



**Politecnico  
di Torino**

# POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Biomedica

A.A. 2020/2021

Sessione di Laurea dicembre 2021

## **Sviluppo di un tool per la normalizzazione della colorazione di biopsie istologiche digitalizzate**



Relatore:

Prof. Filippo Molinari

Candidato:

267207

Correlatore:

Ing. Massimo Salvi

Ramona Polcari

# Indice

<b>ELENCO DELLE FIGURE .....</b>	<b>3</b>
<b>ABSTRACT .....</b>	<b>6</b>
<b>CAPITOLO 1 .....</b>	<b>8</b>
<b>Introduzione .....</b>	<b>8</b>
1.1 Patologia digitale e WSI imaging systems.....	8
1.1.1 Preparazione dei vetrini .....	8
1.1.2 Colorazioni.....	10
1.2 Deep learning in patologia digitale .....	11
1.2.1 Algoritmi di pre-processing .....	13
1.2.2 Individuazione del tessuto e degli artefatti .....	13
1.2.3 Normalizzazione degli stain.....	14
1.2.4 Selezione dei patch significativi .....	18
1.3 Stain normalization: approcci utilizzati.....	18
1.3.1 Algoritmi di Color Deconvolution.....	20
1.4 Pipeline di Stain Normalization.....	25
<b>CAPITOLO 2 .....</b>	<b>27</b>
<b>Materiali e metodi.....</b>	<b>27</b>
2.1 Datasets .....	28
2.2 Tissue Detection.....	32
2.2.1 Architettura degli algoritmi.....	34
2.2.2 Primo algoritmo: Tissue_normalization .....	36
2.2.3 Secondo algoritmo: Tissue_puzzle .....	43
2.3 Puzzle Image .....	48
2.4 Stain normalization .....	51
2.4.1 Algoritmo di Macencko .....	51
2.4.2 Algoritmo di normalizzazione dell'intera WSI.....	54
2.4.3 Selezione delle immagini target.....	55
2.4.4 Sliding window e normalizzazione della finestra corrente .....	56

2.5 Validazione degli algoritmi.....	60
2.5.1 Metriche per tissue detection.....	60
2.5.2 Metriche per la normalizzazione degli stain .....	62
<b>CAPITOLO 3 .....</b>	<b>65</b>
<b>Risultati.....</b>	<b>65</b>
3.1 Validazione della tissue detection .....	65
3.2 Validazione della stain normalization.....	75
<b>CAPITOLO 4 .....</b>	<b>86</b>
<b>Conclusioni.....</b>	<b>86</b>
<b>BIBLIOGRAFIA .....</b>	<b>88</b>

# Elenco delle figure

Figura 1- Protocollo da applicare al campione per ottenere la relativa immagine digitale.....	10
Figura 2- Principali colorazioni oggetto dello studio.....	11
Figura 3- Architetture tipiche di framework di deep learning.....	12
Figura 4- Matrice degli stain per un'immagine Ematossilina- Eosina .	15
Figura 5- Strategie di normalizzazione degli stain.....	17
Figura 6- Distribuzione delle concentrazioni degli stain per la colorazione Ematossilina (Viola) - Eosina (Rosa chiaro) nello spazio RGB e nello spazio OD.....	17
Figura 7- Segmentazione guidata per l'estrazione delle patches significative. ....	18
Figura 8-Esempio di color deconvolution.....	20
Figura 9 – Flowchart del metodo di stain normalization proposto da Zheng et al.....	22
Figura 10-Risultati degli algoritmi di stain normalization applicati a diverse tipologie di tessuto.....	24
Figura 11- Plot dei valori di rSE per i metodi citati in funzione dei tempi computazionali per la normalizzazione .....	25
Figura 12- Pipeline di tissue detection e di stain normalization. ....	26
Figura 13- Biopsia tricromica di rene. ....	27
Figura 14-Visualizzazione della biopsia di mammella (325 - 2021-03-14 16.25.11. ndpi) con il software NDP.view. ....	31
Figura 15-Visualizzazione della biopsia di una mammella con il software Aperio Image Scope (142I17 H&E.svs).....	31
Figura 16- Formato piramidale della Whole Slide Image (WSI) .....	32
Figura 17- Segmentazione manuale di tutto il tessuto presente in una biopsia di mammella (2016I005002.ndpi) in colorazione immunoistochimica.....	33
Figura 18- Segmentazione manuale del tessuto di interesse istologico in una biopsia di mammella (2016I005002.ndpi) in colorazione immunoistochimica, nel riquadro rosso si evidenzia una parte di tessuto sfocata.....	34
Figura 19- Segmentazione ridefinita della biopsia di mammella di fig. 14. ....	34

Figura 20 – Schema generale dell’algoritmo tissue_normalization.....	35
Figura 21 - Schema generale dell’algoritmo tissue_puzzle.....	36
Figura 22 - Segmentazione di tessuto effettuata su biopsia tricromica di rene (TRIC 1) con un thresholding globale.....	38
Figura 23 – Maschera degli artefatti .....	39
Figura 24 – Immagine originale e maschera finale per una biopsia di mammella in colorazione HE (11825). .....	42
Figura 25 - Biopsia di mammella in colorazione HE (_1006649), maschera ottenuta dalla combinazione di filtro laplaciano e gaussiano. ....	43
Figura 26 - a) WSI originale di una biopsia di mammella in colorazione HE, relative maschere ed output finale. ....	43
Figura 27- Filtro bilaterale. ....	45
Figura 28 - Segmentazione di una biopsia PAS di polmone con i due algoritmi, ISSUE_NORMALIZATION e TISSUE_PUZZLE.....	46
Figura 29 - Risultati degli algoritmi di tissue detection.....	48
Figura 30- Schema dell'algoritmo di puzzle image.....	49
Figura 31 – Patches di tessuto (biopsia di rene PAS 467). ....	50
Figura 32 – Creazione della puzzle image per la biopsia di mammella 1173.....	51
Figura 33 – Scomposizione SVD.....	53
Figura 34- Puzzle image e stain. ....	53
Figura 35 - Flowchart dell'algoritmo di normalizzazione. ....	55
Figura 36 - (1) Puzzle images ricavate da una biopsia di mammella in colorazione HE ed una di polmone in colorazione PAS.....	57
Figura 37 – Individuazione della griglia sulla quale scorrerà la sliding window.....	58
Figura 38 – Zoom della finestra evidenziata in giallo in figura 37.....	58
Figura 39 - Biopsia di rene (PAS 31).....	60
Figura 40- Confronto fra le maschere ottenute dai vari algoritmi.....	69
Figura 41- WSI di una biopsia di mammella in colorazione immunostochimica (2016I004557;A;1;3 - 2020-11-27 16.13.46) e relative maschere di tessuto. ....	70

Figura 42- Sovrapposizione della maschera automatica ottenuta dall'algorithmo tissue normalization sovrapposta alla maschera segmentata manualmente per una biopsia di rene (PAS 166).....	70
Figura 43- Medie e deviazioni standard dei valori di recall.....	71
Figura 44 - Medie e deviazioni standard del valore di precision.....	72
Figura 45-Medie e deviazioni standard dei valori di F1-score. ....	72
Figura 46 - Medie e deviazioni standard dei valori del Jaccard Index.....	73
Figura 47- Scelta puzzle target. ....	74
Figura 48- WSI originale di una biopsia di prostata in colorazione HE (69.18.B1), puzzle images ed analisi degli stain. ....	75
Figura 49- Dettaglio dell'immagine 48a, WSI originale, dettaglio dell'immagine 48h, WSI normalizzata. ....	75
Figura 50- WSI originale di una biopsia di rene in colorazione PAS (1004761), puzzle images ed analisi degli stain ....	76
Figura 51- Dettaglio dell'immagine 50a, WSI originale e dettaglio dell'immagine 50h, WSI normalizzata. ....	76
Figura 52- Valori di Chi-Square per le biopsie renali. ....	78
Figura 53 - Valori della distanza di Bhattacharyya per biopsie di rene.....	79
Figura 54- Valori dell'indice di correlazione per biopsie di rene.....	79
Figura 55- Valori di Chi-Square per le biopsie di prostata.....	80
Figura 56- Valori della distanza di Bhattacharyya per biopsie di prostata.....	80
Figura 57- Valori dell'indice di correlazione per biopsie di prostata. ....	80
Figura 58 - Valori di Chi-Square per le biopsie di mammella. ....	81
Figura 59- Valori della distanza di Bhattacharyya per biopsie di mammella.....	81
Figura 60 – Valori dell'indice di correlazione per biopsie di mammella.....	82

# Abstract

L'introduzione commerciale degli scanner per imaging da vetrino istologico completo (WSI) permette di ottenere un rendering digitale della biopsia e di visualizzare le strutture ad altissime risoluzioni, proporzionalmente alla risoluzione selezionata aumenta il peso computazionale dell'output finale.

Le fasi di preparazione ed assemblaggio di un vetrino istologico sono completamente manuali, il posizionamento della cover può generare artefatti a causa della presenza di polvere, bolle d'aria o contaminazioni di altri microrganismi mentre la digitalizzazione può imporre delle intensità di colori diverse a seconda dello scanner utilizzato, della luce circostante e della tecnologia di utilizzo. Se il campione non è perfettamente allineato con il piano focale dello scanner l'immagine può risultare addirittura sfocata.

L'analisi e il processing di una WSI in formato nativo sono dunque dei processi computazionalmente onerosi, l'obiettivo di questo lavoro di tesi è sviluppare un'intera pipeline di normalizzazione degli stain dell'immagine che mantenga elevata la qualità dell'output riducendo i tempi computazionali. Le WSI digitalizzate in analisi sono state ricavate da biopsie di diversi tessuti in diverse tipologie di colorazione. La procedura di normalizzazione degli stain implica la trasformazione di un'immagine  $I$  in un'immagine normalizzata  $I_{NORM}$ .

L'immagine finale è morfologicamente identica a quella di input a meno dei colori che vengono normalizzati sulla base di un'immagine template scelta accuratamente in una fase preliminare.

Dopo un'attenta analisi delle diverse metodologie utilizzate per effettuare la normalizzazione degli stain di un'immagine, il metodo selezionato è la color deconvolution di Macencko che offre un ottimo trade off tra stabilità dei risultati e tempi computazionali. I tool sviluppati si soffermano, oltre che sulla normalizzazione, sull'implementazione di una serie algoritmi di pre e post-processing che migliorano la qualità della normalizzazione riducendo i tempi di

processing. In primis si individuano solo le porzioni di tessuto presenti, ottenendo delle maschere che permettono di escludere gli artefatti e lo sfondo.

Per rendere più veloce e performante la procedura di normalizzazione dell'intero vetrino viene implementato un algoritmo che permette di estrarre delle patches quadrate (di dimensioni predefinite) significative e rappresentative di ogni sezione di tessuto. Queste patches vengono assemblate insieme a formare un'immagine rappresentativa degli stain della WSI, definita "puzzle image".

Le puzzle image vengono estratte per ogni immagine del dataset e per ogni cartella viene selezionata una puzzle image che verrà utilizzata come target per la normalizzazione.

La puzzle image target è scelta in base alla bontà della stain separation sulla stessa. La normalizzazione avviene scorrendo la WSI originale con una sliding window ed ogni finestra viene normalizzata utilizzando le matrici di normalizzazione calcolate sulla puzzle image scelta come target e su quella relativa alla WSI in analisi.

Questa tecnica permette di non caricare l'intera WSI in memoria per poi normalizzarla ma normalizza la singola finestra ricomponendo poi l'immagine normalizzata che avrà la stessa struttura morfologica dell'immagine originale con i colori degli stain della puzzle image target. In questo modo i tempi computazionali vengono ridotti al minimo.

# Capitolo 1

## Introduzione

Nel primo capitolo vengono descritti i processi per ottenere un preparato istologico e le principali colorazioni utilizzate per le immagini digitalizzate (Whole Slide Image, WSI). Ci si sofferma poi sull'utilizzo dei framework di deep learning per l'analisi delle WSI e sull'importanza degli steps di pre-processing per migliorare i tool a sostegno del lavoro dell'anatomo patologo evidenziando gli obiettivi principali del lavoro di tesi.

### 1.1 Patologia digitale e WSI imaging systems

La patologia, come la maggior parte delle specialità mediche richiede sempre più un miglioramento della qualità e dell'accuratezza diagnostica per la sicurezza del paziente. L'introduzione commerciale degli scanner per imaging da vetrino istologico completo (Whole - Slide Imaging o WSI) ha permesso di aprire la strada all'uso della patologia digitale come risorsa per applicazioni cliniche, didattico-formative e di ricerca. La tecnologia odierna permette di ottenere un rendering digitale dell'intero vetrino e di visualizzare le strutture ad altissime risoluzioni su uno schermo senza la necessità di utilizzare un microscopio, ciò permette al patologo di navigare liberamente sulle sezioni istologiche a diverse risoluzioni facilitando e velocizzando la diagnosi [16].

#### 1.1.1 Preparazione dei vetrini

Per ottenere le immagini istologiche digitalizzate il tecnico di laboratorio esegue manualmente diversi steps (Fig. 1):

- Prelievo;
- Fissazione;
- Deidratazione e pulizia;

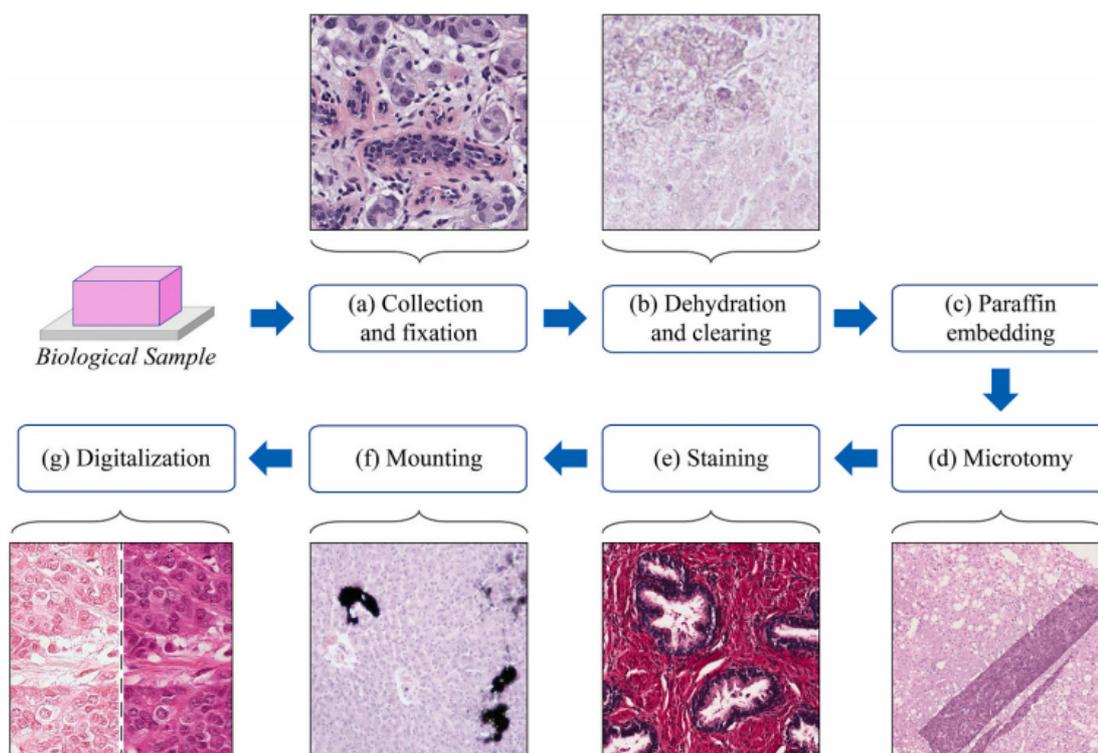
- Inclusione in paraffina;
- Sezionamento;
- Colorazione;
- Montaggio;
- Digitalizzazione.

Per preservare il tessuto da batteri ed enzimi cellulari, il campione prelevato tramite biopsia è trattato con un fissativo per prevenire le alterazioni chimico-fisiche. I tessuti prelevati perdono la loro consistenza e quindi devono essere inclusi in materiali più resistenti come la paraffina. La paraffina non è miscibile in acqua, dunque, la fase dell'inclusione è preceduta da una di disidratazione. Il blocchetto di paraffina ottenuto è sagomato e sezionato con microtomo in fettine molto sottili. Il giusto spessore è 5  $\mu\text{m}$  poiché permette di visualizzare correttamente sia l'architettura cellulare che la morfologia della cellula (nuclei e citoplasma), le strutture cellulari non possono essere correttamente visualizzate su sezioni di spessore maggiore ai 10  $\mu\text{m}$  mentre spessori inferiori ai 2  $\mu\text{m}$  permettono di visualizzare solo le strutture subcellulari come i nucleoli. Le sezioni estratte sono poi riposte sul vetrino portaoggetti per l'incisione delle informazioni generali riguardanti il campione. Dopo la microtomia il vetrino diventa traslucido e si passa alla colorazione, propedeutica all'analisi istologica, in questa fase si aumenta il contrasto delle diverse componenti cellulari e tissutali migliorando la leggibilità del vetrino. Il pH, la concentrazione della soluzione e il tempo di immersione nel colorante influenzano notevolmente la colorazione finale del campione, inoltre anche uno scorretto posizionamento del tessuto sul vetrino si traduce talvolta in pieghe di tessuto distinguibili sull'immagine digitalizzata con un colore molto intenso.

I coloranti si classificano in sintetici, naturali, acidi e basici, generalmente si usano in soluzione e affinché penetrino nei tessuti, questi ultimi devono essere idrofobi e quindi la paraffina idratata e sopraffinata.

Le fasi finali sono montaggio e digitalizzazione, il campione è inserito in una cover che lo protegge dagli agenti esterni e viene digitalizzato con scanner moderni [1]. Il posizionamento della cover può generare artefatti a causa della presenza di polvere, bolle d'aria o contaminazioni di altri microrganismi mentre la digitalizzazione può imporre delle intensità di colori diverse a seconda dello scanner utilizzato, della luce circostante e della tecnologia di

utilizzo. Inoltre, se il campione non è perfettamente allineato con il piano focale dello scanner l'immagine può risultare sfocata.



*Figura 1- Protocollo da applicare al campione per ottenere la relativa immagine digitale. (a) Prelievo e fissazione, la scelta del fissativo è fondamentale per mantenere un elevato contrasto fra le strutture; (b) Deidratazione e pulizia, permettono l'attacco corretto della paraffina nella fase successiva; (c) Inclusione in paraffina; (d) Microtomia, sono evidenziate delle pieghe di tessuto causate dallo scorretto posizionamento del campione sul vetrino; (e) Colorazione; (f) Montaggio con artefatti dovuti alla presenza di polvere; (g) Digitalizzazione dello stesso tessuto con due scanner differenti [1].*

### 1.1.2 Colorazioni

Esistono diverse tipologie di colorazione: istologica, istochimica, immunohistochimica ed immunofluorescenza.

Lo studio in oggetto interessa la maggior parte delle colorazioni: istologiche, istochimiche ed immunohistochimiche (Fig. 2).

Le colorazioni istologiche si basano principalmente sull'affinità dei coloranti per un determinato tessuto o per le sue condizioni di pH fornendo un'informazione per lo più di tipo

morfologico, i coloranti acidi colorano componenti basici sia cellulari (ad esempio il citoplasma) che tissutali (tessuto osseo, connettivo) mentre quelli basici colorano le strutture acide (ad esempio nucleo, ribosomi, cartilagine e secreti). Le istochimiche sviluppano una reazione chimica vera e propria che darà vita ad un precipitato colorato direttamente sulla sezione dando informazioni sulla natura delle sostanze chimiche contenute all'interno. L'esame immunoistochimico è un altamente specifico in quanto prevede l'utilizzo di anticorpi che colorano determinati antigeni presenti sul tessuto.

Le colorazioni istologiche si dividono in bicromiche, due coloranti (uno acido e uno basico) e tricromiche (due coloranti acidi e uno basico). In questo studio sono trattate tutte le colorazioni presentate in Fig. 2.

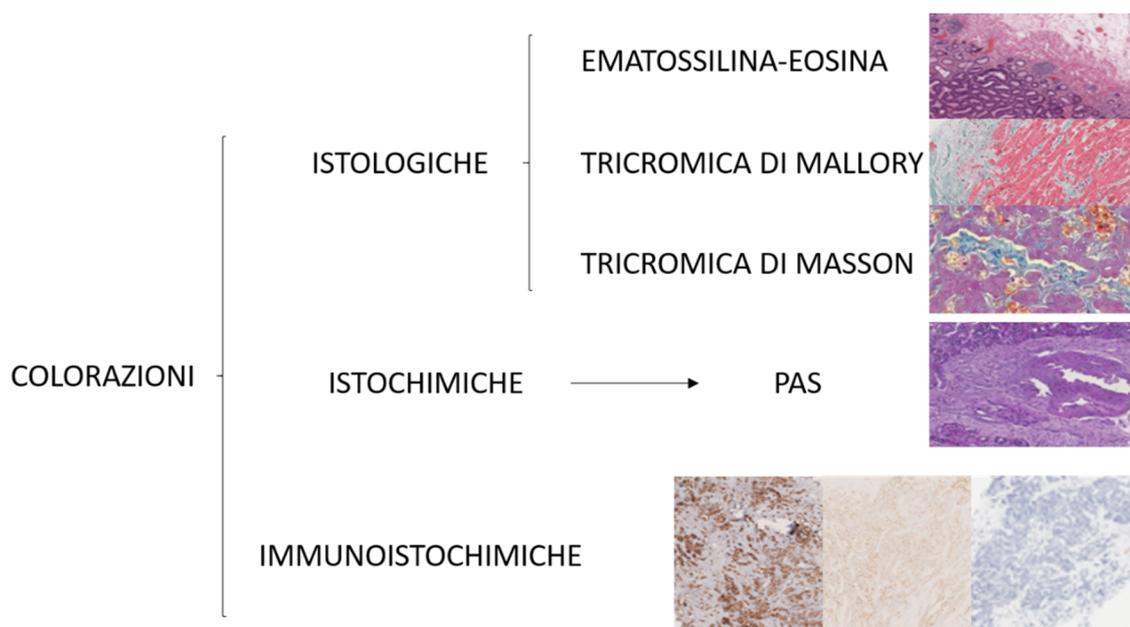


Figura 2- Principali colorazioni oggetto dello studio.

## 1.2 Deep learning in patologia digitale

I framework di deep learning sono rapidamente diventati la metodologia principale per l'analisi delle immagini mediche grazie alla loro potente capacità di apprendimento e capacità di lavorare con pattern complessi. Nell'ambito della patologia digitale i tool di deep learning lavorano su problemi di classificazione (Fig. 3a), detection (Fig. 3b) e segmentazione (Fig. 3c). Per reti di deep learning si intendono reti neurali artificiali che hanno diversi layers tra quello di input e di output. La complessità dell'architettura è direttamente proporzionale

all'aumento del numero dei parametri di training, in questo modo si riesce a lavorare con modelli di apprendimento sempre più performanti. Il recente incremento delle dimensioni dei dataset e della potenza computazionale ha permesso di applicare le reti neurali convoluzionali (CNN - Convolutional Neural Network) all'analisi della singola immagine applicando convoluzioni sequenziali a partire dall'immagine di input. Le CNN sono le reti più utilizzate, sono allenate sull'intera immagine o su patch significativi della stessa, le features più significative vengono apprese ottimizzando la funzione di costo. Durante il processo di training, i pesi per ogni neurone vengono aggiornati continuamente e dopo la fase di apprendimento il network è pronto per essere utilizzato su nuove immagini.

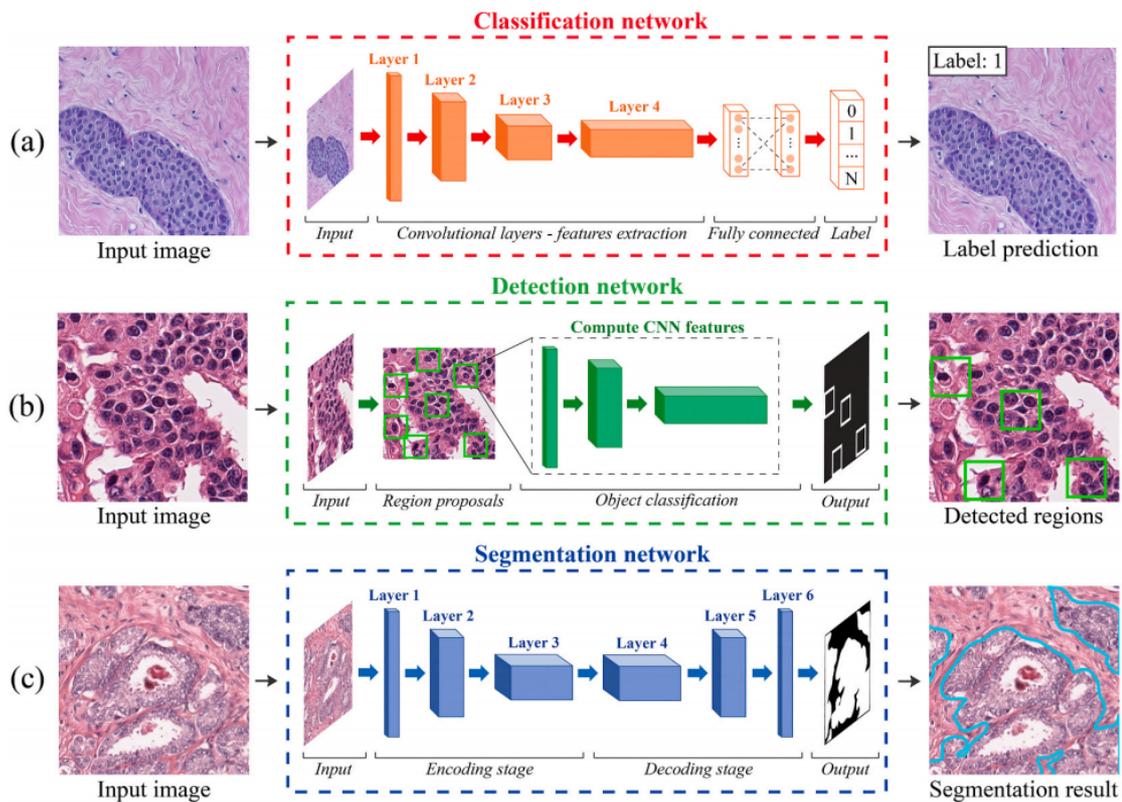


Figura 3- Architetture tipiche di framework di deep learning. (a) Classificazione, partendo da un'immagine di input individua una categoria (label prediction), (b) Detection, individuazione di strutture di interesse, (c) Segmentazione, la rete individua i contorni dell'oggetto [1].

### **1.2.1 Algoritmi di pre-processing**

La patologia digitale lavora con delle immagini molto grandi, computazionalmente onerose e cariche di artefatti quindi degli steps di pre-processing sono necessari per rendere stabile un qualunque algoritmo di deep learning che lavori successivamente sulle immagini.

Con il termine pre-processing si intendono tutte le strategie applicate ai dati grezzi (ad esempio l'intera WSI o parte di essa) finalizzate all'ottenimento di un network iniziale che renda più robusto il modello di processing completo.

I modelli di pre-processing proposti in questo lavoro sono:

- Individuazione del tessuto ed esclusione degli artefatti;
- Selezione dei patch significativi;
- Normalizzazione degli stain.

### **1.2.2 Individuazione del tessuto e degli artefatti**

Nelle WSIs la qualità delle immagini scannerizzate è il risultato del trade off tra la qualità del campione ottenuto e le caratteristiche tecniche dello scanner. Artefatti come pieghe di tessuto o bolle d'aria possono interferire con la corretta scansione del vetrino a causa di una scorretta selezione dei punti di messa a fuoco dando come risultato un'immagine sfocata.

I vetrini contengono in genere un'area di tessuto di 15mm x 15 mm. L'intera immagine è catturata ad una risoluzione incredibilmente alta, dunque, l'output finale ha un peso computazionale elevato, anche di alcuni gigabyte. Per ridurre i costi computazionali nel processing delle immagini è pratica comune individuare le sezioni di maggior interesse clinico prima di procedere ad un'analisi dettagliata. Difatti solitamente sono presenti delle sezioni molto ampie che non contengono tessuto istologico e che possono essere rimosse tramite un'accurata selezione del tessuto. Talvolta nemmeno una buona segmentazione è sufficiente per avere dei risultati soddisfacenti, quando i patologi analizzano visivamente le WSIs escludono naturalmente le regioni con artefatti (bolle d'aria, segni di pennarelli, capelli, peli etc), dunque un sistema digitale dovrebbe essere in grado sia di individuare il tessuto che di escludere gli artefatti e le sfocature distinguendo le aree in-focus e quelle out-of-focus.

### 1.2.3 Normalizzazione degli stain

La procedura di normalizzazione degli stain implica la trasformazione di un'immagine  $I$  in un'immagine normalizzata  $I_{NORM}$  attraverso la trasformazione  $I_{NORM} = f(I, \theta)$ , dove  $\theta$  è un set di parametri estratto da un'immagine template e  $f(\cdot)$  è la funzione che crea una corrispondenza tra l'immagine in output e il template.

L'immagine finale è morfologicamente identica a quella di input a meno dei colori che vengono normalizzati sulla base dell'immagine template, dunque ogni immagine normalizzata avrà la sua mappa di distribuzione di intensità per matchare la distribuzione di colori dell'immagine template.

I metodi attualmente utilizzati si dividono in:

- *Normalizzazione di colore globale*, si effettua dopo aver separato le informazioni di colore ed intensità utilizzando diversi spazi colore e si ottiene il risultato finale lavorando con gli istogrammi (Fig. 3A). Ad esempio, la media dei colori dell'immagine target può essere trasferita a quella dell'immagine di input in modo tale che le sue variazioni di intensità siano preservate ed il contrasto ottenuto sull'immagine processata sia quasi lo stesso dell'immagine target. Un'altra tecnica solitamente utilizzata consiste nel mappare sull'istogramma dell'immagine target quello dell'immagine sorgente per fare in modo che le caratteristiche di luminosità e colore delle due immagini si somiglino. Lavorare con gli istogrammi risulta però scomodo in quanto adattando l'istogramma dell'immagine sorgente si ottengono degli effetti poco naturali [1][17][18][19].

- *Normalizzazione di colore dopo la separazione degli stain*, viene effettuata separando i contributi dei singoli coloranti utilizzati (Fig. 3B). L'intensità raggiunta dal singolo componente cellulare dipende dall'assorbimento del singolo stain secondo la legge di Lambert-Beer [20].

