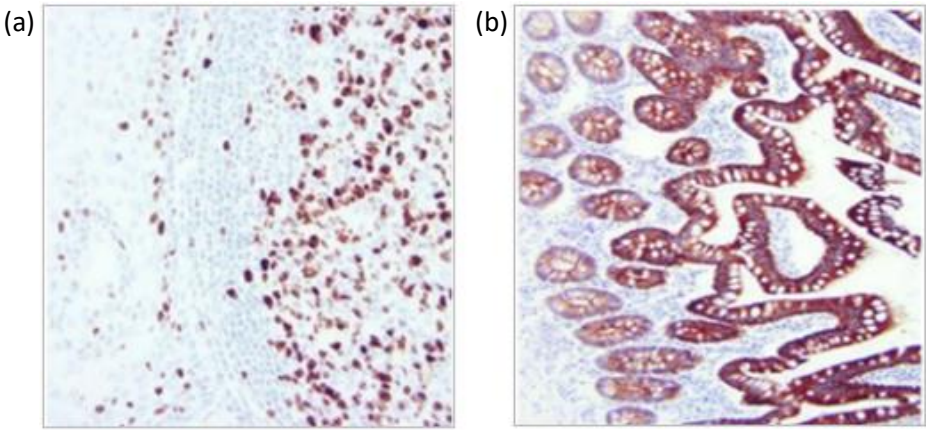


Figure 1.6. IHC-stained sections of (a) tonsil and (b) intestine [18].

On the bases of the different anatomical regions analyzed, different types of antibodies may be used, specific for different receptors [18].

1.3 Artifacts in histology

In cytology and histology an artifact is an artificial structure or tissue alteration visible on a microscopic slide that does not occur on living tissue. Artifacts may alter normal morphological and cytological features or may even lead to complete uselessness of the tissue, thus compromising an accurate diagnosis. It is fundamental to recognize the presence of artifacts and to distinguish them from normal tissue components or pathological changes.
Following the steps normally required to produce digitalized histological images, artifacts may be classified into some main categories [21–23]:

- prefixation artifacts;
- fixation artifacts;
- tissue-processing artifacts;
- artifacts during embedding;
- artifacts related to microtomy;
- staining artifacts;
- mounting artifacts;
- digitalization artifacts.

Prefixation artifacts

Prefixation and fixation are the first steps in the generation of digital slides, whose goal is to preserve the tissue as faithfully as possible compared to the living state.

Prefixation artifacts can be deposits like tattoo pigment, can be generated by surgical procedures like for the laser knife damage and the crush artifact, or may be contaminants introduced into the tissue for example when handling it. Fall within these cases cellulose contamination and specimen marking dyes.

Cellulose contamination can be recognized by the characteristic appearance of plant cells, which are squared and have intensely stained walls. This artifact can result through the preparation of the specimen with cotton gauze, paper, or a cork board, or it can arise when the tissues of the gastrointestinal tract are not cleaned properly before prefixation (Figure 1.7 – a).

Specimen marking dyes are normally employed to mark the marginal parts of the tissue to be analyzed in order to be able to observe them under the microscope and to ensure a correct orientation of the sample. Artifact can be generated when these dyes penetrate deeper levels of the tissue (Figure 1.7 – b).

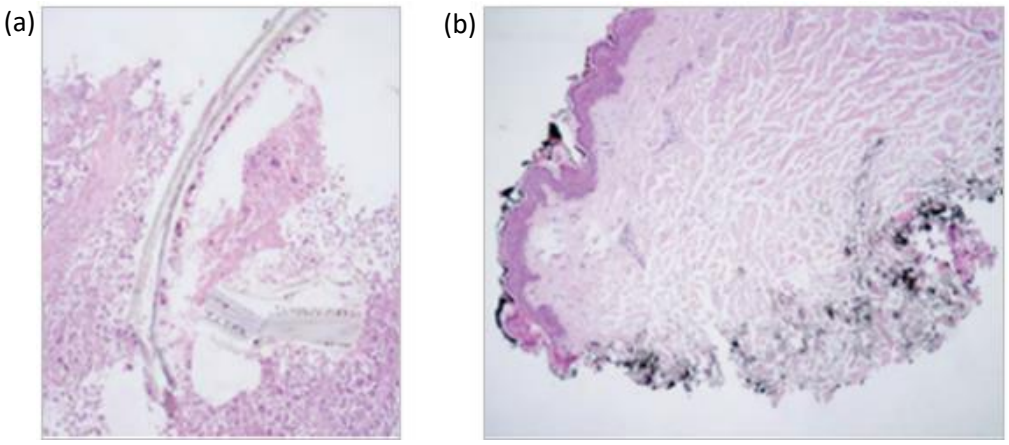


Figure 1.7. Examples of prefixation artifacts. The figure shows (a) cellulose contamination, (b) artifact produced by a specimen marking dye [19].

Fixation artifacts

Fixation artifacts are produced if the tissues are not adequately penetrated by the fixative, when fixation is performed in suboptimal conditions, or because an inappropriate reagent is used. Examples of fixation artifacts are formalin and mercury pigments.

Formalin pigments are typically found near red blood cells, which is why they are typically found in tissues that are rich in blood, or in tissues that have been fixed for a long time. They have the appearance of birefringent and finely granular deposits which are brown to black in color (Figure 1.8 - a). These pigments are generated when hemoglobin and acid formalin react with each other to form acid formaldehyde; in particular, formic acid is produced as a result of the natural breakdown of formalin. By immersing the sample in a saturated alcoholic solution of picric acid prior to the staining process, formalin pigments can be removed. Alternatively, it can be prevented by using shorter times of fixation or buffered solutions.

Mercury pigments appears as brown to black granular deposit distributed randomly throughout the tissues and may be produced during fixation in presence of fixatives which contain mercuric chloride (Figure 1.8 – b). The artifact is extracellular and can be removed by treating the section with alcoholic iodine solution and then bleaching it with sodium thiosulphate.

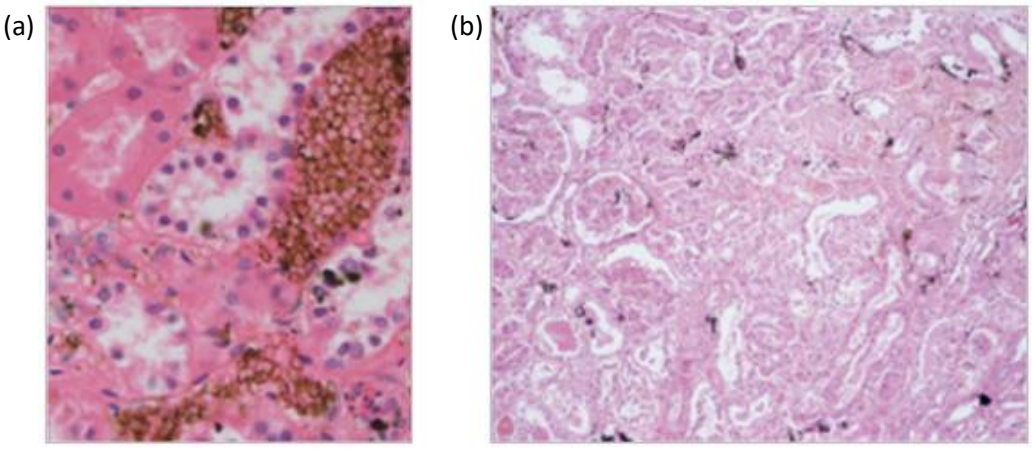


Figure 1.8. Examples of fixation artifacts. The figure shows (a) formalin pigment, and (b) mercury pigment [19].

Tissue-processing artifacts

Tissue-processing artifacts can arise as a result of dehydration and clearing procedures improperly performed. In particular, in the first step substances such as alcohol are used to remove the aqueous fixative fluids from the sample, while subsequently the substance used in the first step is replaced with a miscible one both with the embedding medium and with the dehydrating fluid in order to clean the sample.

An artifact that can be generated in this phase is the presence of water in the sample following improper dehydration. In this case the sample is opaque because the water is not miscible with the clearing agents, and furthermore some microscopic details can be hidden due to the presence of small drops of water visible under the microscope (Figure 1.9 - a).

Another artifact that may occur during tissue processing is the contamination of the slide with squamous epithelia. They lie on the tissue's surface, are to large extent transparent, and usually do not impair focus quality during scanning (Figure 1.9 – b).

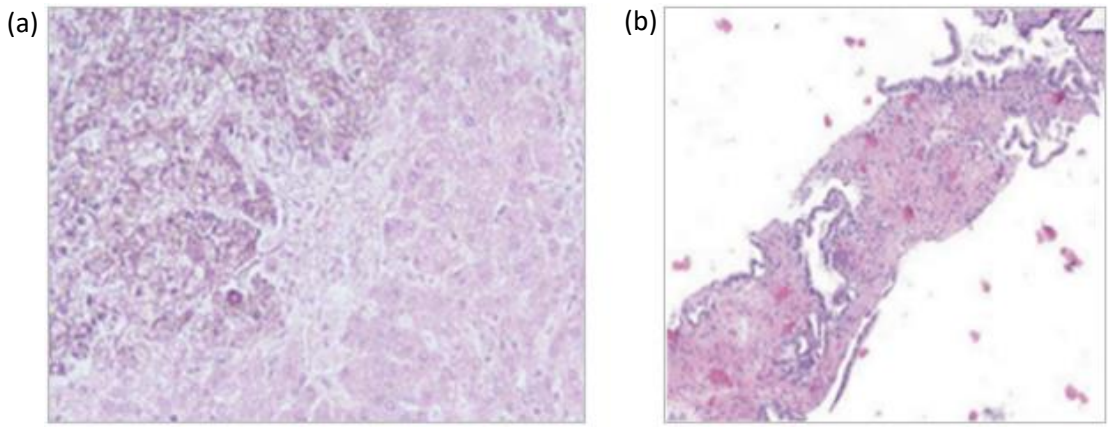


Figure 1.9. Examples of tissue-processing artifacts. The figure shows (a) small water droplets on the tissue caused by inadequate dehydration, and (b) contamination with squamous epithelia [21,23].

Artifacts during embedding

Even the paraffin embedding step can give rise to some artifacts. For example, cracks can occur in tissues during sectioning if the specimen becomes too hard due to excessive exposure during embedding. Additionally, diagnostically relevant tissue parts may be missing or damaged following microtomy if the tissue is not oriented correctly.

These artifacts can be prevented by adequately exposing the specimen to the embedding medium and taking care to orient the specimen correctly in the mold.

Artifacts related to microtomy

Microtomy, the means by which tissues are sectioned, involves some artefacts that can get incorporated if proper technique is not followed. Among the artifacts that can occur in this step there are wrinkling, curling, nicks in tissue, knife lines, alternate thick and thin sections.

The knife lines in the sections can cover the entire area of the tissue or can appear as single lines, and typically manifest when the edge of the knife is damaged. If the artifact is severe it can also be seen at a macroscopic level (Figure 1.10 - a).

Folding and wrinkles of sections of tissue (Figure 1.10 - b) are artifacts that occur when there are structures in the sample with different consistencies around which very thin paraffin

sections must stretch irregularly. Folds can be removed in two ways: either by moving the sections to another high-temperature water bath, or by the careful use of forceps.

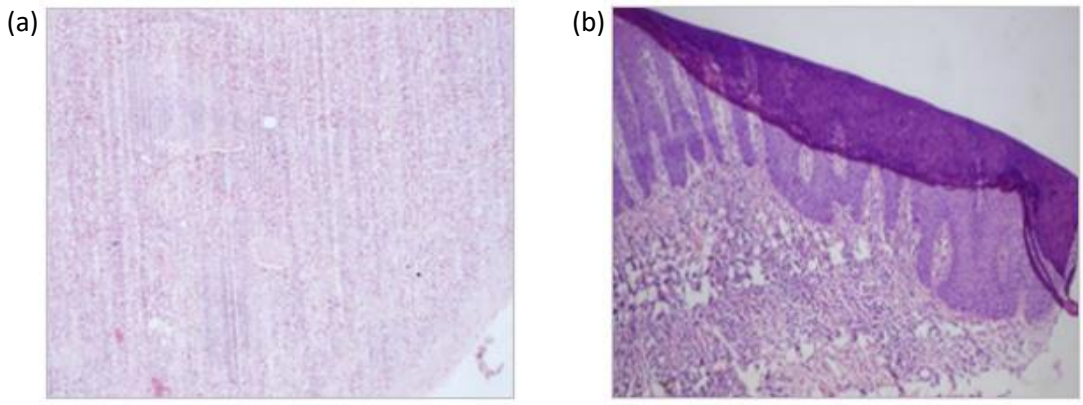


Figure 1.10. Sections of tissue showing (a) knife lines and (b) curling artifact [21,22].

Staining artifacts

In the staining phase two main types of artifacts can be generated: precipitates or contaminants which derives from the staining solution(s) and partial or patchy staining in otherwise successfully stained preparations. These can be prevented by using temperature and times suitable for the stain used and filtrating the staining solution to remove the impurities.

A staining artifact is the one caused by residual wax. If the wax is not removed from some parts of a section before staining, they will result colorless as the staining solutions cannot penetrate within that region (Figure 1.11 - a). The problem is that due to the final clearing of the sample before the next step, the wax residues are removed and at that point it becomes impossible to trace the cause of the artifact. Treating the sample with xylene for a longer time or repeating the staining can solve the problem.

Another artifact that falls into this category is due to incomplete staining, which can be caused simply by not having enough dye in the staining dish (Figure 1.11 - b). This is more frequent when automatic machines are used to carry out the staining.

A third artifact is that of stain deposits that can occur due to solutions that have not been filtered and may contain precipitates of dye, undissolved dye or in general other solid components (Figure 1.11 - c). In particular, precipitation can be generated as result of long staining times or processes that require heat, or when volatile solvents are employed. The majority of precipitate-related artifacts can be removed by using slides sealed staining jars that vertically hold slides.

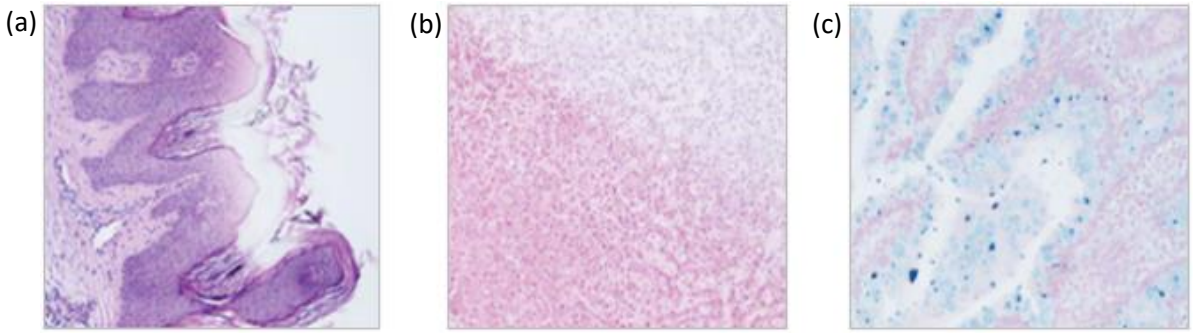


Figure 1.11. Sections of tissue showing artifacts caused by (a) residual wax, (b) incomplete staining, and (c) stain deposit [19].

Mounting artifacts

Mounting artifacts are those that arise when a mounting medium is used to close the section with the coverslip and protect it from possible damage. These artifacts can alter the appearance of the stained tissue. Examples of mounting artifacts are air bubbles and contamination of mounted sections.

The artifact of air bubbles can be generated when the mounting medium used is too thin, because in this case more air can be drawn in at the edges which traps a bubble under the coverslip (Figure 1.12 – a). This artifact can be prevented by using a not too small thickness for the mounting medium. Alternatively, bubbles can be removed from the slide.

Other artifacts can be generated following contamination of mounted sections with materials such as cellulose fibers, microorganisms, airborne fibers or dust (Figure 1.12 – b).

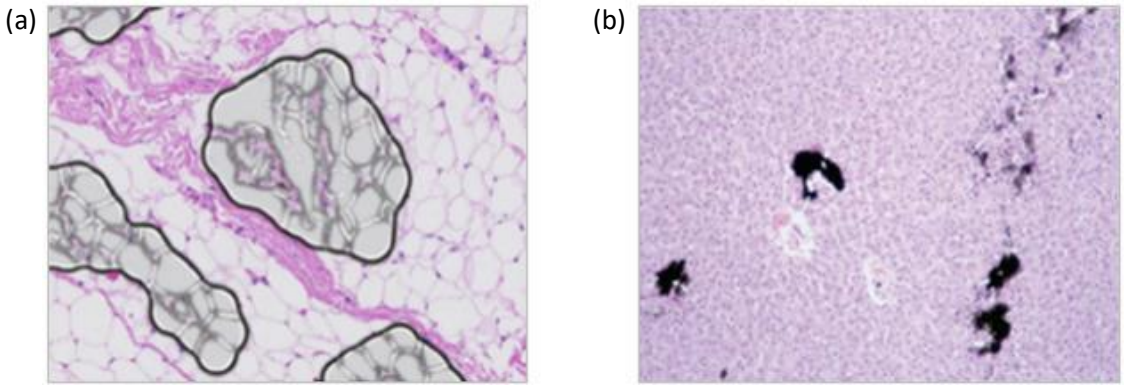


Figure 1.12. Sections of tissue showing (a) air bubbles and (b) dust contamination [21,38].

Digitalization artifacts

Digitalization artifacts arise during the scanning phase. Regarding color scheme, brightness, and contrast, different scanner systems produce images with a highly variable quality. In

addition, artifacts such as image blurring if the histological sample is not aligned with the focal plane of the scanner (Figure 1.13) or striping may occur.

These artifacts, particularly areas out of focus, can hinder the rendering of accurate diagnoses by pathologists, or impact the accuracy of automated image analysis.

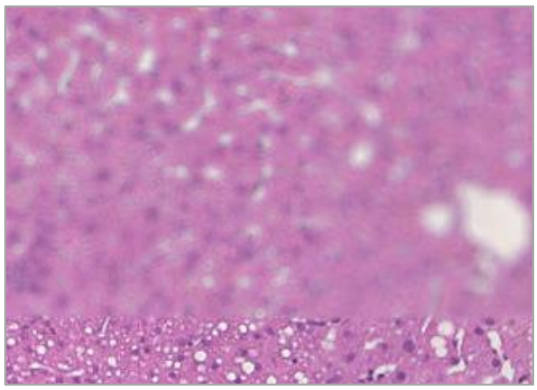


Figure 1.13. Out of focus artifact caused by digitalization.

Other artifacts

Digital WSIs can be affected by other artifacts introduced on the slide after placement of the coverslip. These artifacts include pen marks, fingerprints, or coverslip crack.

Pen marks (Figure 1.14 - a) are often left by pathologists to indicate a pathological feature of interest, like the presence of cancerous cells. Ideally, slides should be scanned before any physical markup, but this may cause prohibitive delays and disruptions to the current pathology workflow and cannot be applied to a large number of historical slides.

Pen marks can also be chemically removed, but doing so runs the risk of damaging the tissue underneath. Instead, a DL approach may be used to digitally remove the on-slide annotations.

Fingerprints (Figure 1.14 - b) are common and appear as small transparent fat drops in digitized histological slides [23,24].

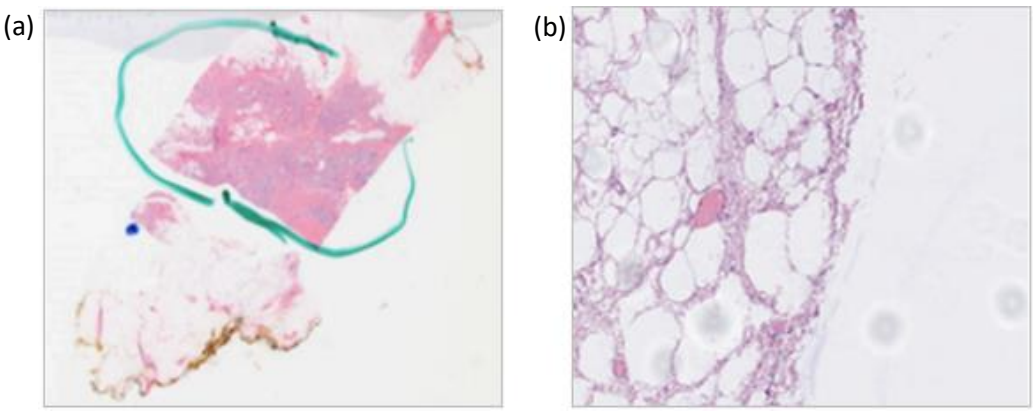


Figure 1.14. Digitized histological slides showing (a) pen mark and (b) fingerprints [23,25].

1.4 Quality controls in digital pathology

During the process carried out to obtain digitized histological images, various artifacts can be generated, due to which tissue regions that are important for diagnosis could be unusable, unclear, or completely missing. Poor-quality tissue presentation may results in delays in pathology reporting, usually caused by poor-quality glass slides needing to be reproduced or WSIs rescanned. Therefore, a quality control mechanism is needed to ensure that whole slide images are of good enough quality to be further analyzed by pathologists and to identify slides that need to be reproduced or regions that should be avoided during computational analysis, since an artificial intelligence system can fail to make a correct diagnosis in those regions.
Nowadays, quality control processes in clinical and research settings are mostly carried out by manual operators, but they are demanding processes, prone to human error and which exhibit some inter-operator variability. Furthermore, there may be artifacts that are not an obstacle for the pathologist to make a diagnosis, but which can negatively affect the analysis performed by automatic algorithms.
For these reasons, automated approaches for quality control of digitalized histological samples are needed which should subsequently be implemented in the daily workflow of the pathologists and technicians. However, quality assessment in relation to histopathology images is complex and the availability of quality assessment tools developed specifically for histopathology slides is currently limited.

1.4.1 State of the art

Janowczyk et al. [23] introduced HistoQC, an open source tool through which it is possible to perform automatic quality controls of histological slides. HistoQC allows to identify artifacts that may be hindering for clinical analysis or computation analysis, and it does so by calculating specific metrics on the image such as contrast, brightness and color histograms, through the use of specific functions as the edge detector, and through the use of supervised classifiers that allow for example to recognize pen markers.
In HistoQC a pipeline of modules is sequentially applied to a WSI. Each module acts on the image to either quantify visual characteristics associated with a digital pathology image, allowing for identification of heterogeneity within a population of images, or detect different artifacts that could be on a WSI. The user supplies a configuration file that defines the parameters of the quality control pipeline, such as which modules to execute and in what order, and relevant output images are created, with metadata and metrics being saved in a file.

