POLITECNICO DI TORINO

Laurea Magistrale in Ingegneria Biomedica

Sviluppo di un algoritmo automatico per l'analisi quantitativa di immagini istologiche in immunoistochimica

Relatore

Prof. Filippo Molinari

Candidato

Fabrizio Scano

Correlatore

Ing. Massimo Salvi

Anno Accademico 2019/2020

Alla mia famiglia

INDICE

1.INTRODUZIONE

Il lavoro di tesi riguarda l'elaborazione di immagini di immunoistochimica provenienti da biopsie di tessuto mammario. L'analisi di queste immagini a scopo diagnostico viene attualmente effettuata dal personale medico in modo qualitativo e pertanto si registra un'elevata variabilità sia intra che inter-operatore. L'obiettivo principale della tesi, che si inserisce in questo contesto, è proprio quello di realizzare un algoritmo automatico in grado di analizzare queste immagini e fornire dei risultati numerici oggettivi e ripetibili a supporto dell'anatomopatologo. L'analisi automatica delle immagini sarà eseguita tramite il software Matlab partendo dal file originale nel formato proprietario del microscopio, da qui verrà riconosciuto il tessuto istologico e su quest'ultimo verranno applicate diverse tecniche di elaborazione per distinguere le cellule positive al marcatore immunoistochimico (colorate in marrone) da quelle negative (colorate in blu). L'algoritmo sviluppato va ad identificare le singole cellule utilizzando criteri morfologici e cromatici, fino ad arrivare ad un risultato quantitativo finale espresso come numero di cellule risultate positive sul totale di cellule presenti nel campione. La validazione dell'algoritmo sarà effettuata tramite il confronto dei risultati ottenuti dall'algoritmo con le segmentazioni manuali effettuate da un operatore esperto.

1.1 Tumore al seno

Il seno, situato tra cute e parete del torace, è prevalentemente composto da tessuto ghiandolare e tessuto adiposo. Un insieme di strutture ghiandolari viene chiamato lobulo, diversi lobuli formano un lobo, all'interno di un seno si trovano dai 15 ai 20 lobi. La moltiplicazione incontrollata delle cellule che formano le ghiandole e la loro conseguente trasformazione in cellule maligne causa il tumore al seno, malattia potenzialmente molto grave a meno di una tempestiva individuazione e cura. Le cellule maligne sono in grado di staccarsi dal tessuto in cui si sono generate e, migrando, andare ad attaccare altri tessuti o

organi del corpo umano. I tumori al seno più frequenti nascono dai lobuli o dalle cellule delle pareti dei dotti galattofori (i dotti che portano il latte materno al capezzolo)[1].

I dati epidemiologici riguardanti il tumore al seno sono fondamentali per riconoscere la rilevanza di questa patologia nella società moderna e soprattutto per mettere in evidenza l'importanza di una pronta e precisa diagnosi. In Italia sono stati diagnosticati 52'300 tumori alla mammella solo nel 2018, questa cifra rappresenta il 14% della totalità di tumori maligni diagnosticati nello stesso anno tra uomini e donne e il 30% di quelli riscontrati tra sole donne. Qualsiasi sia la fascia d'età presa in analisi il tumore al seno rimane quello più frequente nelle donne, infatti rappresenta il 41% dei tumori diagnosticati entro i 50 anni, il 35% tra i 50 e i 69 anni e il 22% nelle donne over 70.

	Età		
Rango	0-49	50-69	70+
1	Mammella (41%)	Mammella (35%)	Mammella (22%)
2	Tiroide (15%)	Colon-retto (11%)	Colon-retto (16%)
3	Cute (melanomi) (7%)	Polmone (7%)	Polmone (8%)
4	Colon-retto (4%)	Utero corpo (7%)	Pancreas (6%)
5	Utero cervice (4%)	Tiroide (5%)	Stomaco (5%)

Tabella 1: Incidenza dei tumori nel sesso femminile suddivisi per fascia di età (Pool AIRTUM 2008-20014)