Prima della separazione degli stain è necessario esprimere l'intensità di luce assorbita in termini di densità ottica, ossia nello spazio OD [21] in modo da ottenere una dipendenza lineare (Fig.9).

I valori di intensità RGB non possono essere usati direttamente per la separazione degli stain poiché la relazione tra la concentrazione e l'intensità di ogni stain non è lineare (Eq. 1a). In figura 5 viene mostrata la concentrazione degli stain per la colorazione istologica Ematossilina-Eosina. Nello spazio RGB (Fig. 5A) le due concentrazioni non sono linearmente separabili,

potrebbero essere separabili tramite una relazione non lineare difficilmente stimabile mentre nello spazio OD sono facilmente separabili tramite una relazione lineare (Fig. 5B).

L'intensità dell'immagine nello spazio OD ( $V$ ) può essere definita come il logaritmo del rapporto tra luce incidente e trasmessa:

$$\text{a) } V = \log \frac{I_0}{I} = \varepsilon Lc \qquad \text{b) } V = W \cdot H \qquad (1)$$

Dove:

- $\varepsilon$  è il coefficiente di assorbimento;
- $L$  è lo spessore del campione;
- $c$  è la concentrazione della sostanza assorbente.

Il valore di densità ottica di ogni pixel può essere definito come il prodotto tra la matrice degli stain ( $W$ ) e la matrice di concentrazione degli stain ( $H$ ) (Eq. 1b).

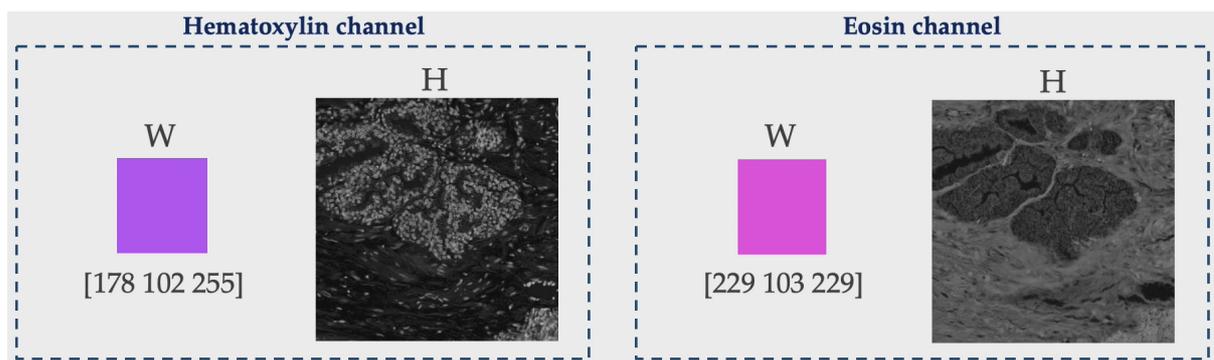


Figura 4- Matrice degli stain per un'immagine Ematossilina- Eosina (2x3), ogni riga rappresenta le intensità dello stain e la sua occorrenza.

La matrice  $W$  ha dimensione  $N \times 3$ , dove  $N$  corrisponde al numero degli stain presenti mentre  $H$  ha dimensioni  $M \times 2$ , dove  $M$  rappresenta il numero di pixel e rappresenta la densità dello specifico stain.

Ogni valore all'interno di  $W$  indica lo stain corrente e l'occorrenza dello stesso nell'immagine come mostrato in figura 4.

Tramite color deconvolution si può scomporre la matrice  $W$  ottenendo i singoli stain dell'immagine, mentre per inversione di matrice si ottiene  $H$  dall'Eq. 1:

$$H = ((W^T \cdot W)^{-1} \cdot W^T) \cdot V \quad (2)$$

-Trasferimento di colori tramite reti di deep learning, si sviluppa tramite l'utilizzo di reti generative avversative (GANs) con il metodo del transfer learning (Fig. 3C).

Lo "style transfer" permette di trasferire lo stile di un'immagine in un determinato dominio a quello di un'altra in un altro dominio. Le GANs vengono addestrate in maniera competitiva tramite un discriminatore e un generatore all'interno di un framework di gioco *minmax*.

Lo scopo è di addestrare la GAN a ricreare immagini che corrispondano il più possibile alle immagini realistiche da una distribuzione casuale  $z$ , il generatore  $G$  genera una distribuzione di campioni  $P_{DATA}$ , dove  $DATA$  rappresenta concettualmente le features latenti dell'immagine che la GAN individua da sola in fase di training ed è addestrato sul feedback del discriminatore tramite back propagation.

Il compito del discriminatore  $D$  è di lavorare in maniera avversativa fornendo in output la probabilità che l'immagine generata sia realistica per migliorare le performance della rete [1][17].

Il framework alle base della rete è il seguente:

$$\minmax V(G, D) = E_{x \sim P_{DATA}} [\log(D_x)] + E_{z \sim P_{NOISE}} [\log(1 - (D(G_z)))] \quad (3)$$

Il generatore prova a minimizzare la funzione espressa nelle Eq. 3 mentre in maniera avversativa il discriminatore prova a massimizzarla fino a che entrambi i network non raggiungono una condizione di ottimo.

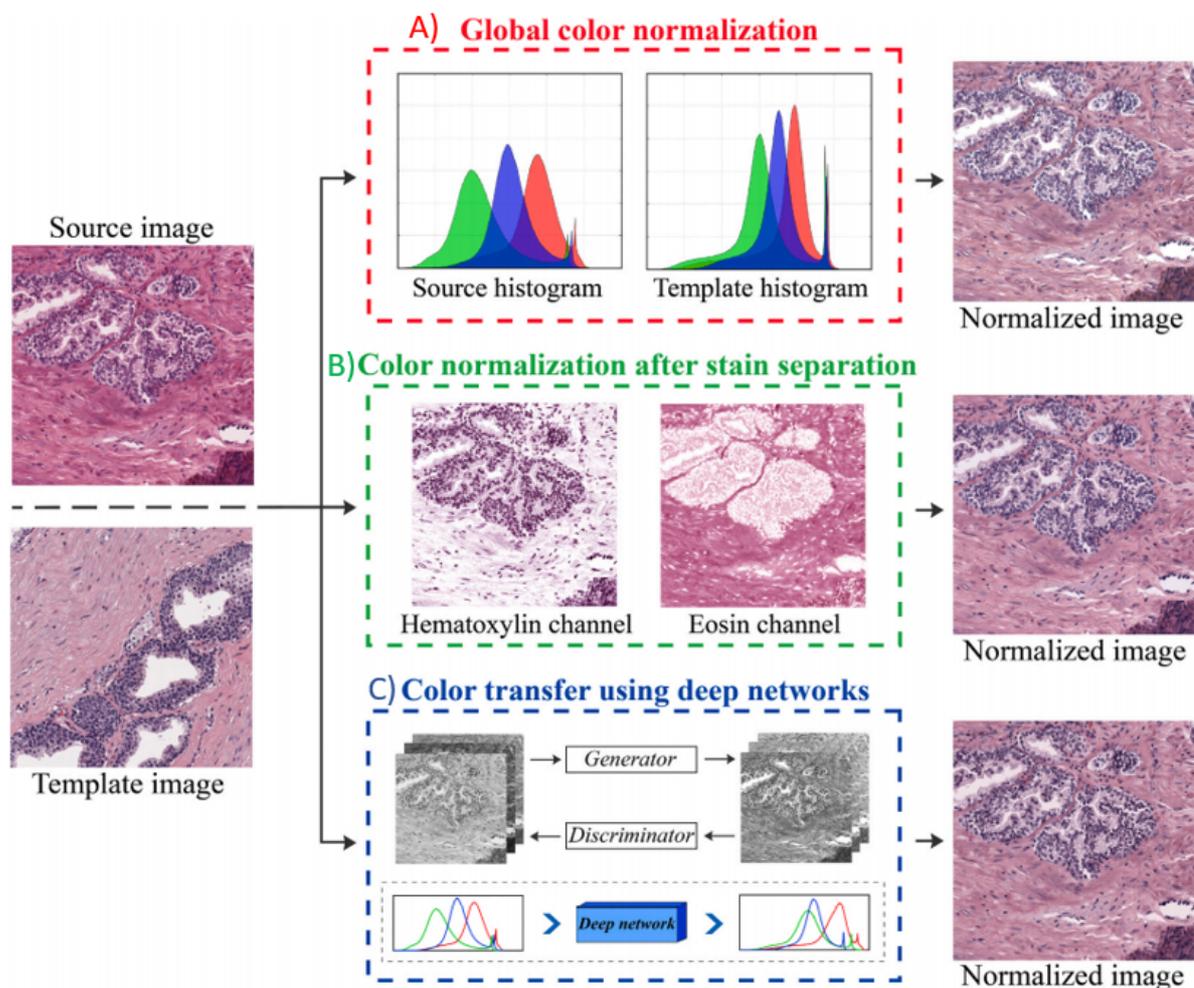


Figura 5- Strategie di normalizzazione degli stain [1].

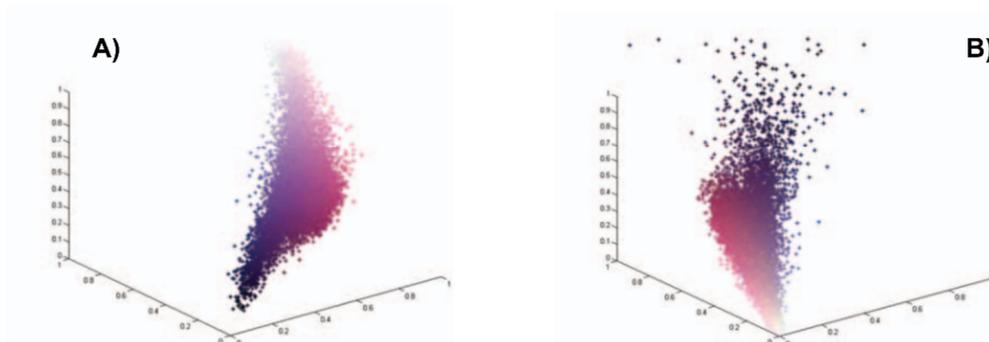


Figura 6- A sinistra la distribuzione delle concentrazioni degli stain per la colorazione Ematossilina (Viola) - Eosina (Rosa chiaro) nello spazio RGB, a destra lo stesso nello spazio OD. È evidente come a sinistra le due concentrazioni non siano linearmente separabili. [2]

### 1.2.4 Selezione dei patch significativi

Gli algoritmi di deep learning generalmente vengono applicati alle intere WSIs, in questo modo si rischia di perdere informazioni discriminanti ai fini della diagnosi a causa dell'eccessivo e necessario sottocampionamento e si limita l'apprendimento ad un unico pattern. Nelle immagini istologiche, l'informazione di interesse è contenuta nelle patches ad elevata risoluzione quindi l'idea migliore è quella di allenare la rete proprio su queste patches. Durante gli anni sono state suggerite diverse tecniche per effettuare una selezione sulle patches significative, dalla selezione casuale a quella guidata, la più comune consiste nel suddividere l'immagine con una griglia e selezionare le patches con una sliding window.

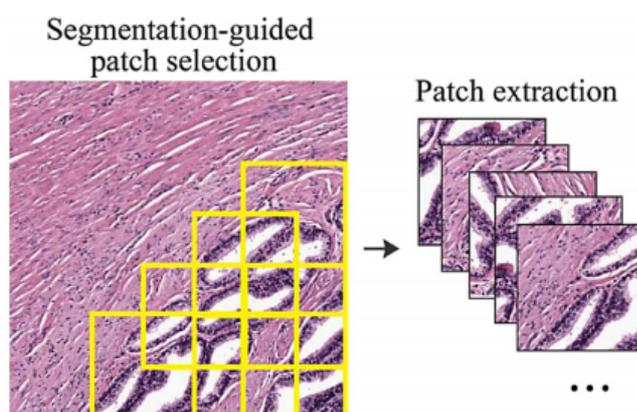


Figura 7- Segmentazione guidata per l'estrazione delle patches significative [22].

### 1.3 Stain normalization: approcci utilizzati

In questo paragrafo viene descritto lo stato dell'arte attuale riguardo gli approcci utilizzati per la normalizzazione dei colori delle immagini. Come spiegato nel paragrafo 1.2.3 queste tipologie di algoritmi lavorano sulla combinazione dell'informazione data dalla matrice degli stain ( $W$ ) e da quella di densità degli stain ( $H$ ) dell'immagine sorgente per modificare l'intensità di colore dell'immagine target. L'immagine sorgente presenta generalmente una distribuzione di colori di difficile interpretazione mentre l'immagine target, che funge da immagine template è selezionata attentamente da un patologo esperto. I metodi di normalizzazione attuali possono essere raggruppati in:

- Metodi basati sulla color deconvolution;

- Metodi basati sulla manipolazione e trasformazione dell'istogramma: Rehinard et al. hanno sviluppato un algoritmo che implementa una trasformazione tra l'immagine sorgente e l'immagine target nello spazio  $l\alpha\beta$  computando la media e la deviazione standard per ogni canale riconvertendo poi l'immagine di output nello spazio RGB [3]
- Metodi basati sull'ottimizzazione iterativa: Vahadane et al. presentano un algoritmo basato sulla NMF (Non Negative Matrix Factorization) della matrice dell'immagine nella densità ottica (V) cercando di calcolare allo stesso tempo la coppia di matrici W e H che minimizza l'errore rispetto a V [8], ossia la funzione riportata nell'equazione sottostante:

$$f(H, W) = \frac{1}{2} \|V - WH\|_F^2 \quad (2)$$

Questo lavoro di tesi si focalizzerà su uno degli approcci più noti basato sulla color deconvolution.

### 1.3.1 Algoritmi di Color Deconvolution

Gli algoritmi di color deconvolution si differenziano in base alle modalità di calcolo della matrice  $W$  (prf 1.2.3). L'obiettivo della color deconvolution è quindi quello di individuare gli stains che colorano il tessuto, stimarne le concentrazioni e separarli come illustrato in Fig. 7.

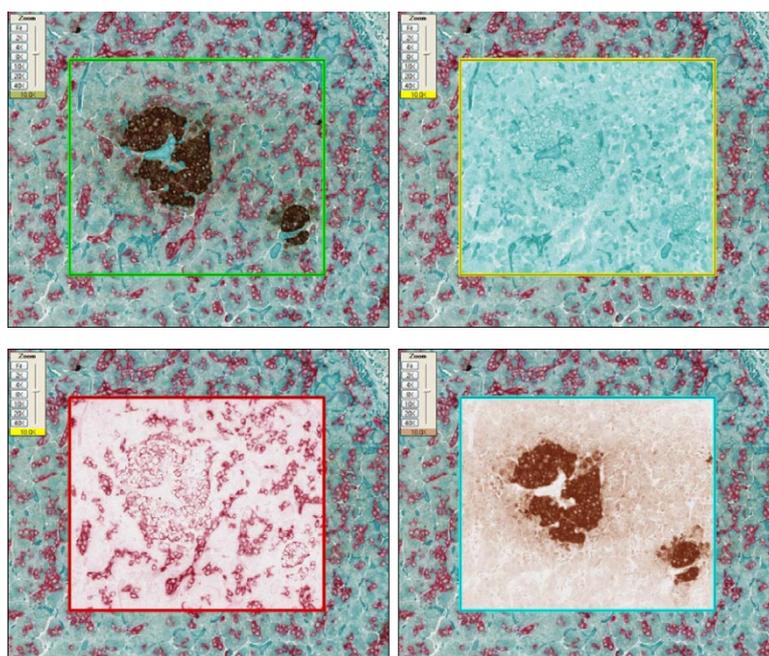


Figura 8-Esempio di color deconvolution, in alto a destra l'immagine di riferimento, a sinistra il canale 1 (Crystal Light Green), in basso a sinistra il canale 2 (Fast Red) ed in basso a destra il canale 3 (DAB). [Color Deconvolution Algorithm, User's Guide Aperio, [Color Deconvolution Algorithm User's Guide \(jhmi.edu\)](#)]

Ruifrok et al. hanno sviluppato un algoritmo basato sulla color deconvolution per migliorare l'individuazione ed il conteggio dei nuclei nelle immagini con colorazione immunoistochimica (Ki67) che normalmente colora i nuclei di marrone e lo stroma coi toni dell'azzurro. Lavorando sull'analisi dell'immagine in formato HSI i nuclei molto scuri vengono individuati come blu piuttosto che marroni invece grazie alla color deconvolution si riescono a separare correttamente più di tre canali di colore nell'immagine basandosi su una trasformazione ortonormale dell'immagine RGB originale [6]. Il metodo fornisce la possibilità di individuare non solo la densità di staining anche nelle aree multi-stain ma anche di calcolare numericamente l'area occupata da ogni colorazione e quindi il rapporto fra le aree di colorazione diversa. Tale metodo utilizza dei valori fissi di  $W$  per ogni tipologia di colorazione, così facendo se l'intensità degli stain dell'immagine da normalizzare si discosta molto da quella

rappresentata dai valori prefissati di  $W$  si ha una stima del vettore degli stain approssimativa peggiorando le performance della color deconvolution. Si configura quindi come un metodo non adattativo [1][6].

Macenko utilizza invece un metodo adattativo basato sulla Single Value Decomposition (SVD) per il calcolo di  $W$ . Lo scopo di questa tecnica matematica è di trovare un piano dove proiettare i pixels trasformati (OD) e trovare i valori estremi per l'identificazione dei vettori degli stain. Questo approccio è applicabile a diverse tipologie di colorazione e ha tempi computazionali poco onerosi, il più grande svantaggio è rappresentato dalla stima errata dei vettori degli stain quando delle regioni non colorate hanno un'intensità (nello spazio OD) più alta di una soglia preimpostata ( $\beta=0.15$ ). In più la presenza di variazioni di intensità molto forti nelle strutture può causare il fallimento dell'intero processo di color deconvolution.

Un altro algoritmo di tipo adattativo è stato proposto da Zheng et al. Per la singola WSI vengono campionati un certo numero di pixels e convertiti nello spazio OD (Fig. 9). Gli stain normalizzati dell'immagine sono ottenuti tramite una matrice ADC e una matrice pesata sugli stains. Le componenti rappresentative degli stain dell'immagine sono ricomposte tramite la matrice SCA di un'immagine template. La color deconvolution è alla base del calcolo della matrice ADC (Adaptive Color Deconvolution), i valori RGB vengono tradotti nello spazio OD, per l' $i$ -esimo pixel:  $o_i = -\ln(x_i/I_{max})$  e  $s_i = D \cdot o_i$ ,  $o_i$  rappresenta il valore OD corrispondente al valore RGB ( $x_i$ ),  $D$  viene definita matrice di color deconvolution e si ricava dalla matrice SCA ( $M$ ) ricavata a sua volta sperimentalmente,  $I_{max}$  è l'intensità dello sfondo, ossia il valore dei pixel non colorati.. Lo studio riguarda la colorazione istologica H&E,  $s_i$  è l'output che rappresenta la matrice degli stain e può essere rappresentata come  $s_i = (h_i, e_i, d_i)^T$  dove  $h_i$  ed  $e_i$  rappresentano le matrici degli stain di ematossilina ed esosina e  $d_i$  rappresenta il residuo della separazione. La matrice ADC è ottenuta ottimizzando tutte le variabili in gioco [23].

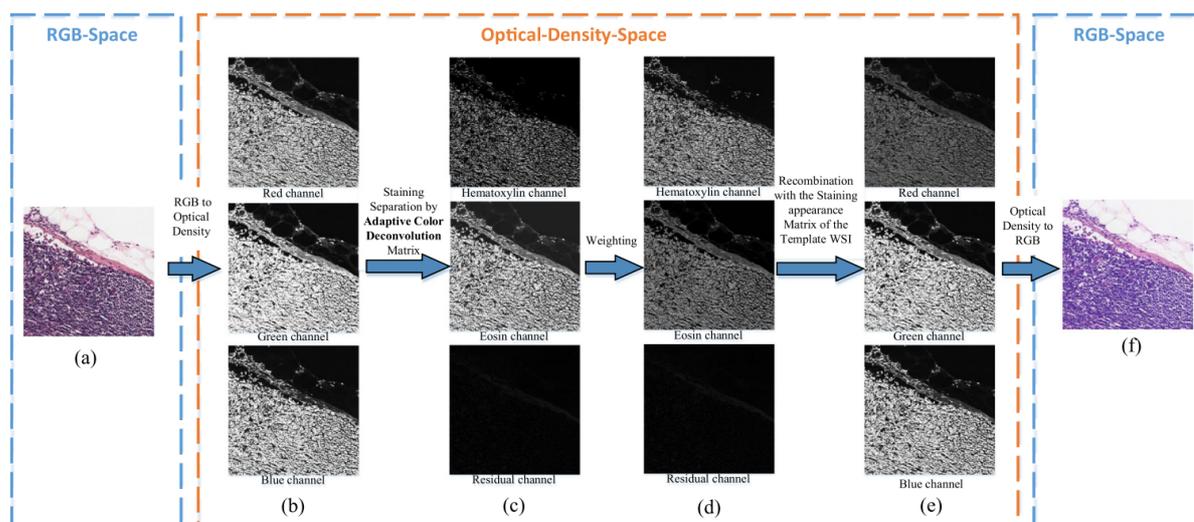


Figura 9 – Flowchart del metodo di stain normalization proposto da Zheng et al (a) Immagine originale; (b) Canali dell'immagine RGB convertiti nello spazio OD. (c) Separazione degli stain nello spazio OD. (d) Ottimizzazione dei parametri. (e) Ricombinazione degli stains. (f) Riconversione dell'immagine di output nello spazio RGB [23].

Khan et al. hanno proposto un nuovo descrittore del colore (SCD) per quantificare la concentrazione degli stain nell'immagine mentre la matrice degli stain ( $W$ ) della specifica immagine è calcolata con un classificatore color-based.

L'algoritmo esegue la mappatura non lineare dei canali dell'immagine sorgente sui canali dell'immagine target utilizzando funzioni spline-based regolarizzate ricavate da parametri statistici dell'immagine stessa [4].

Uno degli algoritmi più recenti riguardo la normalizzazione delle immagini è lo SCAN presentato da Salvi et al. che consta di quattro moduli: separazione degli stain preliminare, segmentazione delle strutture, separazione definitiva degli stains e normalizzazione dell'immagine.

L'analisi in oggetto è effettuata su 270 immagini H&E. Preliminarmente viene effettuata una white detection per escludere dinamicamente le regioni bianche, successivamente si applica la separazione degli stain di Macenko e si segmentano le strutture presenti nell'immagine tramite una versione migliorata dell'algoritmo MANA.

La normalizzazione finale dell'immagine non è effettuata invertendo direttamente l'Eq. 1b ma per non perdere il significato fisico della matrice delle densità degli stains ( $H$ ) la prima riga di

tale matrice ( $H(1, :)$ ) è computata come la mediana delle triplette RGB del primo canale ottenuto dalla color deconvolution di Macenko.

Lo stesso procedimento è poi effettuato per il secondo canale della matrice  $H$ , sia per l'immagine sorgente che per l'immagine target [1].

La matrice delle densità di stain normalizzata è poi ottenuta come segue:

$$H_s^{norm}(j, :) = \frac{H(j, :)_s}{H(j)_s^{RM}} H_T^{RM}(j), j = 1, 2 \quad (3)$$

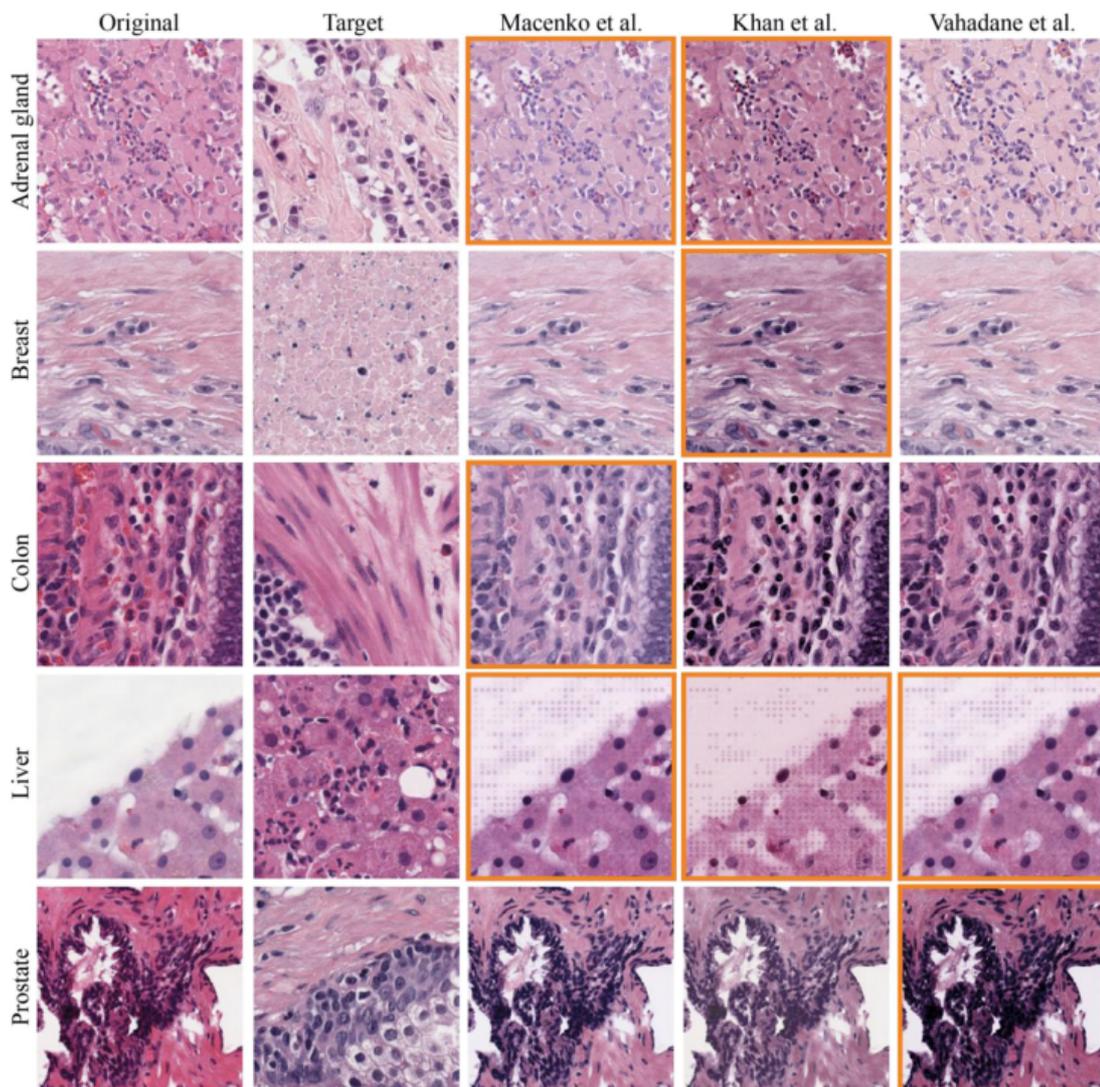


Figura 10-Risultati degli algoritmi di stain normalization applicati a diverse tipologie di tessuto. La prima colonna rappresenta le ROI delle immagini originali, la seconda le ROI dell'immagine target. Le boxes arancioni segnalano output con artefatti visibili [1].

Alcuni dei metodi citati nel paragrafo corrente e nel paragrafo 1.2.3 sono stati comparati in uno studio di ricerca che analizzava i risultati grafici riportati in Fig. 9 insieme ai tempi computazionali e ad alcune valutazioni numeriche (fig. 11). La valutazione quantitativa degli algoritmi è stata effettuata considerando l'rSE (Relative Square Error) tra la matrice degli stains dell'immagine originale  $W_{GT}$  e quella dell'immagine normalizzata ( $W$ ), più basso è l'rSE migliori sono le performance degli algoritmi [1]:

$$rSE = \frac{\|(W - W_{GT})^T(W - W_{GT})\|}{\sqrt{\|(W)^T(W)\| \cdot \|(W_{GT})^T(W_{GT})\|}} = \frac{\|W - W_{GT}\|_F^2}{\|W\|_F \cdot \|W_{GT}\|_F} \quad (4)$$

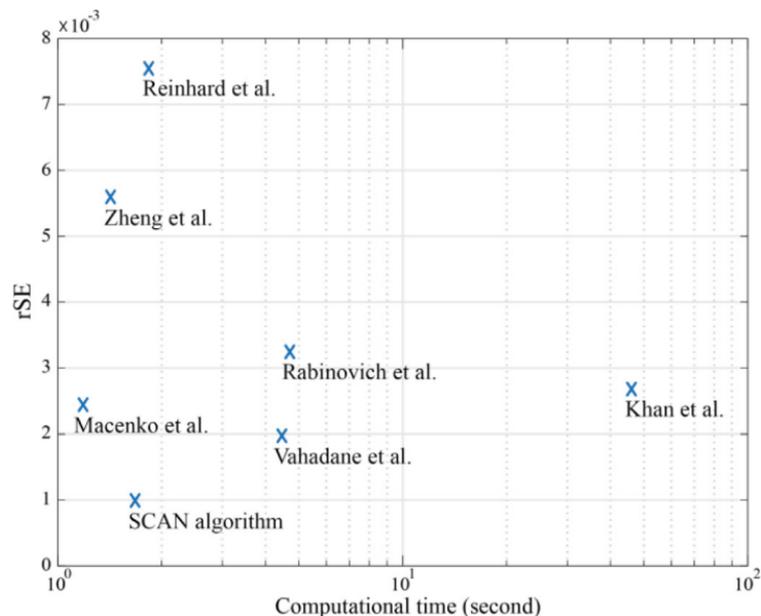


Figura 11- Plot dei valori di rSE per i metodi citati in funzione dei tempi computazionali per la normalizzazione dell'immagine [1].

Le performance elevate dell'algoritmo SCAN sono dovute alla combinazione di segmentazione e normalizzazione.

Nell'analisi oggetto di questa tesi verrà utilizzata invece la color deconvolution di Macenko, un algoritmo stabile che fornisce un giusto trade off tra qualità dei risultati e tempi computazionali.