Furthermore, it is possible to identify slides that have outliers by sorting columns, it is possible to represent the graphs of the metrics and it is possible to make qualitative comparisons between the generated masks and the original images.
To evaluate the ability of HistoQC to identify regions of artifact-free tissue, 450 randomly selected slides form the TCGA breast cancer cohort, stained with H&E, at magnification of 40x, were used.
For the validation phase, a comparison was made with the evaluations of the pathologists. In particular, an overlap between the evaluation of HistoQC and the evaluation of the pathologist on good quality tissue of at least 85% of the evaluated area was established as an acceptability threshold, and two pathologists with experience were asked to evaluate in independently 250 masks and determine if they were acceptable or not. In addition, both pathologists were asked to evaluate 50 images of the TGCA to establish agreement between operators on the masks produced by HistoQC. The results establish an agreement between HistoQC and expert 1 of 94%, between HistoQC and expert 2 of 97%, and an interobserver agreement of 96%.

Chen et al. [24] begin evaluation of the effects of quantitative quality controls via the integration of HistoQC into routine processes. A dataset consisting of 1814 WSIs from 512 renal biopsies was used. The slides were stained with H&E, PAS, silver and trichrome.
To quantify improvements in curation reproducibility, a comparison in inter-reader concordance between HistoQC aided and unaided curation was made. The concordance in the identification of these WSIs among computational pathologists rose from moderate to excellent agreement when aided by HistoQC.

Shakhawat et al. [25] proposed a method to asses WSI quality by distinguishing the origins of quality degradation that are the focus-error or noise caused by the scanner and the artifact occurred in the slide preparation.
While in a previous work by Hashimoto et al. [26] the quality of the scanned image was evaluated only on the base of sharpness and noise measurement, which are the main factors for the quality failure in the WSI scanner, Shakhawat et al. introduced also the artifact detection.
In this method, the machine learning technique is used to detect artifacts first, and then quality is evaluated on focus error and noise with excluding artifact regions. They used low magnification images (i.e., 1x or 2x) for tissue artifact detection, and higher magnification images (i.e., 20x or 40x) for quality estimation because the focusing error cannot be detected from the low-magnification image. The high-resolution WSI was divided into fixed-size nonoverlapped image blocks and image blocks containing more than 75% white pixels were

detected as glass area and eliminated for quality estimation. The algorithm of this method is illustrated in Figure 1.15.

The dataset used consisted of 52 slides from major tissue organs produced from human and rat and stained with H&E, IHC and PAS.

For the artifact detection two separate support vector machine (SVM) binary classifier were used, one for air bubble detection and the other for tissue fold, using the texture information and other physical properties. Then, the quality of the WSI was estimated from the high-resolution image based on the sharpness and noise measurements using referenceless quality evaluation method (RQM).

This method was compared with RQM method proposed by Hashimoto et al. and in the presence of artifacts showed better results in terms of the number of detected poor blocks, WSI quality and evaluation time. In the absence of artifacts, evaluation time was higher but WSI quality was same.

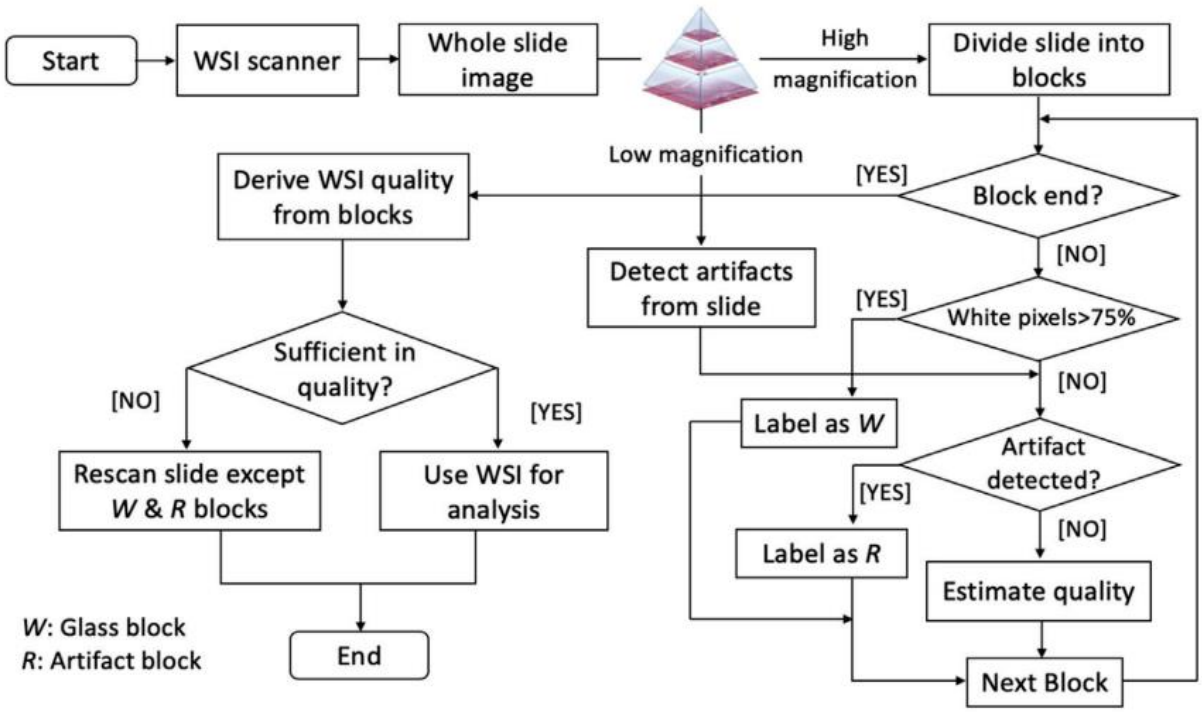


Figure 1.15. Algorithm of the quality evaluation method proposed by Shakhawat et al. [25].

Smit et al. [27] proposed an approach based on a multi-class deep learning model trained on whole-slide images covering multiple tissue and stain types for semantic segmentation of the artifacts caused by out of focus, tissue folds, ink, dust, marker, and air bubbles. The dataset used included 142 whole-slide images, consisting of 9 tissue and 8 stain types, digitized using 7 scanners. Overview of this method is shown in Figure 1.16.

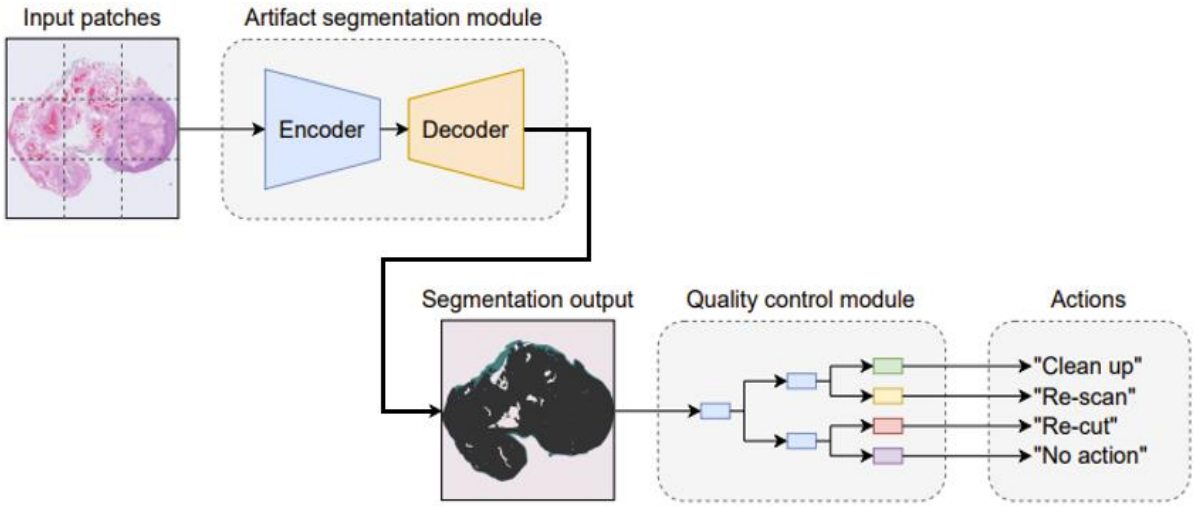


Figure 1.16. Overview of the quality control method proposed by Smit et al. [27].

The framework proposed by Smit et al. contained an artifact segmentation module in which a tissue segmentation network was first used to pre-process an input whole-slide image to eliminate the white spaces around tissue block, and a trained artifact segmentation network was then applied on the pre-processed image. It also contained a quality control module in the form of a decision tree to transform the artifact segmentation output into one of the four actions: clean up, re-scan, re-cut, and do nothing.

This approach reaches a Dice score of 0.91, on average, across the six artifact types. The slide-level quality performance was compared to the one of HistoQC on an external test set, showing better results.

Haghighat et al. [28] developed PathProfiler, an artificial intelligence tool to automate the quality assessment of WSIs of a retrospective cohort of prostate cases. This tool was designed to indicate simultaneously the 'usability' of an image and the presence of artefacts. This facilitates the identification of the image within the pipeline for consideration of re-scanning or re-staining in order to improve the quality and render it 'usable', and facilitates the user to identify when re-scanning or re-staining would not resolve the quality issue such as in presence of a significant number of intrinsic artefacts, dirt, ink or bubbles in the tissue area.

The quality assessment pipeline of PathProfiler is illustrated in Figure 1.17.

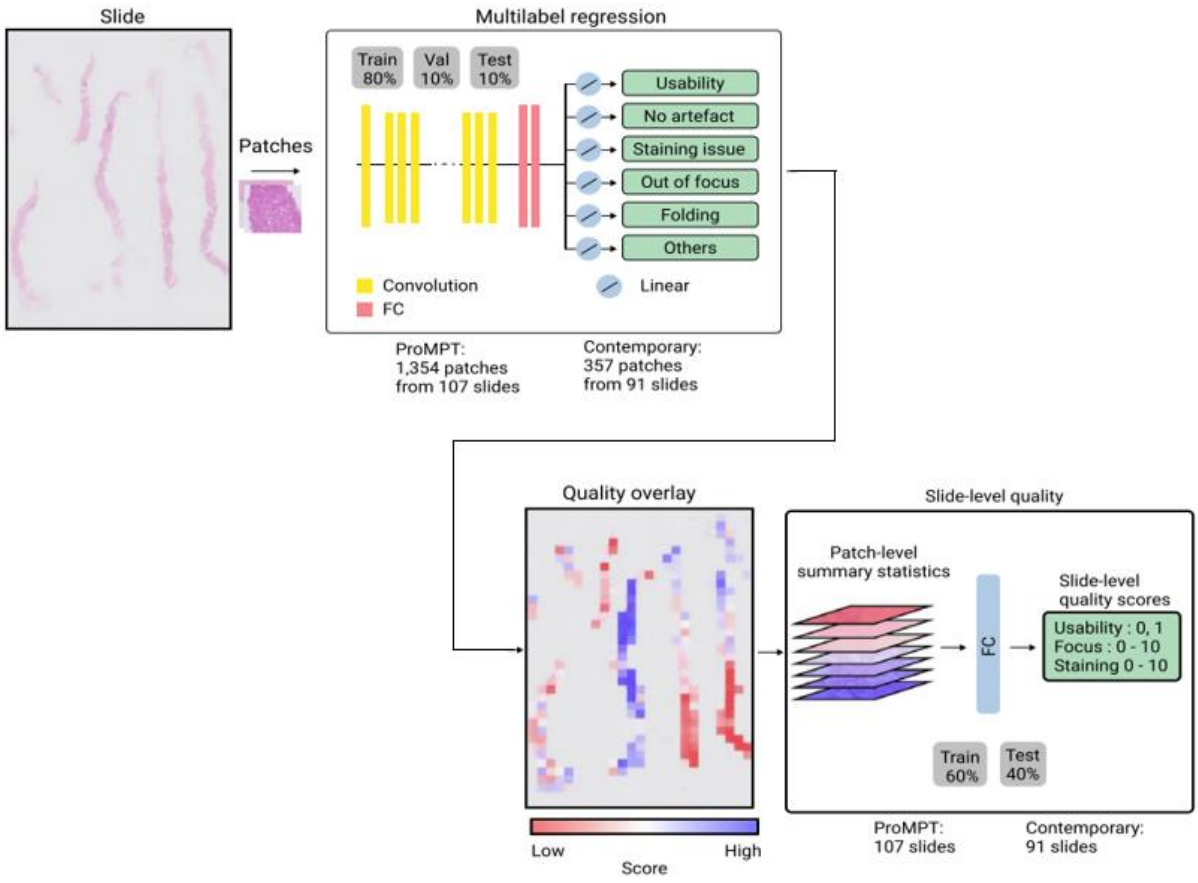


Figure 1.17. PathProfiler quality assessment pipeline [28].

After tissue segmentation, patches of 256x256 are extracted at 5x magnification and resized to 244x244 to accommodate for the ResNet18 CNN model. For each patch, a multi-label pre-trained model predicts the presence of an artefact, and for focus and H&E staining artefacts it also predicts a quality score for each patch, where 0 stands for no quality issue, 0.5 stands for slight quality issue, and 1 stands for severe quality issue. For each output category, a quality overlay is generated. In the last step, using statistical parameters they map the predicted quality overlays to the slide-level standardized scoring system. These slide-level quality scores include usability of the WSI (binary 0 or 1), and a score 0–10 for quality of focus and H&E staining from the lowest quality to highest quality, where the cut-off score for acceptable quality for diagnostic purposes is 4.

The dataset used for the algorithm development included 198 H&E stained WSIs of prostate. The performance of the proposed multivariate model on the test dataset of image patches were evaluated through the ROC-AUC value. For out of focus and staining artifacts the ROC-AUC were 0.85 and 0.84 respectively, with an higher model accuracy when these artifacts were severe. As regards the slide-level quality assessment, the performance were evaluated through the Pearson correlation coefficient and the ROC-AUC of the predicted usability, focus and

staining scores (versus reference standard), with results of 0.889, 0.869 and 0.824 respectively for Pearson correlation coefficient, and results of 0.987, 0.826 and 0.751 respectively for ROC-AUC. This showed that the quality measures as predicted by PathProfiler closely align with the reference standard. Moreover, the calculated accuracy of binary focus and staining scores was 1, with a cut-off threshold of 5, while the accuracy of the predicted usability score was 0.987, with a cut-off threshold of 0.5.

Chapter 2

Materials and Method

2.1 Dataset

The dataset used for this study is composed of 3400 WSIs from different centers. In particular, some of them were provided by health centers and hospitals, while other were extracted from the open access datasets TCGA (The Cancer Genome Atlas) and TCIA (The Cancer Imaging Archive).

The histological images were captured at 20x or 40x magnification factor. They were collected in different formats: SVS, NDPI, TIF, BIF.

The WSIs used contain a representative selection of different organs and anatomical structures and different histological stains, so as to allow the algorithm to better generalize on new data. In particular, the organs and anatomical structures of origin are adrenal gland, bone, breast, colon, liver, lung, prostate, kidney, and myocardium. The stains included are H&E, PAS, IHC and trichrome (TRIC). Figure 2.1 shows the subdivision by stain of the 3400 WSIs considered.

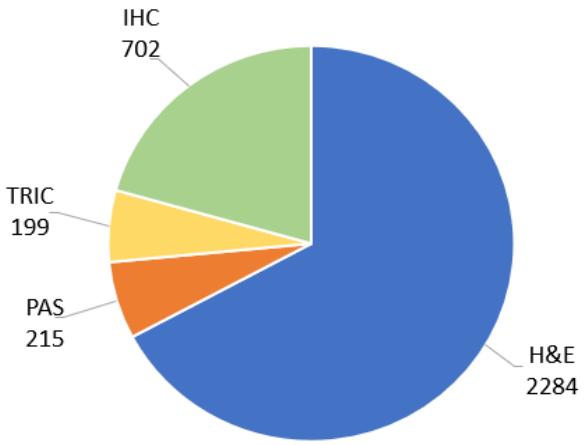


Figure 2.1. Subdivision by stain of the 3400 WSIs included in the dataset.

The dataset includes WSIs of different quality, ranging from excellent quality images without artifacts to low quality images with various type of artifacts. The artifacts considered for the purposes of this study were divided into three categories: tissue folds, out of focus regions,

other (i.e., bubble, pen mark, edge of the coverslip on the tissue). The whole dataset was manually annotated distinguishing between tissue and artifacts present.

The 3400 WSIs have been divided into four groups for the development of the algorithm: training set, validation set, test set 1 and test set 2. For the creation of the validation set, 10% of the total images available for training + validation were used, taking care to create a validation set that was proportional to the training set in terms of distribution of artifacts and stains. In particular, the dataset was split as follows:

- Training set: 2700 WSIs;
- Validation set: 300 WSIs;
- Test set 1: 225 WSIs;
- Test set 2: 175 WSIs.

The composition of each subset in terms of number of WSI relative to a certain tissue or stain is shown in Figure 2.2.

TRAINING SET

	H&E	IHC	PAS	TRIC
ADRENAL	8	-	-	-
BONE	5	7	-	-
BREAST	403	485	-	-
COLON	272	-	-	-
KIDNEY	11	-	102	91
LIVER	248	-	-	-
LUNG	136	-	84	1
PROSTATE	528	132	-	-
MYOCARDIAL	-	-	-	87

VALIDATION SET

	H&E	IHC	PAS	TRIC
ADRENAL	2	-	-	-
BONE	1	2	-	-
BREAST	50	43	-	-
COLON	30	-	-	-
KIDNEY	1	-	10	8
LIVER	29	-	-	-
LUNG	16	-	13	-
PROSTATE	61	23	-	-
MYOCARDIAL	-	-	-	11

TEST SET 1

	H&E	IHC	PAS	TRIC
ADRENAL	19	-	-	-
BONE	33	-	-	-
BREAST	56	10	-	-
COLON	38	-	-	-
KIDNEY	-	-	-	-
LIVER	30	-	-	-
LUNG	32	-	6	1
PROSTATE	-	-	-	-
MYOCARDIAL	-	-	-	-

TEST SET 2

	H&E	IHC	PAS	TRIC
ADRENAL	-	-	-	-
BONE	-	-	-	-
BREAST	-	-	-	-
COLON	-	-	-	-
KIDNEY	-	-	-	-
LIVER	-	-	-	-
LUNG	-	-	-	-
PROSTATE	175	-	-	-
MYOCARDIAL	-	-	-	-

Figure 2.2. Composition of each subset.

2.2 Pipeline

The general pipeline of the method is shown in Figure 2.3.

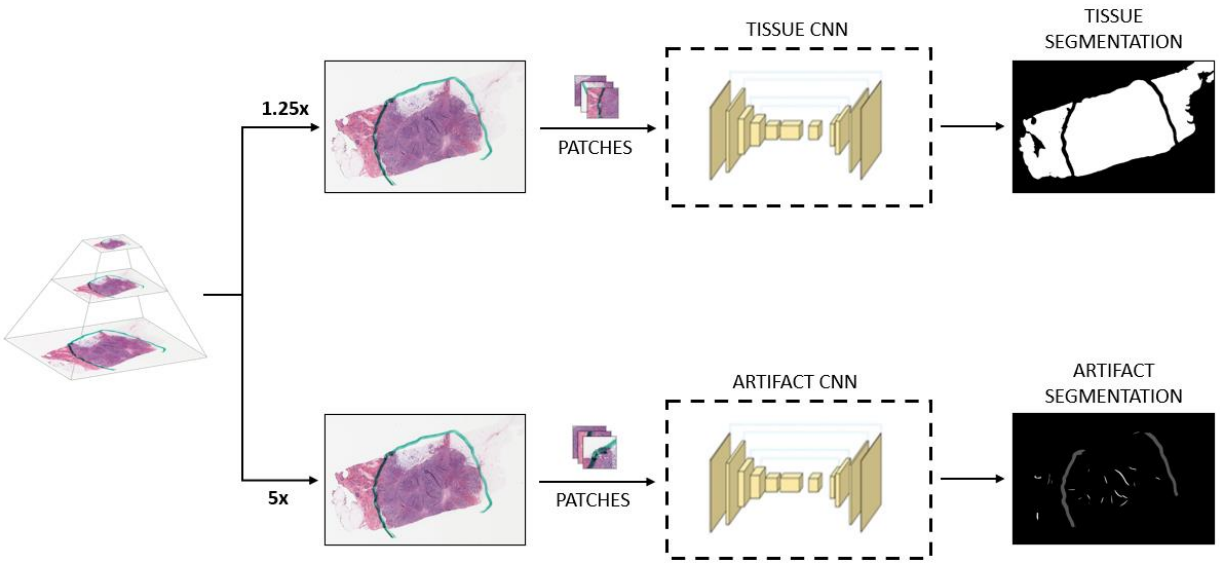


Figure 2.3. Pipeline of the proposed method for tissue and artifact segmentation.

First of all, the images at 1.25x magnification for the tissue network and 5x for the artifact network are extracted from the starting pyramidal WSI. For the creation of network inputs, patches of 768x768 pixels are extracted from these images. These patches are then fed into the trained neural networks, which are DeepLabV3+ models with a ResNeSt as backbone network, in order to generate the automatic predictions relating to tissue segmentation and artifact segmentation. These predictions will subsequently be compared with the manual predictions for the validation of the algorithm.

2.3 Manual mask generation

Figure 2.3 shows the steps followed for the creation of the manual masks.