L'età gioca un ruolo importante nell'incidenza di questa patologia, infatti nelle donne under 25 è pari a 1-2 casi su 100'000 ogni anno, 10 su 100'000 tra i 25 e i 29 anni, 150 su 100'000 tra i 40 e i 44 anni che portano ad un totale di 5000 nuovi casi ogni anno in donne under 45 anni. Tra i 45 e i 49 anni ci si attesta su 200-250 diagnosi positive su 100'000 e si ritrovano gli stessi dati anche tra i 50 e i 59 anni. Ciò porta a 23'000 tumori alla mammella riscontrati in donne di età compresa tra 50 e 69 anni in un anno, mentre invece nello stesso arco di tempo ne vengono diagnosticati 18'000 in donne che superano i 70 anni. Anche la mortalità aumenta con l'età, il 60% dei decessi di donne a causa di questa patologia si verifica infatti oltre i 70 anni.

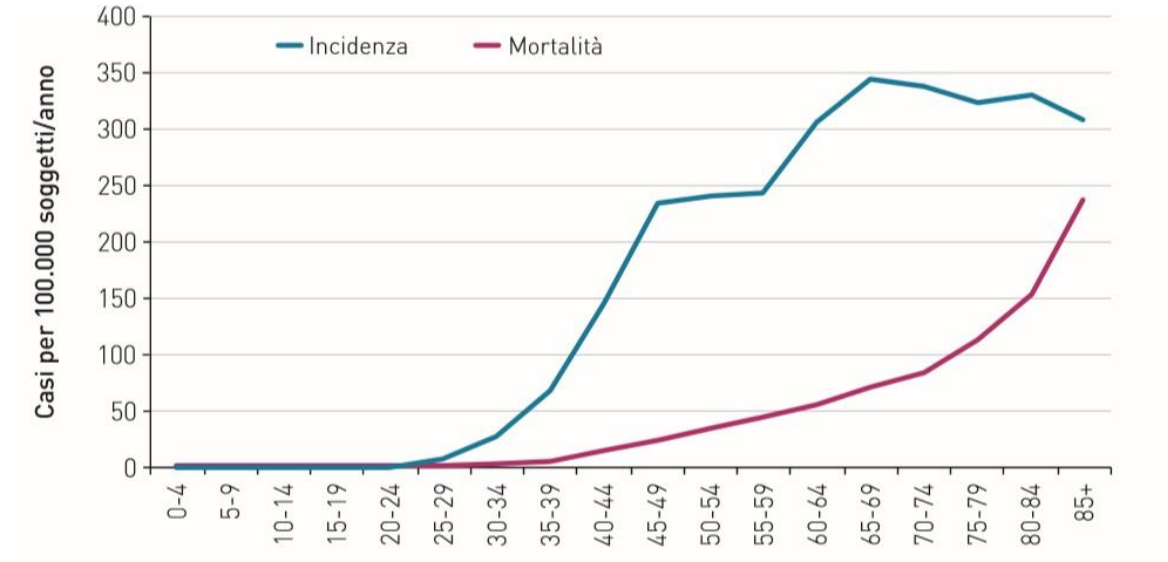


Figura 1: Tassi età-specifici del tumore della mammella (Pool AIRTUM 2008-2014)

Secondo i dati ISTAT nel 2015 in Italia ci sono stati 12'274 decessi causati da tumore alla mammella, in particolare questa cifra rappresenta l'8% delle morti totali causate da tumore e il 17% delle morti oncologiche femminili.

Questi dati risultano fondamentali per avere un quadro chiaro ed oggettivo della situazione in Italia riguardante il tumore al seno ed è altrettanto importante sottolineare che ad oggi grazie alla grande possibilità di diagnosticare precocemente la neoplasia e a tutto il processo di screening la prognosi è considerata ottima, come riportato anche dall'AIOM (Associazione Italiana di Oncologia Medica)[2].