## 1.4 Pipeline di Stain Normalization

Il lavoro di questo studio è focalizzato sull'implementazione degli steps di pre e post processing che ruotano intorno alla normalizzazione delle immagini e che permettano di rendere più stabile e performante l'intera pipeline. Di seguito la pipeline generale:

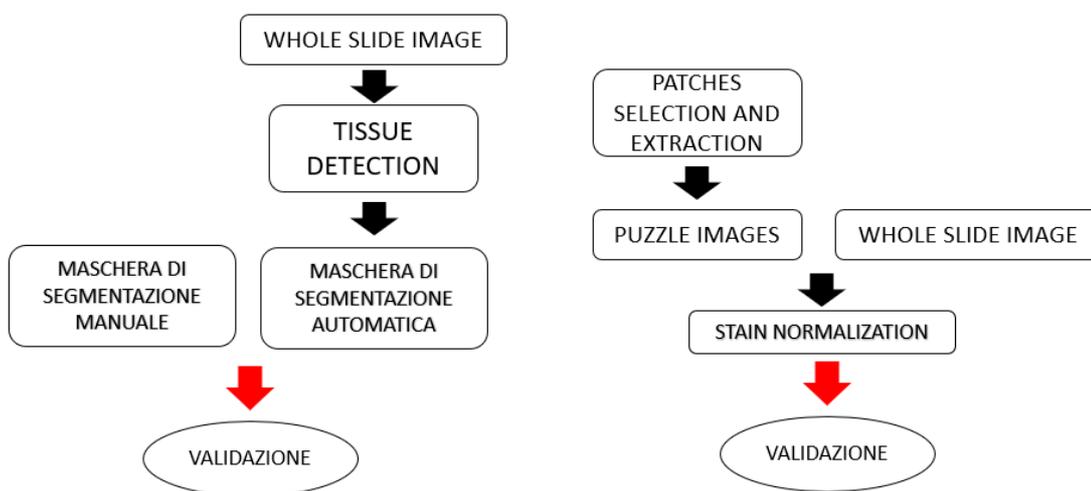
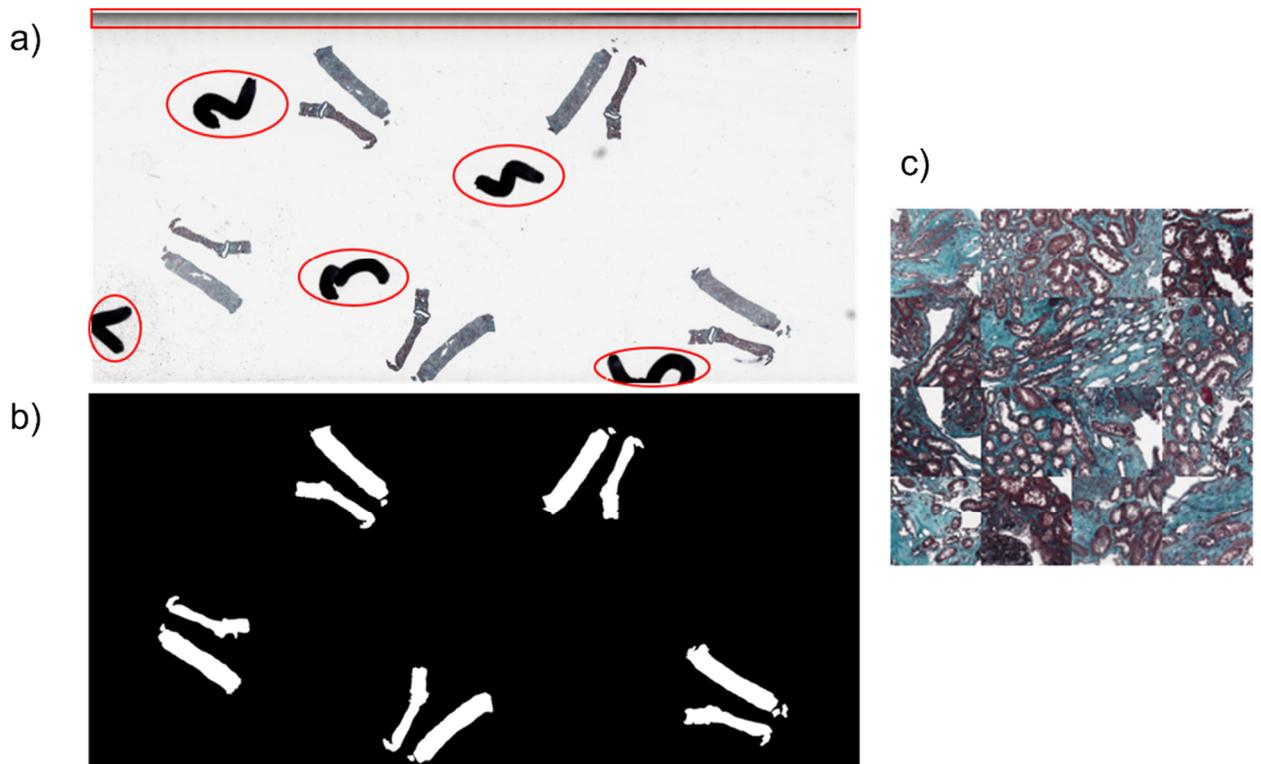


Figura 12- A destra la pipeline di tissue detection, a sinistra quella di stain normalization.

Il primo step prevede l'individuazione del solo tessuto istologico di interesse escludendo sfondo ed artefatti (Figg. 13a e 13b). Ottenute le maschere che segmentano il tessuto istologico si procede all'individuazione delle patches più significative per la biopsia in esame, tali patches sono raggruppate in un'unica immagine che verrà definita "puzzle image" (Fig. 13c). Le puzzle images sono necessarie per normalizzare l'intera WSI, difatti le matrici degli stains (W) e di densità degli stains (H) vengono calcolate sulle puzzle images ed in seguito le caratteristiche della normalizzazione vengono estese all'intera WSI con l'ausilio di una sliding window.



*Figura 13- Biopsia tricromica di rene. (a) Immagine originale con evidenziati gli artefatti. (b) Maschera automatica che individua unicamente il tessuto. (c) Esempio di puzzle image ottenuta unendo le patches significative.*

# Capitolo 2

## Materiali e metodi

Come accennato nel capitolo precedente l'obiettivo di questo lavoro di tesi è riuscire ad ottenere una pipeline di normalizzazione degli stain veloce e computazionalmente poco onerosa.

La pipeline è stata sviluppata tramite quattro algoritmi relativi ai vari steps presentati e due algoritmi di validazione dei risultati.

Gli algoritmi relativi agli steps descritti in Fig. 12 comprendono:

- due algoritmi di tissue detection;
- un algoritmo di patch selection and extraction per creare la puzzle image;
- un algoritmo di normalizzazione dell'intera WSI.

I due scripts di tissue detection si distinguono per livello di specificità e sensibilità, nello specifico il primo algoritmo, che da qui in poi è indicato come *algoritmo\_normalization*, ha elevata sensibilità ed è in grado di individuare tutto il tessuto presente, comprese le zone di adipe (regioni tissutali di scarso interesse clinico) mentre il secondo algoritmo, *algoritmo\_puzzle*, è maggiormente specifico, individua le parti significative di tessuto, comprendenti tutti gli stains che colorano l'immagine ed è utilizzato per ottenere dalle segmentazioni dalle quali poter ricavare una puzzle image più completa possibile, priva di artefatti e sezioni bianche. Tutti gli algoritmi sono stati sviluppati sia in programmazione Matlab che Python.

## 2.1 Datasets

Sono stati forniti da due diverse aziende ospedaliere piemontesi, Ospedale di Alba ed ospedale “Le Molinette”, dei vetrini digitalizzati di intere biopsie prelevate da diversi pazienti.

Le biopsie sono state effettuate su diversi tessuti: mammella, fegato, rene, surrene, colon, prostata, osso, polmone e tessuto miocardico. Le WSIs sono state raggruppate in due datasets differenti, che per semplicità verranno di seguito indicati come *dataset A* e *dataset B* (Tab 1).

Il dataset A consta di 714 WSIs mentre il dataset B, più ampio, di 2647 WSIs.

<b>DATASET A</b>	<b>OSPEDALE DI ALBA</b>	<b>OSPEDALE MOLINETTE</b>	<b>COLORAZIONE</b>
<i>MAMMELLA</i>	3	0	COLORAZIONE ISTOLOGICA (HE)
	40	209	COLORAZIONE IMMUNOISTICHIMICA
<i>CANCRO ALLA MAMMELLA</i>	0	40	COLORAZIONE ISTOLOGICA (HE)
<i>RENI</i>	0	68	COLORAZIONE ISTOCHIMICA (PAS)
	0	68	COLORAZIONE ISTOLOGICA (TRICROMICA)
<i>POLMONI</i>	0	50	COLORAZIONE ISTOCHIMICA (PAS)
<i>TESSUTO MIOCARDICO</i>	0	50	COLORAZIONE ISTOLOGICA (TRICROMICA)
<i>PROSTATA</i>	0	186	COLORAZIONE ISTOLOGICA (HE)

<b>DATASET B</b>	<b>OSPEDALE DI ALBA</b>	<b>OSPEDALE MOLINETTE</b>	<b>COLORAZIONE</b>
<i>SURRENE</i>	0	29	COLORAZIONE ISTOLOGICA (HE)
<i>OSSO</i>	0	41	COLORAZIONE ISTOLOGICA (HE)
	0	41	COLORAZIONE IMMUNOISTOCHEMICA
<i>MAMMELLA</i>	11	0	COLORAZIONE ISTOLOGICA (HE)
	124	0	COLORAZIONE IMMUNOISTOCHEMICA
	0	433	COLORAZIONE IMMUNOISTOCHEMICA
	247	0	COLORAZIONE ISTOLOGICA (HE)
<i>COLON</i>	0	3	COLORAZIONE ISTOLOGICA (HE)
<i>RENE</i>	0	119	COLORAZIONE ISTOLOGICA (TRICROMICA)
	0	119	COLORAZIONE ISTOCHEMICA (PAS)
<i>FEGATO</i>	0	308	COLORAZIONE ISTOLOGICA (HE)
<i>POLMONI</i>	0	101	COLORAZIONE ISTOCHEMICA (PAS)
<i>TESSUTO MIOCARDICO</i>	0	100	COLORAZIONE ISTOCHEMICA (TRICROMICA)
<i>PROSTATA</i>	132	278	COLORAZIONE ISTOLOGICA (HE)

---

0	561	COLORAZIONE ISTOCIMICA (PAS)
---	-----	---------------------------------

---

*Tabella 1-Descrizione dei due datasets.*

Le WSIs sono fornite in formato. svx, digitalizzate con AperioScanscope XT Slide Scanner ed in formato. ndpi, digitalizzate con NanoZoomer.

I files in formato. svx possono essere visualizzati ed elaborati utilizzando il software ImageScope di Aperio (Fig.14) mentre per i files. ndpi occorre utilizzare il software NDP.view (Fig. 15).



*Figura 14-Visualizzazione della biopsia di mammella (325 - 2021-03-14 16.25.11. ndpi) con il software NDP.view.*

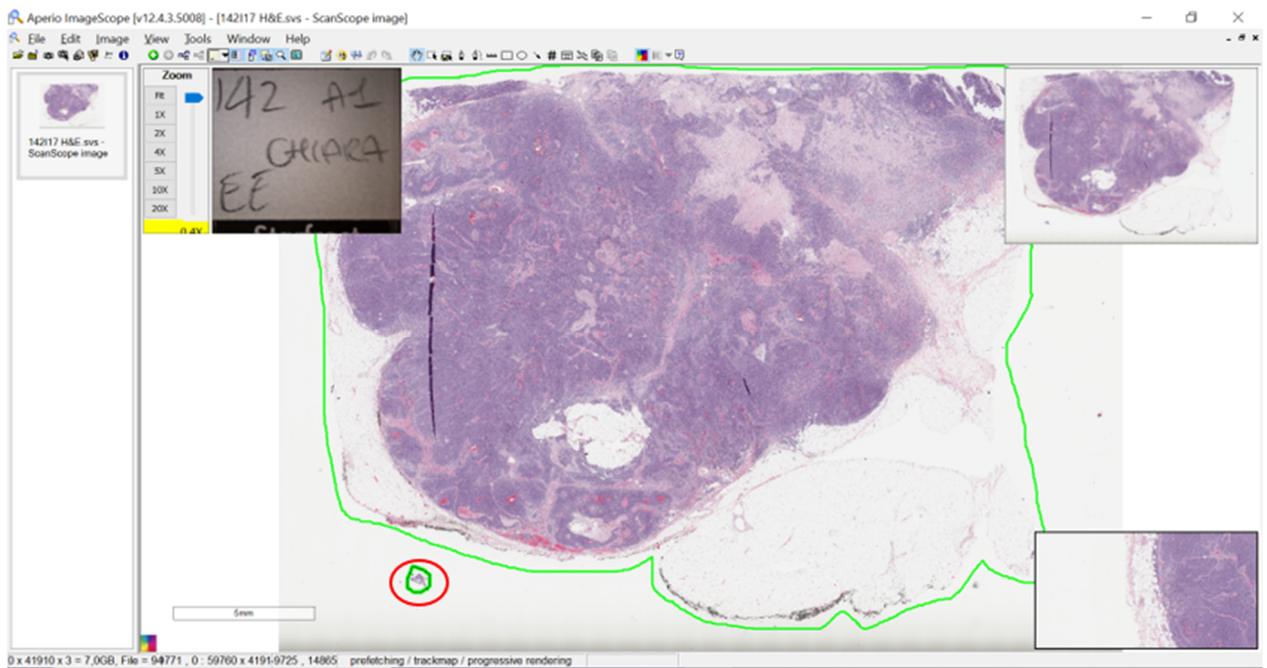


Figura 15-Visualizzazione della biopsia di una mammella con il software Aperio Image Scope (142117 H&E.svs).

Ogni vetrino digitalizzato presenta una struttura piramidale che contiene le informazioni necessarie per visualizzare ed estrarre l'immagine alla risoluzione desiderata (Fig. 15). Questi software sfruttano tale struttura per visualizzare la biopsia a partire da una risoluzione minima (ingrandimento 1X) fino alla massima risoluzione consentita (ingrandimento 20X).

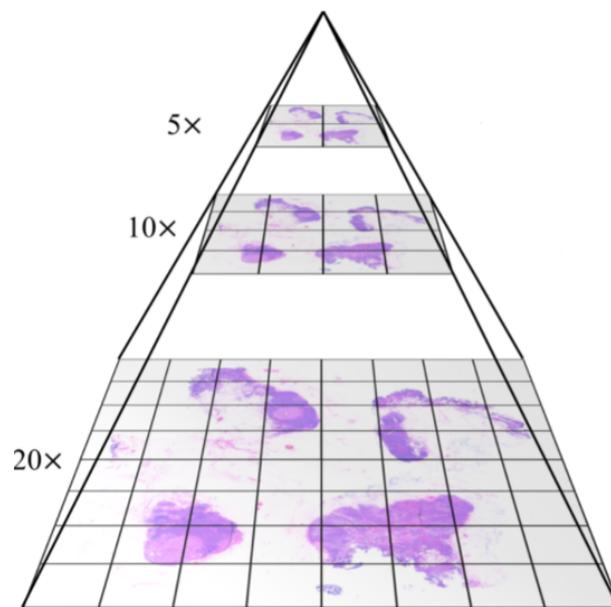


Figura 16- Formato piramidale della Whole Slide Image (WSI) [9].

## 2.2 Tissue Detection

Il primo step della pipeline è l'individuazione del tessuto istologico presente.

I due algoritmi sviluppati sono l'*algoritmo\_normalization* e l'*algoritmo\_puzzle* e si differenziano per la parte di post processing sulla maschera grezza iniziale.

Prima di sviluppare gli algoritmi sono state effettuate sul *dataset A* delle annotazioni manuali con i software descritti nel paragrafo 2.1. Le annotazioni sono state effettuate manualmente per segmentare tutto il tessuto presente escludendo ogni tipologia di artefatto (segni di pennarello, macchie etc.). In figura 15 la segmentazione è stata effettuata con il software di Aperio mentre nelle figure 17 e 18 con il software NDPview.

Le annotazioni effettuate sono state poi riviste per eliminare parti di tessuto troppo piccole per essere considerate istologicamente significative difatti le figure 15 e 19 mostrano lo stesso tessuto istologico segmentato in maniera differente. In figura 19 il pezzo di tessuto evidenziato in rosso in figura 15 non è stato considerato. In figura 17 sono evidentemente presenti dei frustoli sfocati, le segmentazioni ridefinite escludono anche effetti di sfocatura (Fig.18).

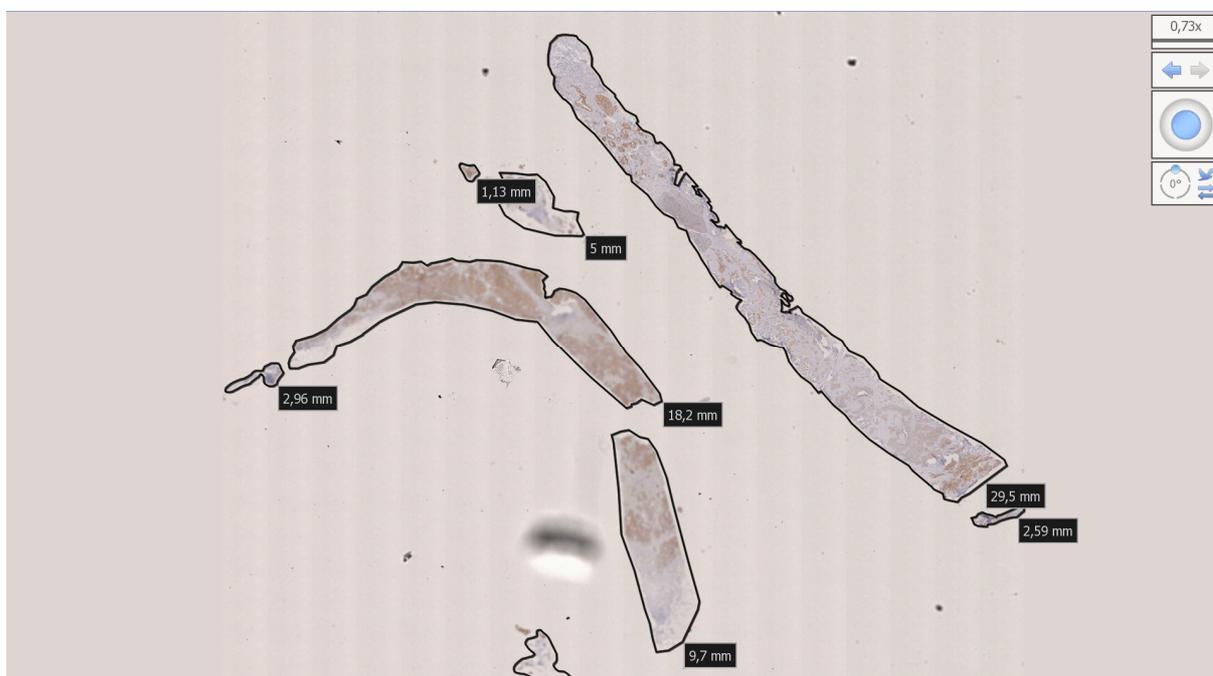
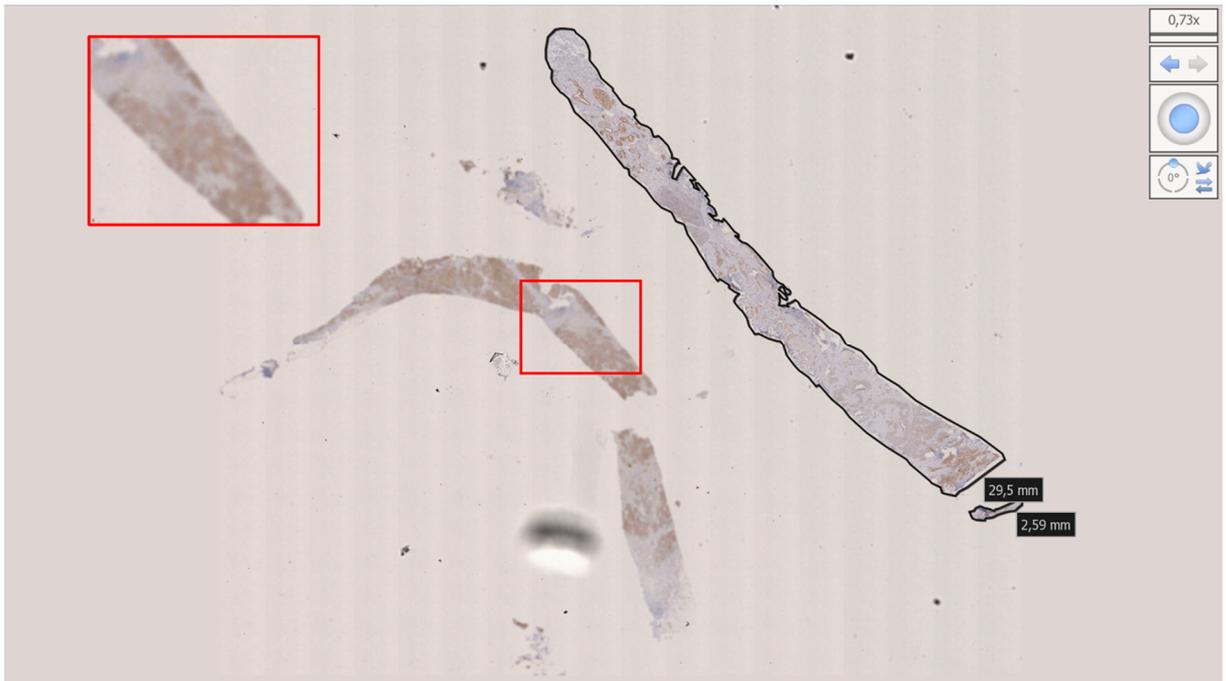
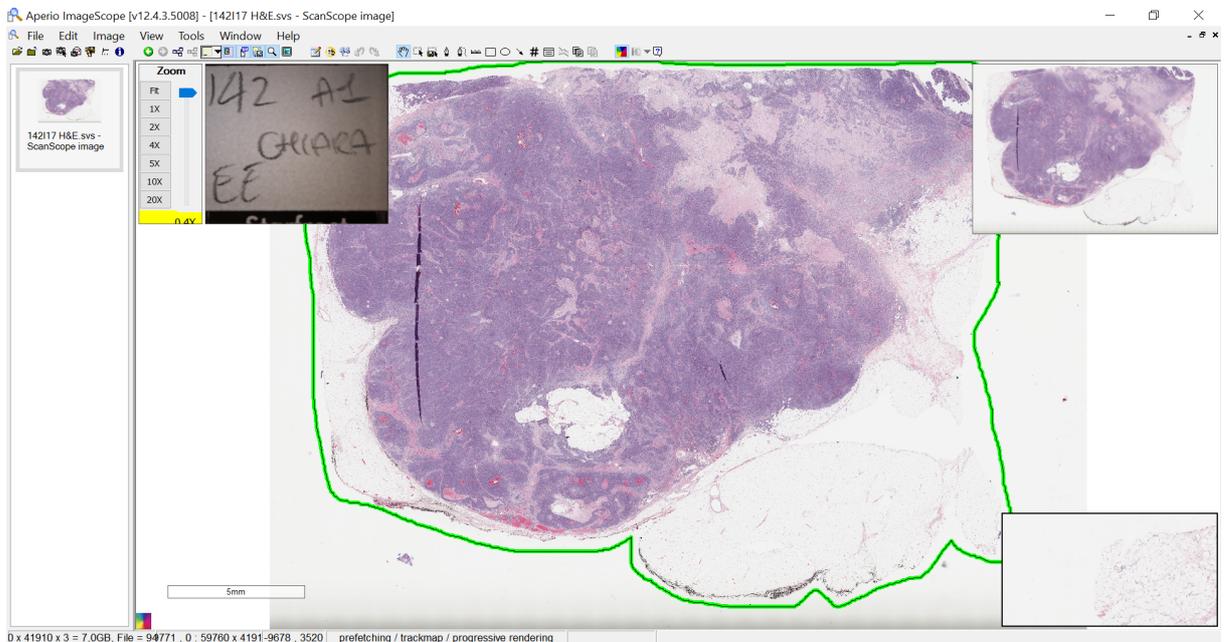


Figura 17- Segmentazione manuale di tutto il tessuto presente in una biopsia di mammella (2016I005002.ndpi) in colorazione immunoistochimica.



*Figura 18- Segmentazione manuale del tessuto di interesse istologico in una biopsia di mammella (2016I005002.ndpi) in colorazione immunoistochimica, nel riquadro rosso si evidenzia una parte di tessuto sfocata.*



*Figura 19- Segmentazione ridefinita della biopsia di mammella di fig. 14.*

## 2.2.1 Architettura degli algoritmi

Gli algoritmi di individuazione del tessuto *algoritmo\_stain* ed *algoritmo\_puzzle* seguono la struttura presentata dai flochart nelle figure 20 e 21:

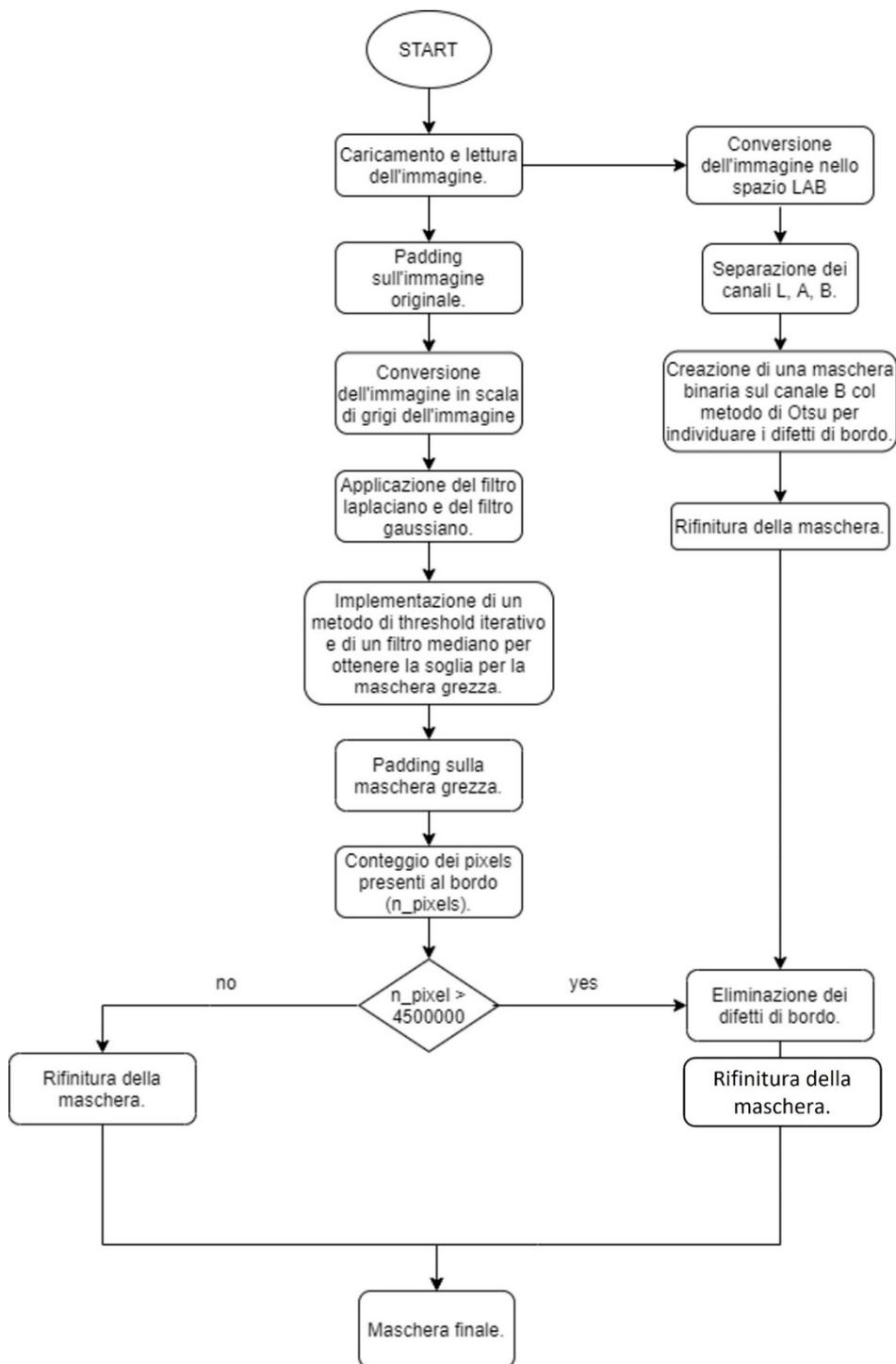


Figura 20 – Schema generale dell'algoritmo *tissue\_normalization*.