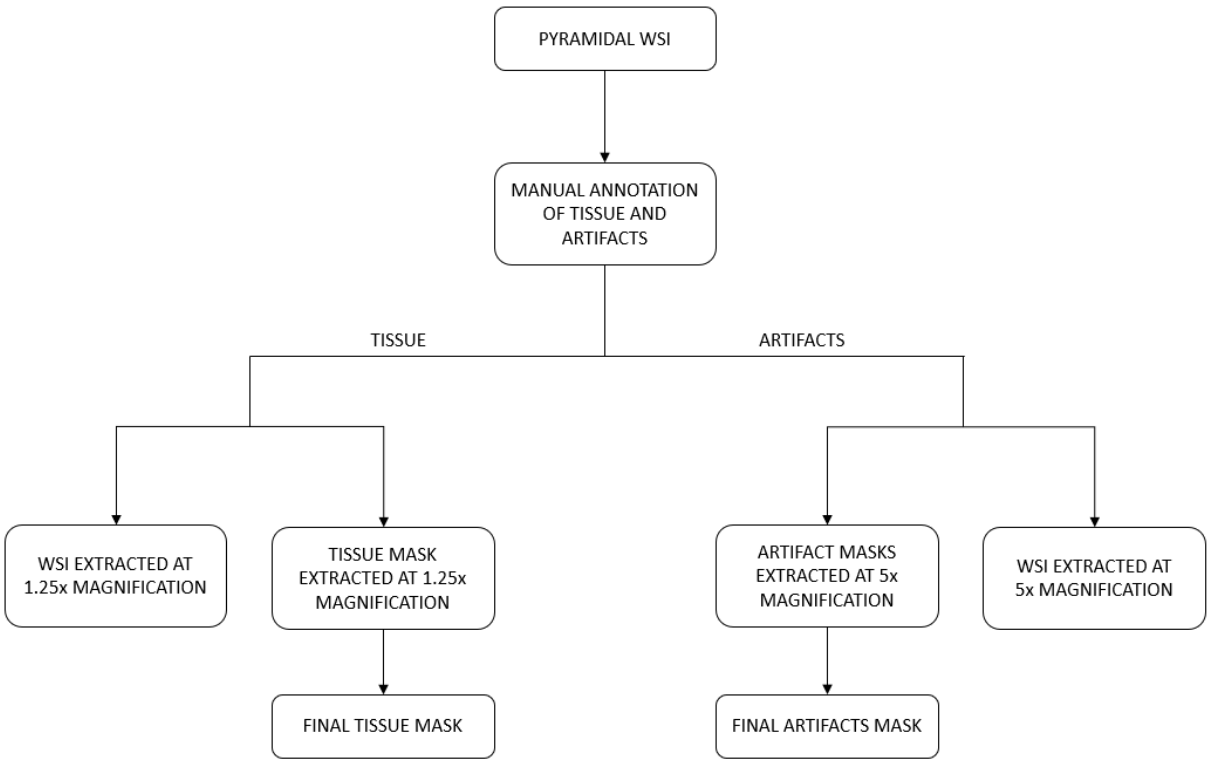


Figure 2.3. Diagram describing the procedure followed for the creation of the tissue mask and the artifacts mask.

In the first step, for each WSI in the dataset, tissue and artifacts were manually annotated on the pyramid image. Three different software were used for the WSIs visualization and for the manual annotation process: ImageScope for the SVS format, NDP.view for the NDPI format, and QuPath for the TIF and BIF formats.

To distinguish the tissue and the different artifacts, four colors were used for the manual annotations (Figure 2.5):

- Yellow for the tissue;
- Blue for the tissue fold artifact;
- Red for the out of focus (OOF) regions of the image;
- Green for other artifacts that include bubbles, pen marks and the presence of the edge of the coverslip on the tissue.

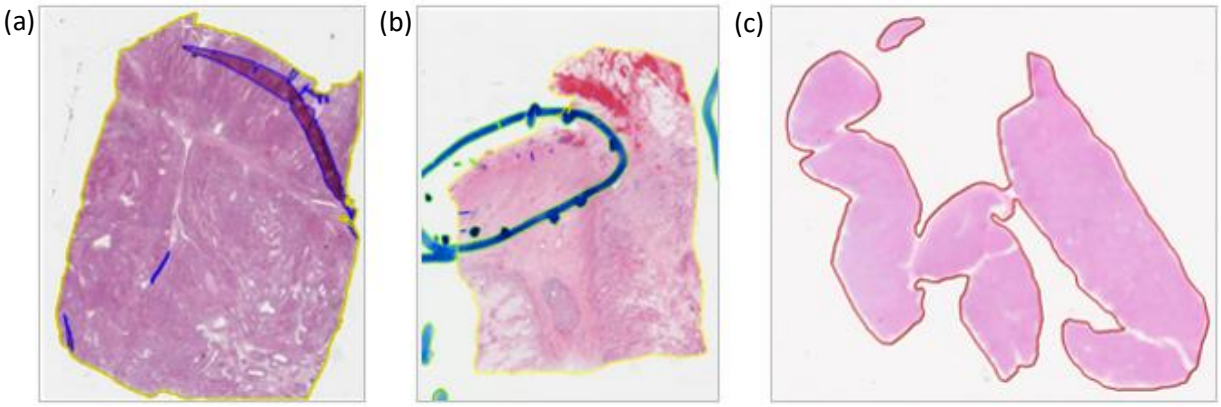


Figure 2.5. Examples of manual annotations on histological slides. The figure shows (a) the annotation of tissue in yellow and tissue fold artifacts in blue, (b) the annotation of tissue in yellow and pen marks in green, and (c) the annotation of out of focus regions in red.

After annotating the entire dataset, the image and four binary masks, one for each color used in the annotation step, were extracted at 1.25x and 5x magnifications for each WSI. In this way, the binary masks of the tissue, the out of focus regions, the tissue folds and the areas with artifacts falling into the "other" category were created. In particular, at this stage the tissue mask is a binary mask in which the background is black and the entire area of the tissue, including the artifacts found on it, is white.

In the last step the labels corresponding to the different annotated elements were created and for each chosen magnification the two final masks were generated:

1. the final tissue mask, which has two classes: tissue, out of focus regions on it and tissue fold artifacts are set to 255 (white), while the white background, artifacts in the “other” category and dark artifacts in the white background are set to 0 (black);
2. the final artifact mask, which is a grayscale image in which each gray level indicates a class, for a total of four classes: tissue fold artifacts are set to 255 (white), out of focus regions are set to 170 (light gray), artifacts in the “other” category are set to 85 (dark gray), and the white background and the tissue are set to 0 (black). The areas of the image that have more overlapping artifacts are also black.

Figure 2.6 shows an example of the process followed to obtain the final masks of tissue and artifacts.

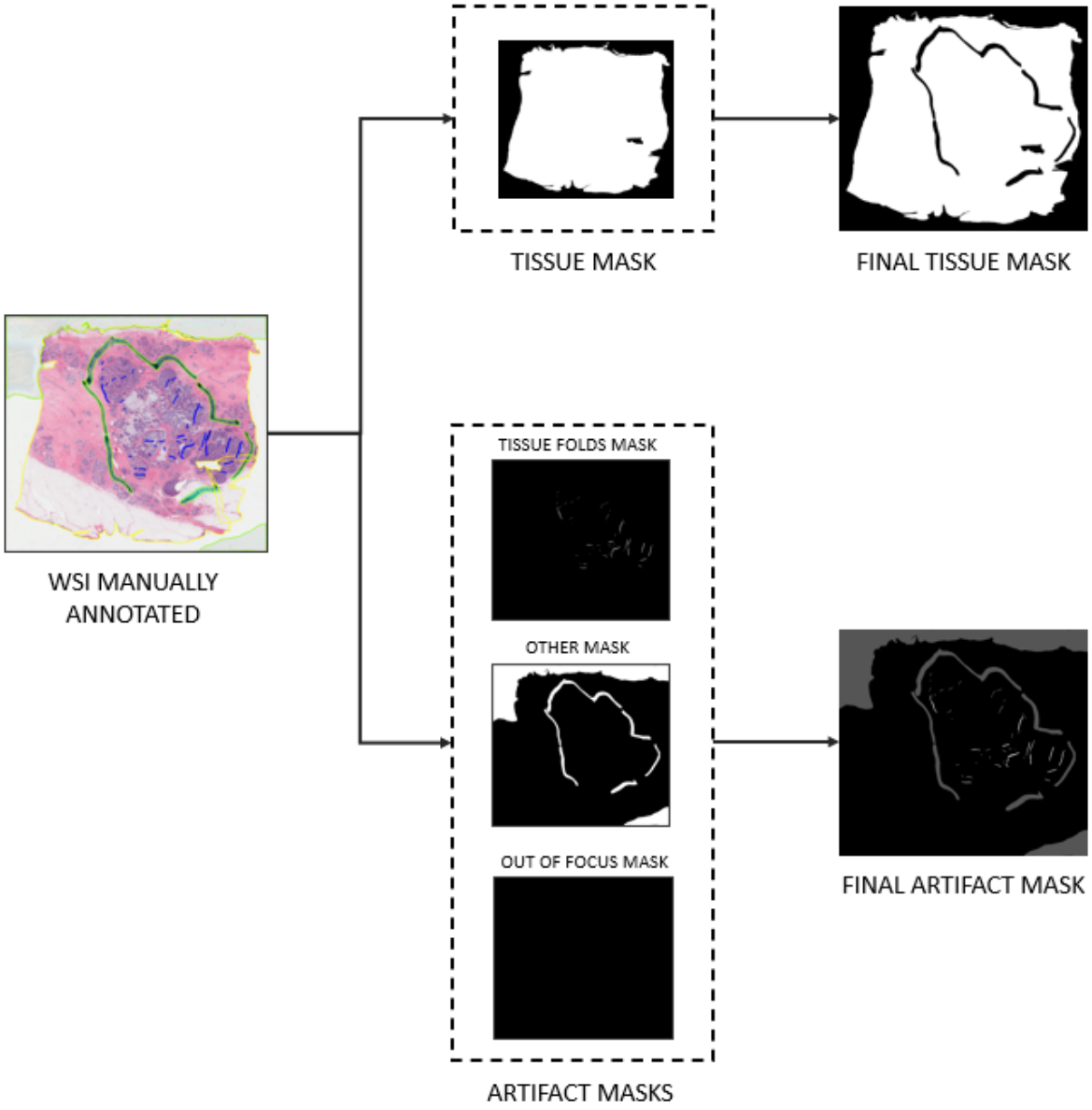


Figure 2.6. Example of the manual masks generation process. Firstly, the pyramidal WSI is annotated using different colors for the different elements to be segmented. After that, four binary masks are generated, one for the tissue and the other for the three different categories of artifact. In this specific example, the mask of the out of focus regions is completely black because there are no regions of this type in the starting image. In the last step, the labels corresponding to the different annotated elements are created and the final masks of tissue and artifacts are generated.

2.4 CNNs for tissue and artifact detection

Once the final automatic masks of tissue and artifacts were produced, they were used, together with the corresponding images, to create the dataset used for the development of the algorithms for tissue and artifact segmentation. Patches were extracted from each manual mask and image, appropriately selected on the basis of specific criteria, in order to obtain two balanced datasets, one to train the tissue network and the other to train the artifact network. The two networks are based on the same architecture, a DeepLabV3+ model that uses a ResNeSt as a backbone network in order to capture cross-channel feature correlations while preserving a multi-path architecture to learn independent features. After training, an overlapping sliding window was used in the inference phase to generate automatic predictions for both tissues and artifacts, keeping only the central part of each prediction.

2.4.1 Data preparation

The first step is the data preparation. For each WSI, 768x768 pixels size patches were extracted from the starting image and from the manual mask at the two different magnifications, 1.25x and 5x, for the tissue and the artifacts. A "smart" approach was used for the extraction of the patches, i.e., specific criteria were followed for the two networks in order to avoid extracting a large number of non-informative patches. The extracted patches were analyzed to evaluate the percentage of area of interest present (i.e., tissue or artifacts, based on the network to be trained), and a patch selection procedure was performed for tissues and artifacts in order to avoid a bias in the network caused by an imbalance in the number of patches related to a single WSI.
For the patch extraction the percentage of background contained in the patches was evaluated on the white mask, which is a mask obtained by identifying what in the starting image is white (i.e., mainly the background) and putting the pixels corresponding to the maximum value of intensity, leaving everything else black.
Figures 2.7 and 2.8 show the flowcharts relating to the extraction of patches for tissues and artifacts respectively, and describe the criteria followed to evaluate whether or not to extract the patch considered.

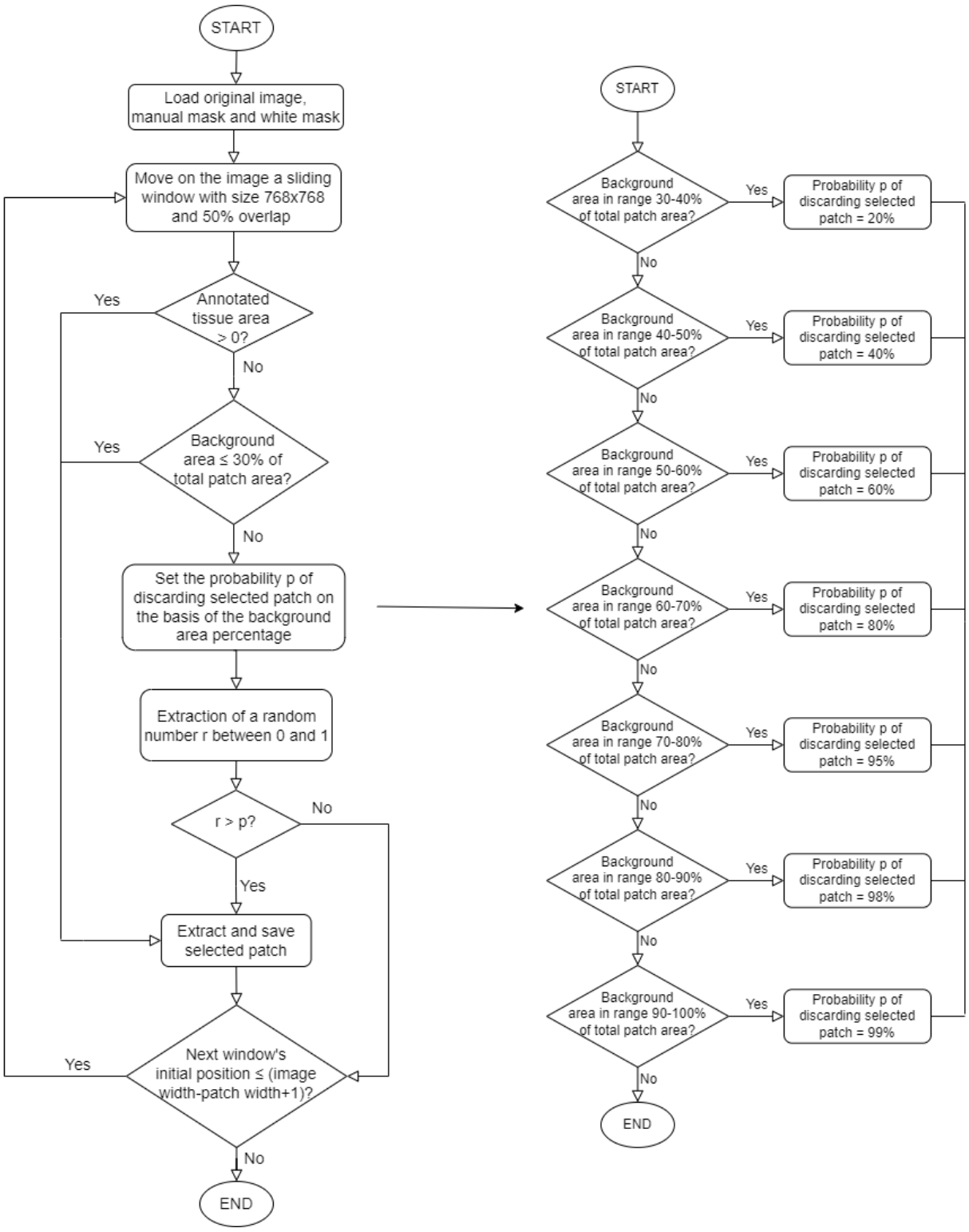


Figure 2.7. Flowchart relating to patch extraction from the images of the tissue dataset with 1.25x magnification.

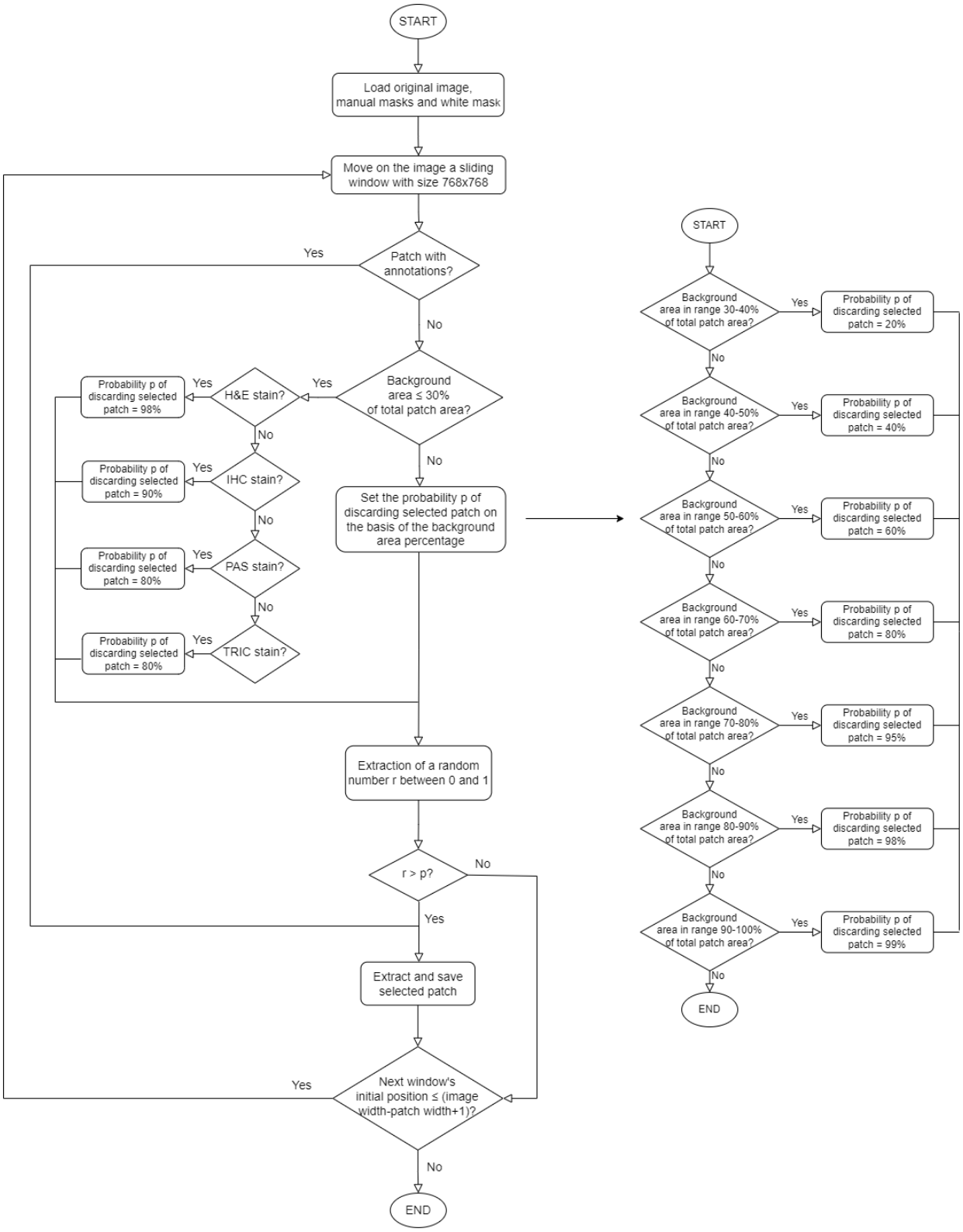


Figure 2.8. Flowchart relating to patch extraction from the images of the artifact dataset with 5x magnification.

For the preparation of the tissue dataset, consisting of images and masks extracted at a magnification of 1.25x, for each image a sliding window of 768x768 pixel dimensions was scrolled on the input, setting a priori an overlap of 50% between a window and the next one. The main criteria followed are two:

1. Presence or not of annotation, i.e. manual segmentation of the tissue, in the selected patch. In the presence of annotation the patch is extracted regardless, in the absence of annotation it is evaluated whether or not to extract the patch on the basis of the percentage of background.
2. Percentage of background in the patch, evaluated on the white mask. If the percentage of the background in the selected patch is less than 30% of the total area of the patch, then it is extracted, otherwise it is attributed a decreasing probability of extraction based on the amount of white present. In particular:

    - Percentage of white between 30% and 40% of the total area of the patch → 20% probability of discarding the patch;
    - Percentage of white between 40% and 50% of the total area of the patch → 40% probability of discarding the patch;
    - Percentage of white between 50% and 60% of the total area of the patch → 60% probability of discarding the patch;
    - Percentage of white between 60% and 70% of the total area of the patch → 80% probability of discarding the patch;
    - Percentage of white between 70% and 80% of the total area of the patch → 95% probability of discarding the patch;
    - Percentage of white between 80% and 90% of the total area of the patch → 98% probability of discarding the patch;
    - Percentage of white between 90% and 100% of the total area of the patch → 99% probability of discarding the patch.

    At this point a random value between 0 and 1 is extracted and it is compared with the probability of discarding the patch: if it is higher, the patch is extracted.

This process is iterated until the sliding window has examined the entire image area.

Also for the preparation of the artifact dataset, consisting of images and masks extracted at a magnification of 5x, for each image a sliding window of dimensions 768x768 pixels was scrolled on the inputs, but the presence or absence of overlap was chosen on the basis of the area occupied by the annotation. In this case, the main criteria followed are three:

1. Presence or not of annotation, i.e. manual segmentation of artifacts, in the selected patch. In the presence of annotation, the patch is extracted regardless, in the absence of annotation it is evaluated whether or not to extract the patch based on the percentage of background present and the stain of the image.
2. Percentage of background in the patch, evaluated on the white mask. If the percentage of background in the selected patch is greater than 30% of the total area of the patch, it is attributed a decreasing probability of extraction based on the amount of white present in the patch in the same way as described for the tissues, otherwise it is attributed a different probability of extraction by evaluating the stain of the image. In particular:

    - H&E stain → 98% probability of discarding the patch;
    - IHC stain → 90% probability of discarding the patch;
    - PAS stain → 80% probability of discarding the patch;
    - TRIC stain → 80% probability of discarding the patch.

    Then, a random value between 0 and 1 is extracted and it is compared with the probability of discarding the patch: if it is higher and if the patch is not totally black, the patch is extracted and it is chosen whether or not to impose an overlap with the next evaluating the third criterion.
3. Annotation area. If the annotation area is small, no overlap is imposed, if it is large, an overlap of 50% is imposed with the next patch.

This process is iterated until the sliding window has examined the entire image area.

After the patch extraction, a patch selection was performed to balance the dataset. Since WSIs have different sizes, the number of patches related to a WSI can vary from one image to another. Using a dataset of this type, the network will be induced to perform better on the WSIs to which a greater number of patches are associated. For this reason, a patch selection criterion has been followed whereby if more than 30 patches are associated with a WSI, 30 of these are randomly selected and the others are discarded, otherwise all are taken. This allows to reduce the imbalance of the dataset.

2.4.2 Training

The architecture used for the two networks is the DeepLabV3+, a Convolutional Neural Network with a ResNeSt as a backbone network.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of Artificial Neural Networks mainly used for image recognition tasks, which use the backpropagation mechanism to extract features with different levels of complexity from the input. The fundamental structure has different layers, of which the convolution layers and the pooling layers extract the features from the input, while the fully connected layer produces the final output by mapping the extracts features. Advancing from one level of the network to the other, the degree of complexity of the extracted features can increase.