Il cancro al seno può essere di due tipi, invasivo o non invasivo, per quanto riguarda le forme non invasive queste sono la DIN (neoplasia duttale intraepiteliale) e la LIN (neoplasia lobulare intraepiteliale), mentre le forme invasive sono: il carcinoma duttale, che rappresenta tra il 60 e il 70% delle forme di cancro al seno diagnosticate, il carcinoma lobulare, 10-15%, e altre con incidenza minore ovvero il carcinoma tubulare, papillare, mucinoso, cribriforme. Il tumore è soggetto anche ad una classificazione in uno dei seguenti stadi in base alla gravità:

- Stadio 0, detto anche carcinoma in situ, può essere lobulare o duttale;

- Stadio I, fase iniziale, il tumore si presenta in dimensioni ridotte e i linfonodi non sono coinvolti;
- Stadio II, fase iniziale, in cui il tumore può avere o dimensioni maggiori dello stadio I senza coinvolgimento dei linfonodi o le stesse dimensioni ma con coinvolgimento;
- Stadio III, tumore localmente avanzato che coinvolge i linfonodi o i tessuti adiacenti (pelle)
- Stadio IV, cancro con presenza di metastasi con coinvolgimento di altri organi interni.

Nel momento in cui il tumore viene diagnosticato e classificato allo stadio 0 la sopravvivenza a cinque anni si attesta al 98%, nel momento in cui invece i linfonodi risultano positivi questa scende al 75%. Inoltre, per ciò che concerne il cancro con metastasi, le pazienti che subiscono il trattamento chemioterapico hanno una sopravvivenza media di due anni[1].

1.2 Metodi di diagnosi

Nella maggior parte delle patologie la diagnosi rappresenta una fase cruciale per il buon esito di una eventuale terapia, in particolar modo per il tumore al seno, come confermato anche dai dati statistici visti prima. Gli esami che vengono effettuati sono i seguenti:

- Visita senologica, si tratta di una visita specialistica nella quale il medico raccoglie dati familiari e personali in modo da creare una precisa anamnesi, viene eseguito anche un esame visivo e tattile su seno e ascelle al fine di scegliere al meglio gli esami successivi;
- Mammografia, questo esame viene fatto utilizzando un macchinario apposito, il mammografo, ed sfrutta i raggi X per creare un'immagine del seno che può essere impressa su una lastra o digitalizzata nei mammografi digitali. Attraverso queste immagini è possibile individuare la presenza di masse di densità maggiore a quella del tessuto adiposo del seno e che quindi risultano più chiare in un'immagine a toni di grigio. La mammografia è l'esame d'elezione per la diagnosi precoce del tumore alla mammella;

Figura 2: Apparecchiatura per mammografia

- Ecografia, viene utilizzata come supporto alla mammografia nel caso in cui quest'ultima abbia dato risultati dubbi, sfrutta gli ultrasuoni e può servire da guida durante altri esami come la biopsia o durante interventi chirurgici;

Figura 3: Ecografo

- Tomosintesi mammaria, si tratta di un esame in cui il seno viene scomposto in tante immagini prese da angolazioni diverse che permettono la ricostruzione di un volume tridimensionale utile ad analizzare le lesione più piccole o posizionate in punti che ne rendono difficile l'individuazione tramite mammografia;
- Risonanza magnetica nucleare, è un esame che viene definito di seconda integrazione rispetto alla mammografia, utilizza campi magnetici ad alta intensità per scansionare il volume corporeo selezionato e, in questo caso, necessita dell'iniezione di un mezzo di contrasto per poter distinguere le aree della mammella sane da quelle malate[3];

Figura 4: Risonanza magnetica nucleare

- Agoaspirato e agobiopsia, sono tecniche che permettono di prelevare campioni di tessuto biologico dall'area interessata tramite un ago cavo, leggermente più spesso di quelli utilizzati per le normali iniezioni. Sul tessuto prelevato vengono fatti esami citologici (agoaspirato) e istologici (biopsia)[4].

1.3 Immunoistochimica

L'immunoistochimica è una tecnica complessa e specifica atta a rilevare determinati antigeni nei tessuti o nelle cellule che devono essere analizzate. Tale tecnica sfrutta una reazione immunitaria tra l'antigene, eventualmente presente sul campione opportunamente preparato, e l'anticorpo specifico posto su di esso. La rivelazione sarà successivamente effettuata tramite da un anticorpo secondario coniugato ad un enzima catalizzatore che reagisce con un substrato colorando le cellule in modo che possano essere analizzate tramite un microscopio ottico[5].