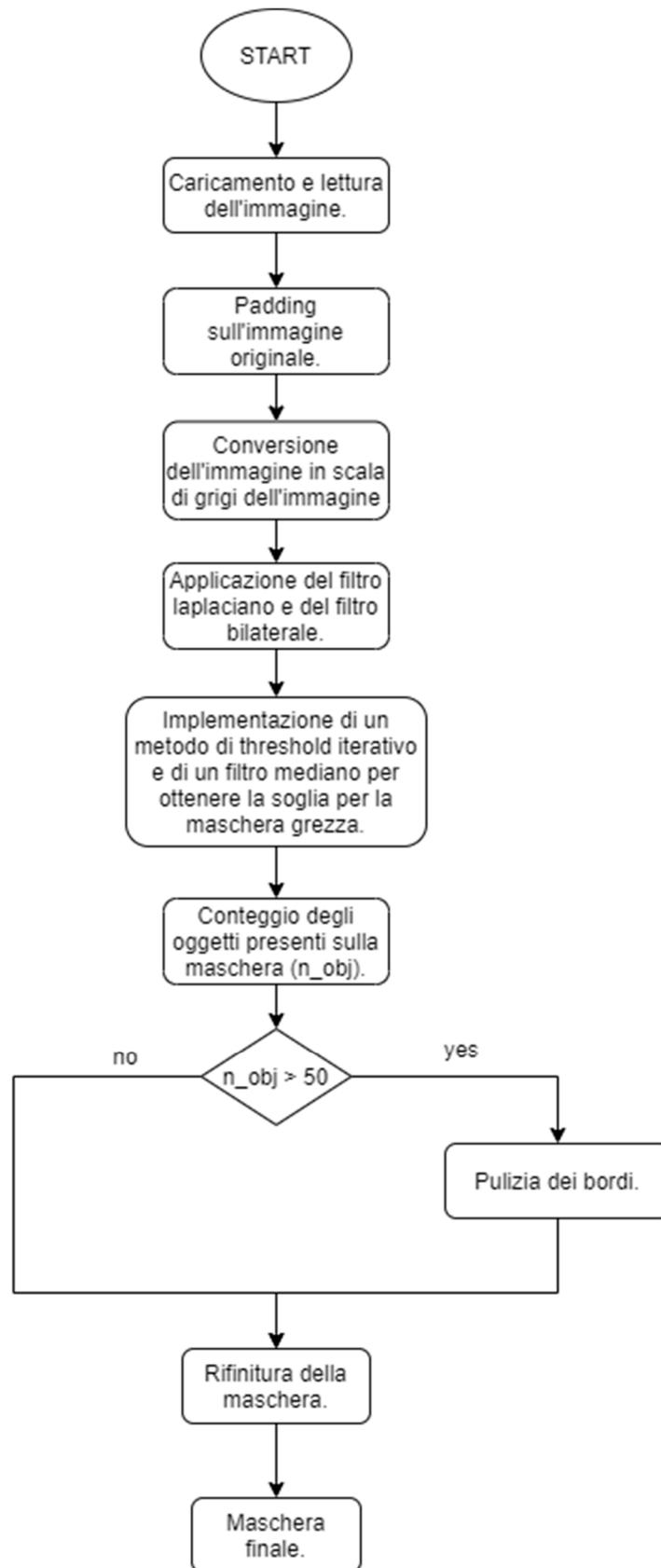
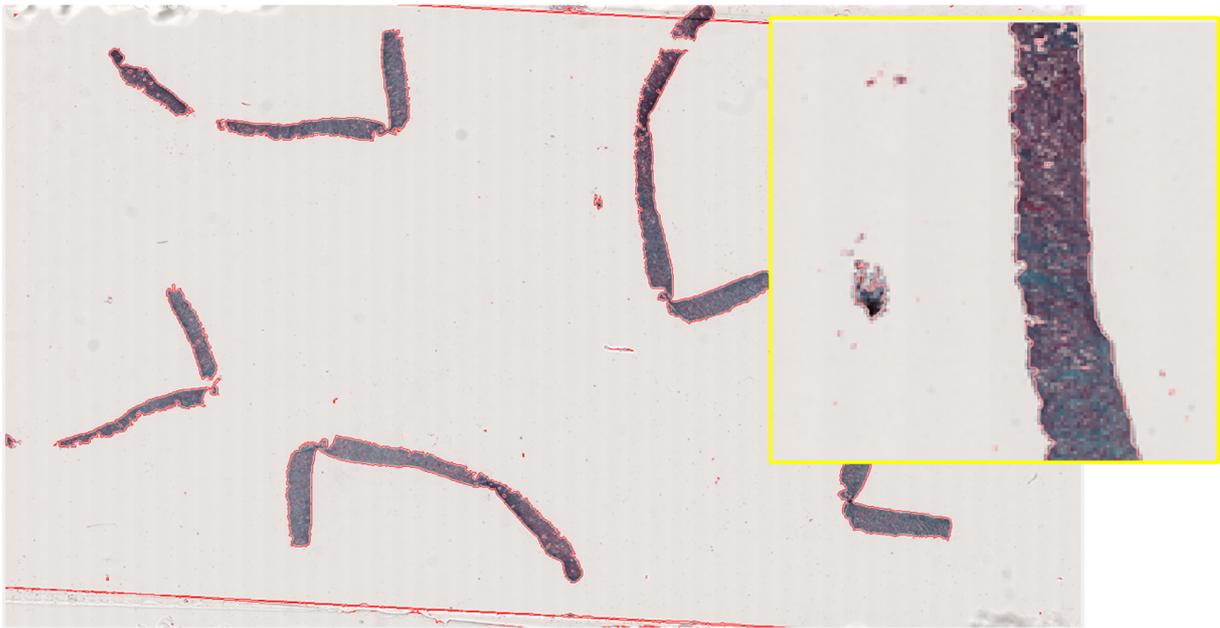


Figura 21 - Schema generale dell' algoritmo tissue\_puzzle.

### **2.2.2 Primo algoritmo: Tissue\_normalization**

Il dataset, suddiviso come mostrato in Tab.1 è caricato interamente e le immagini vengono analizzate una per una. Le colorazioni analizzate sono istologiche, istochimiche ed immunoistochimiche.

Idealmente lo sfondo di una WSI dovrebbe essere acromatico ed uniformemente bianco per facilitare l'analisi del tessuto in esame ma sfortunatamente in condizioni reali lo sfondo è caratterizzato da un'inconsistenza di colore dovuta principalmente al tipo e alla qualità dell'acquisizione dell'immagine, ad esempio la luce della sorgente può avere delle sfumature dal biancastro al giallo, i sensori utilizzati possono produrre più o meno rumore e i metodi di acquisizione possono essere tile-based e line-based. Entrambe le tecnologie utilizzano una camera integrata ed un sistema di movimento che scorre la WSI, le tecnologie tile-based catturano dei field of view quadrati sull'immagine fino a ricoprire l'intera WSI mentre le tecnologie line-based catturano delle strisce ininterrotte di tessuto [11]. Le tecnologie tile-based provocano artefatti di illuminazione più gravi ed amplificati vicino ai bordi. Quando lo sfondo è pressoché uniforme con un thresholding globale, che utilizza una soglia calcolata sulla base delle occorrenze del colore più lontano dallo sfondo, si riesce ad identificare il tessuto presente. Le regioni più problematiche restano quelle ai bordi del tessuto come mostrato in figura 22.



*Figura 22 - Segmentazione di tessuto effettuata su biopsia tricromica di rene (TRIC 1) con un thresholding globale, nel dettaglio evidenziato si nota come i bordi non vengano correttamente individuati dalla maschera escludendo delle piccole parti di tessuto.*

L'algoritmo presentato è il risultato della combinazione di diversi metodi base: filtraggi, thresholding e operazioni di rifinitura dell'immagine [10].

Il codice sviluppato in Python permette, tramite la libreria OpenSlide di leggere le immagini in formato .svs e .ndpi e sfruttare la struttura della WSI (Fig. 16) per lavorare sull'immagine alla risoluzione desiderata. Il primo step dell'algoritmo prevede la conversione nello spazio Lab per creare una maschera che contenga la maggior parte degli artefatti presenti al bordo, ad esempio linee, bolle e segni di pennarelli. Lo spazio Lab è uno spazio in cui L rappresenta la luminosità mentre a e b si riferiscono alle coordinate di cromaticità. Applicando al canale B una soglia calcolata col metodo di Otsu si ottiene una maschera contenente solo gli artefatti al bordo (Fig. 23) che può essere utilizzata per pulire la maschera finale. Tale maschera viene rifinita con operazioni morfologiche per renderla più omogenea.



*Figura 23 - In alto una biopsia di PAS di prostata (1123-17-H1) con evidenti difetti di bordo, in basso la relativa maschera che contiene solo i difetti.*

Gli steps di segmentazione del tessuto si effettuano invece sull'immagine in scala di grigi in modo da ricavare le informazioni sulle strutture presenti nell'immagine dalla componente luminosa.

Per individuare le strutture viene utilizzato un filtro laplaciano.

Tale filtro passa alto evidenzia i rapidi cambiamenti di intensità di colore ed è quindi utilizzato per individuare le discontinuità ed i bordi del tessuto.

L'immagine in input a tale filtro ha subito un padding simmetrico per evitare le problematiche legate ai bordi dell'immagine.

Il laplaciano lavora isotropicamente sulla derivata seconda dell'immagine:

$$L(x, y) = \frac{\partial^2(I)}{\partial(x^2)} + \frac{\partial^2(I)}{\partial(y^2)} \quad (5)$$

L'equazione 5 è risolta tramite convoluzioni, in questo caso per approssimare le derivate seconde è stato utilizzato il kernel 3x3 rappresentato di seguito:

0	1	0
1	-4	1
0	1	0

(6)

Il valore assoluto del filtro laplaciano è calcolato sull'immagine a cui viene applicato un filtro gaussiano molto forte ( $\sigma = 5.5$ ) per ridurre la sensibilità del laplaciano al rumore (Eq. 8).

$$GB[I]_p = \sum_{q \in S} G_\sigma(\|p - q\|) I_q \quad (8)$$

dove  $G_\sigma$  è il kernel gaussiano:

$$G(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (9)$$

e  $\sigma$  è la deviazione standard della distribuzione normale e definisce l'intorno del pixel in analisi. Il filtro gaussiano è un filtro lineare effettua una media pesata delle intensità dei pixels nell'intorno del pixel in posizione (x,y) con effetto passa basso ed utilizza dei pesi che decrescono proporzionalmente alla distanza dai pixels in analisi.

Poiché l'azione del filtro è indipendente da ciò che è presente nell'immagine [12] ciò che conta ai fini del calcolo dell'output è solo la distanza fra i pixels.

In questo modo si ottiene una buona base per evidenziare ed isolare tutte le strutture presenti nell'immagine in quanto all'applicazione dei filtri si ha una buona risposta dalle regioni di tessuto e una risposta molto basse dalle regioni a sfondo omogeneo come mostrato in figura 25.

Per la maschera viene calcolata iterativamente una soglia sull'output ottenuto dagli step precedenti ed infine, ottenendo il vettore contenente i valori delle soglie di ogni iterazione.

Il valore finale della soglia viene scelto applicando un filtro mediano, calcolando l'intorno destro del massimo valore trovato e selezionando il valore più vicino possibile alla variazione della curva del vettore come mostrato in figura 26c.

Dalla soglia ottenuta si ricava la maschera grezza (Fig. 26d) per poi passare alla rifinitura della stessa.

Alla maschera grezza viene effettuato un padding simmetrico per contare i pixels presenti al bordo.

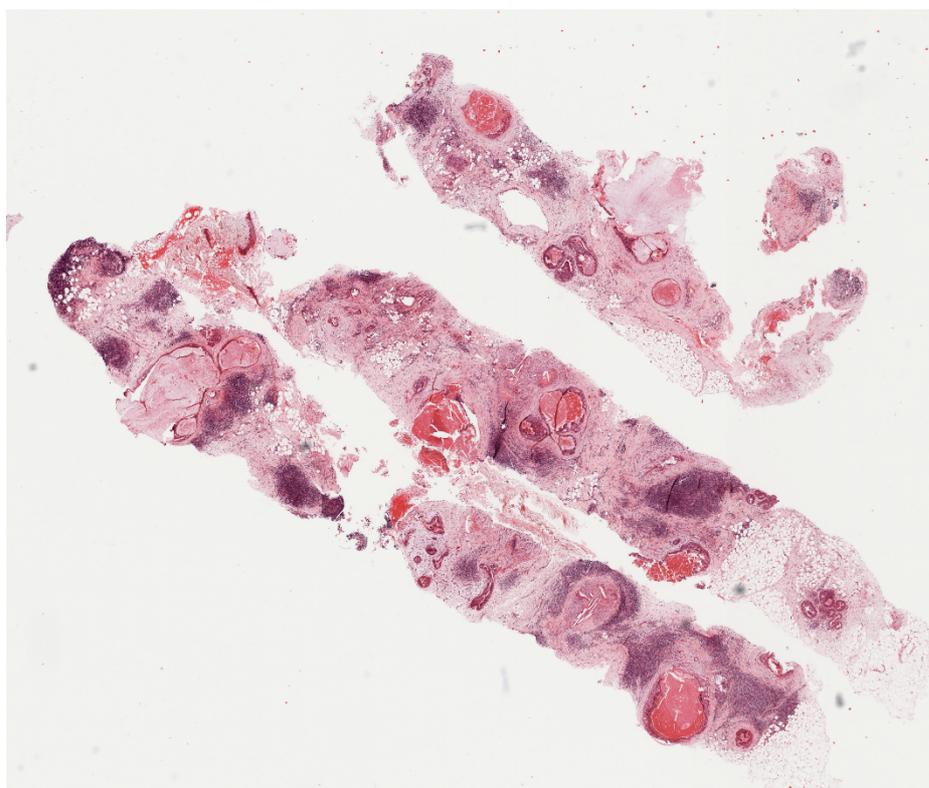
I pixels di bordo si configurano come artefatti nella maggior parte dei casi (Fig. 23), viene impostata una soglia manuale per determinare quando i pixels al bordo vanno eliminati. Le soglie ricavate empiricamente sono fissate a 200000 pixels per le immagini in colorazione istologica e istochimica e a 4500000 per la colorazione immunoistochimica.

Il controllo sui pixels di bordo è necessario per non rischiare di eliminare tessuto significativo qualora un frustolo della biopsia fosse vicino al bordo dell'immagine come mostrato in figura 24.

Quando il numero di pixels al bordo supera la soglia predefinita la maschera grezza viene ripulita utilizzando la maschera ottenuta dal canale B prima del processing.

Infine, la maschera è ridefinita tramite operazioni morfologiche elementare ad ottenere l'output finale (Fig. 26e):

- Chiusura morfologica e riempimento di tutti i buchi presenti nella maschera (presenti nella maschera grezza, in figura 26d);
- Apertura morfologica;
- Eliminazione delle linee sottili;
- Riempimento dei buchi rimasti alla fine del processing.



*Figura 24 – Immagine originale e maschera finale per una biopia di mammella in colorazione HE (11825).*

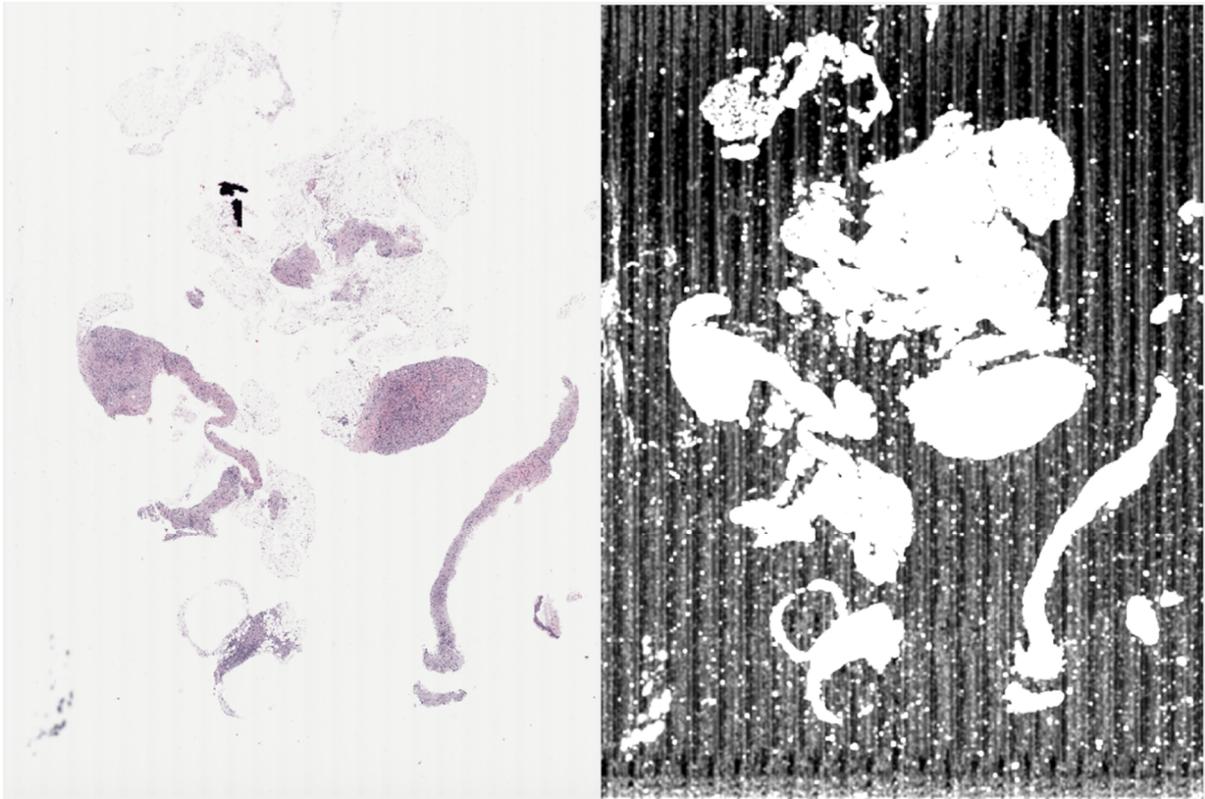


Figura 25 - Biopsia di mammella in colorazione HE (\_1006649) a destra. A sinistra la maschera ottenuta dalla combinazione di filtro laplaciano e gaussiano.

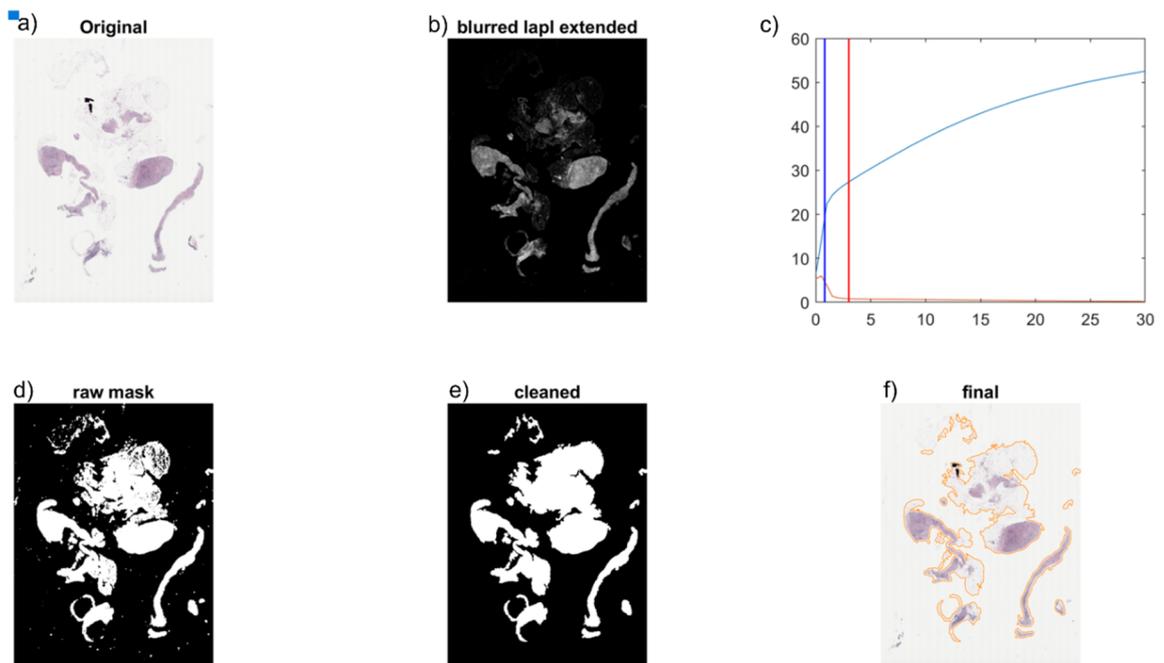


Figura 26 - a) WSI originale di una biopsia di mammella in colorazione HE. b) Maschera della biopsia dopo l'applicazione del filtro laplaciano e gaussiano. c) In rosso la soglia calcolata con un global thresholding. d) Maschera grezza. e) Maschera rifinita. f) Tessuto segmentato.

L'algoritmo riesce ad individuare correttamente tutte le parti di tessuto, comprese le zone di adipe (Fig. 26f) e a mantenere i bordi ben definiti senza perdere tessuto (Fig. 28).

Rispetto ad un qualunque algoritmo che lavora su un canale specifico dell'immagine, ad esempio il canale Q dopo la conversione dallo spazio RGB a quello YIQ, l'algoritmo presentato funziona su tutte le tipologie di colorazione in maniera corretta (prf. 2.5).

Questo algoritmo viene utilizzato per ricreare la maschera per isolare il tessuto da normalizzare con il tool di stain normalization.

### 2.2.3 Secondo algoritmo: Tissue\_puzzle

La prima versione dell'algoritmo presentata nel paragrafo precedente è finalizzata alla segmentazione di tutto il tessuto presentato tarando il giusto trade off tra percentuale di tessuto individuato ed artefatti individuati.

Una seconda versione di questo algoritmo è stata necessaria per individuare delle maschere che segmentassero in maniera corretta il tessuto, escludendo però sia gli artefatti che le parti di tessuto poco significative o che non contenessero in maniera evidente gli stains presenti nell'immagine. Queste maschere sono utilizzate infatti per estrarre delle patches significative per ricreare un'immagine target che racchiuda tutta l'informazione colorimetrica della biopsia.

Come descritto in figura 21 anche in questo caso la maschera viene individuata lavorando sull'immagine in scala di grigi.

Lo step successivo prevede l'individuazione delle strutture presenti nell'immagine, in questo caso il filtro laplaciano è stato applicato insieme ad un filtro bilaterale (Eq. 9).

$$BF [I]_p = \frac{1}{W_p} \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(I_p - I_q) I_q \quad (9)$$

dove  $W_p$  è un fattore di normalizzazione:

$$W_p = \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(I_p - I_q) \quad (10)$$

Il filtro bilaterale è un filtro non lineare che effettua una somma pesata dei pixels nel vicinato del pixel in analisi in posizione  $(x,y)$ ; i pesi dipendono sia dalla distanza spaziale che dalle intensità [14]. A differenza del filtro gaussiano questa media pesata è effettuata tenendo

conto delle variazioni di intensità che permettono di preservare i bordi delle strutture dell'immagine.

I parametri  $\sigma_r$  e  $\sigma_s$  gestiscono il filtraggio da applicare all'immagine, nell'equazione 9, che rappresenta una media pesata, i termini  $G_{\sigma_r}$  e  $G_{\sigma_s}$  sono rispettivamente un range gaussiano e una gaussiana spaziale (Fig.27) e le relative deviazioni standard gestiscono rispettivamente l'influenza dei pixels ad una certa distanza e dei pixel con un'intensità differente da  $I_q$  [15]

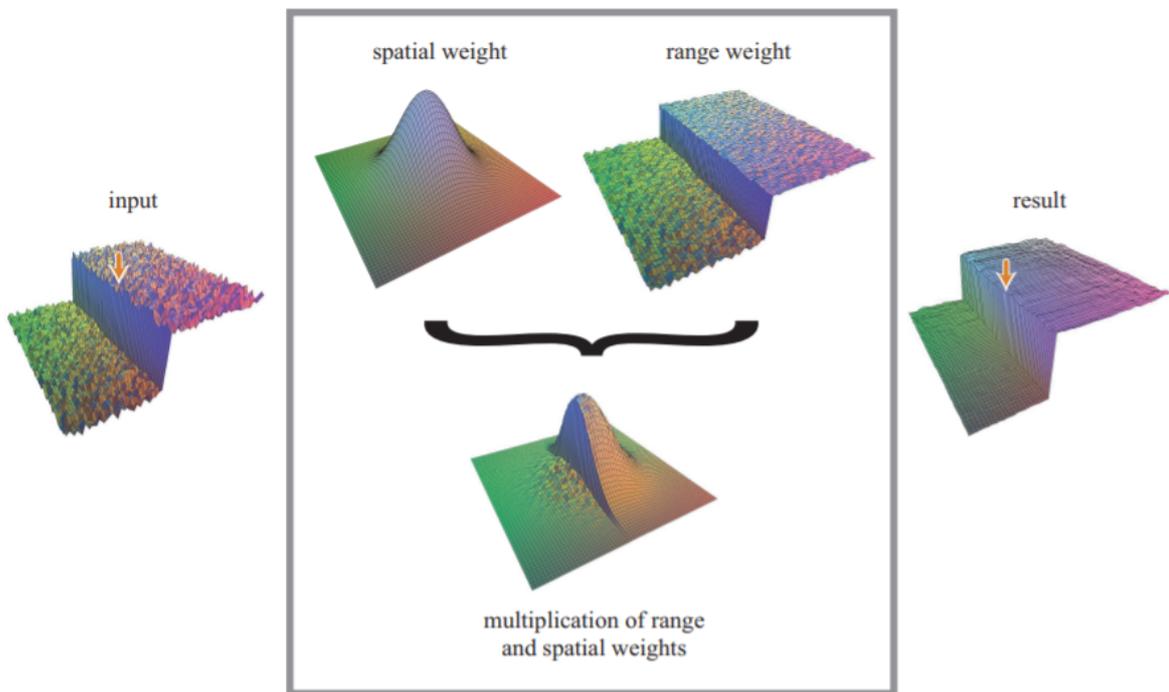
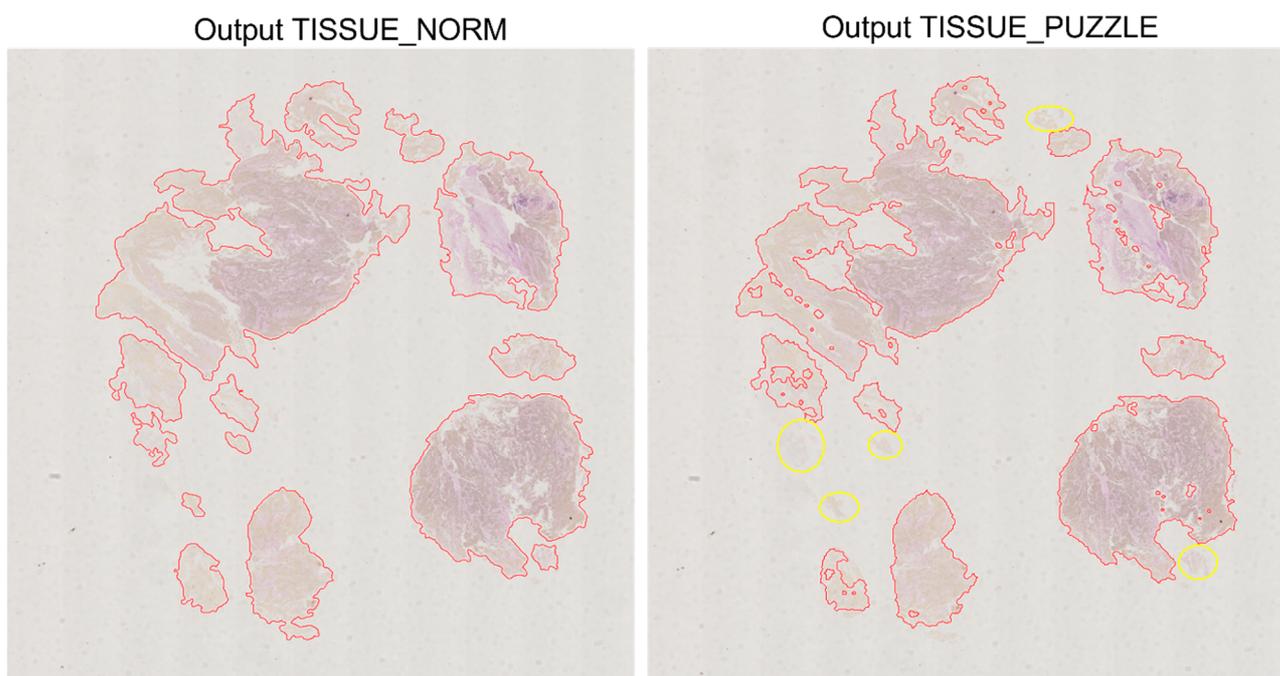


Figura 27- Il filtro bilaterale sfoca un'immagine di input preservando le discontinuità di bordo. Ogni pixel è sostituito da una media pesata dei pixels adiacenti. Ogni pixel del vicinato è pesato in modo tale da penalizzare i pixels con un'intensità fortemente distante da quella del pixel in analisi. La combinazione della gaussiana col range mostrato in figura fa in modo che solo pixels con le giuste intensità contribuiscano all'output finale [14].

All'aumentare del parametro  $\sigma_r$  il filtro bilaterale si avvicina al comportamento di un filtro gaussiano poiché la deviazione standard gaussiano tende ad essere più piatta e costante sull'intervallo di intensità dell'immagine, mentre all'aumentare di  $\sigma_s$  la sfocatura è amplificata [12].

L'utilizzo di un filtro bilaterale permette di mantenere le discontinuità ai bordi e di segmentare il tessuto ignorando gli spazi bianchi presenti anche all'interno della figura come mostrato in figura 28.



*Figura 28 - Segmentazione di una biopsia PAS di polmone con i due algoritmi, a destra TISSUE\_NORMALIZATION e a sinistra TISSUE\_PUZZLE. Nel primo caso viene detectato tutto il tessuto, nel secondo caso si presta più attenzione ad individuare solo le sezioni significative a livello di presenza degli stain, escludendo gli spazi bianchi.*

Come si può notare dalla figura 28, l'output del secondo algoritmo esclude alcune parti piccole di tessuto evidenziate in giallo, questo perché i due algoritmi si differenziano anche a livello di post processing della maschera grezza.

La soglia è calcolata come per l'algoritmo *tissue\_normalization*.

In questo caso si effettua solo un'operazione morfologica di apertura ed una pulizia dei bordi basata sul conteggio degli oggetti della maschera.

In questo caso risulta più efficace il conteggio degli oggetti presenti nella maschera piuttosto che il conteggio dei pixels poiché nel post processing non viene effettuato un filling della maschera.

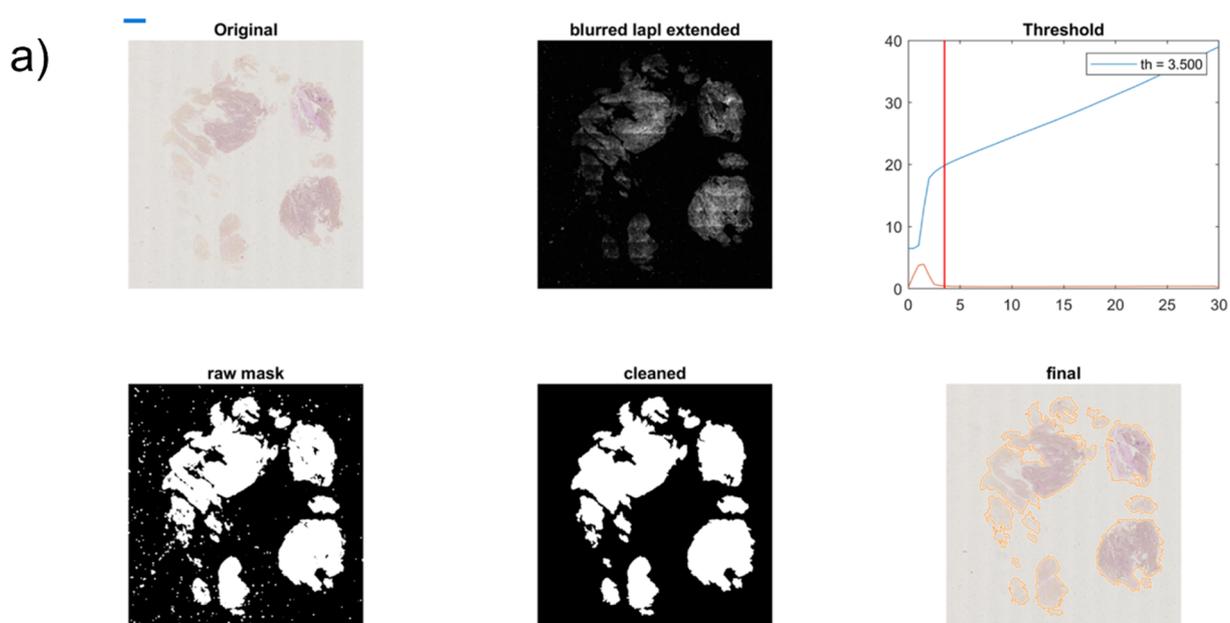
Riempiendo le strutture con un filling ed effettuando delle operazioni morfologiche il numero di strutture presenti sulla maschera si riduce solo a quelle rappresentative del tessuto.