The basic architecture of a CNN is shown in Figure 2.9. Typically, there are two main blocks: the feature extraction block, which can be repeated several times, consisting of several convolutional layers followed by a pooling layer, and the block that produces the final output and consists of one or more fully connected layers [29,30].

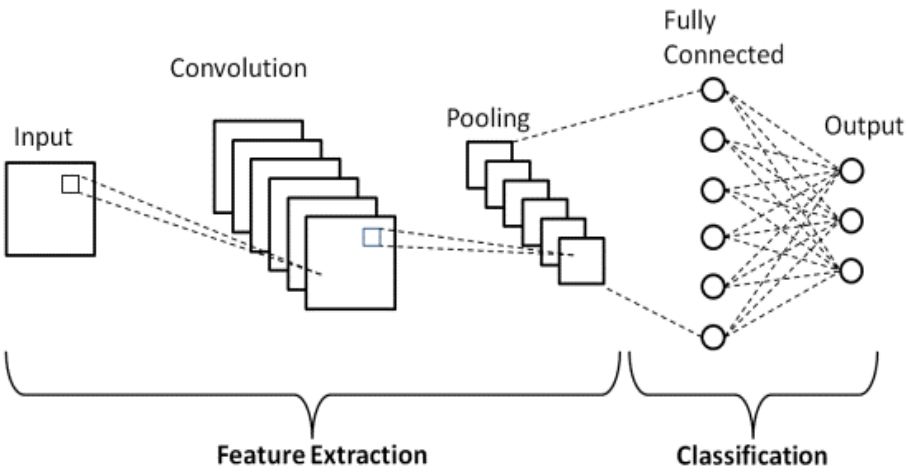


Figure 2.9. Basic architecture of a CNN. The input contains the pixel values of the image.

The *convolutional layer* has the task of extracting features using a combination of linear operations which are the convolutional operations, and non-linear operations, performed by the activation functions.

Convolutional operations consist of applying a *kernel* to the input. Both the input, called *tensor*, and the kernel are arrays of numbers, but the kernel is usually small in size. The output of the operation is obtained through the element-wise product between the input tensor and each kernel element; by summing all the results the output tensor, that is the *feature map*, is finally obtained (Figure 2.10). A different kernel can be applied to each convolutional operation to extract different feature maps; the size and the number of kernels applied are two hyperparameters which define the convolution operation.

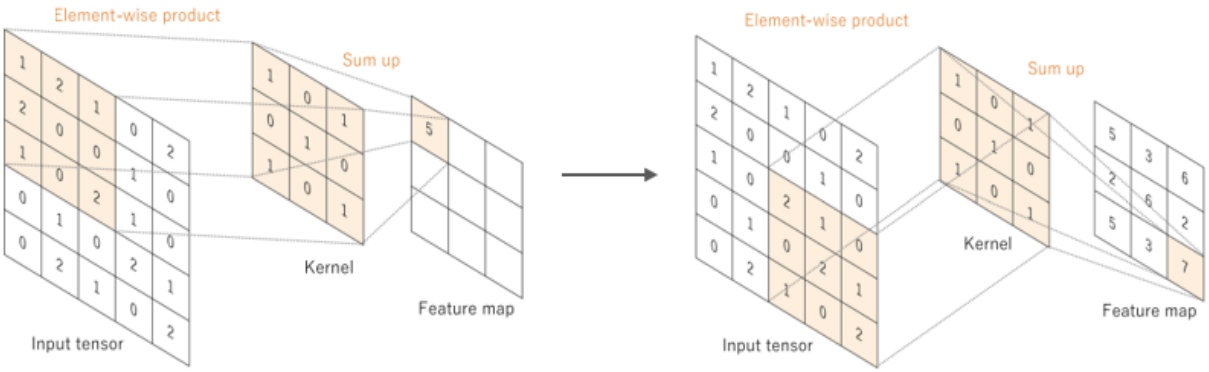


Figure 2.10. The figure shows a convolution operation performed with a 3 x 3 kernel, a stride of 1 and no padding [30].

Model complexity can be significantly reduced by optimally defining three hyperparameters: the depth, the stride and the padding.

The depth of the volume produced as a result of convolutional operations affects the features extracted from the neural network. Using a greater depth can improve network performance, but increases the computational cost as the total number of neurons increases. This parameter should be chosen optimally by evaluating the depth of the input volume and how many kernels are used in each layer.

The stride indicates the distance between two consecutive positions of the kernel. A value of 1 is generally chosen; by choosing a larger stride, the output produced will have lower dimensions as there will be less overlap.

Padding allows to center the kernel on the edges of the image and produce an output that has the same in-plane dimension as the input; to do this, rows and columns are added to the edges of the input tensor. In the case of zero-padding, the added columns and rows consist of zeros (Figure 2.11).

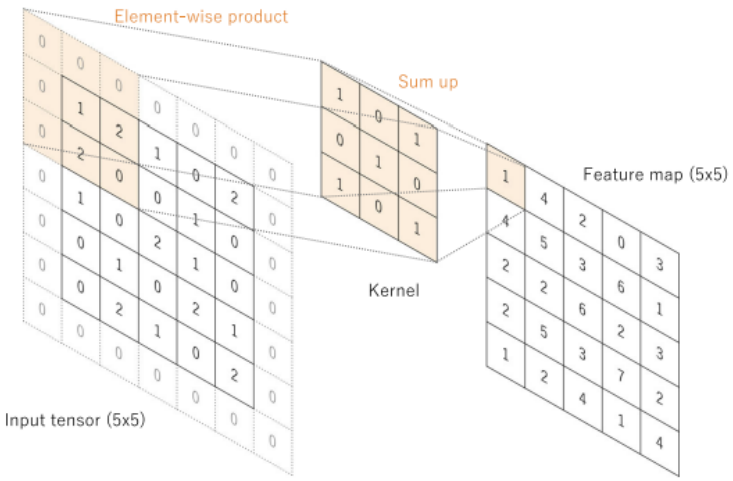


Figure 2.11. Example of a convolution operation with zero-padding [30].

After the convolution operations, non-linear operations are performed by applying activation functions. Most often the ReLU function is used, which is the rectified linear unit, but other types of functions can also be used (Figure 2.12).

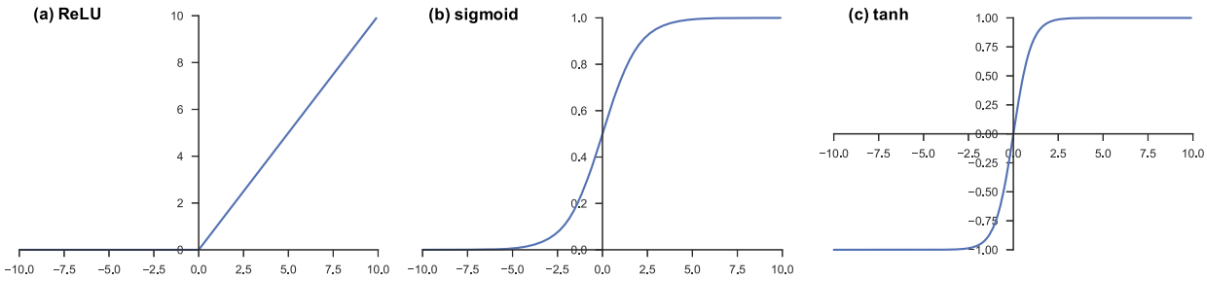


Figure 2.12. The figure shows some activation functions commonly used in neural networks: (a) rectified linear unit (ReLU), (b) sigmoid, and (c) hyperbolic tangent (tanh) [30].

The sigmoid function is limited between 0 and 1, so even if the input to the function is a very large negative or positive number, the output cannot reach very high values, thus making the training process more stable. This function is especially used for binary classification because the goal is to predict the probability as an output, and probability of anything exists only between the range of 0 and 1.

The tanh function is a sigmoidal function limited between -1 and 1. Inputs equal to zero correspond to output equal to 0, positive inputs are as close to 1 as they are positive, and negative inputs are as close to -1 as they are negative. This function is often used in binary classification problems and is generally preferred to sigmoid function.

The ReLU function is used in almost all the convolutional neural networks or deep learning. This function takes only null or positive values, so it is computationally simple while

maintaining sufficient grade of non-linearity and has proved to perform generally better. A modified version of this function is the Leaky ReLU, which has a low positive slope for negative values of the input.

The second type of layer is the *pooling layer*, which performs downsampling operation along the spatial dimensionality of the given input, further reducing the number of parameters and the computational complexity of the model. There are different forms of pooling operation, including max pooling and global average pooling.

The *fully-connected layer* is the last layer of a Convolutional Neural Network. After the input has passed into the last layer of the feature extraction block, the resulting feature map is transformed into a one-dimensional numeric matrix connected to the dense layer, so called because it consists of one or more fully-connected layers. Neurons at this level are directly connected to neurons in adjacent layers and have an associated learnable weight. As a result of this layer the final output of the network is obtained, which in the classification task is represented by the probabilities of belonging to a certain class. Typically, the number of classes defines the number of nodes associated with the last fully-connected layer.
A nonlinear activation function follows each fully connected layer. This function is chosen based on the task and is generally different from those used in the previous layers.

DeepLabV3+ architecture

The DeepLabV3+ is an encoder-decoder that uses Atrous Convolution and Atrous Spatial Pyramid Pooling (ASPP) to do semantic segmentation. The spatial pyramid pooling module captures multi-scale contextual information by pooling features at multiple resolutions and multiple active fields of view, while the encoder-decoder structure is able to capture sharp object boundaries through a gradual recovery of the spatial information performed by the decoder module.

In the encoder module, DeepLabV3+ employs Atrous convolution to extract features at an arbitrary resolution. In Atrous convolution, the dilation rate r defines the active field of view of the convolution. This allows to perform the convolutional operation on a larger area, while maintaining a small kernel size. In fact, the kernel actually expands when the dilation rate changes, and the remaining positions are then filled with zeros. As result, a $k \times k$ kernel with an atrous dilation rate r will produce a kernel $k_e = k + (k - 1)(r - 1)$ and introduces $r - 1$ zeros within consecutive filter values. Standard convolution is a particular case of atrous convolution with $r = 1$ (Figure 2.13).

Using the Atrous convolution it is therefore possible to have denser features even without using large kernels, which would be operationally expensive and time-consuming [31,32].

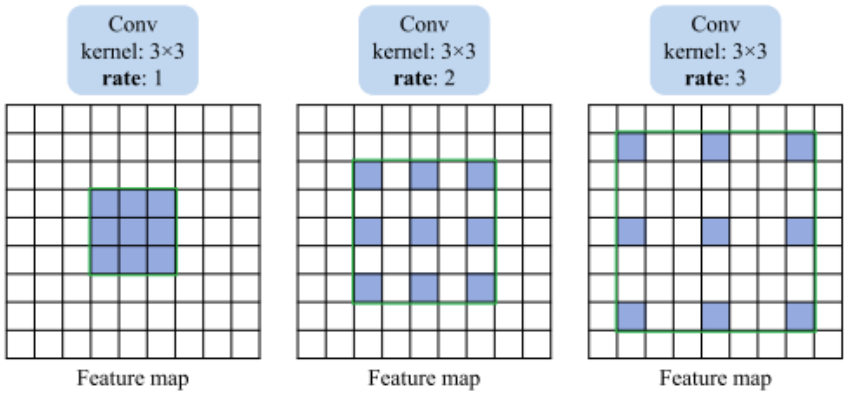


Figure 2.13. Atrous convolution with kernel size 3x3 for different dilations rates *r*. With r = 1, atrous convolution works as a standard convolution; with r = 2, it works as a convolution with a kernel size of 5x5; with r = 3, it works as a convolution with a kernel size of 7x7 [31].

Atrous Spatial Pyramid Pooling (ASPP) involves applying several atrous convolution layers with different rates to the image, in order to capture multi-scale data. The ASPP feature map is obtained through one 1×1 convolution, three 3×3 atrous convolutions with rate parameters 6, 12, and 18, and image pooling to extract image-level features. The resulting features are then concatenated together and passed through a 1×1 convolution. As result, the encoder networks in the output feature map encode features which are rich in semantic information. In the decoder module, these features are then bilinearly upsampled with a factor of 4 and, after being reduced using 1×1 convolution, concatenated with the corresponding low-level features that come from a backbone network with the same shape. The 1×1 convolution is applied so that the corresponding low-level features cannot outweigh the encoded features, since they usually contain a large number of channels. After the concatenation, the features are passed through a few 3×3 convolutions to refine the features before undergoing a final bilinear upsampling by a factor of 4. At the end, a softmax layer is used to obtain a segmentation map [31].

The architecture of DeepLabV3+ with backbone network is shown in Figure 2.14.

The use of a backbone network allows to implement transfer learning. This concept allows models that are trained on general datasets to be specialized for specific tasks by using a considerably smaller dataset that is problem-specific. In this case, the network does not perform feature extraction, but uses the feature extraction carried out by another pre-trained network on a much larger dataset, since transfer learning is based on the idea that even very different dataset can share the same generic features if these have been extracted from sufficiently large datasets. The backbone network architecture used for the development of this algorithm is a ResNeSt101.

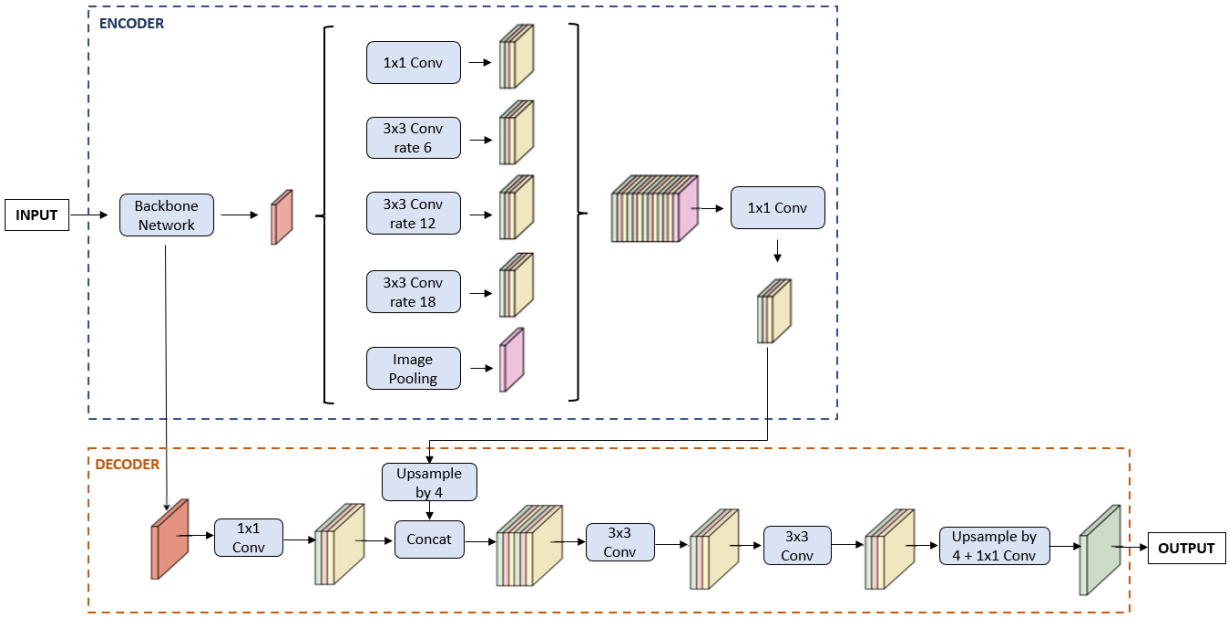


Figure 2.14. DeepLabV3+ architecture with backbone network [31].

ResNeSt backbone network

The ResNeSt is a Split-Attention Network which combines the channel-wise attention strategy with multipath network layout, so it captures cross-channel feature correlations, while preserving independent representation in the meta structure. This architecture has shown to have better performance on the classification tasks than the classic architectures generally used. The network is made up of several modules stacked together, the Split-Attention Blocks, whose structure is shown in Figure 2.15.

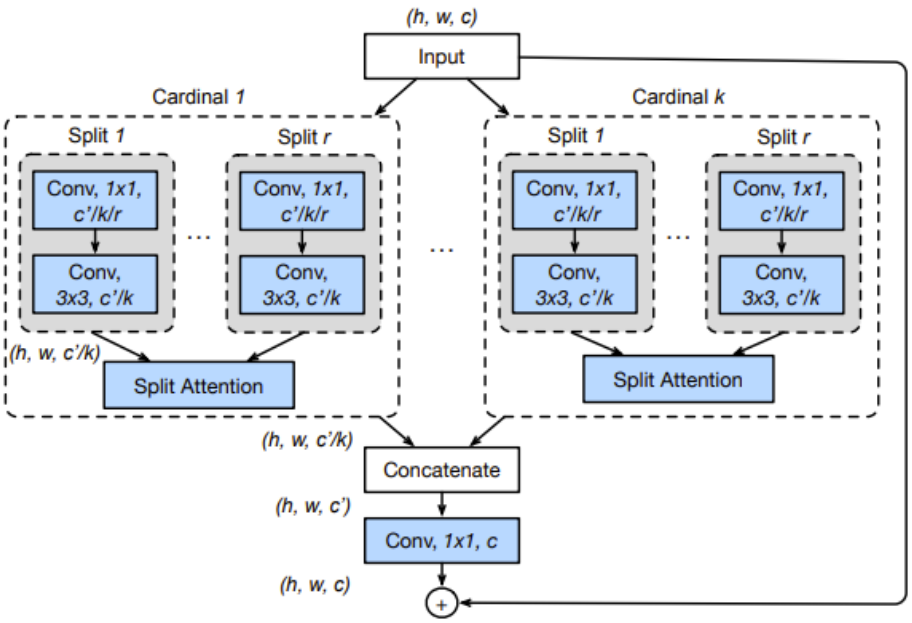


Figure 2.15. ResNeSt block represented in cardinality-major view [33].

This block is a computational unit which performs a set of transformations on low dimensional embeddings and concatenates their outputs as in a multi-path network. Each transformation uses a channel-wise attention technique to capture interdependencies of the featuremap [33]. Each Split-Attention Block consists of:

- Featuremap group operations. The number of featuremap groups that can be created from the features is determined by the *cardinality* hyperparameter K; these featuremap groups are known as *cardinal groups*. The *radix* hyperparameter R indicates the number of splits in a cardinal group, hence the total number of feature groups is $G = K*R$. After each split is possible to apply a series of transformations F_i to each individual group, and the intermediate representation of each group is indicated as $U_i = F_i(X)$.
- Split-Attention operations (Figure 2.16). For each cardinal group it is possible to obtain a combined representation by merging its splits through an element-wise summation. This sum is fed into a global pooling layer to extract certain global statistics denoted as s^k, which are then fed into another network to calculate numbers that are subsequently transformed into weights by applying a softmax function. These weights at the end will be used to compute weighted summation of the initial inputs. In fact, in the final step a weighted fusion V^k of the cardinal group representation is aggregated using channel-wise soft attention, where each featuremap channel is created using a weighted combination over splits.

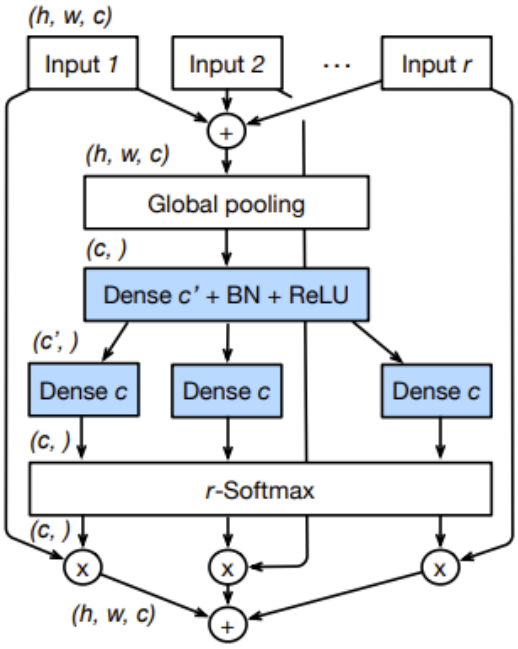


Figure 2.16. Description of Split-Attention operation within a cardinal group [33].

After the Split-Attention operation, the cardinal group representations are concatenated along the channel dimension: $V = Concat\{V^1, V^2, ...V^K\}$. If the input and output featuremap have the same shape, the Split-Attention block uses a shortcut connection to create the final output, that is: $Y = V + X$. For blocks with a stride, to align the output shapes an appropriate transformation T is applied to the shortcut connection: $Y = V + T(X)$.

The configuration shown in Figure 2.15, where the featuremap groups with the same cardinal index are physically near to one another, is the *cardinality-major implementation*. Although this approach is simple and intuitive, it is challenging to modularize and speed it using standard CNN operators. For this reason, an equivalent *radix-major implementation* (Figure 2.17) can be used.