In questo algoritmo non viene effettuato il filling sulla maschera grezza, quindi, risulta più facile ripulire i bordi contando le strutture presenti nella maschera, che in questo caso saranno in numero maggiore ma non saranno di interesse istologico.

D' altra parte in questo caso si può accettare una pulizia dei bordi meno accurata in quanto lo scopo ultimo della variazione dell' algoritmo è solo quello di individuare parti significative di tessuto che contengano tutti gli stains presenti.

Di seguito sono riportati i risultati dei due algoritmi per la stessa immagine, una biopsia polmonare in colorazione PAS (Fig.29).

In ordine in figura 29 sono rappresentate l'immagine originale, la maschera in cui vengono applicati rispettivamente i filtri, il calcolo della soglia, la maschera grezza, ripulita ed infine il risultato finale della segmentazione.



b)

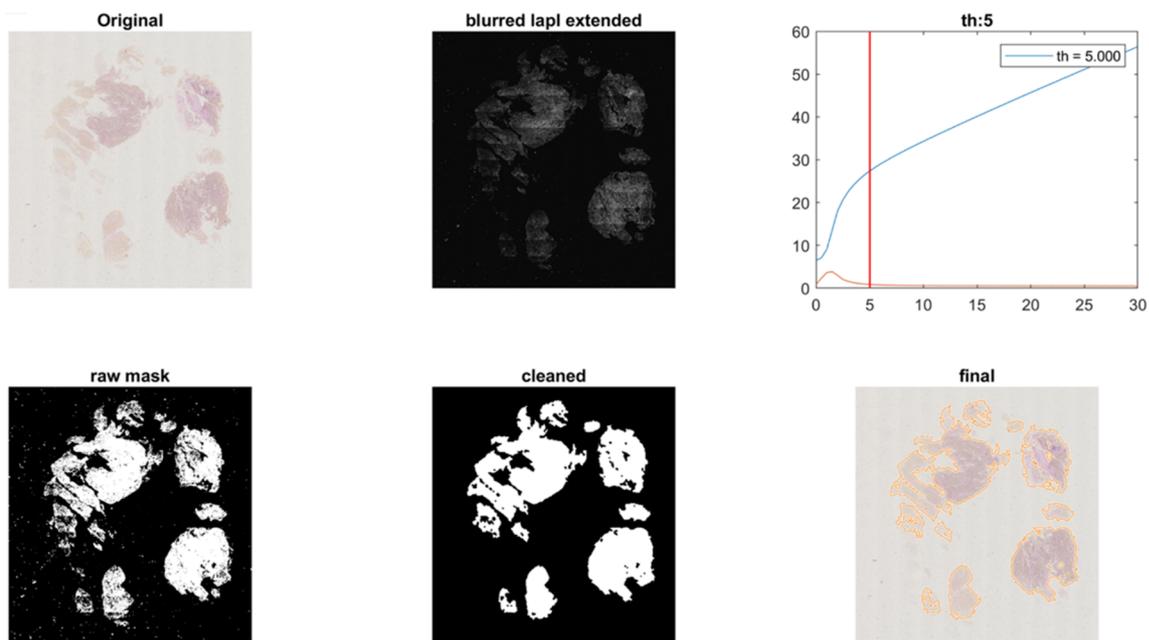


Figura 29 - Risultati degli algoritmi di tissue detection: a) tissue\_normalization; b) tissue\_puzzle.

## 2.3 Puzzle Image

Le WSIs sono estremamente onerose da processare, un modo più rapido è quello di estrarre delle patches significative che abbiano determinate caratteristiche.

Tali patches vengono assemblate a formare un'immagine, definita puzzle image come quella in figura 13c.

L'algoritmo è descritto in figura 30.

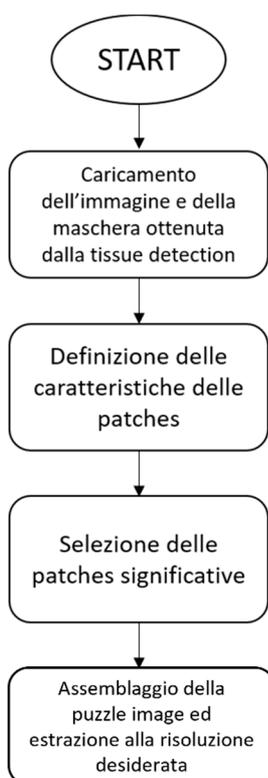


Figura 30- Schema dell'algoritmo di puzzle image.

Per lo sviluppo di questo algoritmo vengono utilizzate le maschere di tessuto ottenute con l'algoritmo *tissue\_puzzle*.

L'obiettivo è creare delle immagini puzzle rappresentative:

- che contengano entrambi gli stains;
- che non contengano artefatti;
- che contengano una percentuale bassa di spazi bianchi.

Come descritto nel paragrafo 2.2.3 le maschere ottenute individuano le parti significative di tessuto e permettono di individuare delle patches significative a livello istologico.

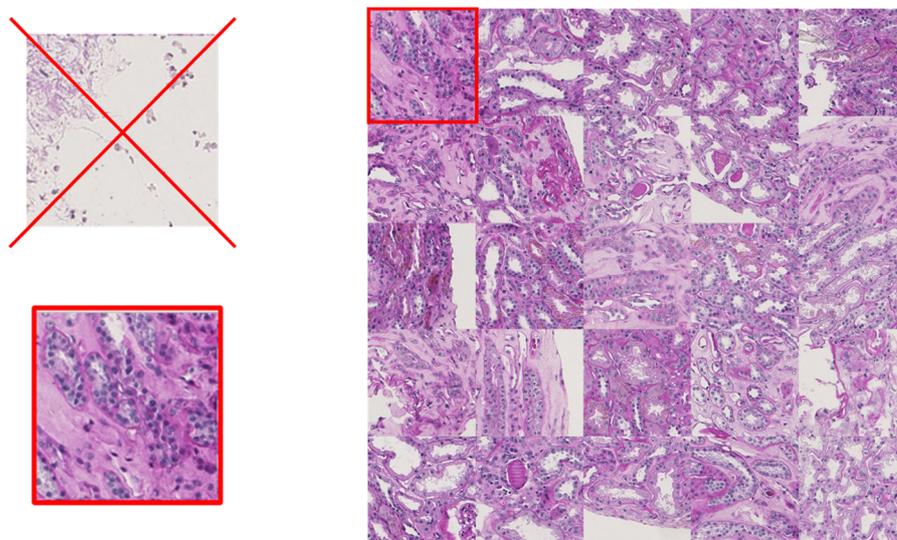
La maschera binaria contiene un numero di oggetti pari al numero di frustoli o sezioni di tessuto presenti sulla WSI, si parte calcolando l'area di ogni oggetto presente e quindi l'area totale di tessuto presente.

Vengono poi settate le caratteristiche principali necessarie per la selezione:

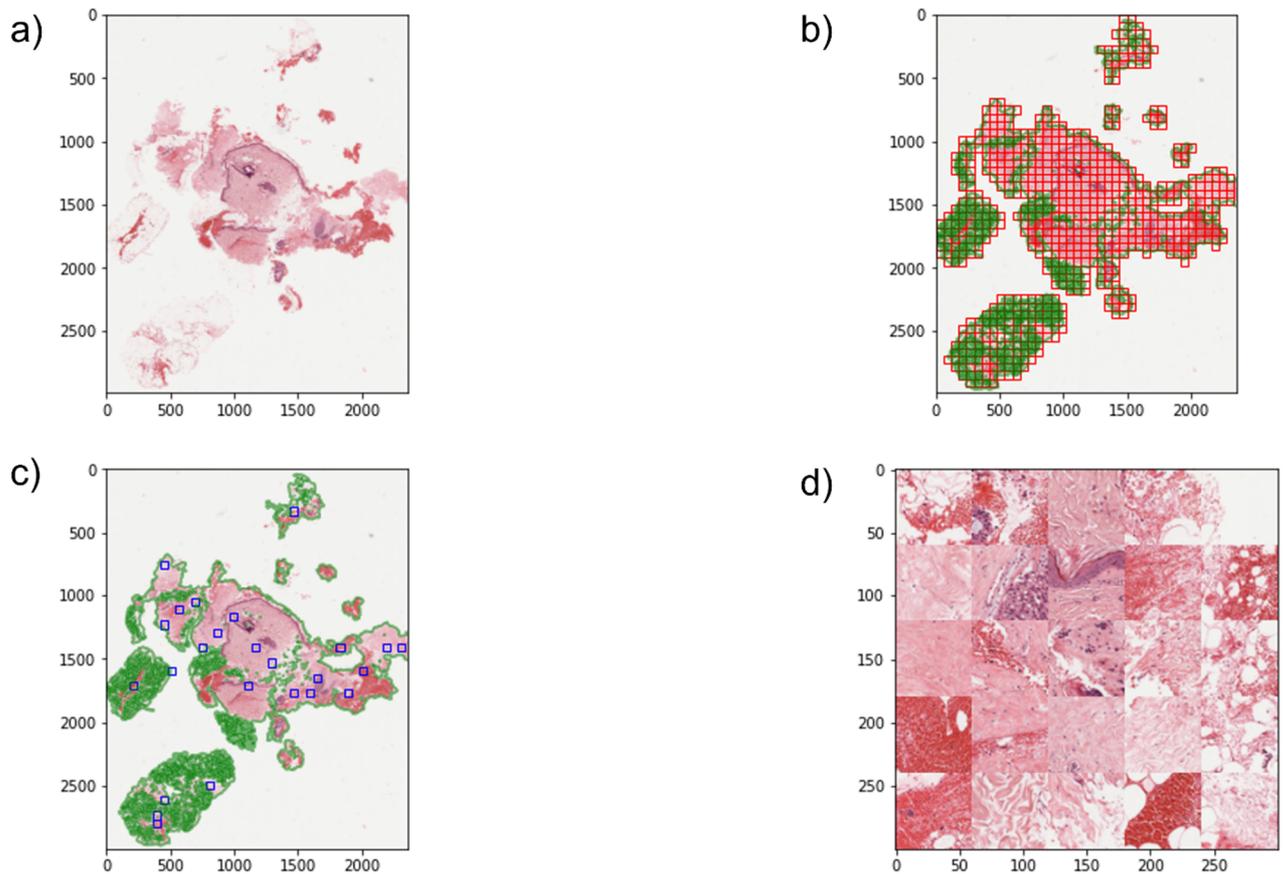
- dimensioni delle patches;
- percentuale di tessuto presente;
- numero massimo di patches selezionabili in base all'immagine in analisi.

La dimensione scelta per le patches è di 60x60, partendo dall'area di tessuto presente si individua il numero massimo di patches di quelle dimensioni che può ricoprire tutto il tessuto presente, la griglia ottenuta è mostrata in figura 32b.

Tra le patches che ricoprono tutto il tessuto, vengono selezionate solo quelle significative, tale selezione è effettuata selezionando solo le patches rappresentate al 90% da tessuto istologico rilevante (Fig. 31).



*Figura 31 - Le patches con una percentuale di tessuto troppo basse vengono escluse e non entrano a far parte della puzzle image finale (biopsia di rene PAS 467).*



*Figura 32 – Creazione della puzzle image per la biopsia di mammella 1173. a) Immagine originale; b) individuazione di tutte le patches selezionabili; c) patches selezionate; d) estrazione delle patches e assemblaggio della puzzle image.*

Come si può notare da figura 32c, le patches non sono state selezionate unicamente sulla base della percentuale di tessuto presente.

Basandosi solo su tale percentuale le patches sarebbero state selezionate su una sola sezione di tessuto (quella con area maggiore) per cui è stato aggiunto un criterio di selezione che permette di selezionare  $n$  patches da ogni sezione in proporzione all'area della stessa ed uno basato sulla distanza per evitare di selezionare patches troppo vicine tra loro e quindi rappresentative della stessa parte di tessuto.

Il secondo criterio citato si basa sul calcolo della distanza minima più grande tra la patch corrente e la prima patch selezionata.

Per l'assemblaggio della puzzle image si calcolano il numero di patches da inserire per riga e per colonna a partire dal numero massimo di patches selezionabili a formare immagini che contengano 1, 2, 3, 4, 6, 8, 9, 12, 15, 16, 20, 24 o 25 patches. Ad esempio, l'immagine in figura 32d contiene il massimo numero di patches assemblabili, ossia 25.

Le immagini ottenute sono pienamente rappresentative del tessuto e compatibilmente alla bontà della maschera utilizzata non contengono artefatti, possono essere estratte alla risoluzione desiderata sfruttando la struttura piramidale del vetrino.

## 2.4 Stain normalization

Il metodo di normalizzazione utilizzato in questo lavoro di tesi è quello implementato da Macencko et. al, un algoritmo stabile e computazionalmente poco oneroso (Fig. 11) che sarà unicamente la base di normalizzazione.

Utilizzando l'algoritmo di Macencko sull'intera WSI in formato nativo i tempi computazionali aumentano esponenzialmente, l'obiettivo di questo lavoro è estendere correttamente la normalizzazione all'intera WSI con dei tempi computazionali ridotti al minimo.

### 2.4.1 Algoritmo di Macencko

Come discusso nel paragrafo 1.2.3 l'immagine viene convertita nello spazio colore OD:

$$OD = -\log_{10}(I) \quad (11)$$

dove  $I$  è l'immagine RGB.

Questa trasformazione permette di ottenere una relazione lineare (fig. 6) tra i due stains, c'è un vettore specifico per ogni stain presente nell'immagine ed il colore risultante è la combinazione lineare di questi due.

Si inizializza la tolleranza per i percentili del valore pseudo-minimo e del valore pseudo-massimo,  $\alpha^{th}$  e  $(100 - \alpha)^{th}$  (), sperimentalmente il valore migliore per  $\alpha$  è pari ad 1.

Un pixel con valore OD pari a 0 è interamente bianco, per ragioni di stabilità i pixels con un valore di OD molto basso, quindi caratterizzati da assenza di stain, vengono eliminati tramite

una soglia  $\beta = 0.15$  ed i restanti valori di OD ( $OD$ ) sono divisi in due matrici (Eq. 12 e 13):  $H$ , valori di densità degli stain e  $W$ , matrice degli stain.

$$OD = W * H \quad (12)$$

$$H = W' * OD \quad (13)$$

Si calcola poi il singolo valore della decomposizione SVD dei valori nello spazio OD.

La distanza geodetica può essere usata per trovare la direzione dove proiettare i pixels trasformati nello spazio OD con lo scopo di trovare gli end points che individuino i vettori degli stains.

Con l'aiuto del piano individuato dagli autovettori relativi agli autovalori più grandi ricavati dalla scomposizione SVD si proiettano i valori OD e si normalizzano alla lunghezza unitaria. La linea proiettata è una curva (Fig. 33a).

L'angolo rispetto alla prima o alla seconda direzione della SVD è calcolato per ogni punto mappando le direzioni nel piano come scalari. In figura 33b è mostrato l'istogramma di questi valori.

L'intensità dell'output finale è proporzionale alle intensità dei vettori degli stains presenti. I percentili  $\alpha^{th}$  e  $(100 - \alpha)^{th}$  inizializzati in precedenza possono essere utilizzati per ottenere un'approssimazione robusta degli estremi [2].

Data un'immagine RGB nello spazio  $R^{m,n}$ , dove  $m$  è il numero di stain presenti ed  $n$  il numero di pixels per canale si otterrà una matrice  $W$  nello spazio  $R^{m,r}$ , dove  $r$  indica il numero delle colorazioni principali ed ogni colonna rappresenta una base colore.

La matrice di colorazione  $H$  di ogni colorante apparterrà invece allo spazio  $R^{r,n}$ .

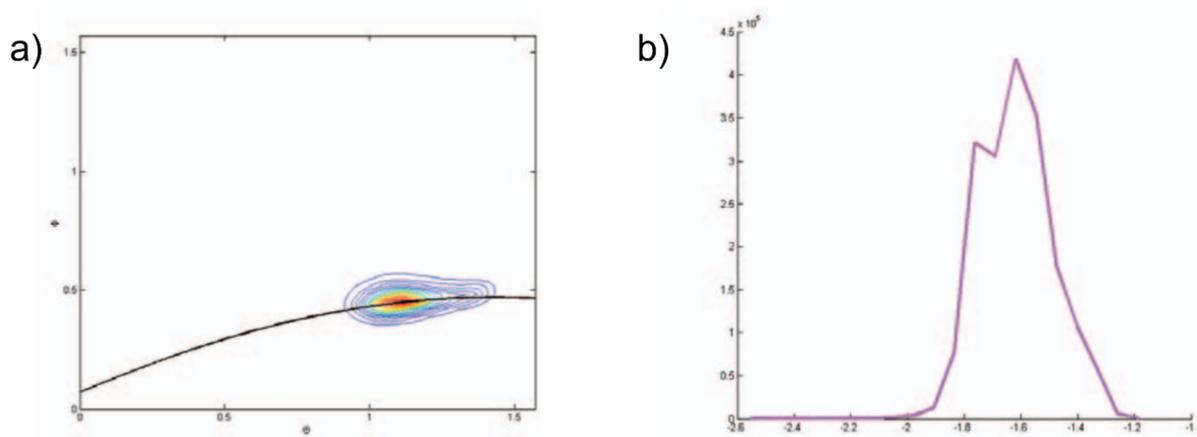


Figura 33 - a) Contorni dell'istogramma dei pixels, la linea di cresta è calcolata mediante la decomposizione SVD ed è rappresentata in nero come una geodetica; b) Istogramma degli angoli che i punti formano con la linea geodetica in figura 33a.

Macencko calcola una matrice  $W$  fissa dalla quale poi si ricava la matrice  $H$  d'interesse per la separazione degli stains (Fig. 34).

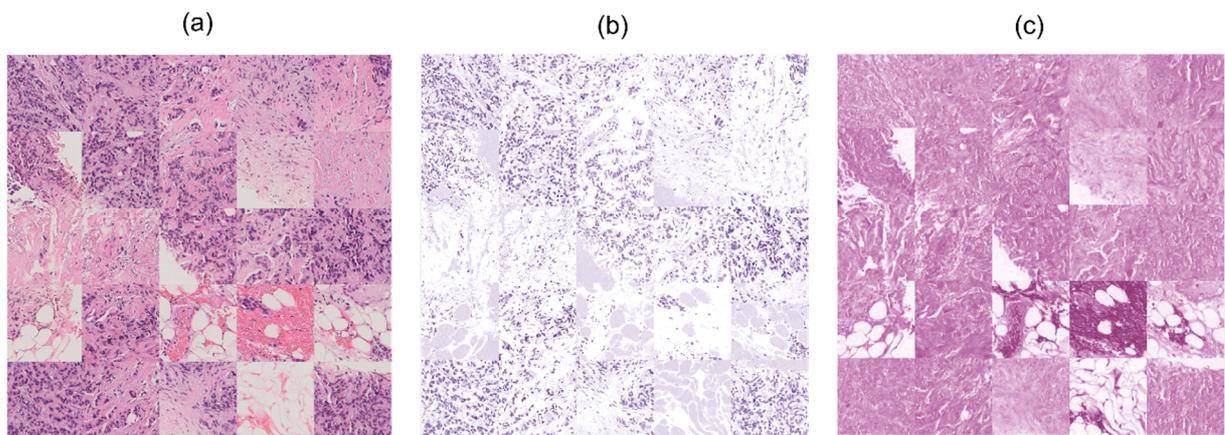


Figura 34- (a) Puzzle image originale in colorazione HE; (b) Stain corrispondente all'ematosilina; (c) Stain corrispondente all'eosina.

## 2.4.2 Algoritmo di normalizzazione dell'intera WSI

L'algoritmo di normalizzazione descritto nel paragrafo 2.4.2 è il cuore di un algoritmo più articolato che permette di normalizzare l'intera WSI.

Di seguito è riportato il flowchart logico dell'intero algoritmo:

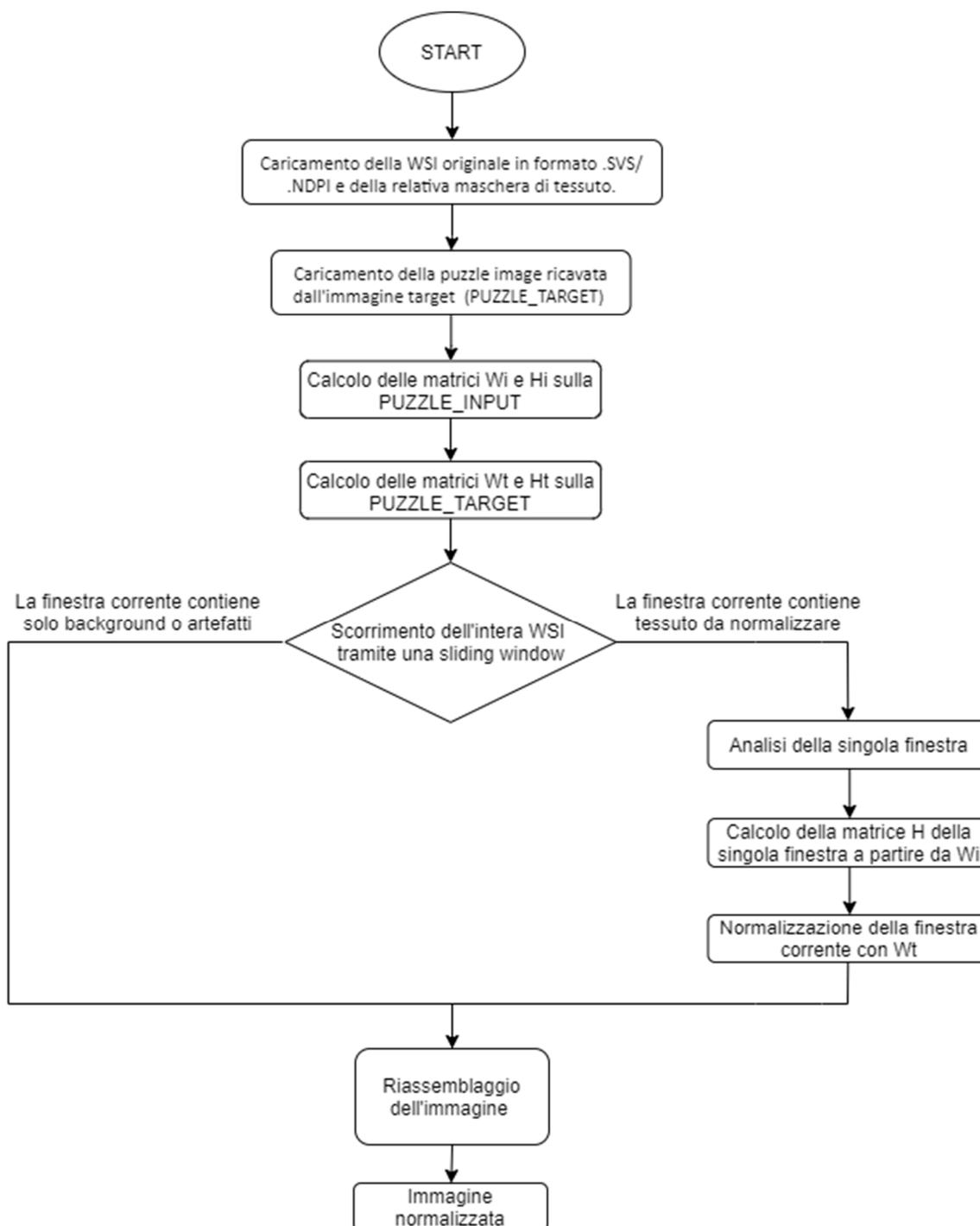


Figura 35 - Flowchart dell'algoritmo di normalizzazione.

L'algoritmo carica l'immagine in formato nativo e permette l'analisi della stessa alla risoluzione desiderata sfruttando la struttura della WSI mostrata in figura 16.

Le puzzle image sono estratte tramite l'algoritmo descritto al paragrafo 2.3 alla risoluzione desiderata, tale risoluzione può essere compatibile o meno con la risoluzione dell'immagine originale.

### **2.4.3 Selezione delle immagini target**

Per effettuare una normalizzazione di colori sono necessarie due immagini, un'immagine sorgente da normalizzare ed un'immagine target dalla quale mappare le caratteristiche colorimetriche per trasferirle all'immagine sorgente.

Preliminarmente vengono scelte delle immagini target per ogni colorazione analizzata ed ogni tipologia di tessuto.

Un'immagine target deve contenere delle informazioni colorimetriche più definite possibili ai fini di normalizzare correttamente l'immagine di input, nello specifico:

- deve contenere entrambi gli stain in maniera evidente e definita;
- non deve contenere artefatti e sfocature;
- gli stains presenti devono potersi separare in maniera corretta;
- la dinamica degli stain deve essere pressoché a metà strada tra la colorazione più scura e la colorazione più chiara presenti nel subset di riferimento.

In figura 36 sono riportate due immagini puzzle su cui è stata effettuata la separazione degli stains.

Per l'immagine (1) si nota come gli stains siano correttamente quantificati e separati in quanto le due immagini 36 (1b) e 36 (1c) contengono strutture diverse e perfettamente complementari tra loro.

In figura 36 (2b) e 36 (2c) gli stains non sono separati in maniera ottimale, difatti in figure 36 (2c) vengono individuati alcuni nuclei sul colore di background.

Per la corretta normalizzazione dell'immagine è strettamente necessario che la separazione degli stain avvenga in maniera corretta poiché le matrici W ed H calcolate sull'immagine target

vengono utilizzate come base per il calcolo della matrice delle densità degli stain dell'immagine normalizzata.

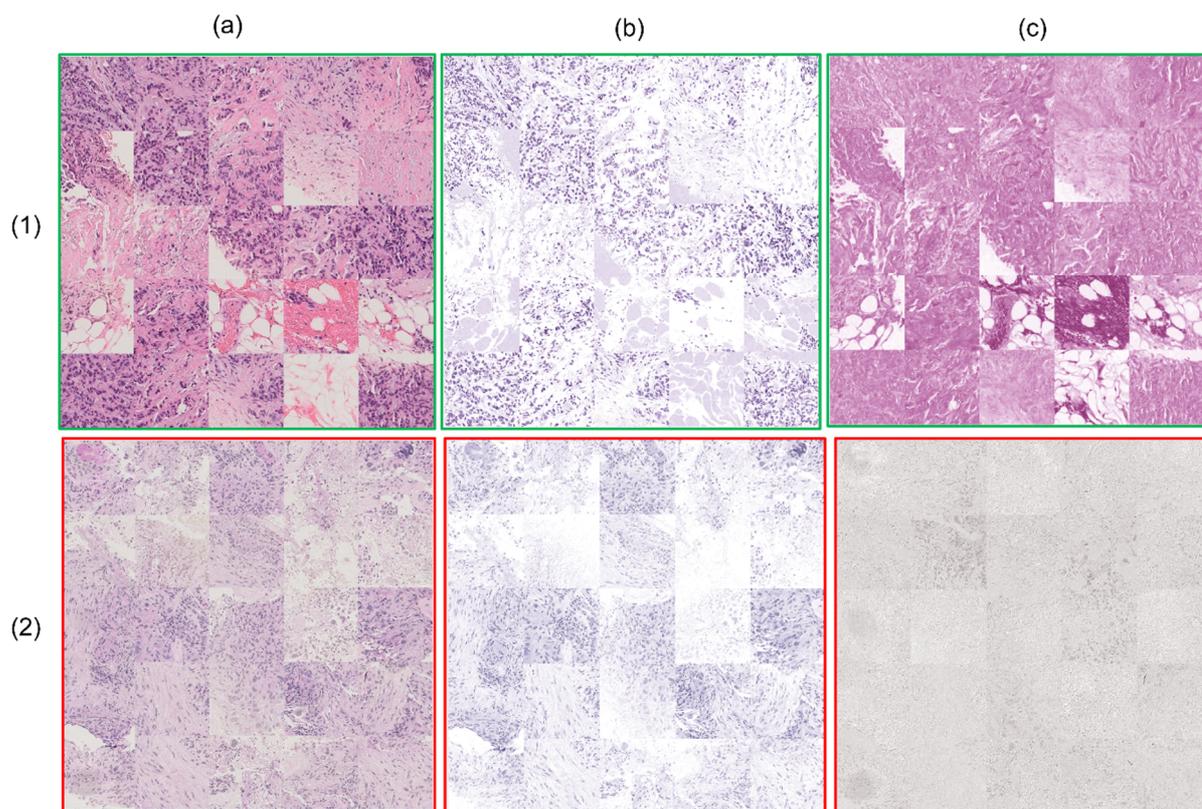


Figura 36 - (1) Puzzle image ricavata da una biopsia di mammella in colorazione HE - (a) Puzzle image originale; (b) Primo stain individuato; (c) Secondo stain individuato. (2) Puzzle image ricavata da una biopsia di polmone in colorazione PAS - (a) Puzzle image originale; (b) Primo stain individuato; (c) Secondo stain individuato.

#### 2.4.4 Sliding window e normalizzazione della finestra corrente

Dopo aver selezionato le immagini target si procede alla normalizzazione delle WSI.

Una WSI in formato nativo è dell'ordine dei gigabyte, in questo caso la normalizzazione sull'intera immagine è computazionalmente onerosa e richiede delle tempistiche estremamente lunghe.

L'idea implementata in questo algoritmo è quella di suddividere l'immagine alla risoluzione scelta in più finestre con l'ausilio di una sliding window (Fig. 37).

L'algoritmo permette di individuare finestre di grandezza variabile sulla WSI.

Per permettere lo scorrimento della finestra delle dimensioni scelte, viene inserito un controllo che individua la dimensione del padding da effettuare sulle righe e sulle colonne dell'immagine per ottenere una matrice di ordine  $N \times N$  dove  $N$  è proporzionale alla dimensione della finestra.