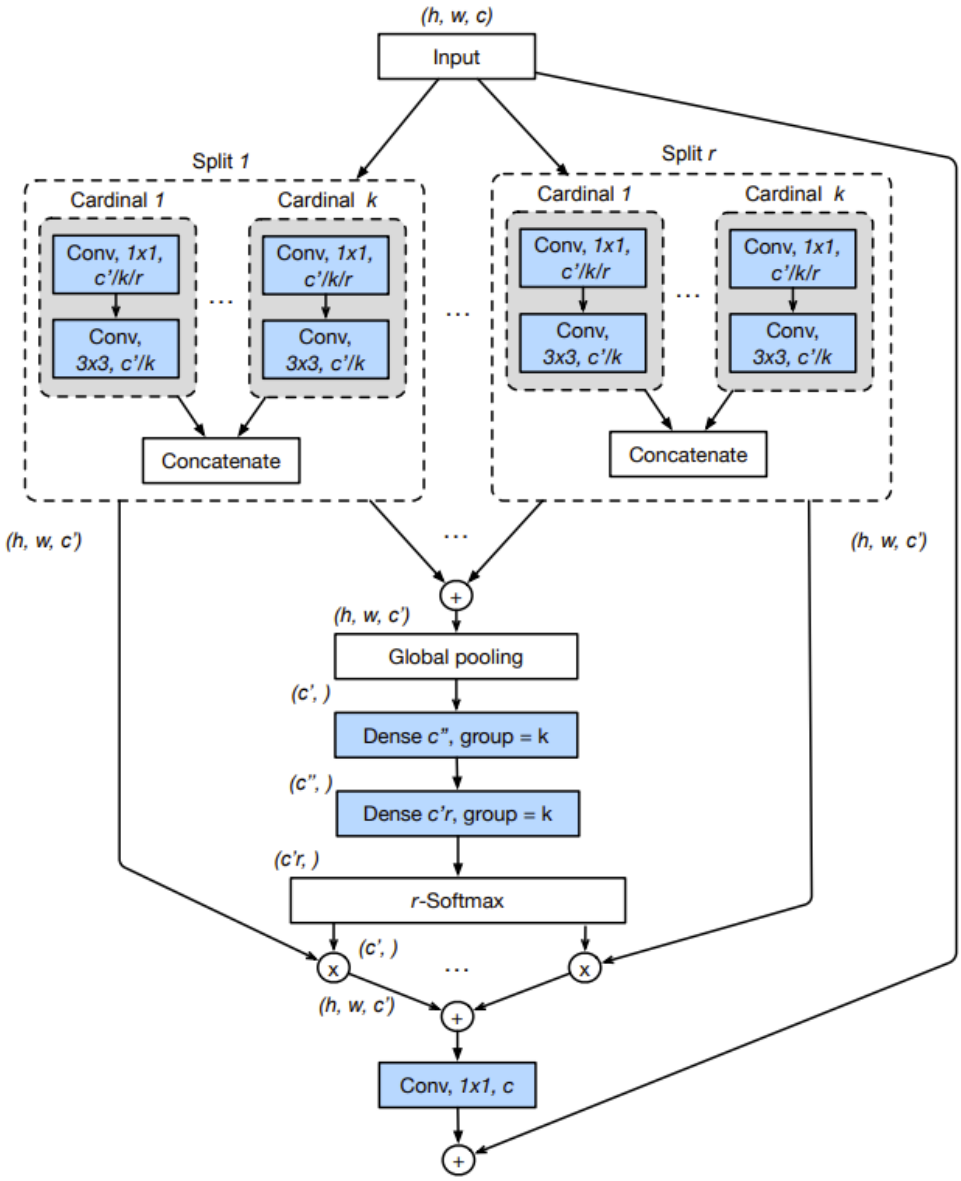


Figure 2.17. ResNeSt block represented in radix-major view [33].

In this configuration, the input featuremap is first divided into *R*K* groups, each of which has a cardinality-index and radix-index. The featuremap groups with the same cardinality-index but different radix-index can then be fused together using a summation across different splits. At this point, as done in the previous configuration, a global pooling layer is used which aggregates over the spatial dimension while keeping the channel dimension separated. After this, two consecutive fully connected (FC) layers are added, each with a number of groups equal to the cardinality, which allow to predict the attention weights for each split.
The radix-major implementation allows for the unification of the initial 1×1 convolutional layers into a single layer and the implementation of the 3×3 convolutional layers using a single grouped convolution with the number of groups of *R*K*. With this configuration standard CNN operators can be used to modularize the Split-Attention block [33].

Characteristics of the backbone network used

The Split-Attention network model used as backbone network is the ResNeSt-101, a convolutional neural network that is 101 layers deep. For this configuration, the radix hyperparameter was set to 2. The output stride is equal to 8 and indicates the ratio between the spatial resolutions of the input and the output; a value of 8 allows for denser feature extraction, and this is typically done for semantic segmentation tasks.
The dataset used to pre-train the network is the ADE20K dataset, which contains more than 20000 indoor and outdoor images densely annotated with pixel-level objects and object parts labels, for a total of 150 classes of objects [34].
The crop size is 512x512 pixels, and indicates the size of the inputs. This means that the patches extracted, with size 768x768 pixels, have been sub-sampled to match with the input size, in order to reduce the computational cost without particularly varying the performance.
The samples per GPU value is equal to 4 and indicates the batch size of each GPU, i.e., how many samples per GPU to load during training, while the workers per GPU value, still equal to 4, indicates how many subprocesses to use for data loading for each GPU. In this case, 4 GPUs are needed to train the model, and that means that the batch size is 16, since this is the result of the product between number of GPUs and samples per GPU.

Normalization layer

The normalization layer can be used to adaptively normalize each layer's input so as to mitigate the risk of overfitting. In the model developed was used the method of Batch Normalization (BN), which makes the neural network faster and more stable.

The mini-batch is a subset of the training set. In BN each input in the mini-batch considered is normalized by subtracting the batch mean and dividing it by the standard deviation, and then is scaled and shifted in order to obtain the final input of the next layer. Normalizing the input features is important because the fact that they are defined in the same range reduces the variation in the distribution of the input data and speeds up network learning.
Batch Normalization also works as regularization since noise is added to each input within the mini-batch, and adding noise at the level of the hidden units prevents the model from relying heavily on a given hidden unit, reducing generalization error.

Data augmentation

Data augmentation consist of using random transformations to create surrogate images of the input data, for example through translation, rotation, cropping, scaling, flipping, variation of brightness or variation of contrast operations, so as to create different images to use in the training phase. In this way is possible to expand the training set and reduce overfitting.
In the model developed, the data augmentation techniques used are scaling and flipping. Scaling was performed outward, setting a new image size of 640x640 after scaling. For flipping a random flipping has been set with a rotation probability of 50%.

Hyperparameters

For training the networks, the same hyperparameters were set for the tissue network and the artifact network, with the exception of the number of epochs which varies between the two networks (Table 2.1).

LEARNING RATE	0.005	
ITERATIONS FOR EPOCH	675	
EPOCHS	TISSUE	ARTIFACTS
	30	50

Table 2.1. Hyperparameters set for training.

The learning rate is an hyperparameter which determines the step size with which to update each learnable parameter. For both networks the learning rate of the optimizer was set to 0.005 and this defines how much weights are updated during training. As optimization algorithm was used the Stochastic Gradient Descent with a modification, the Online Hard Example Mining pixel sampler.

The number of iterations for epoch was determined by dividing the length of the training set for the number of samples per GPU, so in this case is equal to 675 since the length of the training set is 2700 and the samples per GPU are 4.

The number of epochs was set to 30 for the tissue network and 50 for the artifacts network. This number is defined by the sum of one forward pass and one backward pass for all training samples. It must be chosen so as not to be too high, because this would make the model lose generalization capability. In this case, the number of epochs is greater for the artifact network since the segmentation task is more complex. After each epoch, metrics were refreshed and intermediate results were saved.

OHEM pixel sampler

The technique used for training the model is called Online Hard Example Mining (OHEM) and is a modification to Stochastic Gradient Descent (SGD), used in presence of a large imbalance between the annotated objects and the background examples.

SGD is an iterative method for the optimization of loss functions, which at each iteration replaces the current value of the gradient of the loss function with an estimate obtained by evaluating the gradient on the mini-batch. After each iteration the parameters are updated with a step that depends on the chosen learning rate [30].

Even if each mini-batch contains a small number of images, there are thousands of example RoIs in each image, so it can be useful to create a different mini-batch by selecting the hard examples among thousands of candidates examples: this is what is done in OHEM. In fact, considering all the RoIs in each image, there would be many easy examples with low loss that would not contribute much to the gradient. Using OHEM makes training more efficient, leads to better training convergence and improves detection accuracy [36].

To train the developed model it is chosen to implement OHEM pixel sampler using only pixels with confidence score under 0.8 and keeping at least (640 * 640) / 4 pixels during training.

Loss functions

For the training a multiple-loss function was used given by the linear combination of three loss functions: Cross Entropy loss, Lovász-Softmax loss and Dice loss, with loss weights of 0.75, 1.25 and 1 respectively.

The loss function is a function that compares the target and predicted output values since during training the aim is to minimize the loss between the desired output and the predicted output. In presence of multiple-loss functions backpropagation is used to simultaneously optimize all loss functions while training the network.

The *Cross Entropy loss* is one of the most used loss function and allows to quantify the difference between the true distribution (one-hot distribution) and the predicted distribution. Its formula is a summation of the true probability *P'* multiply by the log predicted probability *P* over all classes *i* in the distribution:

$$H(P'|P) = -\sum_{i} P'(i)\, logP(i)$$

The *Lovász-Softmax loss* is a method developed in the context of semantic image segmentation for optimizing the performance with respect to the Intersection over Union (IoU) value. The IoU index of a class *c* is defined as:

$$IoU_c(y', \tilde{y}) = \frac{|\{y' = c\} \cap \{\tilde{y} = c\}|}{|\{y' = c\} \cup \{\tilde{y} = c\}|}$$

where y' is a vector of ground truth labels and $\tilde{y}$ is a vector of predicted labels.

The corresponding loss function to minimize is:

$$\Delta_{IoU_c}(y', \tilde{y}) = 1 - IoU_c(y', \tilde{y})$$

For multiclass semantic segmentation the output scores of the model can be transformed into probabilities through the Softmax function and these are then used to construct a vector of pixel errors *m*(*c*) for each class *c*. By averaging over all the classes *C*, the Lovász-Softmax loss is obtained:

$$loss(f) = \frac{\mathbf{1}}{|C|} \sum_{c \in C} \overline{\Delta_{IoU_c}}(m(c))$$

where *f* is the normalized vector of all network outputs $f_i(c)$.

The Lovász-Softmax loss cannot optimize the mean IoU (mIoU) of the entire dataset since it consider only a small number of pixel predictions at each optimization step. In order not to degrade the dataset-mIoU, this loss must to be optimized taking into account only the classes present in each batch when averaging over the classes [37].

The *Dice loss* was developed to overcome the problem of predictions strongly biased towards background when in the dataset there are a lot of images in which the area of interest occupies only a small region of the entire image [38]. The aim is to maximize the dice coefficient D, which for the class *c* can be written as:

$$D_c(y', \tilde{y}) = \frac{2\,|y' = c \cap \tilde{y} = c|}{|y' = c| + |\tilde{y} = c|}$$

where y' is a vector of ground truth labels and $\tilde{y}$ is a vector of predicted labels.
The corresponding loss function to minimize is:

$$\Delta_{D_c}(y', \tilde{y}) \;=\; 1 \;-\; D_c(y', \tilde{y})$$

Using multiple different loss functions works as a regularization technique which allow to prevent overfitting.
Moreover, an auxiliary head was used in the method developed. This means that was add an auxiliary loss whose segmentation result has no influence on the final segmentation result of the network, but can help to accelerate training and improve accuracy.

2.4.3 Testing

Once the model was trained, the tissue network and the artifact network were tested.
For the test phase, an inference strategy was adopted whereby all pixels at 0 were set to a value of 245 (i.e., white) in order to standardize the appearance of all WSIs, since there were some WSIs with a black background. This leads the network to have better performance.
After that, the predictions were generated using a sliding window approach. This means that a window with dimensions of 768 x 768 pixels has been scrolled on the starting image in order to generate the predictions of each patch on which the window is positioned. The procedure is repeated on the whole image, imposing a 50% overlap between one window and the next. From the resulting softmax, which is the prediction of the starting image, the final automatic mask is obtained.

Figure 2.18 shows an example of the procedure performed by the tissue network to generate the prediction on a WSI.

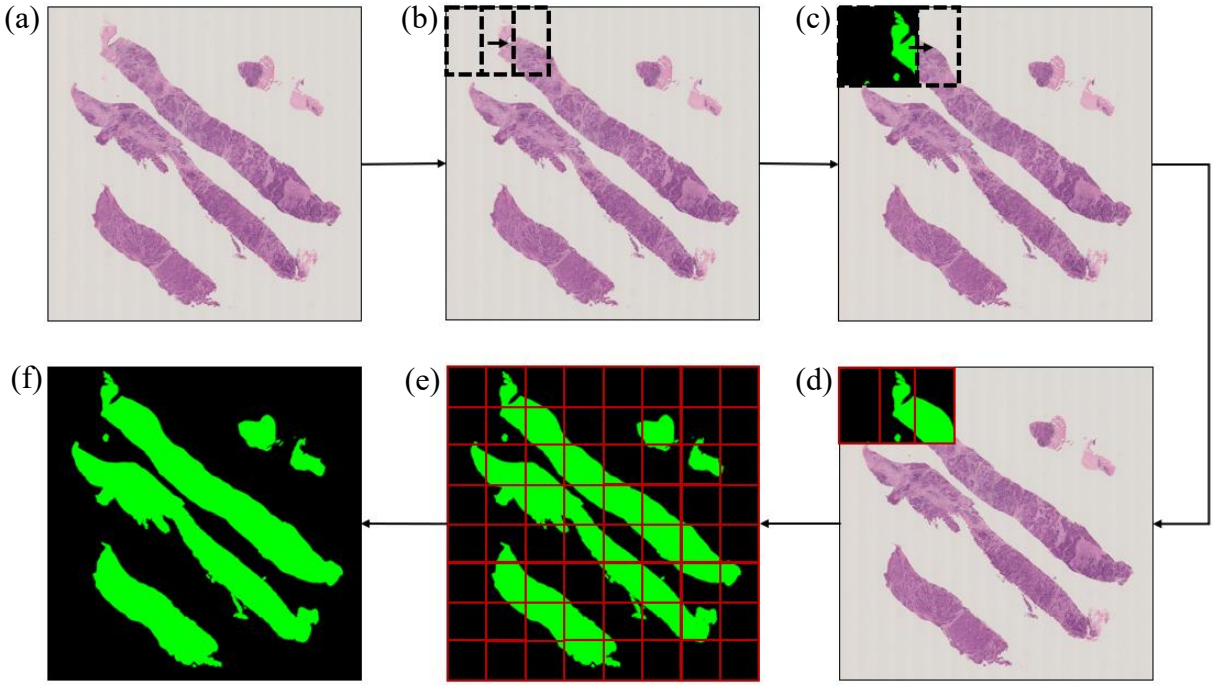


Figure 2.18. Example of the procedure followed in the test phase. The figure shows (a) the original image, (b) the sliding window passed over the original image, (c) the prediction of the neural network on the first patch, (d) the prediction of the neural network on the first and the second patch, (e) the prediction on all patches after the window has scrolled over the whole image and (f) the final softmax.

2.5 Validation metrics

For the validation of the developed method, the results were calculated using pixel-wise metrics, i.e. evaluating the segmentation performance of the networks on the classification of each individual pixel.

The metrics were calculated on the entire dataset. In fact, specifically for the dataset of the artifact network, there are many masks that do not show segmentation, as there are also good quality images that do not present artifacts. This, in the calculation of metrics based on the segmented area such as the Dice Similarity Coefficient, would give invalid values. Conventionally, in these cases the maximum value is given to the result of the metric when both the manual mask and the automatic mask do not present a segmentation in order to indicate that the network is right in not identifying objects to be segmented, but for this specific problem this would greatly unbalance the results. In fact, on a WSI it is unlikely that all three categories of artifacts are present at the same time. When only one artifact is present on a WSI, the mask

corresponding to that artifact will have segmentation, but the masks for the other two categories of artifacts will be completely black. Cases of this type would greatly affect the segmentation results evaluated for each single category of artifact if these were averaged over the entire dataset after being calculated on each WSI. For this reason, it was decided to evaluate the overall results on the entire dataset, i.e. by calculating the overall confusion matrix given by the sum of all the true positives, true negatives, false positives, false negatives on all the WSIs and then calculating the metrics on the resulting confusion matrix.
The metrics chosen for the evaluation of the results are the Dice Similarity Coefficient (DSC), the sensitivity and the specificity.

The *confusion matrix* is a table that allows the visualization of the performances of a classification algorithm. Each row represents the actual classification for each class, while each column represents the predicted classification for each class, or vice versa. The classes are ordered in the same way on the rows and on the columns, so that all the correct predictions are on the main diagonal.
Comparing the true classification and the predicted classification four different outcomes are possible:

- True Positive (TP): the actual classification and the predicted classification are both positive, which means that the classifier has correctly identified the class of the sample;
- True Negative (TN): the actual classification and the predicted classification are both negative, which means that the classifier has correctly identified that the sample does not belong to that class;
- False Positive (FP): the actual classification is negative and the predicted classification is positive, which means that the classifier incorrectly identified the sample as belonging to that class;
- False Negative (FN): the actual classification is positive and the predicted classification is negative, which means that the classifier incorrectly identified the sample as not belonging to that class.

The confusion matrix for a binary classification problem is shown in Figure 2.20, considering class b as the one whose classification is to be evaluated.

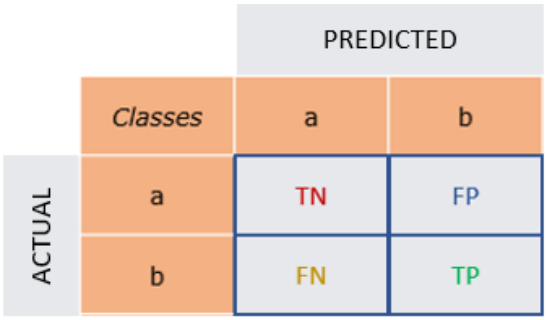


Figure 2.20. Example of confusion matrix for a binary classification problem where class b is that of interest.

The same reasoning can be extended to a multi-class classification problem (Figure 2.21).

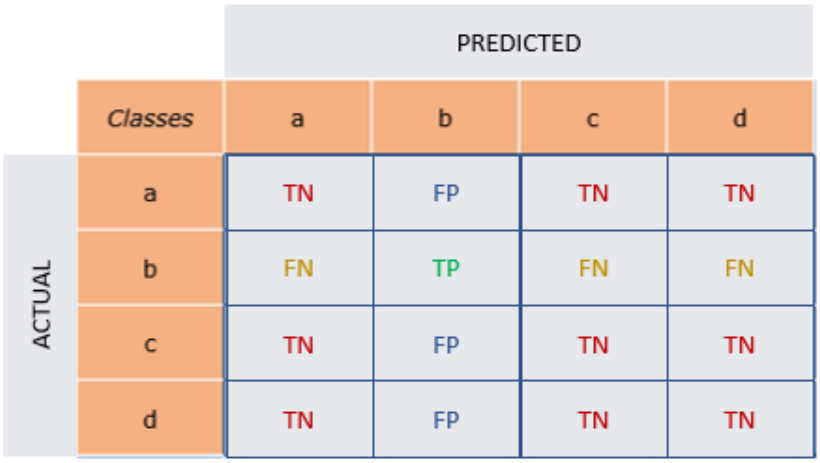


Figure 2.21. Example of confusion matrix for a multi-class classification problem where class b is that of interest.

The *sensitivity*, or true positive rate, is the ability to correctly identify positive values, that is, samples that belong to a certain class. It is given by the ratio between the number of positives correctly identified by the model (TP) and the total number of positives actually present in the dataset, identified (TP) or not (FN) by the model:

$$Sensitivity = \frac{TP}{TP + FN}$$

The *specificity*, or true negative rate, is the ability to correctly identify negative values, that is, samples that not belong to a certain class. It is given by the ratio between the number of

negatives correctly identified by the model (TN) and the total number of negatives actually present in the dataset, identified (TN) or not (FP) by the model:

$$Specificity = \frac{TN}{TN + FP}$$

The *Dice Similarity Coefficient (DSC)* is a metric commonly used for semantic segmentation which allows the evaluation of the overlap between the true object and the predicted object, i.e. the manual segmentation and automatic segmentation. Considering that X is the manual segmentation mask and Y the automatic segmentation mask, the DSC is given by:

$$DSC = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

where $|X|$ and $|Y|$ are the areas in pixels of the manual mask and the automatic mask respectively, while $X \cap Y$ is their intersection. The meaning of the DSC formula is shown graphically in Figure 2.22.

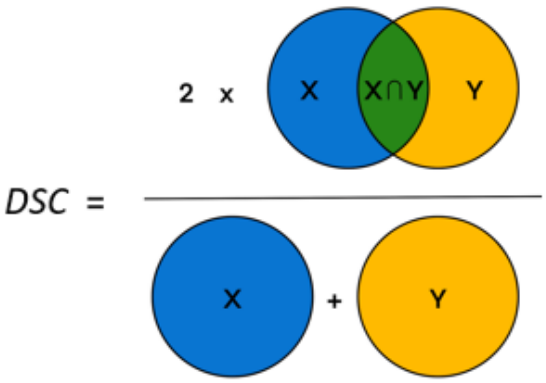


Figure 2.22. Graphical representation of the DSC formula.

The DSC value is between 0 and 1, where 0 indicates that there is no overlap between the two segmentation masks and 1 indicates that there is complete overlap.
This metric can also be calculated on the confusion matrix in terms of true positives, false positives and false negatives. In this case, the value of DSC is given by:

$$DSC = \frac{2\,TP}{2\,TP + FP + FN}$$

Chapter 3

Results

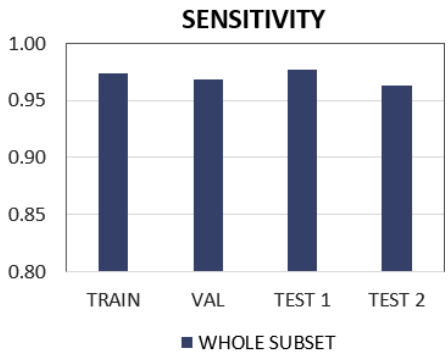
3.1 Tissue network

The results relating to the assessment of tissue network performance on tissue segmentation are reported and discussed below.

The following figures show the results relating to the sensitivity (Figure 3.1), specificity (Figure 3.2) and DSC (Figure 3.3) values calculated first for each subset, without distinction on the type of staining, and then evaluating the performance by type of staining on each subset.
The values of all metrics represented range from 0 to 1. Missing values refer to stains that are not present in that subset.