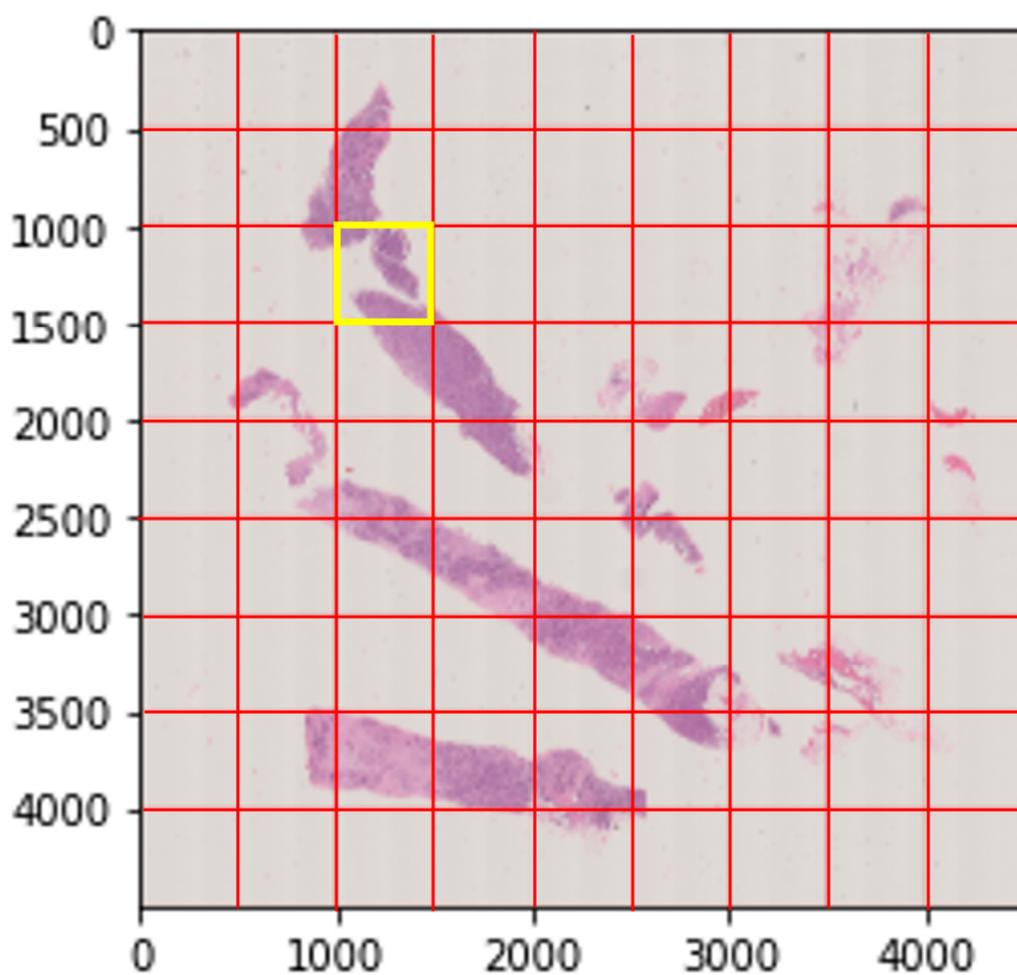


Figura 37 – Individuazione della griglia sulla quale scorrerà la sliding window.

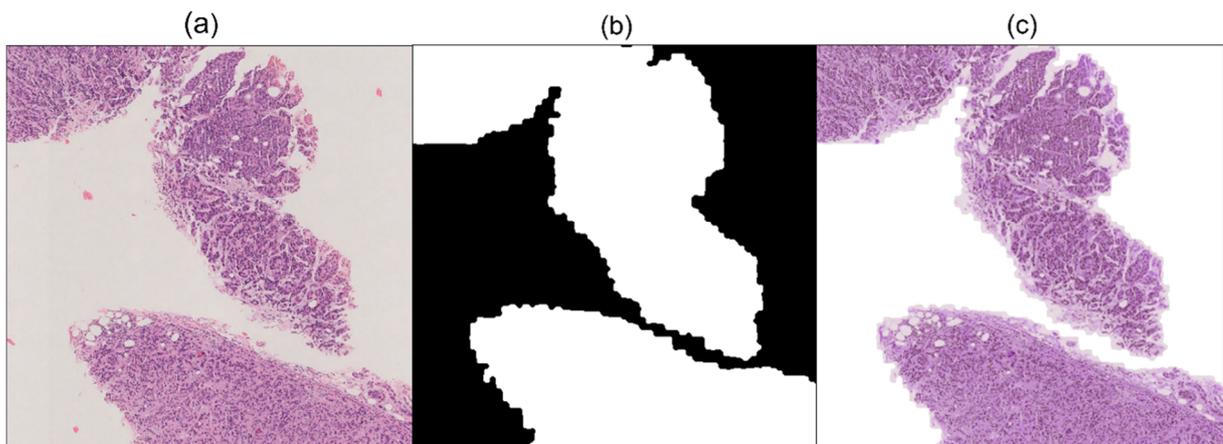


Figura 38 – Zoom della finestra evidenziata in giallo in figura 37. (a) Finestra originale (b) Maschera della finestra. (c) Finestra normalizzata.

La normalizzazione non viene effettuata sull'intera immagine ma sulla finestra corrente, la figura 38 ne riporta un esempio.

L'algoritmo carica in memoria la WSI corrente, la relativa maschera ottenuta con *puzzle\_normalization* (par. 2.2.2) la sua puzzle image e l'immagine target scelta.

Si calcolano le matrici degli stain  $W_i$  e  $W_t$  e di densità degli stain  $H_i$  e  $H_t$ , rispettivamente per l'immagine puzzle di input e per l'immagine puzzle usata come target.

La sliding window inizia a scorrere la WSI e la maschera contemporaneamente individuando di una volta in volta la finestra corrente (Figg. 38b e 38c).

Su tale finestra viene calcolata la matrice di densità degli stains  $H_{norm}$  a partire dalla matrice degli stain  $W_i$ .

Tale scelta è effettuata sulla base del fatto che la matrice  $W_i$  è relativa alla WSI originale e quindi strettamente correlata alla finestra corrente ritagliata dalla WSI.

Non tutte le finestre contengono tessuto, alcune sono parzialmente o totalmente rappresentate da background; prima di effettuare la normalizzazione si valuta la composizione della finestra selezionando grazie alla maschera solo la parte contenente tessuto.

La matrice  $H_{norm}$  è quindi calcolata solo per la parte della finestra rappresentativa del tessuto, qualora la finestra sia parzialmente rappresentata dello sfondo, la parte relativa allo sfondo non viene normalizzata, in output avrò il tessuto normalizzato circondato da uno sfondo

bianco (in figura 38c si nota come la parte rappresentata dallo sfondo in figura 38a sia totalmente bianca).

Se la finestra invece non contiene assolutamente tessuto, nell'output normalizzato è interamente nera.

Nota la matrice  $H_{norm}$ , il tessuto presente viene normalizzato utilizzando la matrice degli stain ricavata dal target  $W_t$ .

La nuova finestra viene inserita in una nuova matrice esattamente nella posizione in cui si trovava nella WSI originale (Fig.39e).

In conclusione, lavorando sulla maschera binaria (Fig. 39b), viene individuato solo lo sfondo originale (Fig 39c) per sostituire il tessuto bianco o nero delle parti che non interessavano la normalizzazione.

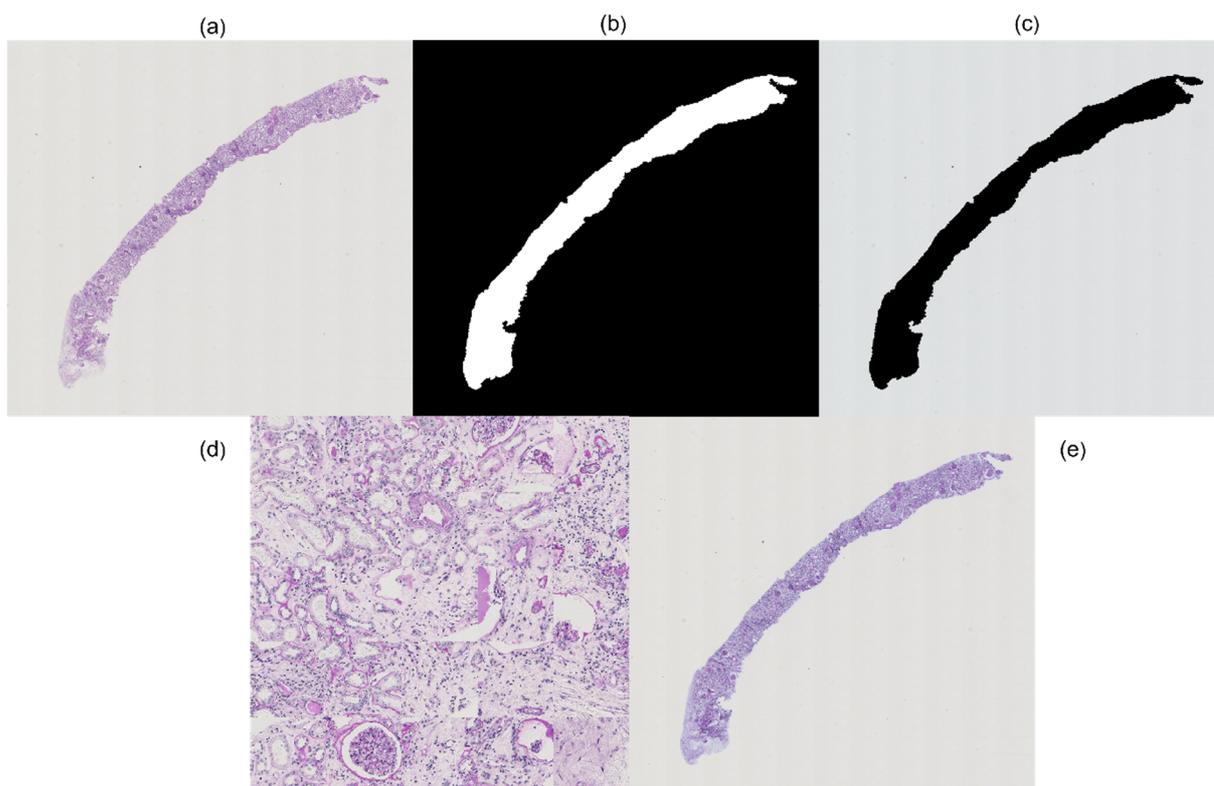


Figura 39 - Biopsia di rene (PAS 31) (a) Immagine originale; (b) Maschera binaria; (c) Maschera dello sfondo; (d) Puzzle image target; (e) Immagine normalizzata.

## 2.5 Validazione degli algoritmi

Gli algoritmi di stain normalization e di tissue detection sono validati non solo visivamente ma anche analiticamente con metriche differenti.

Come descritto nel paragrafo 2.2 con i software Aperio Image Scope e NDP.view si effettuano sulle biopsie delle annotazioni manuali dalle quali si ricavano le maschere di tessuto utilizzate come ground truth per le valutazioni quantitative delle tissue detection.

L'obiettivo di un algoritmo di normalizzazione è mantenere la stessa struttura morfologica delle sezioni di tessuto presenti nell'immagine alterando la colorazione degli stain sulla base di un'immagine target di riferimento.

Gli aspetti che vengono validati sono quindi:

- la similarità morfologica e strutturale dell'immagine originale e dell'immagine normalizzata;
- la variabilità della colorazione degli stain tra le immagini del subset di input e quello di output;
- la similarità degli istogrammi dell'immagine normalizzata e della puzzle target.

### 2.5.1 Metriche per tissue detection

Per valutare la bontà della maschera binaria ottenuta dagli algoritmi di tissue detection vengono utilizzate delle metriche *objects based* che calcolano la percentuale di oggetti correttamente classificati, ossia sezioni individuate e correttamente rappresentative del tessuto.

Un metodo per valutare la corretta segmentazione del tessuto è il calcolo delle *confusion matrices* (Tab.2) che calcolano il numero di veri positivi (TP), falsi positivi (FP), falsi negativi (FN) e veri negativi (VN).

		Classe Predetta	
		Positivo	Negativo
Classe Reale	Positivo	Veri Positivi (TP)	Falsi Positivi (FP)
	Negativo	Falsi Negativi (FN)	Veri Negativi (TN)

Tabella 2 - Confusion matrix.

Per l'algoritmo in esame:

- Veri positivi (TP), sezioni di tessuto classificate correttamente come tale, ossia presenti sia nella maschera automatica che nel ground truth;
- Falsi positivi (FP), oggetti presenti sullo sfondo erroneamente classificati come tessuto, presenti nella maschera automatica e non nel ground truth;
- Falsi negativi (FN), sezioni di tessuto non identificate come tali, presenti nel ground truth e non nella maschera automatica;
- Veri negativi (TN), sfondo ed artefatti che non rappresentano tessuto.

Dalle *confusion matrices* si ricavano due metriche utilizzate per valutare le prestazioni dell'algoritmo:

- *Precision* (Valore Predittivo Positivo), indica la percentuale di oggetti positivi su tutti gli oggetti predetti che hanno reale corrispondenza col ground truth, ossia la percentuale di veri positivi sul numero di elementi totali identificati come positivi dall'algoritmo [24].

$$Precision (VPP) = \frac{TP}{TP + FP} \quad (14)$$

- *Recall* (Sensibilità), descrive in maniera completa la percentuale dei veri positivi, ossia il numero di veri positivi sul numero di elementi totali effettivamente corrispondenti ai positivi sul ground truth [24].

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

- *Indice di Jaccard*, è un coefficiente che esprime il livello di similarità delle due immagini, può essere calcolato come il rapporto tra intersezione ed unione delle maschere (la maschera automatica ed il ground truth) oppure dai valori della confusion matrix [24].

$$J = \frac{mask \cap mask_{gt}}{mask \cup mask_{gt}} = \frac{TP}{TP + FP + FN} \quad (16)$$

Dal calcolo di precision e recall si ricavano:

- *F1-score*, una media armonica fra le due metriche [24].

$$F1 = \frac{precision^{-1} + recall^{-1}}{2} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (19)$$

- *Specificità*, calcolata come il numero di previsioni negative corrette sul totale dei negativi [24].

$$S = \frac{TN}{TN + FP} \quad (20)$$

### 2.5.2 Metriche per la normalizzazione degli stain

Fra le metriche più utilizzate per l'analisi qualitativa delle immagini figura la *Structural Similarity Index Metric (SSIM)* che quantifica la degradazione qualitativa subita dall'immagine sorgente dopo il processing.

La SSIM tiene conto di luminosità, struttura e contrasto e comparando le tre funzioni riportate di seguito tra l'immagine sorgente e l'immagine normalizzata:

$$Luminance: \quad l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \quad (21)$$

$$Contrast: \quad c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (22)$$

$$Structure: \quad s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \quad (23)$$

dove  $\mu_x$  ,  $\mu_y$ ,  $\sigma_x$  ,  $\sigma_y$  sono le medie e le deviazioni standard sui pixels delle due immagini, sorgente e normalizzata;  $\sigma_{xy}$  è il coefficiente di correlazione tra le due immagini mentre  $c_1$ ,  $c_2$  e  $c_3$  sono le costanti utilizzate per stabilizzare il valore della SSIM quando tende a 0.

La formula della SSIM è una combinazione delle equazioni (21), (22) e (23):

$$SSIM(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \cdot \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (24)$$

La SSIM ha valori tra 0 e 1, più è elevato il valore migliore è il risultato della normalizzazione [24].

Questa metrica viene utilizzata per verificare che il processing non applichi modifiche sostanziali all'immagine sorgente, per il caso in esame assicura che le finestre normalizzate vengano riassemblate in maniera corretta poiché di fatto il processo di normalizzazione degli stain altera l'informazione colorimetrica dell'immagine sorgente e non la struttura.

Per valutare l'informazione colorimetrica degli stain vengono implementate altre metriche che misurano quantitativamente la bontà della normalizzazione valutando l'uniformità di colori nei subset normalizzati e la struttura degli istogrammi.

La *NMI (Normalized Median Intensity)* è utilizzata per misurare la costanza colorimetrica nelle immagini di uno stesso subset e avere un'analisi statistica sulle intensità delle intere WSIs, è calcolata come mostrato dall'equazione 23 dove  $U(i)$  è la media dei valori R, G e B per l'i-esimo pixel dell'immagine I.

$$NMI(I) = \frac{Median_{i \in I} \{U(i)\}}{P_{95_{i \in I}} \{U(i)\}} \quad (25)$$

Al numeratore compare la mediana dei valori della distribuzione  $U(i)$  e al denominatore il 95-esimo percentile della stessa, utilizzato per normalizzare il valore e renderlo più robusto rispetto ai pixels rumorosi [25].

La NMI è calcolata per la singola immagine di ogni subset, l'analisi statistica di interesse è ottenuta calcolando la deviazione standard e il coefficiente di variazione (CV) della NMI sul

subset originale e su quello normalizzato per verificare che l'omogeneità di colore sia più elevata dopo la normalizzazione.

L'immagine sorgente è normalizzata sulla base della mappa colorimetrica della puzzle image target, dunque, oltre al confronto sull'omogeneità tra i subset di input e di output è utile analizzare la similarità tra gli istogrammi dell'immagine sorgente e di quella target.

È stata calcolata una percentuale di correlazione tra output e target che indica il livello di similarità tra gli istogrammi dei rispettivi canali R, G e B insieme all'implementazione di due metriche che valutano la distanza fra gli istogrammi.

- *Coefficiente di correlazione*, la correlazione degli istogrammi fornisce un singolo coefficiente che indica il grado di relazione fra due distribuzioni [26]. Misura la covarianza di due variabili sulla deviazione standard delle stesse, oscilla fra -1 e 1.

$$d_c(p, q) = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \quad (26)$$

- *Chi-Square Distance*, una distanza che viene utilizzata per individuare delle dissimilarità fra due istogrammi e tiene conto degli effetti di quantizzazione causati da cambiamenti di luce, forma e deformazione. Il valore minimo è 0 ed è illimitato superiormente [27].

$$\chi^2 = \sum_{i=1}^n \frac{[p(i) - q(i)]^2}{p(i) + q(i)} \quad (27)$$

- *Bhattacharyya distance*, misura la similarità di due distribuzioni di probabilità[pr], è strettamente correlata al coefficiente di Bhattacharyya che esprime la misura di sovrapposizione di due campioni statistici [26].

$$D(A, B) = -\log \sum_{i=0}^{b-1} \sqrt{P_i(A)P_i(B)} \quad (28)$$

# Capitolo 3

## Risultati

In questo capitolo vengono riportati i risultati ottenuti dalla validazione degli algoritmi descritti alle sezioni 2.2 e 2.4.

Dopo aver analizzato visivamente i risultati si valutano quantitativamente ed analiticamente gli algoritmi con le metriche citate al paragrafo 2.5.1 e 2.5.2.

I risultati dell'algoritmo presentato al paragrafo 2.3, ossia le puzzle image, sono state ricavate dal dataset A e dal dataset B ed accuratamente controllate verificando che le immagini in output non contenessero artefatti e fossero correttamente rappresentative della WSI sorgente. L'obiettivo della validazione della tissue detection è di calcolare il grado di specificità e di sensibilità dei due algoritmi su ogni tipologia di colorazione, definendolo così multi stain mentre la normalizzazione degli stain viene validata sulla base dell'analisi delle caratteristiche descritte al paragrafo 2.5.2 su colorazioni PAS ed HE.

### 3.1 Validazione della tissue detection

Entrambi gli algoritmi di tissue detection, *tissue normalization* e *tissue puzzle*, sono validati sul *dataset A* che comprende colorazioni istologiche, istochimiche e immunoistochimiche.

Il confronto è stato effettuato tra gli algoritmi *tissue normalization* e *tissue puzzle* con un semplice algoritmo che individua la maschera di tessuto effettuando un thresholding con il metodo di Otsu sul canale Q dell'immagine convertita nello spazio NTCS. Quest'ultimo algoritmo è indicato nelle tabelle di seguito riportate con il nome di Q algorithm.

	Q algorithm	Tissue normalization	Tissue puzzle
<b>Recall</b>	<i>Mean ± std</i>	<i>Mean ± std</i>	<i>Mean ± std</i>
<i>Mammella</i>	0,8249 ± 0,1253	0,89472 ± 0,0655	0,7380 ± 0,1374
<i>Cancro alla mammella</i>	0,6451 ± 0,1623	0,8942 ± 0,0786	0,7066 ± 0,2572
<i>Rene</i>	0,7617 ± 0,0750	0,9142 ± 0,0635	0,8471 ± 0,1525
<i>Polmoni</i>	0,6981 ± 0,1135	0,9398 ± 0,0570	0,8011 ± 0,1181
<i>Tessuto miocardico</i>	0,7694 ± 0,0741	0,9792 ± 0,0161	0,9098 ± 0,0338
<i>Prostata</i>	0,7200 ± 0,0683	0,8891 ± 0,0756	0,7910 ± 0,1225
<i>Dataset completo</i>	<b>0,7624 ± 0,1158</b>	<b>0,9061 ± 0,070</b>	<b>0,7874 ± 0,1503</b>

Tabella 3 – Valori di recall a confronto.

	Q algorithm	Tissue Normalization	Tissue puzzle
<b>Precision</b>	<i>Mean ± std</i>	<i>Mean ± std</i>	<i>Mean ± std</i>
<i>Mammella</i>	0,7096 ± 0,3153	0,9829 ± 0,02522	0,9949 ± 0,0369
<i>Cancro alla mammella</i>	0,9770 ± 0,1383	0,9634 ± 0,0777	0,9705 ± 0,0966
<i>Rene</i>	0,9449 ± 0,1593	0,9024 ± 0,1641	0,9396 ± 0,1073
<i>Polmoni</i>	0,9966 ± 0,0163	0,9594 ± 0,0309	0,9933 ± 0,0159
<i>Tessuto miocardico</i>	0,9999 ± 0,0004	0,9632 ± 0,0185	0,9964 ± 0,0040
<i>Prostata</i>	0,9663 ± 0,0848	0,9193 ± 0,1091	0,9743 ± 0,0781
<i>Dataset completo</i>	<b>0,8774 ± 0,2407</b>	<b>0,9405 ± 0,1017</b>	<b>0,9776 ± 0,0722</b>

Tabella 4- Valori di precision a confronto.

	Q algorithm	Tissue Normalization	Tissue puzzle
<b>F1-score</b>	<i>Mean ± std</i>	<i>Mean ± std</i>	<i>Mean ± std</i>
<i>Mammella</i>	0,6960 ± 0,1935	0,9351 ± 0,0437	0,8413 ± 0,1079
<i>Cancro alla mammella</i>	0,7522 ± 0,1532	0,9242 ± 0,0659	0,7761 ± 0,2550
<i>Rene</i>	0,8362 ± 0,0951	0,9049 ± 0,1070	0,8848 ± 0,1094
<i>Polmoni</i>	0,8156 ± 0,0825	0,9482 ± 0,03378	0,8814 ± 0,0833
<i>Tessuto miocardico</i>	0,8676 ± 0,0501	0,9709 ± 0,0092	0,9508 ± 0,0186
<i>Prostata</i>	0,8207 ± 0,05692	0,9000 ± 0,0769	0,8682 ± 0,1006
<i>Dataset completo</i>	<b>0,7790 ± 0,1476</b>	<b>0,9227 ± 0,0719</b>	<b>0,8663 ± 0,1198</b>

Tabella 5- Confronto tra i valori di F1-score.

	Q algorithm	Tissue Normalization	Tissue puzzle
<b>Indice di Jaccard</b>	<i>Mean ± std</i>	<i>Mean ± std</i>	<i>Mean ± std</i>
<i>Mammella</i>	0,5652 ± 0,0656	0,8808 ± 0,0656	0,7349 ± 0,1383
<i>Cancro alla mammella</i>	0,6234 ± 0,0996	0,8651 ± 0,0996	0,6859 ± 0,2572
<i>Rene</i>	0,7222 ± 0,1606	0,8340 ± 0,1606	0,8022 ± 0,1679
<i>Polmoni</i>	0,6963 ± 0,0569	0,9033 ± 0,0569	0,7964 ± 0,1167
<i>Tessuto miocardico</i>	0,7693 ± 0,0173	0,9436 ± 0,0173	0,9068 ± 0,1724
<i>Prostata</i>	0,6996 ± 0,1135	0,8240 ± 0,1135	0,7779 ± 0,1241
<i>Dataset completo</i>	<b>0,6572 ± 0,1706</b>	<b>0,8622 ± 0,1086</b>	<b>0,7727 ± 0,1520</b>

Tabella 6 - Confronto tra i valori dell'indice di Jaccard.

Le tabelle 3, 4, 5 e 6 mostrano rispettivamente i valori di precision, recall, F1-score e indice di Jaccard.

Per l'algoritmo *tissue normalization* i valori di precision sono superiori al 90% per ogni subset e dunque per il dataset completo, questo indice fornisce la percentuale di corretti classificati sul numero totale di elementi classificati, compresi i FP; l'algoritmo riesce quindi ad individuare in maniera performante le sezioni presenti sull'immagine che non siano background, ma per ottenere un indice accurato dei corretti classificati occorre osservare i valori di recall.

La recall individua la percentuale di elementi correttamente classificati sul numero totale di elementi che il sistema dovrebbe recuperare se l'algoritmo funzionasse alla perfezione, anche in questo caso otteniamo dei valori elevati per cui la percentuale di FP è molto bassa.

L'F1-score è una media armonica dei valori di precision e recall ed in quanto tale permette di avere un punteggio alto solo quando entrambi i valori sono elevati, per l'algoritmo appena citato sono tutti superiori al 90%.

Confrontando i due algoritmi di tissue detection implementati in questo lavoro si può notare come entrambi siano più performanti dell'algoritmo basato sulla sogliatura di Otsu in quanto possono essere considerati degli approcci multistain che riescono ad individuare correttamente il tessuto anche in colorazioni molto chiare come l'immunoistochimica (Fig. 41).

Una sintesi più immediata dei risultati è mostrata nei bar plot delle figure 43, 44, 45 e 46 dove è evidenziato come la deviazione standard sia notevolmente più elevata per il *Q algorithm* rispetto agli algoritmi implementati, ciò implica un funzionamento ottimale su tutte le immagini del dataset.

Inoltre, i due algoritmi si distinguono tra loro per livello di specificità e sensibilità, in tabella 7 vengono riportati i valori relativi all'intero dataset:

	<b>Sensibilità</b>	<b>Specificità</b>
<b><i>Tissue normalization</i></b>	0,9061	0,0939
<b><i>Tissue puzzle</i></b>	0,7874	0,1503

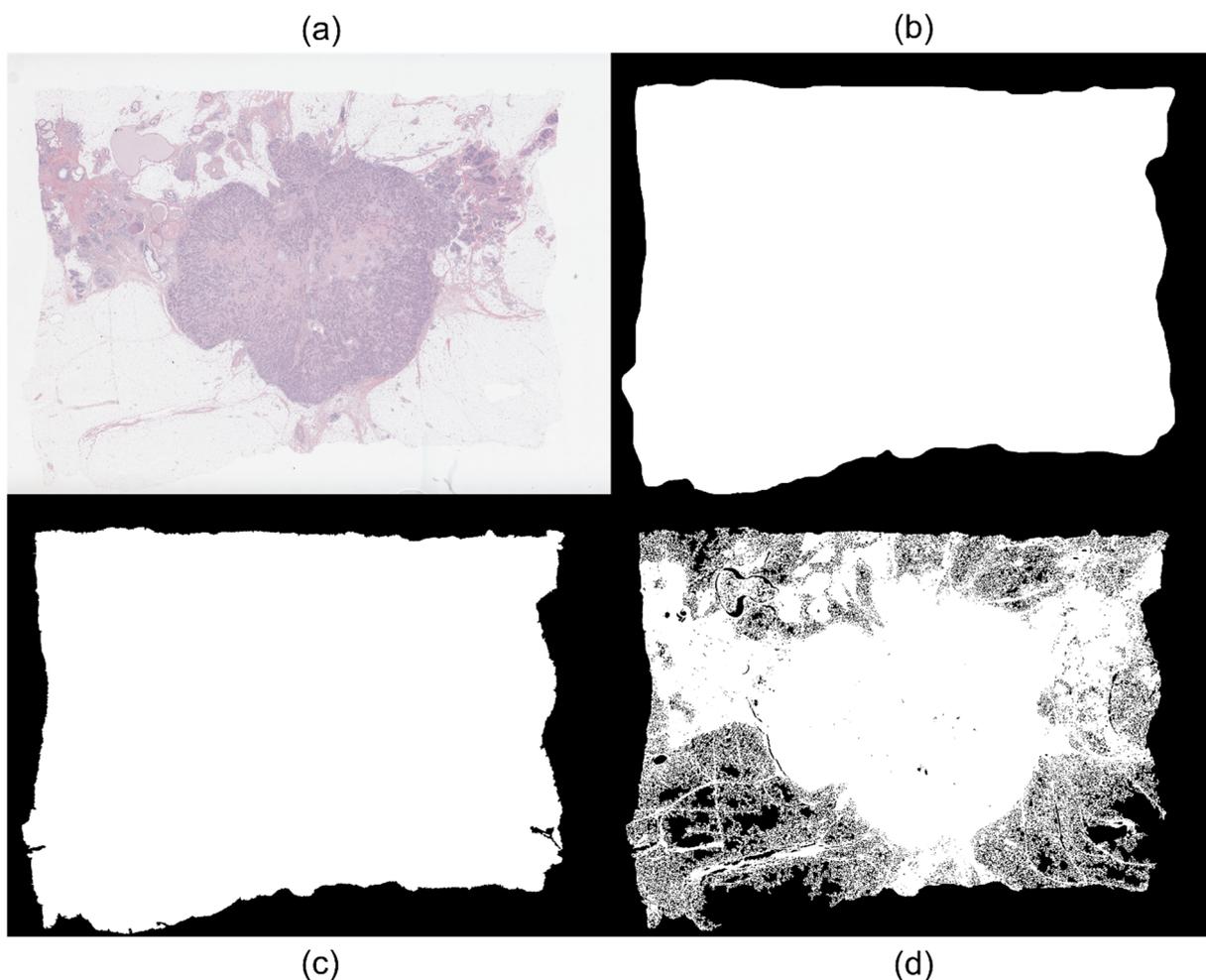
Tabella 7- Valori di specificità e sensibilità calcolati sul dataset A completo.

Per l'algoritmo *tissue normalization* si punta ad ottenere una sensibilità maggiore in quanto l'obiettivo è quello di individuare tutto il tessuto presente (Fig. 40c), comprese zone di adipe o zone istologicamente poco interessanti mentre per l'algoritmo *tissue puzzle* si cerca di individuare in maniera più specifica il tessuto per evitare che le puzzle image siano composte da parti di tessuto poco significative o povere di stain come mostrato in figura 40d.

L'indice di Jaccard fornisce una misura di similarità fra le due maschere, un indice pari ad 1 indica il 100% della sovrapposizione.

In figura 42 è mostrata la sovrapposizione della maschera automatica ottenuta dall'algoritmo *tissue normalization* alla maschera ottenuta dalla segmentazione manuale per una WSI di una biopsia di rene in colorazione PAS.