(a)

	WHOLE SUBSET
TRAINING SET	0.9732
VALIDATION SET	0.9683
TEST SET 1	0.9767
TEST SET 2	0.9629



(b)

	HE	IHC	PAS	TRIC
TRAINING SET	0.9793	0.9390	0.9546	0.9636
VALIDATION SET	0.9787	0.9081	0.9474	0.9726
TEST SET 1	0.9789	0.9171	0.9633	0.9729
TEST SET 2	0.9629	-	-	-

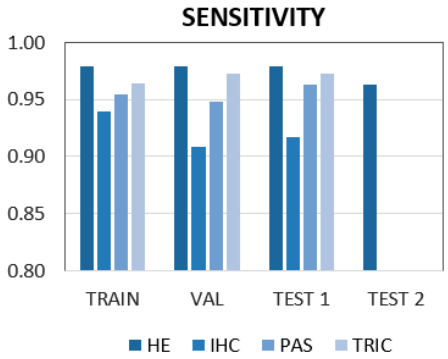
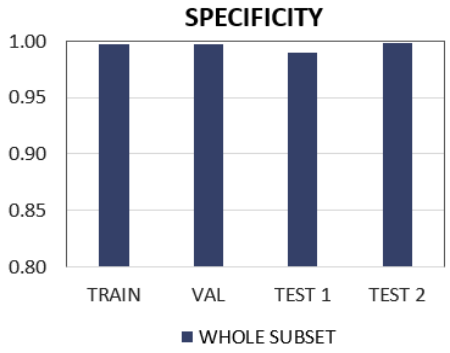


Figure 3.1. Sensitivity results for the tissue network, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

(a)

	WHOLE SUBSET
TRAINING SET	0.9971
VALIDATION SET	0.9976
TEST SET 1	0.9897
TEST SET 2	0.9985



(b)

	HE	IHC	PAS	TRIC
TRAINING SET	0.9970	0.9976	0.9955	0.9990
VALIDATION SET	0.9976	0.9973	0.9977	0.9990
TEST SET 1	0.9914	0.9675	0.9948	0.9909
TEST SET 2	0.9985	-	-	-

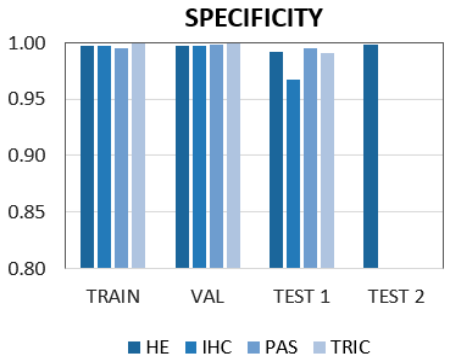
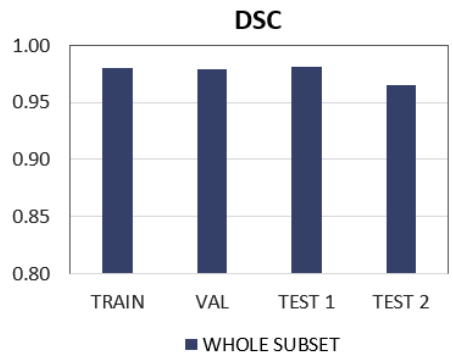


Figure 3.2. Specificity results for the tissue network, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

(a)

	WHOLE SUBSET
TRAINING SET	0.9803
VALIDATION SET	0.9788
TEST SET 1	0.9809
TEST SET 2	0.9651



(b)

	HE	IHC	PAS	TRIC
TRAINING SET	0.9835	0.9637	0.9589	0.9765
VALIDATION SET	0.9841	0.9475	0.9660	0.9810
TEST SET 1	0.9835	0.9123	0.9729	0.9775
TEST SET 2	0.9651	-	-	-

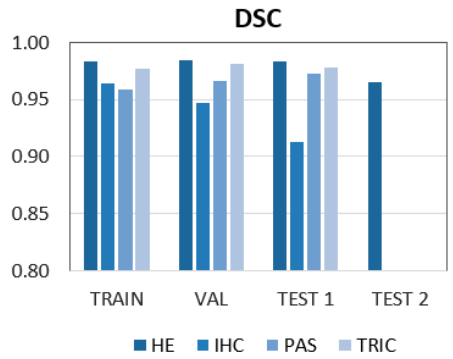


Figure 3.3. DSC results for the tissue network, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

The results show that in general the metrics have high values on all subsets and all stains, always reaching percentages above 90%. The highest results are obtained on specificity and indicate that the number of false positives is very low, so the network has great ability to identify the absence of tissue in the histological slide. The high sensitivity values indicate that the network has very good performance on tissue identification, and the same can be deduced from the high DSC values which indicate that the automatic segmentation produced by the network has a very high percentage of overlap with the manual segmentation.
Looking at the automatic masks produced by the tissue network, it is possible to make a qualitative comparison between automatic segmentation and manual segmentation and identify the cases that most frequently lead the network to make mistakes.

Figure 3.4 shows the comparison between the manual mask and the automatic mask produced by the network for an histological slide with IHC staining. In these cases, the tissue network correctly segments the part of the tissue with darker color, but fails where the tissue is particularly light (in the figure, in the outermost parts of the tissue): these regions of tissue are classified as background. This explains why the sensitivity and DSC metrics tend to assume lower values on the IHC staining.

Figure 3.4. Comparison between the manual mask and the automatic mask relative to a histological slide with IHC staining.

In addition to this specific case for IHC staining, there are other cases where the tissue network recurrently fails the classification.
Figure 3.5 shows an histological slide in which there is an air bubble on the area of the tissue. The manual mask is obtained by subtracting the "other" artifacts present (in this case air bubble) from the total area of the tissue, so that the network does not recognize them as tissue. Nevertheless, the automatic mask shows how the network in these cases also segments the area below the air bubble as tissue.

However, this is plausible, since the tissue is sharp below the air bubble. Furthermore, if the artifact is recognized by the artifact network it is still possible to obtain a segmentation similar to the manual one by subtracting the area of the artifact segmented by the artifact network from the tissue area identified in the automatic mask.

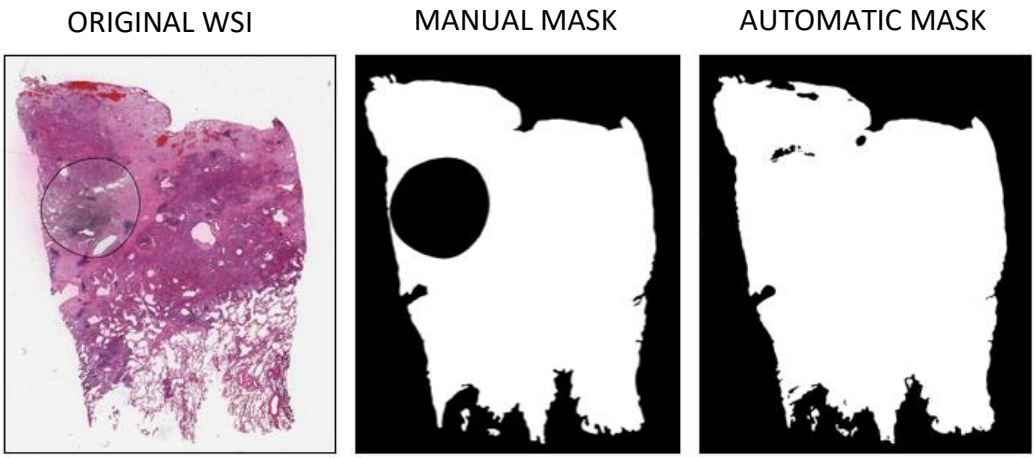


Figure 3.5. Comparison between the manual mask and the automatic mask relative to a histological slide in which there is an air bubble on the tissue.

Another situation in which the tissue network frequently fails is one in which there are bubbles with certain shades of color on the WSI. Figure 3.6 shows a histological slide that has a colored bubble at the edge. In these cases the bubble can take on shades of pink that make the network fall into error in the presence of tissue with HE staining, consequently the area of the bubble is classified as tissue. However, even in this case, if the artifact network identifies the colored bubble as an artifact, by subtracting the segmented artifact area from the automatic tissue mask it is possible to obtain an automatic mask much more similar to the manual one.

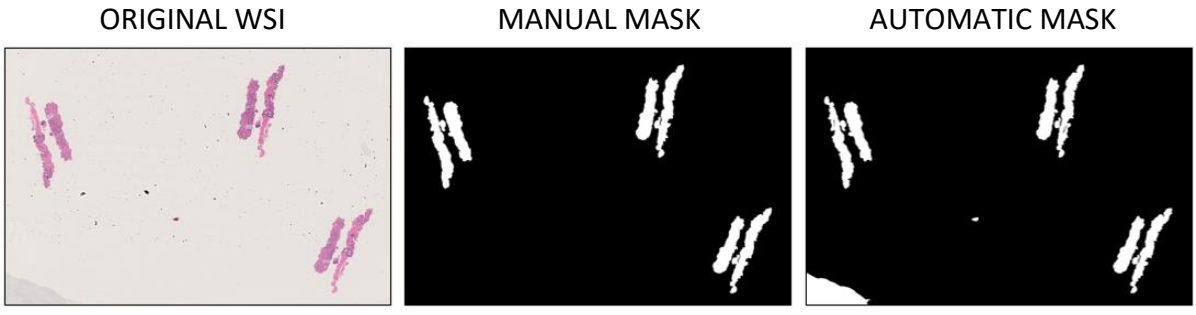


Figure 3.6. Comparison between the manual mask and the automatic mask relative to a histological slide showing a colored bubble on the edge.

Figure 3.7 shows another case in which the network mistakenly segments an artifact as tissue. In some slides there are annotations made by the pathologist with the marker, and when these have a similar color to the tissue, as in the case in the figure, sometimes the network fails and segments the marker as tissue. Again, if the artifact network identifies the pen marker as an artifact it can be subtracted from the automatic mask to obtain the final segmentation of the tissue only.

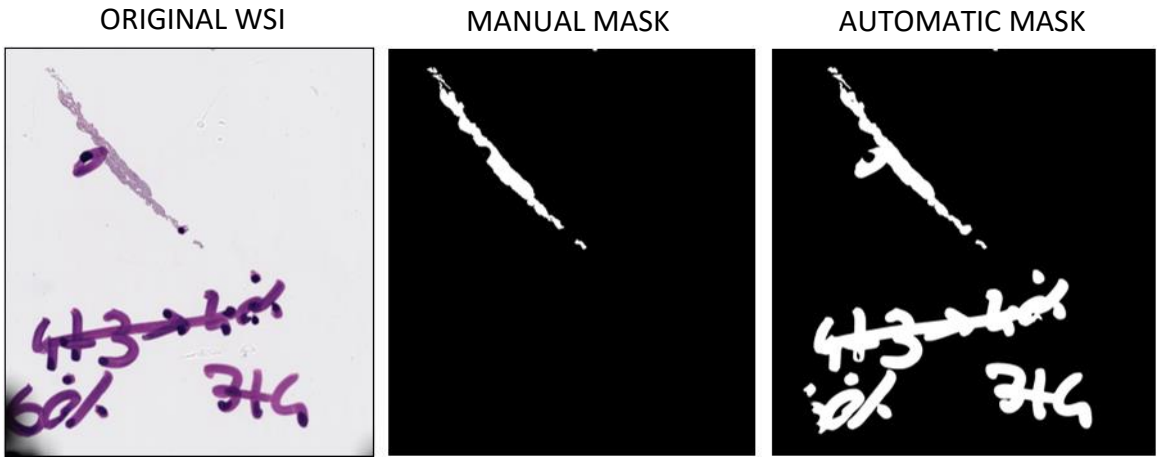


Figure 3.7. Comparison between the manual mask and the automatic mask relative to a histological slide with pen marks.

A final case of recurrent tissue network error is shown in Figure 3.8. Here the histological slide has tissue that is particularly blurred and this causes the network to fall into error, in fact in these cases the presence of tissue is not recognized at all.

Figure 3.8. Comparison between the manual mask and the automatic mask relative to a histological slide with very blurred tissue.

Overall, for the tissue network it can be seen that very high performances have been achieved, that is, the network has a great ability to identify the tissue, despite there are some typical cases in which it often falls into error. The high values of the metrics that occur on test set 1 and test set 2 show that the tissue network has a high generalizability.

3.2 Artifact network

Two different evaluations were made for the validation of the artifact network:

1. Ability to identify specific types of artifact. In this case the validation metrics were calculated on the three different masks relating to the three different categories of artifact to evaluate the performance of the network in recognizing that specific type of artifact;
2. Ability to identify the presence of a generic artifact. In this case the validation metrics were calculated on the mask obtained by combining the predictions for the various artifact classes, in order to evaluate the ability of the network to distinguish between the presence and the absence of artifacts.

The results relating to the assessment of artifact network performance are reported and discussed below.

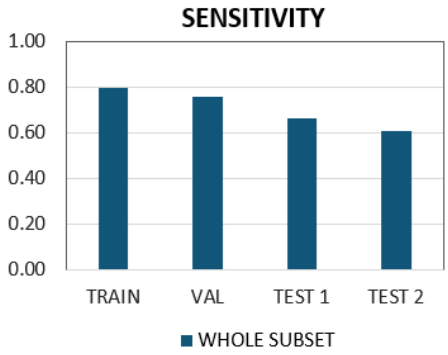
The results are reported first for each category of artifact, and then considering the overall performance on all artifacts. The sensitivity, specificity and DSC values are represented. The specificity values evaluated on the masks of the artifacts are always close to 1 since the artifact covers a rather small area on the WSI and everything else falls into the negative class, therefore errors in the classification of what is not artifact weigh very little, being always few in proportion to the total number of elements of that class. However, to evaluate the ability to recognize the absence of an artifact, the specificity values were also reported, but for a better visualization the range shown in the graph goes from 0.98 to 1. The sensitivity and DSC instead show more variable values within the possible range, so the graph shows the entire range from 0 to 1. In the table, the symbol "-" refers to stains for which there are no WSIs in that subset, while the term NaN refers to stains for which that artifact is not present in that subset, therefore the calculation of some metrics returns indeterminate forms.

All results were calculated first for each subset, and then evaluating the performance by type of staining on each subset.

Tissue fold artifacts

The following figures show the results relating to the sensitivity (Figure 3.9), specificity (Figure 3.10) and DSC (Figure 3.11) values calculated for the tissue fold artifact.

	WHOLE SUBSET
TRAINING SET	0.7999
VALIDATION SET	0.7583
TEST SET 1	0.6636
TEST SET 2	0.6103



	HE	IHC	PAS	TRIC
TRAINING SET	0.8205	0.7511	0.6881	0.6262
VALIDATION SET	0.7542	0.7426	0.8476	0.8816
TEST SET 1	0.6573	0.8268	0.3232	NaN
TEST SET 2	0.6103	-	-	-

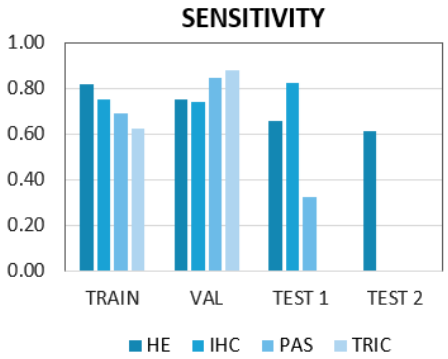
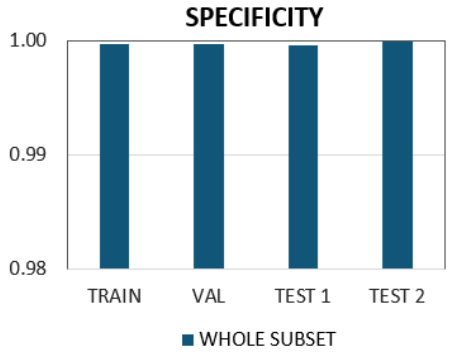


Figure 3.9. Sensitivity results for the artifact network on the tissue fold artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.9997
VALIDATION SET	0.9998
TEST SET 1	0.9996
TEST SET 2	1.0000



	HE	IHC	PAS	TRIC
TRAINING SET	0.9998	0.9997	0.9994	0.9999
VALIDATION SET	0.9998	0.9994	0.9997	0.9999
TEST SET 1	0.9996	0.9997	0.9999	1.0000
TEST SET 2	1.0000	-	-	-

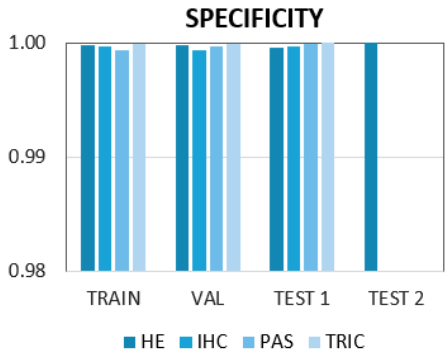
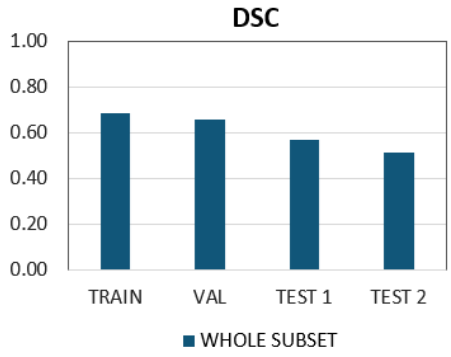


Figure 3.10. Specificity results for the tissue network on the tissue fold artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.6896
VALIDATION SET	0.6570
TEST SET 1	0.5691
TEST SET 2	0.5174



	HE	IHC	PAS	TRIC
TRAINING SET	0.7000	0.7234	0.1370	0.5249
VALIDATION SET	0.6419	0.6590	0.7903	0.7981
TEST SET 1	0.5664	0.6355	0.2266	NaN
TEST SET 2	0.5174	-	-	-

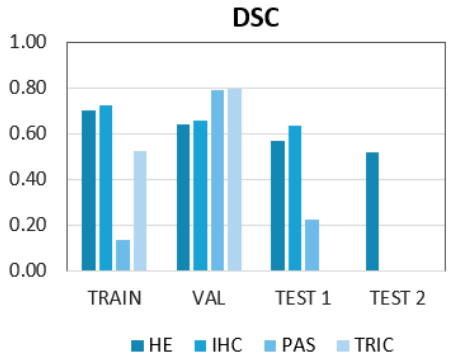


Figure 3.11. DSC results for the tissue network on the tissue fold artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

The results, considering the difficulty of the task, show good values on all subsets, with some variations in performance based on the type of stain.

Looking at the Dice Similarity Coefficient it can be seen that the lowest values are obtained for the PAS staining on the training set and on the test set 1, while the sensitivity shows lower values for the PAS staining only on the test set 1. In particular, low sensitivity values indicate a significant number of false negatives, i.e. cases in which the network does not detect a tissue fold artifact that is present. In this case, the low value of sensitivity for the PAS staining is plausible because considering the low amount of WSI of this type in the dataset (Figure 2.2) and considering that not all the WSIs with this stain in the test set 1 present an artifact of the tissue fold type, it can be deduced that even a small error has a significant impact since it is not compensated by other good predictions.

The specificity values are always close to 1, and this indicates that the network is able to identify very well the absence of tissue fold artifacts.

Through the manual masks and the automatic masks it is possible to make a qualitative comparison to identify the causes of network errors. In particular, it has been noted that in several cases some darker areas of the tissue, typically in correspondence with concentrations of nuclei, are segmented by the network as tissue folds (Figure 3.12), with a negative impact on the DSC value.

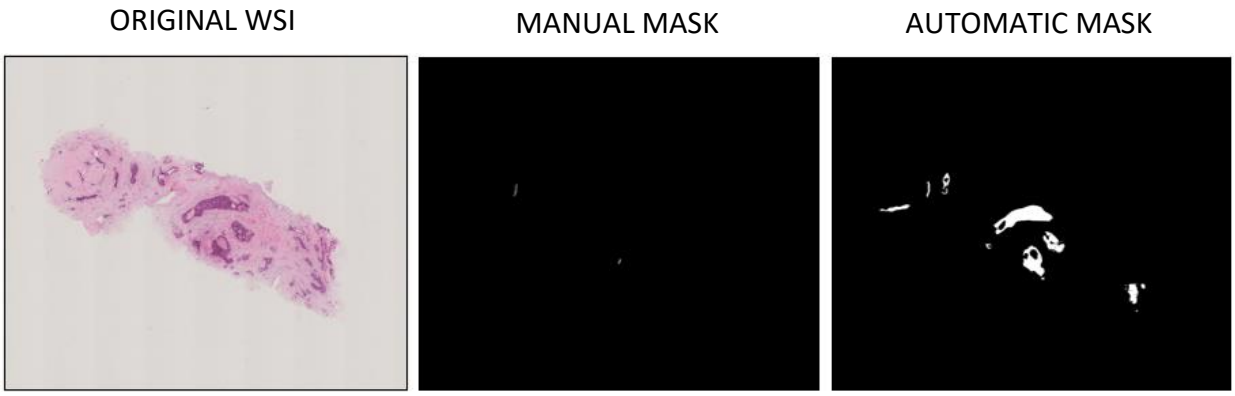


Figure 3.12. The figure shows an incorrect prediction of the network due to nuclei concentrations recognized as tissue folds.

Overall, it can be said that the network has good performance in identifying the tissue fold artifact, it is often able to segment it correctly even if there are some cases in which it falls into error. Furthermore, the fact that there are no major differences between the various subsets indicates that the network is able to generalize well on this task. Figure 3.13 shows an example of correct network prediction on the tissue folds artifact for WSI belonging to test set 1.

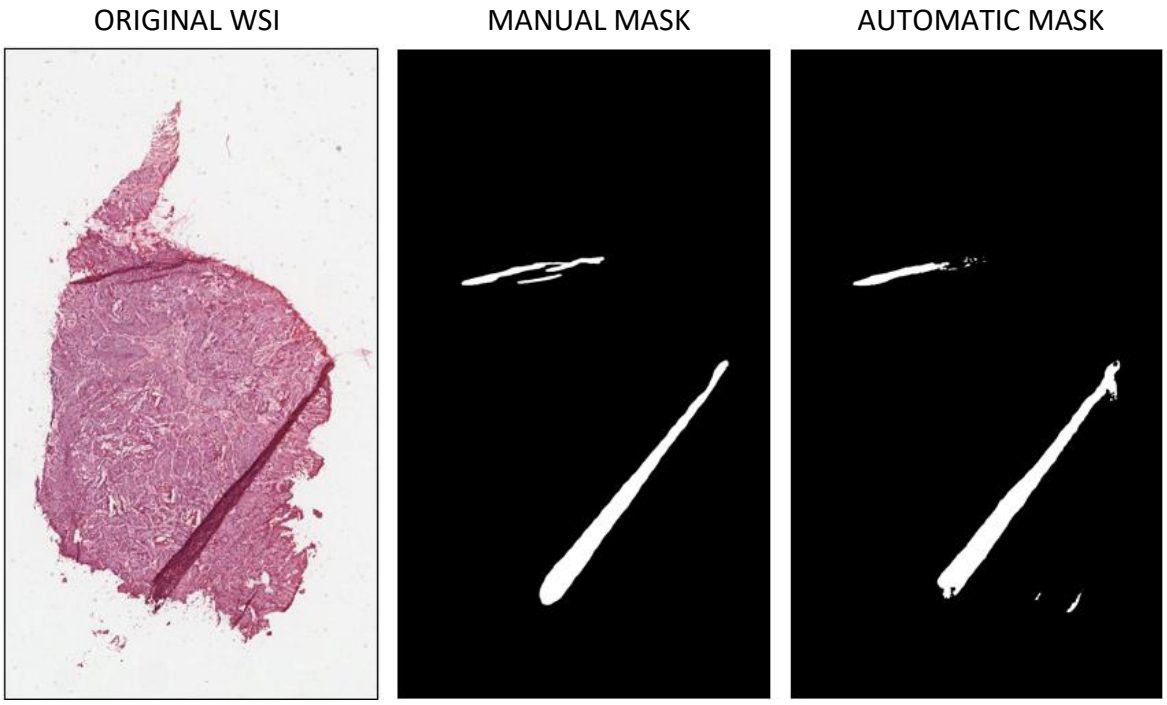
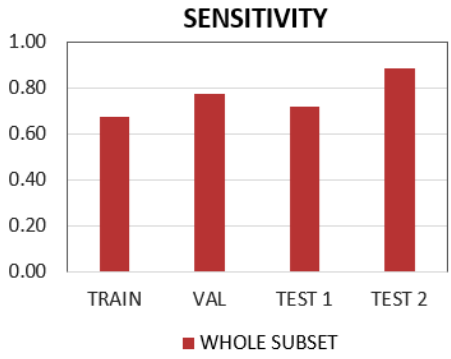


Figure 3.13. Comparison between the manual mask and the automatic mask for a histological slide of test set 1 with tissue fold artifact and HE stain.

Out of focus artifacts

The following figures show the results relating to the sensitivity (Figure 3.14), specificity (Figure 3.15) and DSC (Figure 3.16) values calculated for the out of focus artifact.

	WHOLE SUBSET
TRAINING SET	0.6752
VALIDATION SET	0.7783
TEST SET 1	0.7188
TEST SET 2	0.8872



	HE	IHC	PAS	TRIC
TRAINING SET	0.7432	0.4437	0.9306	0.7400
VALIDATION SET	0.6926	0.8421	0.9869	0.0000
TEST SET 1	0.7261	0.0000	NaN	NaN
TEST SET 2	0.8872	-	-	-

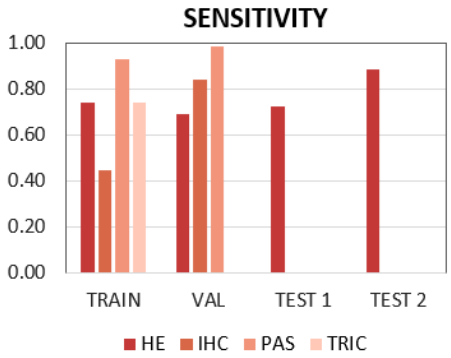
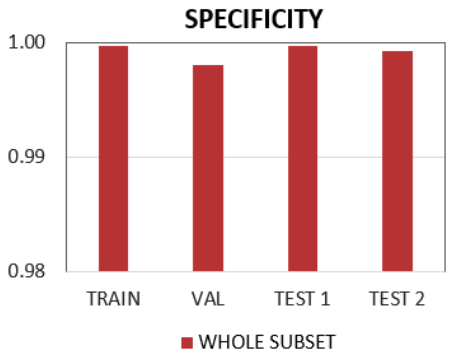


Figure 3.14. Sensitivity results for the artifact network on the out of focus artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.9997
VALIDATION SET	0.9981
TEST SET 1	0.9997
TEST SET 2	0.9993



	HE	IHC	PAS	TRIC
TRAINING SET	0.9998	0.9995	0.9989	0.9999
VALIDATION SET	0.9978	0.9994	0.9994	1.0000
TEST SET 1	0.9997	0.9996	1.000	1.000
TEST SET 2	0.9993	-	-	-

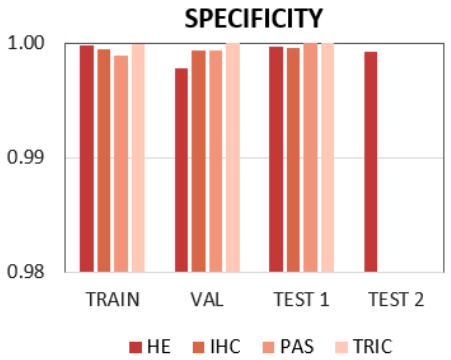
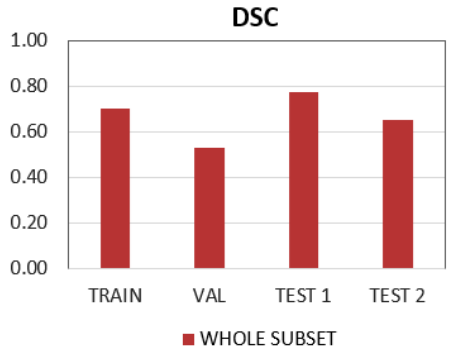


Figure 3.15. Specificity results for the artifact network on the out of focus artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.7032
VALIDATION SET	0.5319
TEST SET 1	0.7781
TEST SET 2	0.6565



	HE	IHC	PAS	TRIC
TRAINING SET	0.7605	0.5303	0.7219	0.6621
VALIDATION SET	0.3900	0.8516	0.9680	0.0000
TEST SET 1	0.7867	0.0000	NaN	NaN
TEST SET 2	0.6565	-	-	-

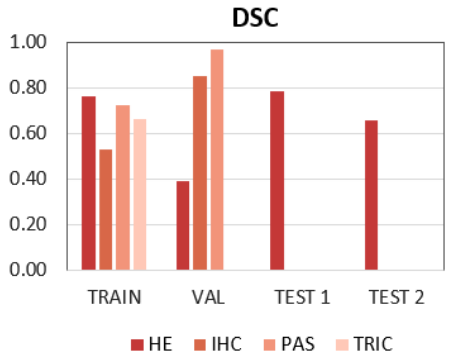


Figure 3.16. DSC results for the artifact network on the out of focus artifact, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

Also for this type of artifact, considering the difficulty of the task, the results are generally good on all the datasets.

Looking at the Dice Similarity Coefficient and sensitivity values, it can be seen that the lowest values are obtained for HE staining on the validation set and for IHC staining on the training set, where there is the lowest overlap between manual mask and automatic mask and the highest number of false negatives, i.e. cases in which the network does not recognize an out of focus artifact. Looking at the specificity values it can be seen that they are all very high, and this indicates that overall the network has good performance in identifying when this artifact is not present; the lowest value, though very high, occurs on the validation set for HE staining, and this indicates that in this case there is a certain number of false positives, i.e. cases in which the network mistakenly identifies an out of focus. Furthermore, it can be noted the presence of sensitivity and DSC values equal to 0 for the TRIC staining on the validation set and for the IHC staining on the test set, to which very high specificity values correspond; in these specific cases the network commits errors in identifying the artifact that are not compensated by other good predictions when calculating the metric considering all the WSIs belonging to that same group, since it must be considered that the number of WSIs with respect to which those values have been calculated is very low (Figure 2.2) and in these cases even small errors can have a great impact on the final result.

Through the manual masks and the automatic masks it is possible to make a qualitative comparison to identify the causes of network errors.
Figure 3.17 shows an error in the prediction of the network committed for the out of focus artifact on a WSI with HE staining with pen marks of a color similar to that of the tissue.

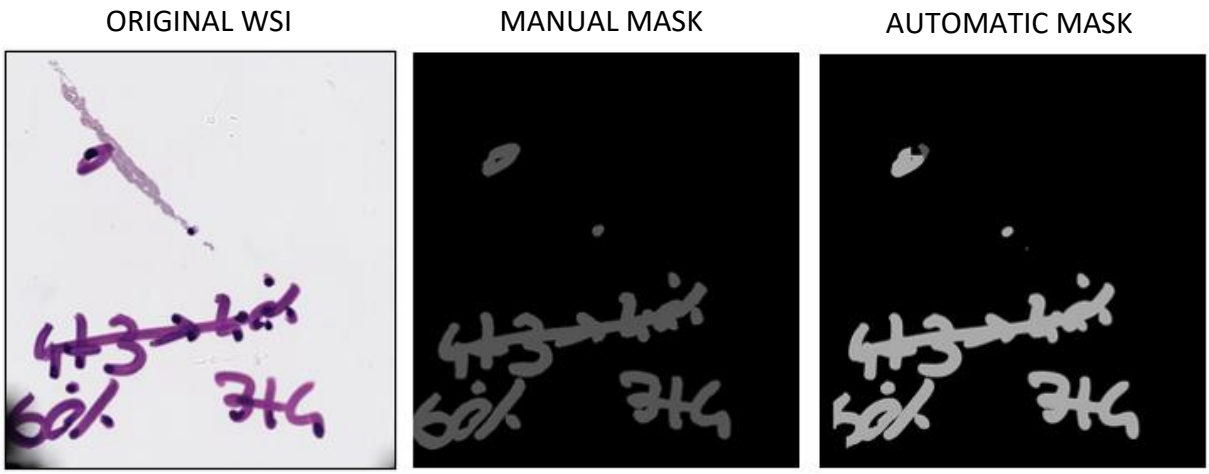


Figure 3.17. The figure shows a WSI with pen marks of a color similar to that of the tissue on which the network fails the prediction.

In this case the network is able to identify the presence of an artifact, but it is wrong on the class since the pen marker, manually annotated as an artifact belonging to the "other" class, is segmented by the network as out of focus. This is probably due to the fact that, in the WSI shown in the figure, the marker has a flat color and similar to that of the tissue, and this causes the network to fall into error. In fact, this does not occur on WSI where the color of the pen marker and the tissue is very different. The cases in which these errors occur are not many, but they have a certain impact on the results since the affected area is quite large compared to that of the generally noted out of focus artifacts.

Figure 3.18 shows a network failure case for this class of artifact on a particular category of images in the dataset. These are some WSI with HE stain in which the area surrounding the tissue tends to a white to yellow color and has a marked discontinuity with respect to the rest of the background which is clearly white. This causes problems with the algorithm and for this reason sometimes on these images the edge is segmented as an artifact even if it is not a blurry region. Cases of this type are not many, but even here, given the extent of the error, they have a significant impact on the results of the metrics.

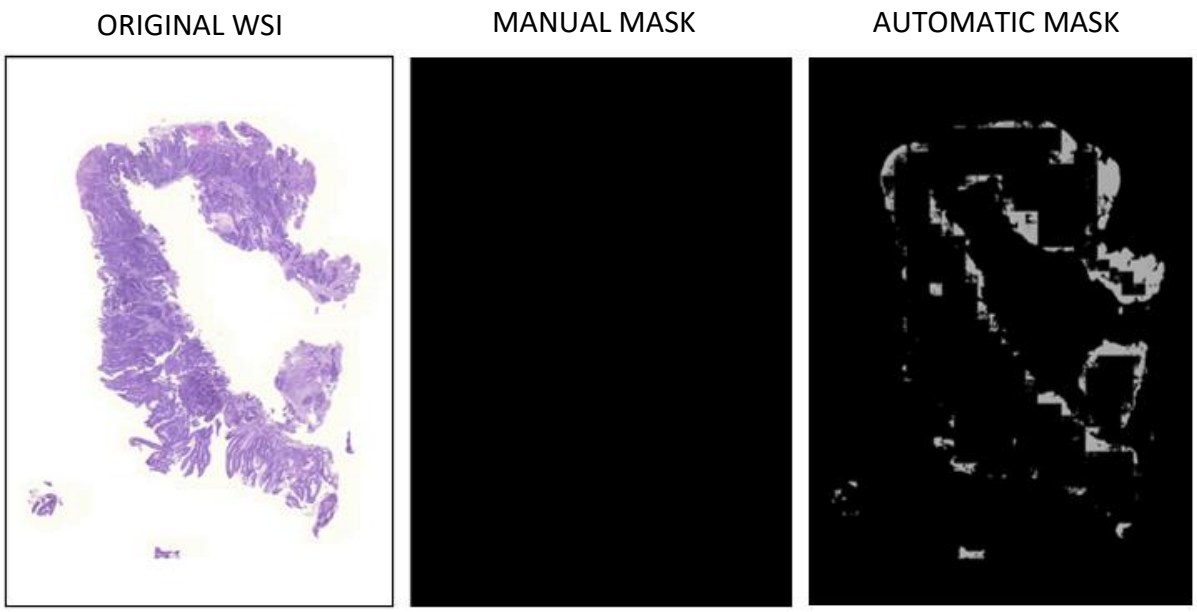


Figure 3.18. The figure shows a particular case of network error on the out of focus artifact.

Another case of network error on the out of focus artifact is the one shown in Figure 3.19. It has been noted that for some HE-stained WSIs where the stained tissue appears very light, the network sometimes makes a mistake and segments the tissue as out of focus even though this type of artifact is not present. This is probably due to the fact that at the zoom level chosen for the images of the artifact dataset, i.e. 5x, WSIs of this type having a particularly light tissue can appear blurry and this causes the network to fall into error.

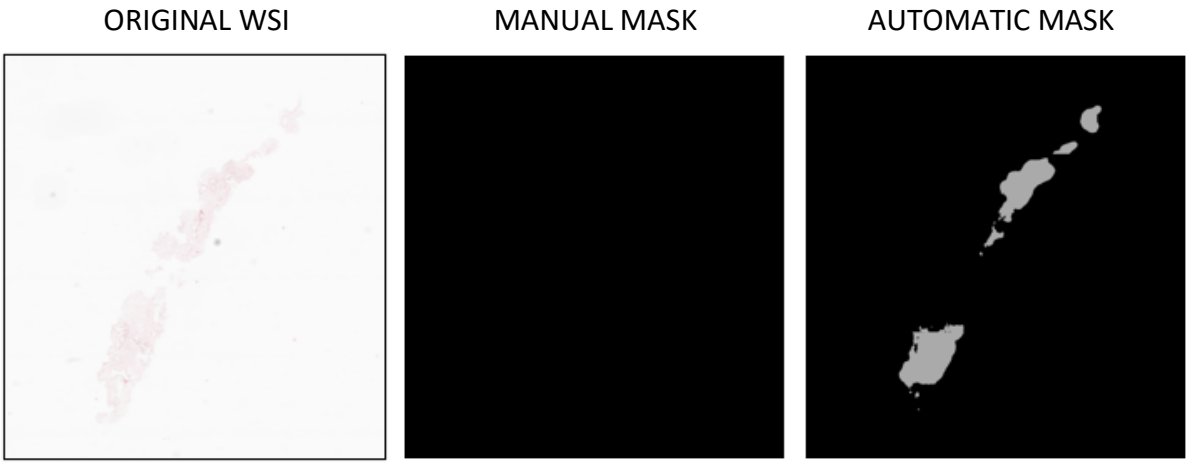


Figure 3.19. The figure shows a network error on the out of focus artifact on a particularly light tissue.

Finally, it has been noted that sometimes the network makes the prediction wrong on WSIs with IHC staining because there are cases in which the tissue, with very light staining, has a particular texture for which it is difficult to distinguish between presence and absence of out of focus even in the manual annotation phase.

Overall, it can be said that the network performs well in identifying out of focus artifacts, especially when the affected areas are particularly blurred. The performances show good values also on the test sets, so it can be said that the network is able to generalize well. Figure 3.20 shows an example of correct network prediction on the out of focus artifact for a WSI belonging to test set 2.

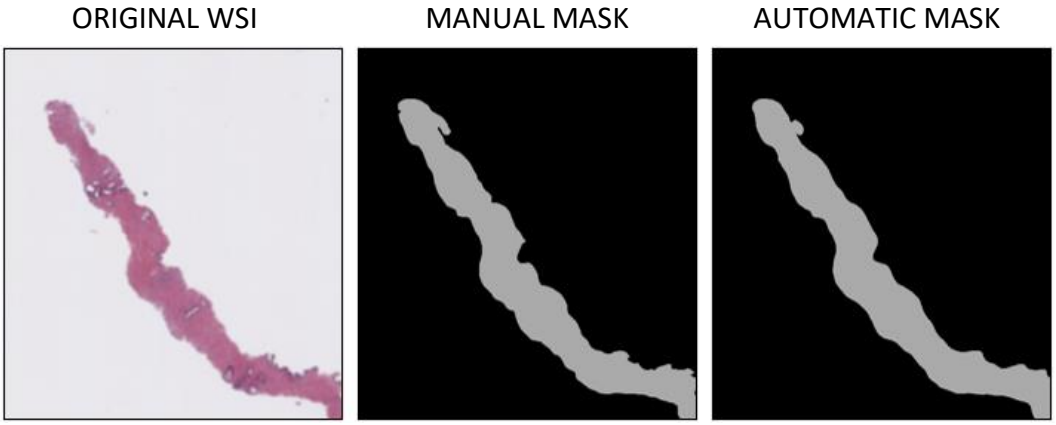
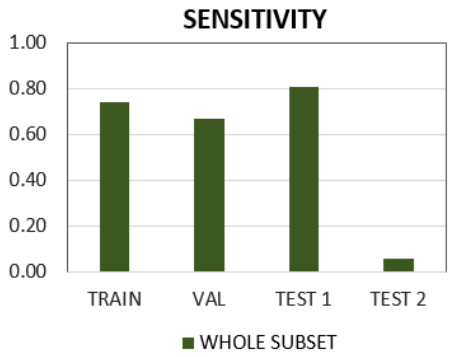


Figure 3.20. Comparison of manual mask and automatic mask for a part of WSI with blurred tissue.

Artifacts in the "other" class

The following figures show the results relating to the sensitivity (Figure 3.21), specificity (Figure 3.22) and DSC (Figure 3.23) values calculated for the artifacts in the "other" class.

	WHOLE SUBSET
TRAINING SET	0.7405
VALIDATION SET	0.6724
TEST SET 1	0.8114
TEST SET 2	0.0584



	HE	IHC	PAS	TRIC
TRAINING SET	0.7574	0.8952	0.5408	0.9052
VALIDATION SET	0.6699	0.7045	0.6623	1.0000
TEST SET 1	0.8007	0.8434	NaN	NaN
TEST SET 2	0.0584	-	-	-

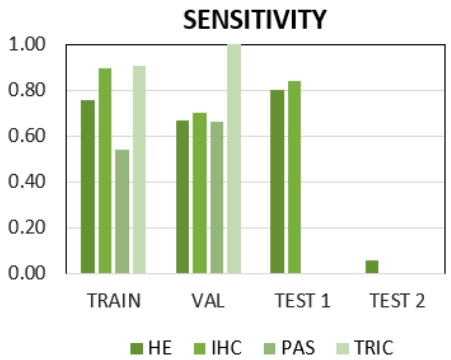
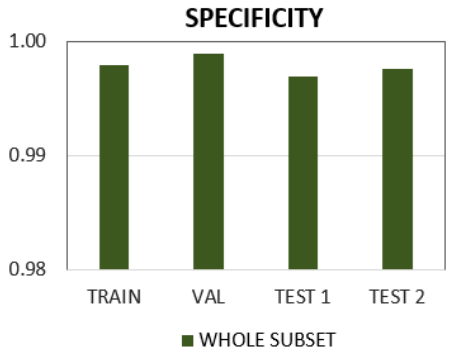


Figure 3.21. Sensitivity results for the artifact network on the artifacts in the "other" class, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.9980
VALIDATION SET	0.9990
TEST SET 1	0.9970
TEST SET 2	0.9976



	HE	IHC	PAS	TRIC
TRAINING SET	0.9983	0.9988	0.9919	0.9962
VALIDATION SET	0.9993	0.9974	0.9979	0.9930
TEST SET 1	0.9969	0.9995	0.9990	1.0000
TEST SET 2	0.9976	-	-	-

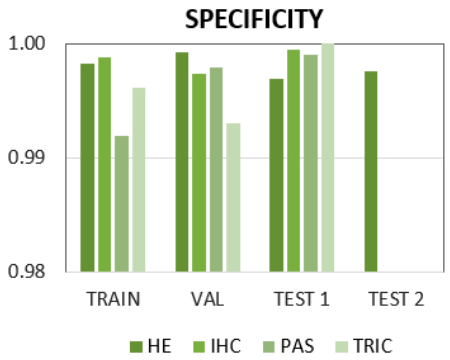
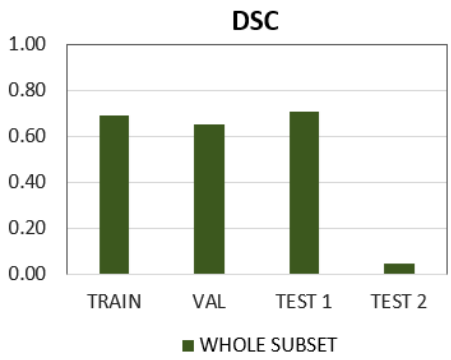


Figure 3.22. Specificity results for the artifact network on the artifacts in the "other" class, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.6953
VALIDATION SET	0.6515
TEST SET 1	0.7091
TEST SET 2	0.0496



	HE	IHC	PAS	TRIC
TRAINING SET	0.7306	0.5897	0.4940	0.5745
VALIDATION SET	0.6998	0.4336	0.0617	0.0519
TEST SET 1	0.6589	0.9065	0.0000	NaN
TEST SET 2	0.0496	-	-	-

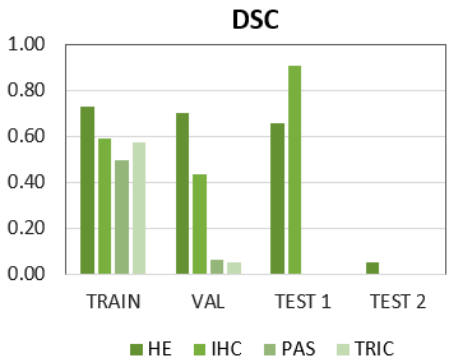


Figure 3.23. DSC results for the artifact network on the artifacts in the "other" class, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

The “other” class of artifacts mainly includes pen markers and air bubbles. The artifact network shows overall good performance on the training set, on the validation set and on the test set 1, while it shows significantly lower values on the test set 2. Particularly low values in DSC terms can be observed on the PAS and TRIC stains for the validation set. Furthermore, observing the values on the PAS staining for the test set 1 it can be deduced that for the WSIs belonging to this group have not been identified false negatives or true positives (NaN sensitivity value), but false positives have been identified (DSC value equal to 0 and a specificity value different from 1), i.e. the network has mistakenly recognized an artifact that is not present. Here too, as for the other artifact classes, this value that may seem anomalous is due to the presence of very few images with this staining for this subset (Figure 2.2), in fact among these there are no artifacts in the "other" category and also small errors made by the network weigh heavily as they are not compensated for by other good predictions. For this same reason there are low values of Dice Similarity Coefficient on the PAS and TRIC colors for the validation set.

By observing the manual masks and the automatic masks it is possible to make a qualitative comparison to identify the main causes of error.
Figure 3.24 shows a recurring network error for the “other” artifact class. It has been noted that in several WSIs the network segments a large part of the background as an artifact and this significantly affects the results of the metrics.

Figure 3.24. The figure shows a recurrent network error on the “other” artifact class caused by the presence of scattered dust on the background.