La sovrapposizione è quasi totale, difatti l'indice di Jaccard è elevato.



*Figura 40- (a) WSI originale di una biopsia di mammella in colorazione HE (3940117 H&E); (b) Segmentazione manuale di tutto il tessuto presente; (c) Maschera automatica ottenuta con l'algoritmo tissue normalization; (d) Maschera automatica ottenuta con l'algoritmo tissue puzzle.*

Come accennato nel paragrafo 2.2.2 in figura 22 alcune WSI presentano dei frustoli interamente o parzialmente sfocati, gli algoritmi tissue puzzle e tissue normalization lavorando sulla struttura dell'immagine riescono ad individuare solo texture non sfocate. In tabella 8 sono riportati il numero di immagini totali analizzate, il numero di immagini per cui l'algoritmo non riesce ad individuare tessuto ed il numero di WSI che presenta frustoli sfocati e il relativo numero di WSI sfocate correttamente processate.

WSIs processate	WSIs non processate correttamente	WSI con tessuto sfocato	WSI con tessuto sfocato correttamente processate
714	0	10	9

Tabella 8 - WSIs processate correttamente.

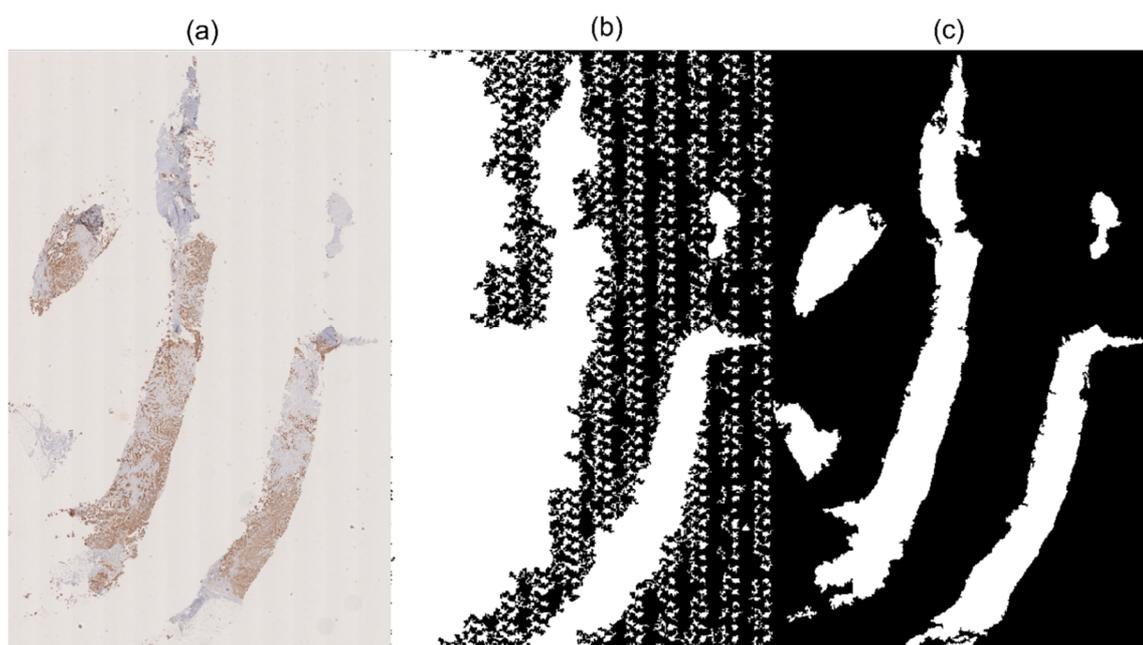
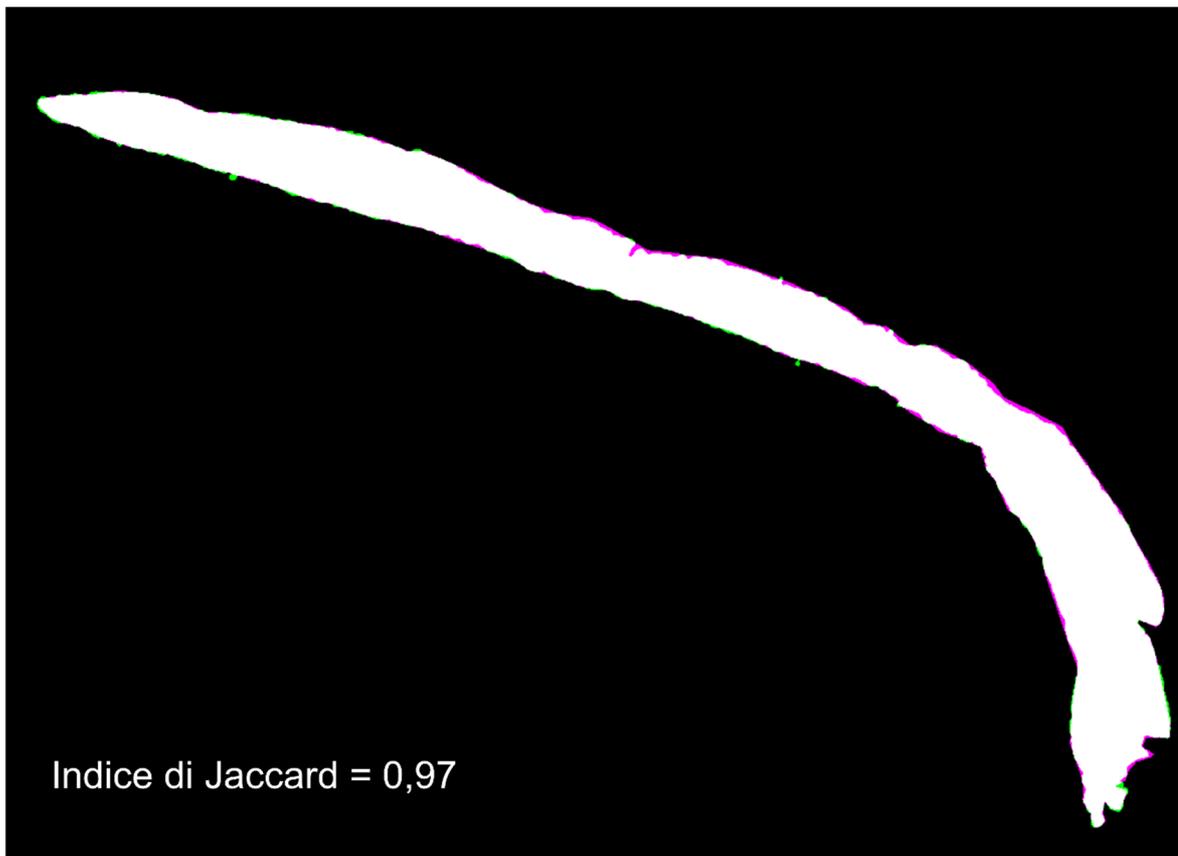


Figura 41- (a) WSI di una biopsia di mammella in colorazione immunoistochimica (2016I004557;A;1;3 - 2020-11-27 16.13.46); (b) Maschera automatica ottenuta con Q algorithm; (c) Maschera automatica ottenuta con l'algorithmo tissue normalization.



*Figura 42- Sovrapposizione della maschera automatica ottenuta dall' algoritmo tissue normalization (in viola) sovrapposta alla maschera segmentata manualmente (in verde) per una biopsia di rene (PAS 166). L'indice di Jaccard è una misura della similarità delle due maschere, sovrapponendole si ottiene un match quasi perfetto e quindi un indice elevato.*

Le prestazioni dell'algoritmo di tissue detection sono influenzate non solo dall'eventuale presenza di falsi positivi ma anche dalla manualità dell'operatore, difatti come si può notare nelle figure 40b e 42 i bordi della maschera manuale sono meno precisi di quelli della segmentazione automatica poiché non si riesce a segmentare in maniera estremamente efficace il tessuto.

Questo purtroppo altera le metriche di valutazione, soprattutto l'indice di Jaccard che ne misura l'overlap.

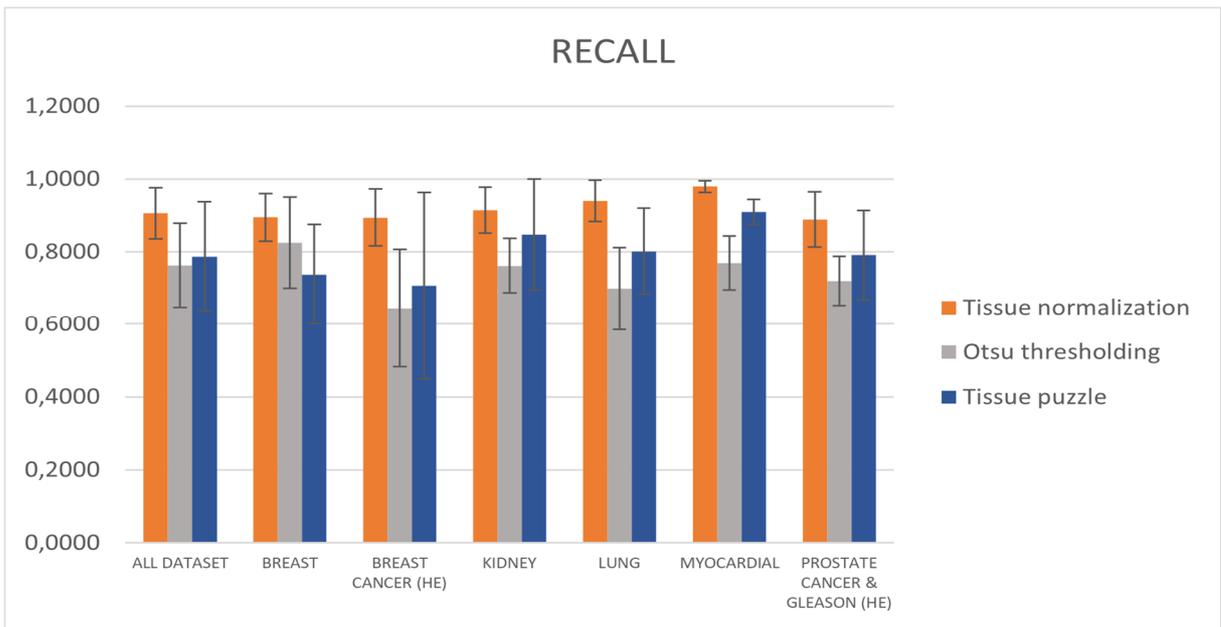


Figura 43- Medie e deviazioni standard dei valori di recall.

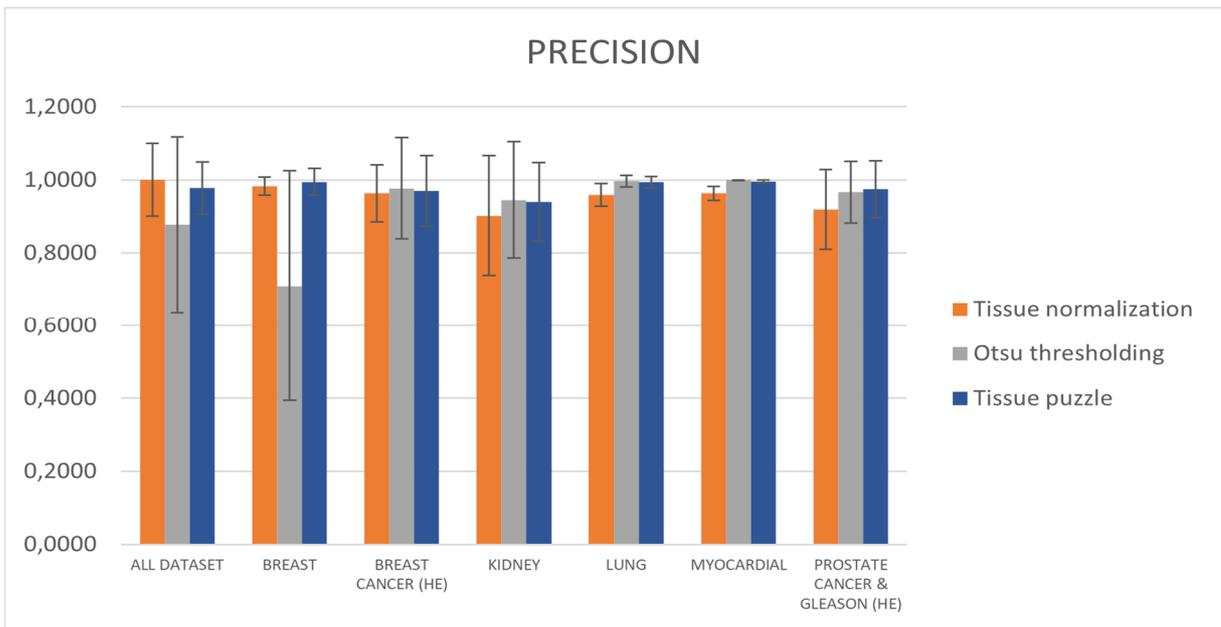


Figura 44 - Medie e deviazioni standard del valore di precision.

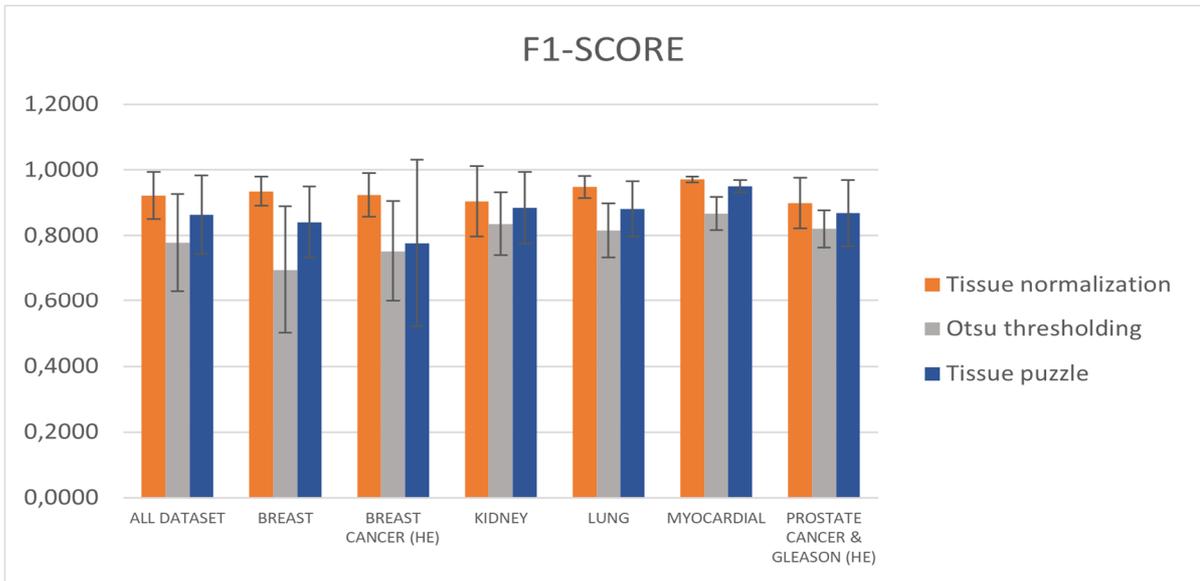


Figura 45-Medie e deviazioni standard dei valori di F1-score.

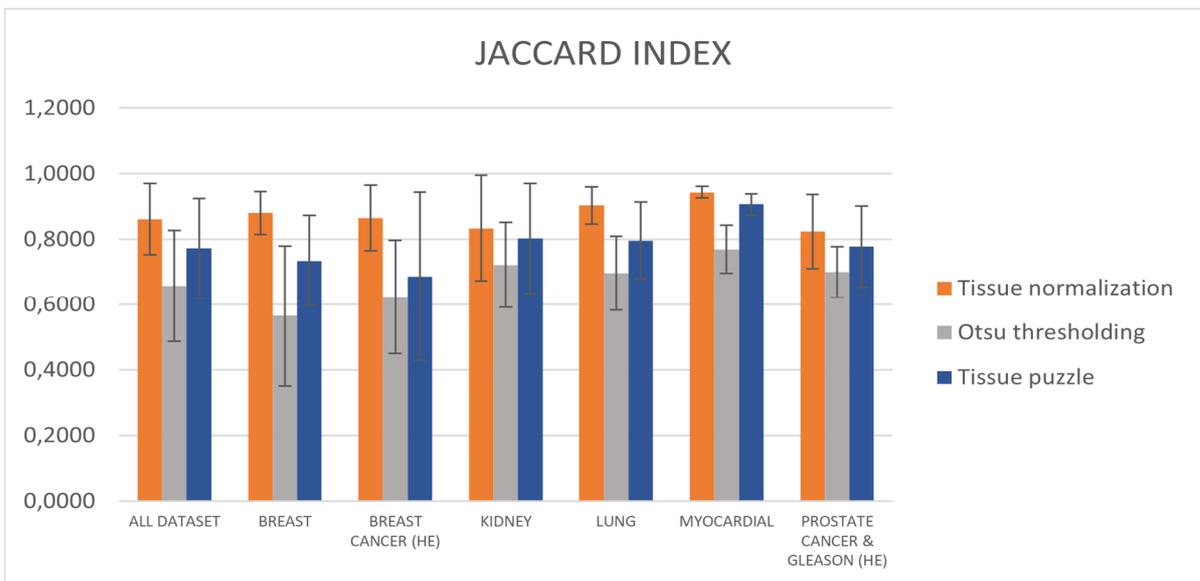


Figura 46 - Medie e deviazioni standard dei valori del Jaccard Index.

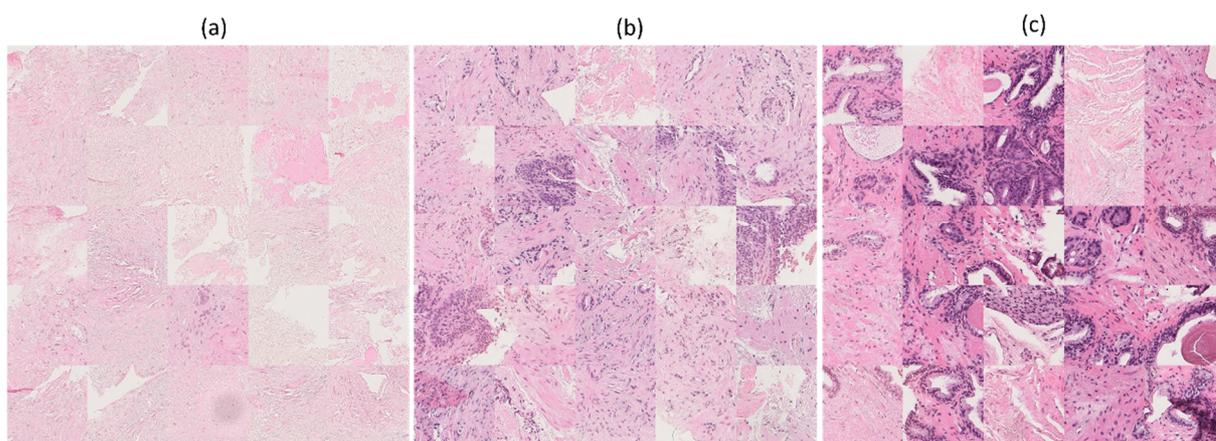
## 3.2 Validazione della stain normalization

La normalizzazione viene effettuata sul *dataset A*, le colorazioni analizzate sono HE su tessuto prostatico e di mammella e PAS su tessuto renale.

Le colorazioni trattate in questo caso sono solo PAS ed HE in quanto la stain normalization di Machenko ha dimostrato risultati migliori nel caso di soli due stain presenti.

La prima fase prevede la scelta accurata delle immagini target come descritto al paragrafo 2.4.3, si scelgono in modo tale che la stain separation sia efficace (Fig. 34) e che la dinamica di colori del target sia un giusto compromesso fra la dinamica dell'immagine più chiara e più scura del subset in esame, come mostrato in figura 47.

La scelta dell'immagine target è fondamentale, un'immagine con colori molto accentuati è generalmente frutto di un'immersione prolungata nei coloranti al momento della colorazione manuale della biopsia e quindi normalizzare il dataset con quei colori sarebbe errato, al contrario un'immagine con una dinamica molto chiara non contraddistingue bene i due stain presenti.



*Figura 47- (a) Immagine con una dinamica di colori molto chiara; (b) Immagine scelta come target; (c) Immagine con una dinamica di colori troppo scura.*

Dopo aver normalizzato l'intero dataset si procede ad un'analisi visiva dei risultati ottenuti in modo da scartare le immagini per le quali la normalizzazione non ha ottenuto risultati ottimali.

Le immagini scartate sono quelle dove gli stain di destinazione non vengono riconosciuti correttamente e quindi vengono erroneamente normalizzati, nelle figure 48, 49, 50 e 51 un esempio di normalizzazione riuscita correttamente per HE (Fig. 48-49) e PAS (Fig. 50-51).

I colori degli stain dell'immagine target sono differenti, in entrambi i casi più chiari e sono evidentemente corrispondenti a quelli dell'output finale.

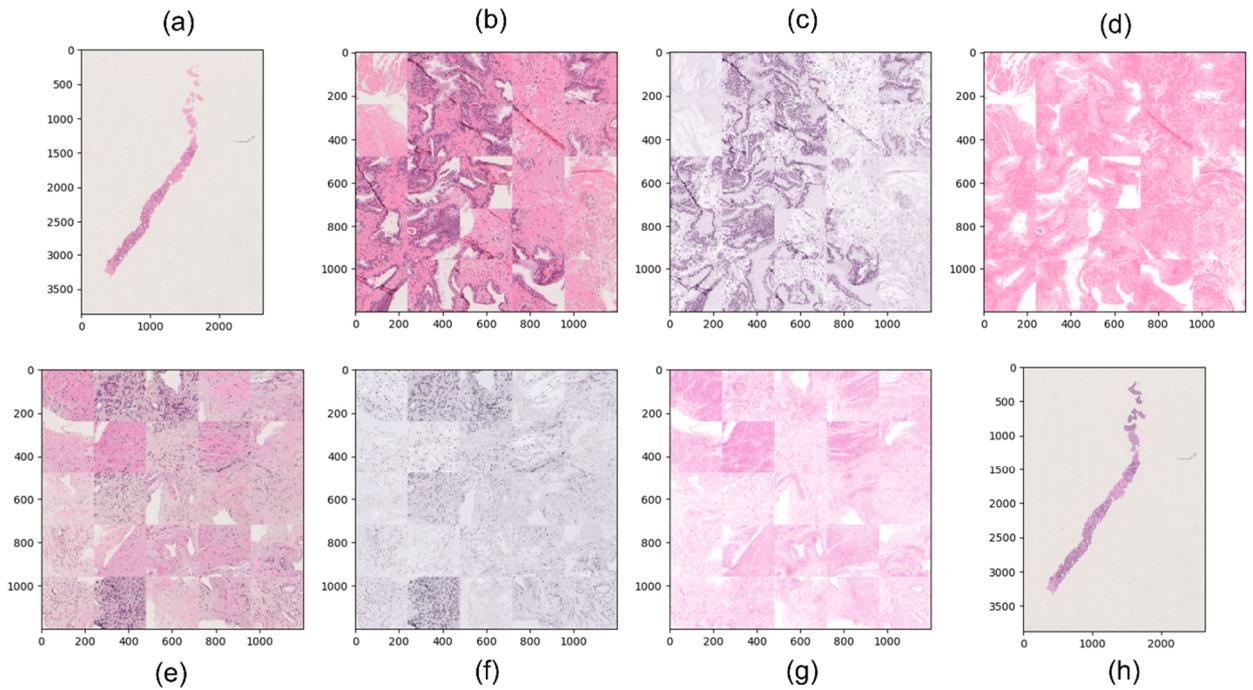


Figura 48- (a) WSI originale di una biopsia di prostata in colorazione HE (69.18.B1); (b) Puzzle image della WSI originale; (c) Canale dell'ematosilina; (d) Canale dell'eosina; (e) Puzzle image scelta come target; (f) Canale dell'ematosilina del target; (g) Canale dell'eosina del target; (h) Biopsia normalizzata.

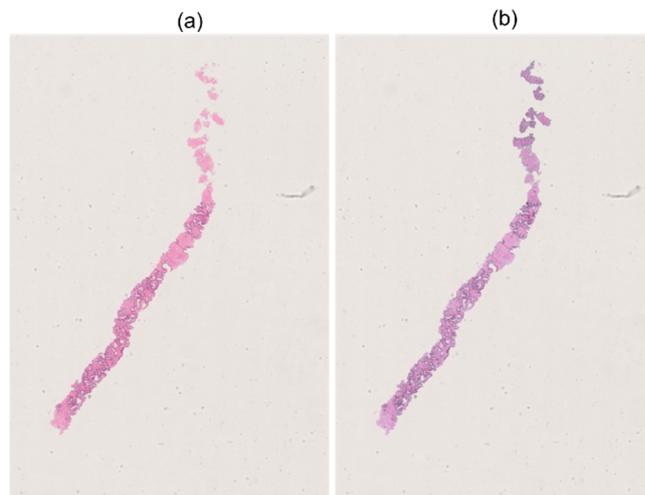


Figura 49-(a) Dettaglio dell'immagine 48a, WSI originale; (b) Dettaglio dell'immagine 48h, WSI normalizzata.

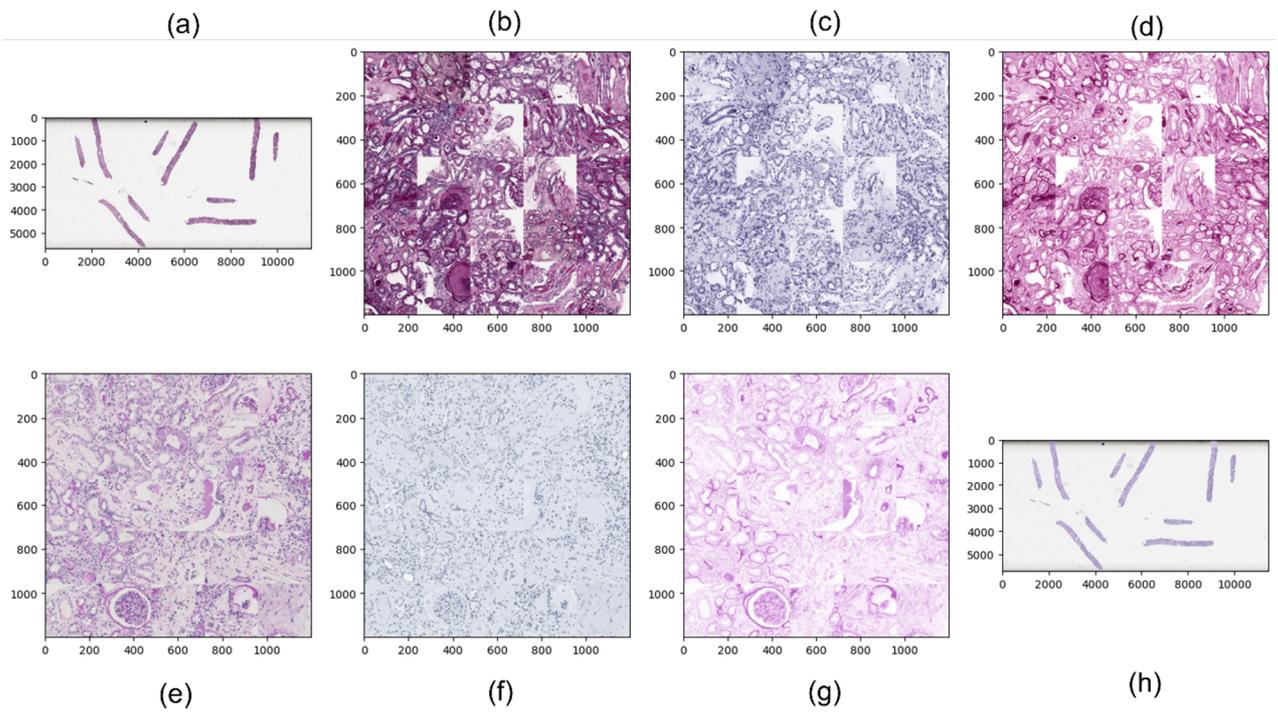


Figura 50- WSI originale di una biopsia di rene in colorazione PAS (1004761); (b) Puzzle image della WSI originale; (c) Canale 1; (d) Canale 2; (e) Puzzle image scelta come target; (f) Canale 1 del target; (g) Canale 2 del target. (h) WSI normalizzata.

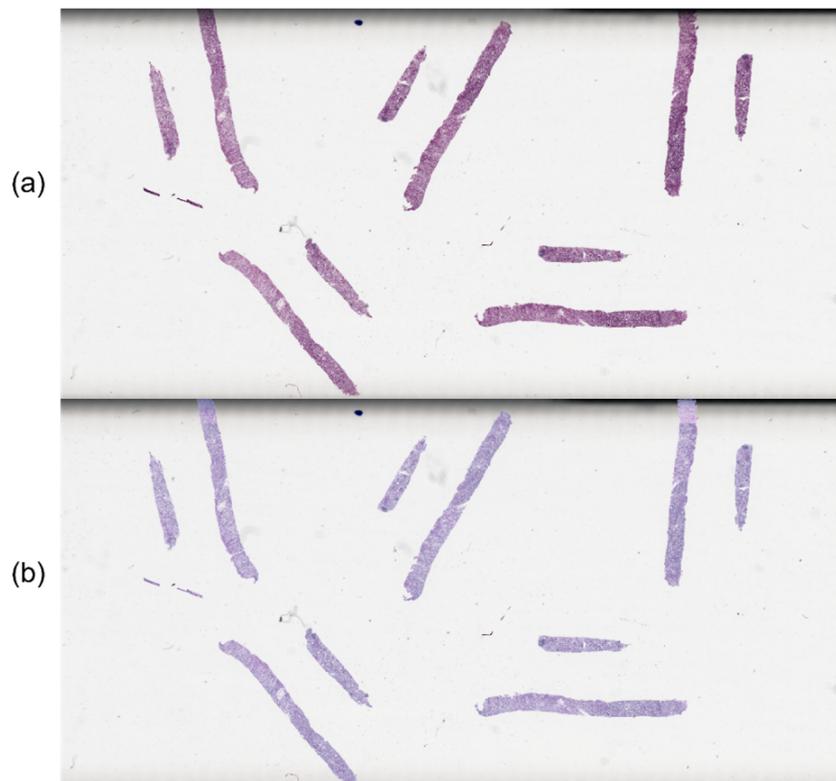


Figura 51- (a) Dettaglio dell'immagine 50a, WSI originale; (b) Dettaglio dell'immagine 50h, WSI normalizzata.

Sulle immagini selezionate viene quantitativamente valutata la bontà della normalizzazione con l'ausilio delle metriche descritte al paragrafo 2.5.2.

La SSIM viene calcolata tra l'immagine di input e di output per confermare che la struttura morfologica dell'immagine sia la stessa e che quindi il riassetto delle finestre normalizzate avvenga in maniera corretta.

In tabella 9 sono riportati i valori di media e deviazione standard, come si può notare i valori di SSIM sono quasi pari all'unità in tutti i subset per cui la morfologia strutturale dell'immagine resta invariata.

<b>SSIM</b>	<b>Mean <math>\pm</math> std</b>
<b>Mammella</b>	0,9819 $\pm$ 0,0098
<b>Rene</b>	0,9916 $\pm$ 0,0103
<b>Prostata</b>	0,9906 $\pm$ 0,0086

Tabella 9-Valori di SSIM.

In seguito sono state valutate le metriche che permettono di valutare la similarità degli istogrammi dell'immagine normalizzata e dell'immagine target.

Le due immagini in analisi hanno dimensioni differenti, inoltre l'immagine normalizzata presenta l'intero background mentre l'immagine puzzle è composta interamente da tessuto.

Per evitare che l'istogramma dell'immagine normalizzata sia compromesso dalla presenza di tutti i pixels dello sfondo sono state utilizzate le maschere ricavate dall'algoritmo *tissue normalization* per valutare quantitativamente solo la parte rappresentativa del tessuto.

Nelle tabelle 10, 11 e 12 sono riportati i risultati di media e deviazione standard per l'indice di correlazione e le distanze chi-square e di Bhattacharyya ottenute dal confronto fra WSI originale ed immagine target e fra WSI normalizzata ed immagine target.

<b>Indice di correlazione</b>	<b>WSI originale – Puzzle target</b>	<b>WSI normalizzata-Puzzle target</b>
<b>Mammella</b>	0,0042 ± 0,2287	0,6644 ± 0,1292
<b>Rene</b>	0,7220 ± 0,1496	0,8025 ± 0,1429
<b>Prostata</b>	0,0803 ± 0,2236	0,4457 ± 0,1336

*Tabella 10- Indice di correlazione.*

<b>Distanza Chi-Square</b>	<b>WSI originale – Puzzle target</b>	<b>WSI normalizzata-Puzzle target</b>
<b>Mammella</b>	13,2839 ± 6,4161	7,6363 ± 2,2844
<b>Rene</b>	20,5092 ± 9,225	13,1369 ± 5,087
<b>Prostata</b>	7,7459 ± 4,2432	4,6730 ± 3.1466

*Tabella 11 - Distanza Chi-Square.*

<b>Distanza di Bhattacharyya</b>	<b>WSI originale – Puzzle target</b>	<b>WSI normalizzata-Puzzle target</b>
<b>Mammella</b>	0,3905 ± 0,1013	0,3041 ± 0,2287
<b>Rene</b>	0,4667 ± 0,8020	0,3321 ± 0,0474
<b>Prostata</b>	0,2823 ± 0,0841	0,2048 ± 0,0679

*Tabella 12 - Distanza di Bhattacharyya*

L'indice di correlazione per come è calcolato oscilla fra -1 e 1 dove l'unità positiva indica la corrispondenza massima.

Dai dati riportati in tabella 10 è evidente come la percentuale di correlazione con l'immagine target aumenti notevolmente per il subset normalizzato.

Inoltre, la deviazione standard tende a diminuire, questo è in altri termini traducibile con una maggior omogeneità del subset che contiene le immagini normalizzate, indice del fatto che le immagini sono state correttamente normalizzate sulla base dello stesso target e hanno colori più simili tra loro rispetto a quelle presenti nel subset di input. Le distanze nelle tabelle 11 e

12 tendono a diminuire, in quanto misure degli overlap tra gli istogrammi , più breve è la distanza, maggior è la corrispondenza, anche ciò indica quindi un match più elevato tra l'immagine target e la WSI normalizzata.

Di seguito vengono riportati dei grafici che permettono di analizzare i valori mediati delle tabelle 10, 11 e 12 in maniera specifica per ogni immagine analizzata nel subset corrente (Figg. 52, 53, 54, 55, 56, 57, 58, 59, 60).

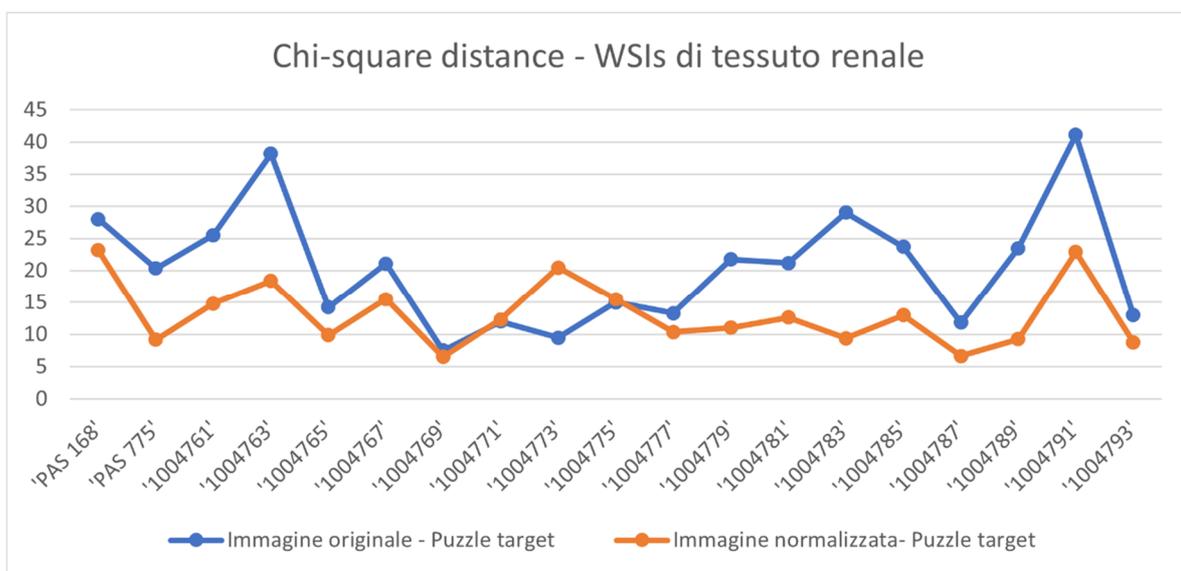


Figura 52- Valori di Chi-Square per le biopsie renali.

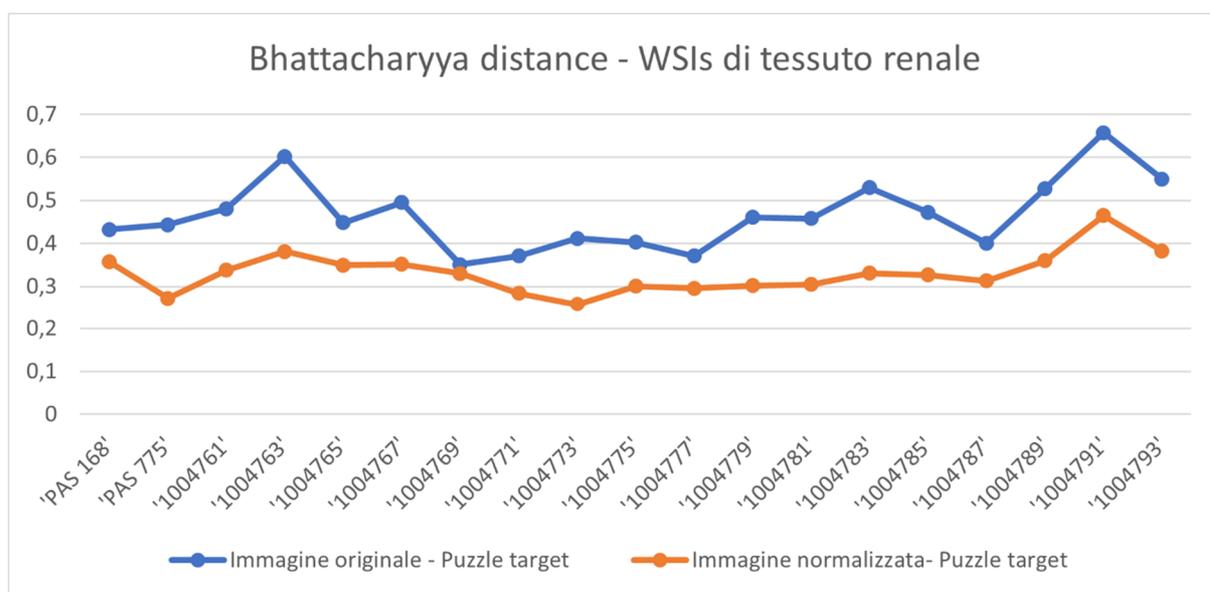


Figura 53 - Valori della distanza di Bhattacharyya per biospie di rene.



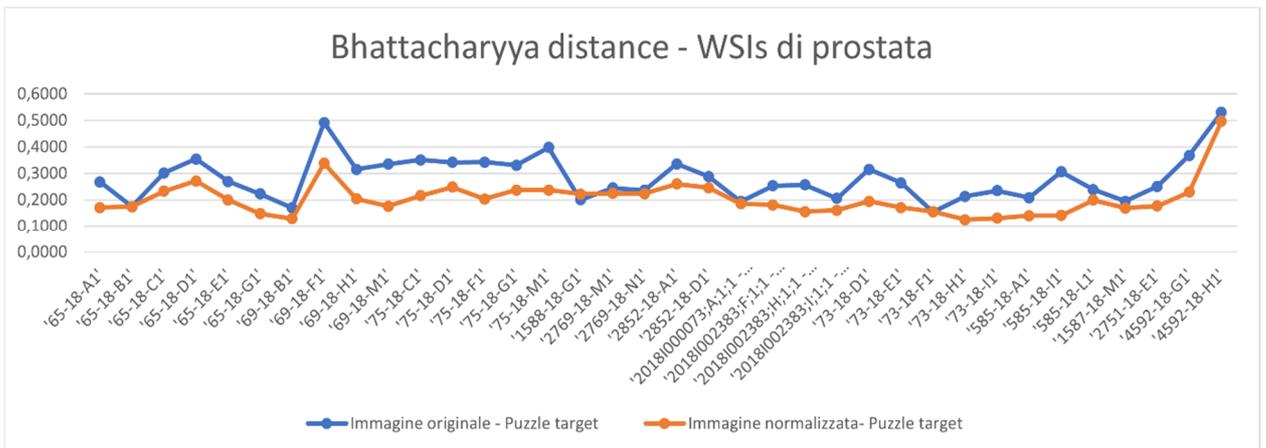


Figura 56- Valori della distanza di Bhattacharyya per biopsie di prostata.

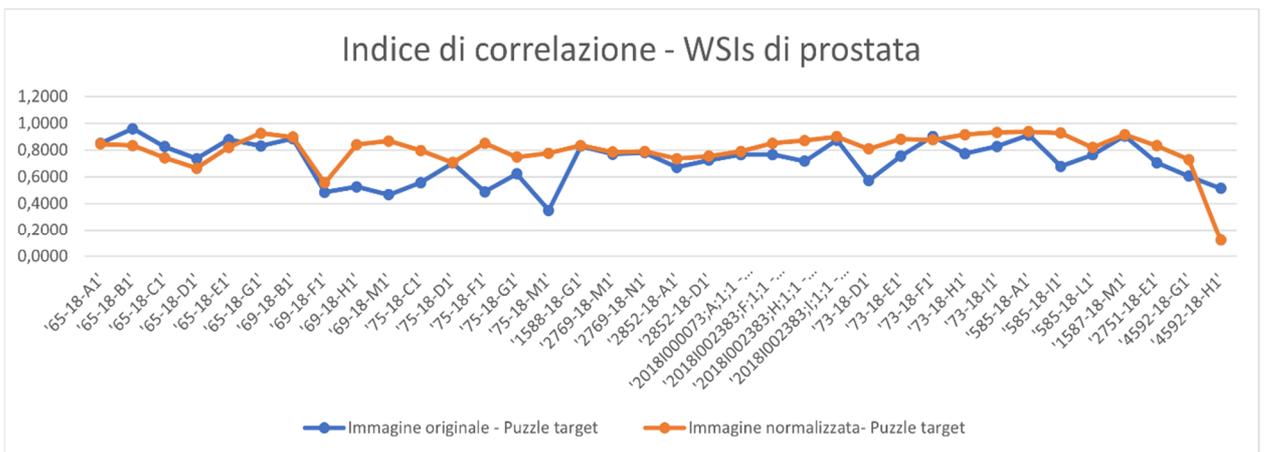


Figura 57- Valori dell'indice di correlazione per biopsie di prostata.

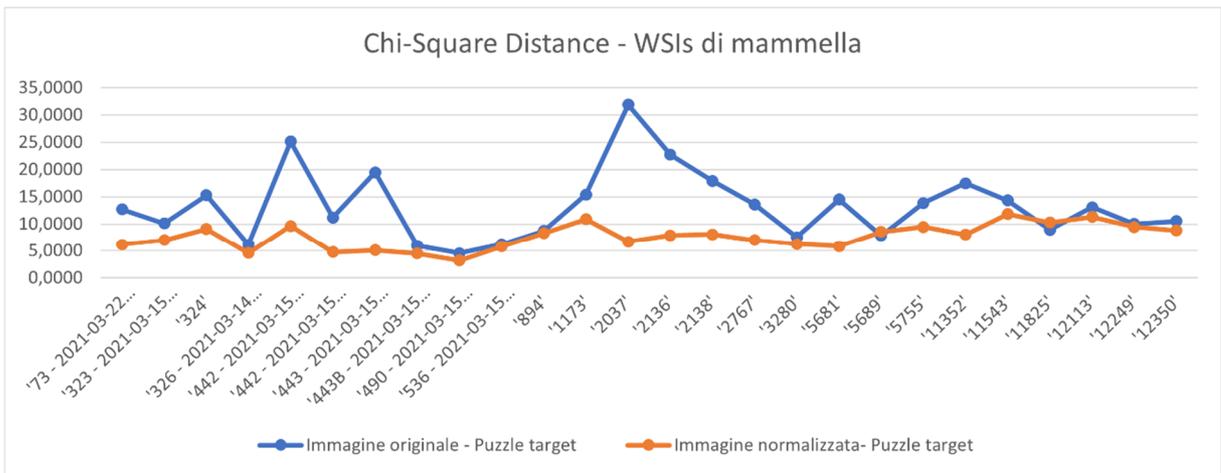


Figura 58 - Valori di Chi-Square per le biopsie di mammella.

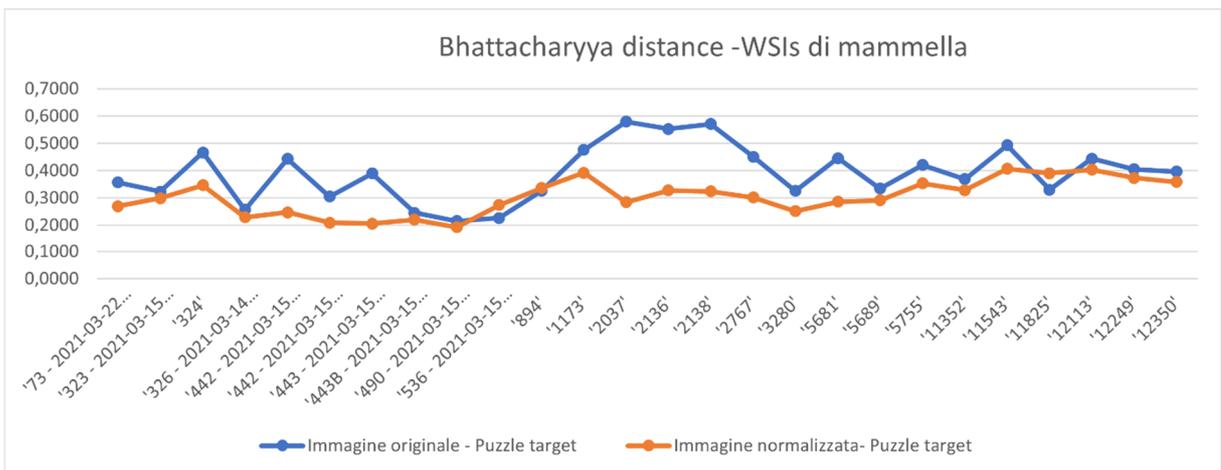


Figura 59- Valori della distanza di Bhattacharyya per biopsie di mammella.

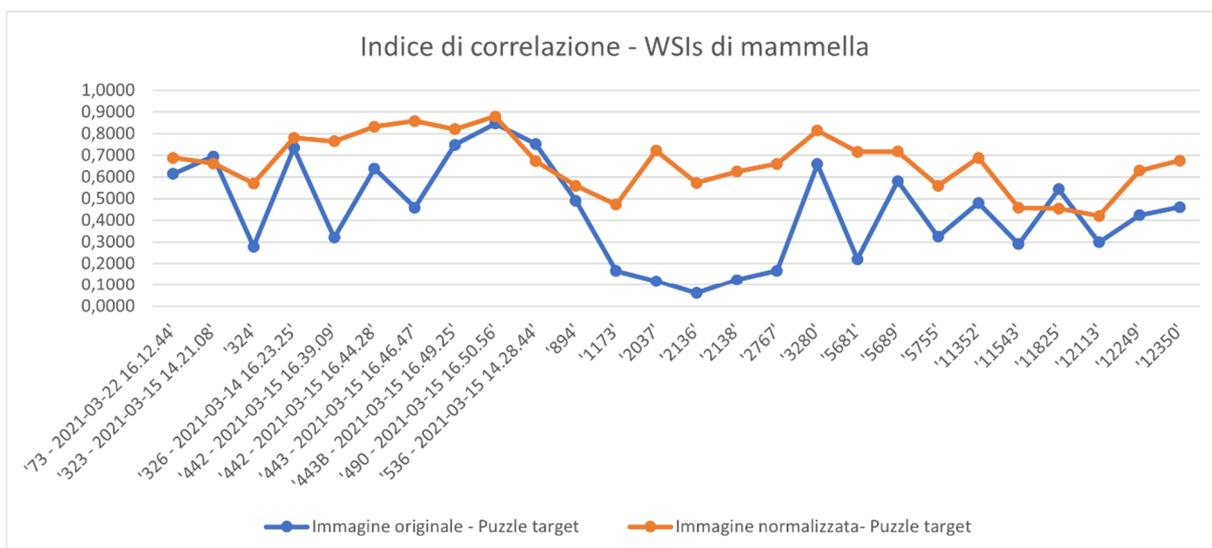


Figura 60 – Valori dell'indice di correlazione per biopsie di mammella.

In conclusione sono stati calcolati i coefficienti di variazioni e le deviazioni standard del valore di NMI per ogni subset pre e post normalizzazione come riportato in tabella 13.

	NMI			
	Subset originale		Subset normalizzato	
	Deviazione standard	CV	Deviazione standard	CV
<b>Mammella</b>	0,1289	0,1556	0,0426	0,0498
<b>Rene</b>	0,1197	0,1838	0,0342	0,0458
<b>Prostata</b>	0,0700	0,0811	0,0321	0,0376

Tabella 13 - Deviazioni standard e coefficienti di variazione della NMI.

La NMI è una metrica che viene utilizzata per verificare la coerenza e la costanza di colori in un subset, calcolando la deviazione standard e quindi il coefficiente di variazione sul subset pre e post normalizzazione si può valutare il grado di disomogeneità colorimetrica tra le immagini.

In questo caso i valori sono più bassi per i subset di output, ossia quelli normalizzati, come volevasi dimostrare si ottiene una maggiore uniformità dopo la normalizzazione.

Infine viene effettuato un t-test per valutare le medie delle due distribuzioni di valori (pre e post normalizzazione) per ogni metrica calcolata.

Il t-test restituisce un risultato positivo per l'ipotesi nulla se le due distribuzioni sono indipendenti, hanno distribuzione normale ed hanno stessa media e stessa varianza.

L'ipotesi nulla è rigettata se hanno diversa media e corrisponde al valore 1, considerando un livello di significatività pari al 5%, mentre l'ipotesi nulla corrisponde al valore 0.

	Distanza Chi-Square		Distanza di Bhattacharyya		Indice di correlazione	
	<i>h</i>	<i>p</i>	<i>h</i>	<i>p</i>	<i>h</i>	<i>p</i>
<b>Reni</b>	1	0,0042	1	$2,768 \cdot 10^{-7}$	1	$4,896 \cdot 10^{-7}$
<b>Mammella</b>	1	$1,000 \cdot 10^{-4}$	1	$5,777 \cdot 10^{-4}$	1	$7,525 \cdot 10^{-5}$
<b>Prostata</b>	1	$8,397 \cdot 10^{-4}$	1	$5,402 \cdot 10^{-5}$	1	0,0224

Tabella 14- T-test per le metriche calcolate, *h* indica il rigetto o meno dell'ipotesi nulla mentre *p* è il *p-value*.

I risultati in tabella 14 confermano la significatività statistica dei risultati, in tutti i casi viene rigettata l'ipotesi nulla, in più con *p* viene riportato il *p-value*, ossia la probabilità di ottenere risultati uguali o meno probabili di quelli osservati, supposta vera l'ipotesi nulla. Come si può notare i valori sono estremamente bassi.

L'ultimo obiettivo del tool era quello di realizzare una normalizzazione qualitativamente ottimale in tempi computazionali ristretti, in tabella 15 sono riportati i tempi computazionali necessari per normalizzare le immagini oggetto dell'analisi.

L'analisi è stata effettuata con un comune pc da 16 gb RAM utilizzando una risoluzione di 2.5x per le WSI e di 10x per le puzzle images.

	Numero di immagini	Tempi computazionali (s)
<b>Reni</b>	19	1120,33
<b>Mammella</b>	26	741,38
<b>Prostata</b>	36	1726,73

Tabella 15-Tempi computazionali dell'algoritmo di normalizzazione.

# Capitolo 4

## Conclusioni

Lo scopo di questa tesi è velocizzare ed automatizzare il lavoro dell'anatomopatologo nell'analisi di biopsie istologiche digitalizzate fornendo un tool che permettesse di normalizzare i colori delle WSIs, escludere gli artefatti e rendere più snella e performante una qualunque pipeline successiva si decida di applicare al vetrino, ad esempio un tool di cancer detection.

L'anatomopatologo esclude visivamente gli artefatti e le zone di tessuto non rilevanti a livello istologico focalizzando la propria attenzione sugli stain che colorano il tessuto discriminando tessuto sano o malato.

L'alterazione degli stains dell'immagine può di fatto rendere complicata l'interpretazione del tessuto.

Gli algoritmi elaborati in questo lavoro di tesi mirano effettuare un pre-processing completo sulla WSI e di fornire in output, nel giro di pochi minuti, una WSIs normalizzata che sia di più immediata interpretazione sia a livello visivo che digitale.

La tissue detection permette non solo di individuare tutto il tessuto presente escludendo gli artefatti ed ottenendo delle maschere che permettano di lavorare solo sulle zone di interesse ma allo stesso tempo di ricreare delle immagini target per la normalizzazione di più WSI.

Generalmente la normalizzazione degli stains si effettua utilizzando come target l'intera WSI opportunamente selezionata sulla base dei colori degli stain, l'innovazione di questo algoritmo consta nell'utilizzare come target un'immagine rappresentativa della WSI selezionata come target.

In questo modo il target sarà “un puzzle” rappresentativo solo ed esclusivamente degli stain della WSI, ciò permette di estrarre le caratteristiche della normalizzazione ed estenderle ad un’intera WSI tramite una sliding window.

Lavorando sulle puzzle images rappresentative del target e dell’immagine corrente si riesce quindi a normalizzare una WSI in pochi minuti.

In questo lavoro le puzzle images sono state estratte ad una risoluzione di 10x mentre la WSI ad una risoluzione di 2.5x, dunque a basso ingrandimento.

La differenza nella scelta della risoluzione è dovuta alla potenza computazionale ridotta dei mezzi a disposizione, nonostante le metriche calcolate forniscano dei risultati ottimali, questi sono fortemente influenzati da questa scelta.

Un obiettivo futuro è normalizzare le immagini ad una risoluzione elevata, 10x o 20x per ottenere delle performance ancora migliori.

Inoltre, come descritto in precedenza, gli algoritmi di tissue detection sono già definibili come multi-stain, ossia permettono l’analisi di tutte le tipologie di colorazioni: istologica, istochimica ed immunohistochemica sia su colorazioni caratterizzate dalla presenza di due stains che su quelle caratterizzate dalla presenza di tre stains.

L’algoritmo lavorando sulla morfologia e sulla struttura del tessuto presente e non su un canale preferenziale di uno spazio colore riesce ad evitare le sfocature ed individuare il tessuto anche su WSI con stain molto chiari, ad esempio sulle immagini di immunohistochemica che tendenzialmente sono caratterizzate da colori come marrone e azzurro chiaro.

In questo lavoro ci si è focalizzati sull’implementare una strategia di normalizzazione veloce e computazionalmente poco onerosa con dei risultati ottimali quindi si sfrutta l’algoritmo di Macencko che lavora preferenzialmente su WSIs con colorazioni caratterizzate dalla presenza di due stain e riesce ad adattarsi alla colorazione PAS.

L’obiettivo finale sarebbe rendere multi stain anche l’approccio di stain normalization per rendere la pipeline più generalizzata possibile.

# Bibliografia

- [1] Salvi M., Michielli N., Molinari F. Stain Color Adaptive Normalization (SCAN) algorithm: Separation and standardization of histological stains in digital pathology. *Comput Methods Programs Biomed.* 2020 Sep;193:105506.
- [2] Macencko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging*. Boston, MA (2009). p. 1107–10.
- [3] E. Reinhard, M. Adhikhmin, B. Gooch, P. Shirley, Color transfer between images, *IEEE Comput. Graph. Appl.* 21 (2001) 34–41.
- [4] A.M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, *IEEE Trans. Biomed. Eng.* 61 (2014) 1729–1738.
- [5] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Jun Shi, Chenghai Xue, Adaptive color deconvolution for histological WSI normalization, *Computer Methods and Programs in Biomedicine*, Volume 170, 2019, Pages 107-120, ISSN 0169-2607.
- [6] A.C. Ruifrok, D.A. Johnston, Quantification of histochemical staining by color deconvolution, *Anal. Quant. Cytol. Histol.* 23 (2001) 291–299.
- [7] Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans Biomed Eng.* 2014 Jun;61(6):1729-38.
- [8] Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, Steiger K, Schlitter AM, Esposito I, Navab N. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Trans Med Imaging.* 2016 Aug;35(8):1962-71.
- [9] B. Li, Y. Li and K. W. Eliceiri, "Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14313-14323.

- [10] Gonzalez, R.C. and Woods, R.E. (2002) Digital Image Processing. Prentice-Hall, Inc. Upper Saddle River, New Jersey.
- [11] Fathima S, Parwani A. WSI fundamentals. PathologyOutlines.com website. <https://www.pathologyoutlines.com/topic/informaticswholeslidefund.html>. Accessed November 15th, 2021.
- [12] Zarella MD. WSI implementation. PathologyOutlines.com website. <https://www.pathologyoutlines.com/topic/informaticswholeslidediag.html>. Accessed November 15th, 2021.
- [13] Bug D., Feuerhake F., Merhof D. (2015) Foreground Extraction for Histopathological Whole Slide Imaging. In: Handels H., Deserno T., Meinzer HP., Tolxdorff T. (eds) Bildverarbeitung für die Medizin 2015. Informatik aktuell. Springer Vieweg, Berlin, Heidelberg.
- [14] Zhang, Ming, "Bilateral filter in image processing" (2009). LSU Master's Theses. 1912. [https://digitalcommons.lsu.edu/gradschool\\_theses/1912](https://digitalcommons.lsu.edu/gradschool_theses/1912).
- [15] Sylvain Paris. 2007. *A gentle introduction to bilateral filtering and its applications*. In *ACM SIGGRAPH 2007 courses (SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 3–es.
- [16] Araùjo ALD, Arboleda LPA, Palmier NR, Fonsêca JM, de Pauli Paglioni M, Gomes-Silva W, Ribeiro ACP, Brandão TB, Simonato LE, Speight PM, Fonseca FP, Lopes MA, de Almeida OP, Vargas PA, Madrid Troconis CC, Santos-Silva AR. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. *Virchows Arch*. 2019 Mar;474(3):269-287.
- [17] Salvi M, Acharya UR, Molinari F, Meiburger KM. The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Comput Biol Med*. 2021 Jan;128:104129.
- [18] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological image analysis: a review, *IEEE Rev. Biomed. Eng.* 2 (2009) 147–171.

- [19] D. Coltuc, P. Bolon, J.-M. Chassery, Exact histogram specification, *IEEE Trans. Image Process.* 15 (2006) 1143–1152.
- [20] D.F. Swinehart, The beer-lambert law, *J. Chem. Educ.* 39 (1962) 333.
- [21] S. Roy, A. kumar Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, *Micron* 114 (2018) 42–61.
- [22] Salvi M, Molinari F. Multi-tissue and multi-scale approach for nuclei segmentation in H&E stained images. *Biomed Eng Online.* 2018 Jun 20;17(1):89.
- [23] Zheng Y, Jiang Z, Zhang H, Xie F, Shi J, Xue C. Adaptive color deconvolution for histological WSI normalization. *Comput Methods Programs Biomed.* 2019 Mar;170:107-120.
- [24] Roy S, kumar Jain A, Lal S, Kini J, A Study about Color Normalization Methods for Histopathology Images, *Micron* (2018), <https://doi.org/10.1016/j.micron.2018.07.005>.
- [25] Bejnordi BE, Litjens G, Timofeeva N, Otte-Höller I, Homeyer A, Karssemeijer N, van der Laak JA. Stain Specific Standardization of Whole-Slide Histopathological Images. *IEEE Trans Med Imaging.* 2016 Feb;35(2):404-15.
- [26] Eric Bazan, Petr Dokládál, Eva Dokladalova. Quantitative Analysis of Similarity Measures of Distributions. *British Machine Vision Conference 2019, BMVC 2019, Sep 2019, Cardiff, United Kingdom.* fahal-01984970.
- [27] V. Asha, N. U. Bhajantri, and P. Nagabhusjan. 2011, GLCM-based chi-square histogram distance for automatic detection of defects on patterned textures. *Int. J. Comput. Vision Robot.* 2, 4 (February 2011), 302-313.
- [28] Md. Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, Tapio Seppänen, Retinex model based stain normalization technique for whole slide image analysis, *Computerized Medical Imaging and Graphics*, Volume 90, 2021, 101901, ISSN 0895-6111.