By zooming in on the original WSIs it can be seen that these errors are due to the presence of artifacts that cannot be annotated manually, for example scattered dust, which make the affected background part not homogenous (Figure 3.25). This creates problems for the network and leads to an incorrect prediction.

Figure 3.25. The figure shows an enlargement on two regions of two different WSIs on which the network makes the prediction wrong. In particular, these are background regions classified by the network as artifacts in the "other" category due to the presence of elements in the background that cannot be annotated manually, which make it non-homogeneous.

Apart from this type of recurring error, it can be said that the network performs well in identifying annotated artifacts in the “other” category and also, considering only test set 1, it has a good generalizability. Figure 3.26 shows the comparison between the manual mask and the automatic mask for a WSI of the training set with bubbles and pen marks.

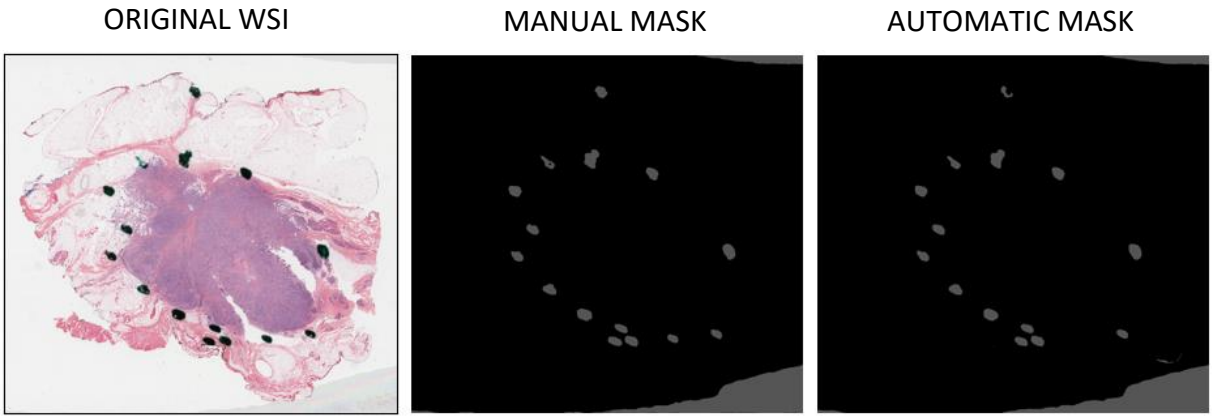
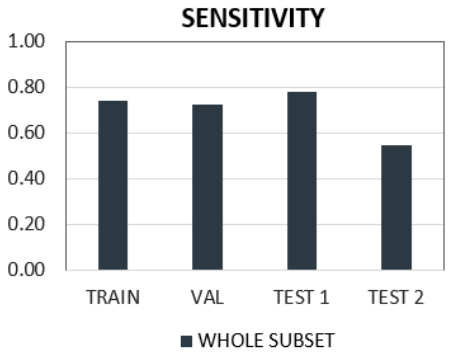


Figure 3.26. Comparison between the manual mask and the automatic mask for a WSI with artifacts in the “other” category, in particular bubbles and pen marks.

All artifact classes

The following figures show the results of sensitivity (Figure 3.27), specificity (Figure 3.28) and DSC (Figure 3.29) calculated on the binary mask containing all artifacts.

	WHOLE SUBSET
TRAINING SET	0.7426
VALIDATION SET	0.7253
TEST SET 1	0.7825
TEST SET 2	0.5485



	HE	IHC	PAS	TRIC
TRAINING SET	0.7649	0.6293	0.6273	0.8904
VALIDATION SET	0.6910	0.7964	0.9749	0.9068
TEST SET 1	0.7720	0.8313	0.3232	NaN
TEST SET 2	0.5485	-	-	-

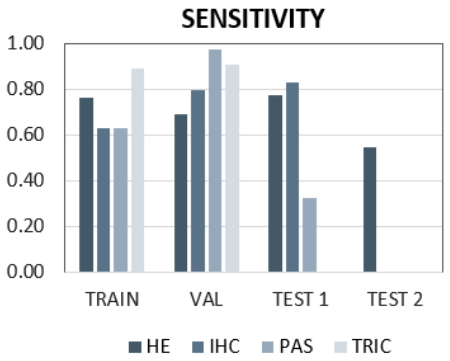
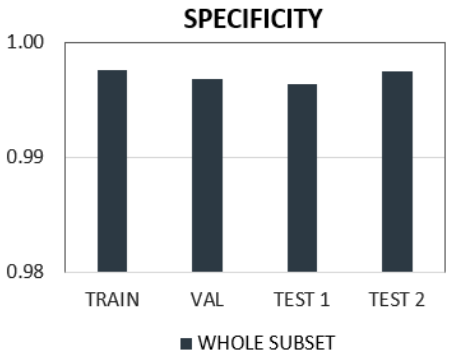


Figure 3.27. Sensitivity results for the artifact network on all artifact classes, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.9976
VALIDATION SET	0.9969
TEST SET 1	0.9964
TEST SET 2	0.9975



	HE	IHC	PAS	TRIC
TRAINING SET	0.9979	0.9980	0.9908	0.9960
VALIDATION SET	0.9970	0.9962	0.9970	0.9929
TEST SET 1	0.9962	0.9988	0.9989	1.0000
TEST SET 2	0.9975	-	-	-

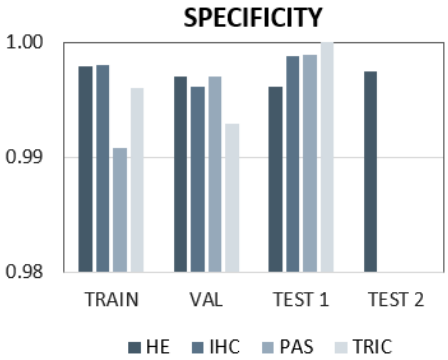
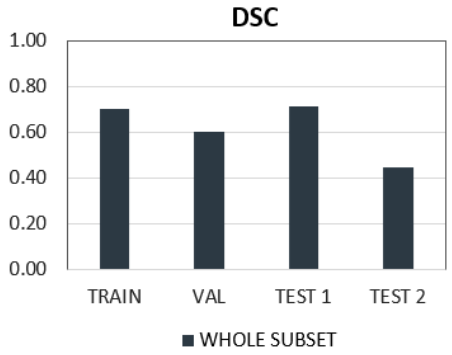


Figure 3.28. Specificity results for the artifact network on all artifact classes, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

	WHOLE SUBSET
TRAINING SET	0.7042
VALIDATION SET	0.6057
TEST SET 1	0.7175
TEST SET 2	0.4461



	HE	IHC	PAS	TRIC
TRAINING SET	0.7373	0.6032	0.5515	0.5770
VALIDATION SET	0.5789	0.6851	0.8831	0.1032
TEST SET 1	0.6870	0.8880	0.0212	NaN
TEST SET 2	0.4461	-	-	-

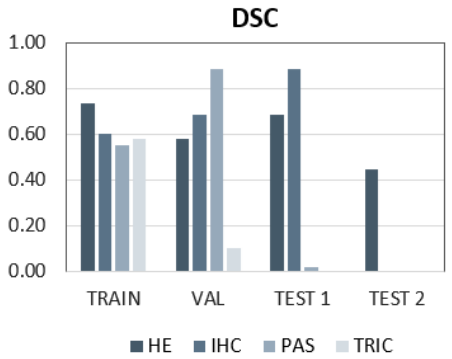


Figure 3.29. DSC results for the artifact network on all artifact classes, reported (a) divided by subset and (b) divided by subset and by stain for each subset.

The second evaluation that was made on the artifact network is aimed at establishing the capacity of the network in distinguishing the presence or absence of a generic artifact on the histological slide. To do this, the network predictions for the various artifact classes were merged into a single binary mask and compared with the manual mask. Here too, considering the difficulty of the task, the results appear very good. Particularly low values of the DSC metric can be noted on the PAS stain for the test set 1 and on the TRIC stain for the validation set, but also here it can be considered that due to the low number of histological slides belonging to these groups present in the dataset (Figure 2.2), of which not all have artifacts, even small errors have a great impact on the final result of the metric if evaluated on these individual groups, as they are not compensated by other good predictions.

Making a general assessment, it can be said that some of the problems previously seen on the individual classes of artefact remain, i.e. darker areas of the tissue segmented as tissue folds (Figure 3.12), particular cases in which the tissue area is mistakenly identified as out of focus (Figures 3.18 and 3.19), and background areas identified as artifacts in the "other" category due to the presence in the background of particular artifacts that cannot be annotated manually (Figure 3.24). However, there are no more network errors due to the classification of a certain type of artifact in a different class of artifact, as in the case of pen markers of a similar color to

the tissue identified as out of focus rather than as artifacts in the "other" category (Figure 3.17), since in this case what is evaluated is the ability of the network to identify an artifact, regardless of the category it belongs to. Figure 3.30 shows an example of correct segmentation of an artifact on which the network had previously shown an error relating to the identified artifact category.

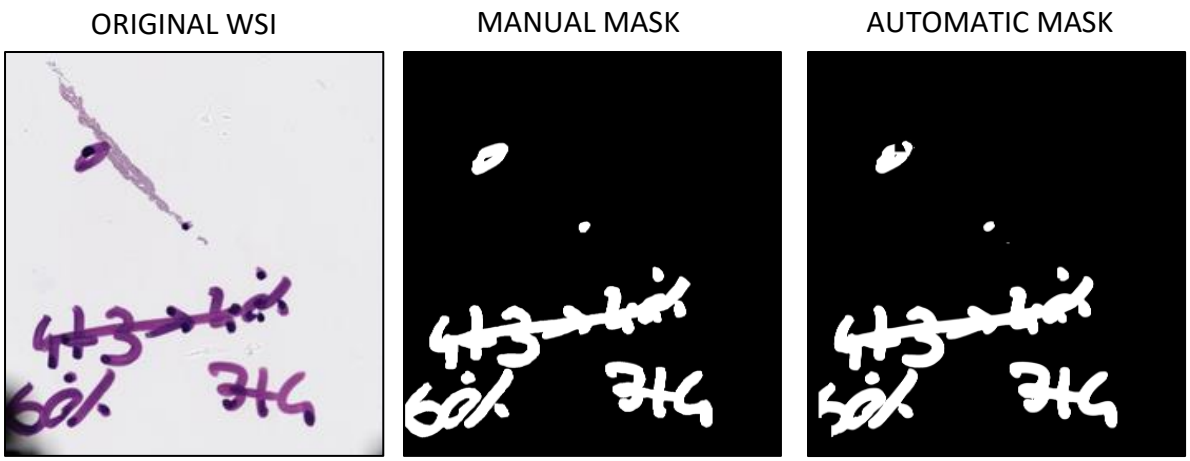


Figure 3.30. Comparison between the manual mask and the automatic mask for a WSI with artifacts.

It can therefore be said that the network is able to distinguish well between the presence and absence of an artifact and shows good generalizability.

Chapter 4

Conclusions and further developments

Looking at the main issues arising in the context of digital pathology, the goal of this thesis was to develop a fully automatic algorithm for the segmentation of tissues and artifacts on digitized histological slides. To achieve this goal, two neural networks based on deep-learning have been developed: the first is based on a binary classification that allows to distinguish the tissue from the background, the second is based on a multi-class classification that allows to identify different categories of artifacts.

The tissue network has shown a very high ability to recognize tissues related to different organs and different stains, even on images never seen before during the training phase. The few mistakes made are mostly due to artifacts, therefore by crossing the information deriving from the segmentation of the artifact network it is possible to eliminate these errors and further increase the performance on tissue segmentation.

Unlike tissues, the artifacts cover a much smaller area and are much less represented since not all of the WSIs included in the dataset have artifacts; moreover, some artifacts are difficult to clearly recognize even for the manual operator. Consequently, considering the difficulty of the task, it can be said that even the artifact network has shown very good performance, with a high generalizability.

Through the qualitative comparison it has been seen that mainly the artifact network commits plausible errors, the cause of which can be identified. In addition, the errors relating to artifacts in the "other" category, which are those on which the network has shown lower performance, are mainly located in the background, where presumably they are less of a hindrance for the analysis of the histological slide.

In conclusion, this work proves useful for a double aspect.

On the one hand, it can be a useful support tool for the pathologist to be introduced in the daily workflow to perform quality controls on histological slides, thus reducing the workload. Evaluating the quality of the slides at the beginning of the working day can greatly speed up the workflow, since the histological slides to be reacquired are immediately identified, thus avoiding possible delays in the processing of a diagnosis.

On the other hand, identifying low-quality WSIs can be useful in the development and testing of new automatic algorithms which, for example, must identify pathological regions of the tissue. In fact, using low quality WSIs the prediction can be negatively influenced by the presence of artifacts that alter the appearance of the tissue. By identifying these images and eliminating them from the dataset, it is possible to have more robust and faster algorithms and it is possible to obtain better performance in the inference phase, without having results distorted by the presence of artifacts.

As for further developments, it is possible to work on various aspects to further improve the performance of this automatic algorithm. The dataset consists of WSIs related to different stains and different organs, of which some are more represented than others. By increasing the number of images on which the network has been trained little due to the reduced number, it is possible to increase the performance of the network. In addition to this, optimization procedures can be performed on the images already present in the dataset in the pre-processing phase, i.e. before they are given as input to the network. For example, being a variegated dataset, the background of the histological slide sometimes appears to have slightly different shades of white between one WSI and the other, and this is due to the fact that the colors are unbalanced, so an optimization procedure aimed at balancing the colors could improve performance. Furthermore, the performance of the network could be evaluated using different magnifications to extract the WSIs from the pyramidal images.

In addition to this, from a clinical point of view it is possible to structure a qualitative score that allows to assign a value to the WSI on the basis of its quality. To do this, it is necessary to have evaluations by the pathologists to understand which artifacts negatively impact the diagnosis in terms of artifact area and localization on the histological slide, because there may be artifacts that, although present, are not an obstacle for the pathologist. From this point of view, taking some WSIs on which the network generated the predictions and asking pathologists to assign a score to the histological slide on the basis of its quality, it would be possible to collect useful data to carry out a correlation study that allows to cross the information deriving from pathologists with predictions on artifacts from the algorithm developed in order to create a quality index that reflects the evaluation made by the pathologist.

In the end, the training times of this model are currently high, so solutions could be sought to reduce them. For example, if the artifacts in the background are not clinically relevant, the background patches could not be included among the elements that are input to the network, so as to reduce the computational weight of the dataset and train the network to identify artifacts only on areas of the WSI where there is tissue.

Bibliography

[1] S. Nam *et al.*, "Introduction to digital pathology and computer-aided pathology," *J Pathol Transl Med*, vol. 54, no. 2, pp. 125–134, Mar. 2020, doi: 10.4132/jptm.2019.12.31.

[2] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, "Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives," *J Pathol Inform*, vol. 9, no. 1, p. 40, Jan. 2018, doi: 10.4103/jpi.jpi_69_18.

[3] V. Baxi, R. Edwards, M. Montalto, and S. Saha, "Digital pathology and artificial intelligence in translational medicine and clinical practice," *Modern Pathology*, vol. 35, no. 1, pp. 23–32, Jan. 2022, doi: 10.1038/s41379-021-00919-2.

[4] M. D. Zarella *et al.*, "A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association," *Arch Pathol Lab Med*, vol. 143, no. 2, pp. 222–234, Feb. 2019, doi: 10.5858/arpa.2018-0343-RA.

[5] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen, "A Multiscale Approach for Whole-Slide Image Segmentation of five Tissue Classes in Urothelial Carcinoma Slides," *Technol Cancer Res Treat*, vol. 19, p. 153303382094678, Jan. 2020, doi: 10.1177/1533033820946787.

[6] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Comput Biol Med*, vol. 128, p. 104129, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104129.

[7] R. Baxter, "Interpretation of histological sections: Stains used in histology," *Kenhub*, Jul. 07, 2022. https://www.kenhub.com/en/library/anatomy/interpretation-of-histologic-sections-stains-used-in-histology (accessed Jul. 29, 2022).

[8] C. Sampias and G. Rolls, "H&E Staining Overview: A Guide to Best Practices," *Leica biosystems*. https://www.leicabiosystems.com/knowledge-pathway/he-staining-overview-a-guide-to-best-practices/ (accessed Jul. 29, 2022).

[9] G. L. Kumar *et al.*, *Special Stains and H & E*, Second. Carpinteria, California: Dako North America, 2010.

[10] M. Wołuń-Cholewa, K. Szymanowski, M. Andrusiewicz, A. Szczerba, and J. B. Warchoł, "Trichrome Mallory's stain may indicate differential rates of RNA synthesis in eutopic and ectopic endometrium.," *Folia Histochem Cytobiol*, vol. 48, no. 1, Jun. 2010, doi: 10.2478/v10042-008-0106-4.

[11] "Gomori's trichrome," *CliniSciences*. https://www.clinisciences.com/en/buy/cat-gomori-s-trichrome-3956.html (accessed Jul. 29, 2022).

[12] “Masson-Goldner’s trichrome Stain,” *BioTrend*. https://www.biotrend.com/kauf/cat-masson-goldner-s-trichrome-stain-5383.html (accessed Jul. 29, 2022).

[13] S. Jackson, D. Grabis, and C. Manav, “Giemsa: The Universal Diagnostic Stain,” *MilliporeSigma*, Billerica, MA, Sep. 2018.

[14] D. Giri, “Wright’s Stain : Preparation, Principle, Procedure and Results,” *Laboratory tests*, Jul. 25, 2019. https://laboratorytests.org/wrights-stain/ (accessed Aug. 03, 2022).

[15] N. Parry, “Verhoeff-van Gieson Stain: A Special Histology Stain for Elastic Fibers,” *BiteSize Bio*, Apr. 15, 2014. https://bitesizebio.com/19952/verhoeff-van-gieson-stain-a-special-histology-stain-for-elastic-fibers/ (accessed Aug. 03, 2022).

[16] J. Anderson, G. Rolls, and S. Westra, “Immunohistochemistry: An Overview + Steps to Better IHC Staining,” *Leica biosystems*. https://www.leicabiosystems.com/knowledge-pathway/immunohistochemistry-an-overview-steps-to-better-ihc-staining/ (accessed Jul. 29, 2022).

[17] J. A. Ramos-Vara, “Technical Aspects of Immunohistochemistry,” *Vet Pathol*, vol. 42, no. 4, pp. 405–426, Jul. 2005, doi: 10.1354/vp.42-4-405.

[18] “Explore the expression profiles of human cancers,” *The human protein atlas*. https://www.proteinatlas.org/humanproteome/pathology (accessed Aug. 03, 2022).

[19] G. O. Rolls, N. J. Farmer, and J. B. Hall, *Artifacts in Histological and Cytological Preparations*. Leica Microsystems, 2008.

[20] V. Rastogi, “Artefacts: A Diagnostic Dilemma – A Review,” *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*, 2013, doi: 10.7860/JCDR/2013/6170.3541.

[21] B. Schömig-Markiefka *et al.*, “Quality control stress test for deep learning-based diagnostic model in digital pathology,” *Modern Pathology*, vol. 34, no. 12, pp. 2098–2108, Dec. 2021, doi: 10.1038/s41379-021-00859-x.

[22] J. Jiang *et al.*, “Image-to-image translation for automatic ink removal in whole slide images,” *Journal of Medical Imaging*, vol. 7, no. 05, Oct. 2020, doi: 10.1117/1.JMI.7.5.057502.

[23] A. Janowczyk, R. Zuo, H. Gilmore, M. Feldman, and A. Madabhushi, “HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides,” *JCO Clin Cancer Inform*, no. 3, pp. 1–7, Dec. 2019, doi: 10.1200/CCI.18.00157.

[24] Y. Chen *et al.*, “Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies,” *J Pathol*, vol. 253, no. 3, pp. 268–278, Mar. 2021, doi: 10.1002/path.5590.

[25] H. M. Shakhawat, T. Nakamura, F. Kimura, Y. Yagi, and M. Yamaguchi, “[Paper] Automatic Quality Evaluation of Whole Slide Images for the Practical Use of Whole Slide Imaging Scanner,” *ITE Transactions on Media Technology and Applications*, vol. 8, no. 4, pp. 252–268, 2020, doi: 10.3169/mta.8.252.

[26] N. Hashimoto, P. A. Bautista, M. Yamaguchi, N. Ohyama, and Y. Yagi, “Referenceless image quality evaluation for whole slide imaging,” *J Pathol Inform*, vol. 3, no. 1, p. 9, Jan. 2012, doi: 10.4103/2153-3539.93891.

[27] G. Smit, F. Ciompi, M. Cigéhn, A. Bodén, J. van der Laak, and C. Mercan, “Quality control of whole-slide images through multi-class semantic segmentation of artifacts,” in *Medical Imaging with Deep Learning*, 2021.

[28] M. Haghighat *et al.*, “Automated quality assessment of large digitised histology cohorts by artificial intelligence,” *Sci Rep*, vol. 12, no. 1, p. 5002, Dec. 2022, doi: 10.1038/s41598-022-08351-5.

[29] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Nov. 2015.

[30] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[31] S. Das, A. A. Fime, N. Siddique, and M. M. A. Hashem, “Estimation of Road Boundary for Intelligent Vehicles Based on DeepLabV3+ Architecture,” *IEEE Access*, vol. 9, pp. 121060–121075, 2021, doi: 10.1109/ACCESS.2021.3107353.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” Feb. 2018.

[33] H. Zhang *et al.*, “ResNeSt: Split-Attention Networks,” Apr. 2020.

[34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene Parsing through ADE20K Dataset,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5122–5130. doi: 10.1109/CVPR.2017.544.

[35] S. Gould, R. Fulton, and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 1–8. doi: 10.1109/ICCV.2009.5459211.

[36] A. Shrivastava, A. Gupta, and R. Girshick, “Training Region-based Object Detectors with Online Hard Example Mining,” Apr. 2016.

[37] M. Berman, A. R. Triki, and M. B. Blaschko, “The Lovàsz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks,” May 2017.

[38] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” Jun. 2016.

[39] “Masson-Goldner Trichrome Staining Kit,” *Diapath*. Masson-Goldner Trichrome Staining Kit (accessed Jul. 29, 2022).

[40] “Mallory Trichrome - kit,” *Histo-Line*. https://www.histoline.com/en/01mt100t (accessed Jul. 29, 2022).

[41] G. Smit, C. Mercan, and F. Ciompi, “Artifact detection in digitized histopathology images,” *Computational Pathology Group*, Feb. 2021. https://www.computationalpathologygroup.eu/projects/artifact_detection/ (accessed Aug. 03, 2022).

[42] M. Grandini, E. Bagli, and G. Visani, “Metrics for Multi-Class Classification: an Overview,” Aug. 2020.