L'immunoistochimica è un importante strumento per la diagnosi del tumore al seno in particolare grazie all'uso di immunomarcatori appropriati e tecniche standardizzate per la valutazione dei risultati. Diversi studi hanno esplorato l'espressione di alcuni immunomarcatori in tumori fibroepiteliali del seno, tra questi è stato messo in evidenza l'indice proliferativo Ki-67 rivelatosi particolarmente utile nell'individuazione del tumore primitivo (PT) e nella classificazione appunto dei tumori fibroepiteliali[6].

Le immagini oggetto di analisi in questo lavoro di tesi sono state trattate proprio con Ki67-ematossilina, rendendo possibile il riconoscimento delle cellule positive a questo indice proliferativo grazie alla loro colorazione marrone, mentre l'ematossilina colora in blu le cellule negative.

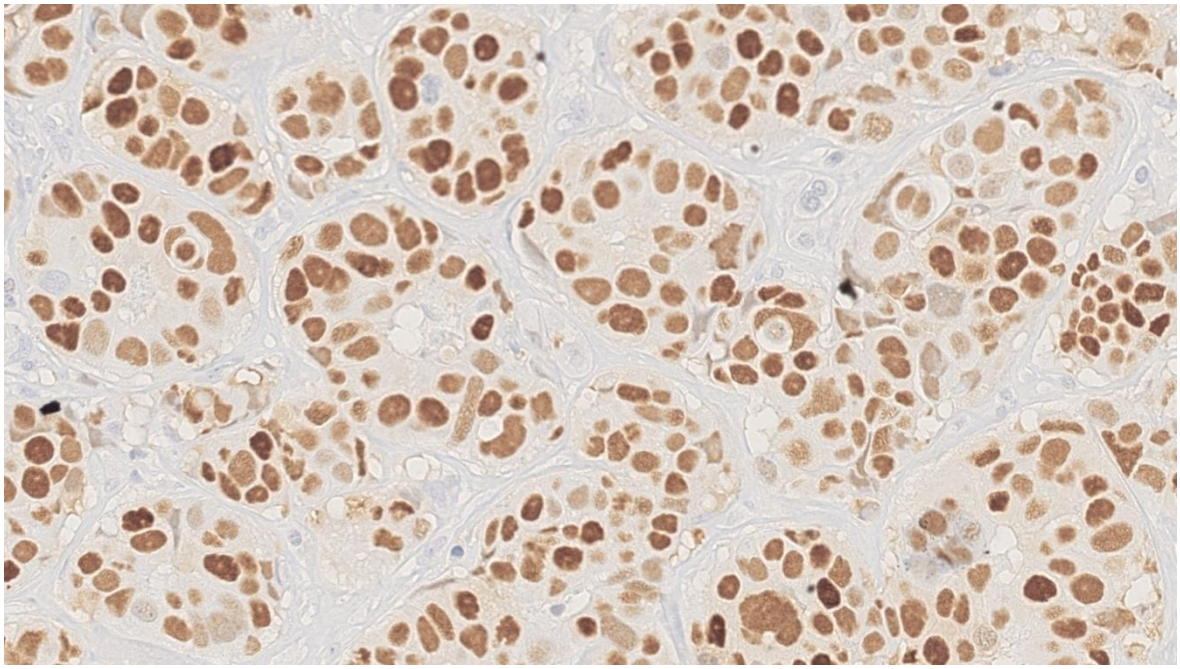


Figura 5: Immagine di biopsia trattata con Ki67-ematossilina

L'analisi di queste immagini avviene tuttora ad opera di personale specializzato (anatomopatologi) che, osservando l'immagine al microscopio, valuta la percentuale di cellule positive sul totale, si tratta quindi di una misura qualitativa. Questo tipo di analisi è dunque strettamente legato all'operatore ed in particolar modo alla sua esperienza; proprio per questo motivo soffre inevitabilmente di una grande variabilità sia intra che inter-operatore.

1.4 Obiettivi del progetto

All'interno del contesto descritto precedentemente si inserisce questo progetto di tesi, in particolare gli obiettivi preposti sono quelli di creare un algoritmo completamente automatico che possa restituire come risultato una misura quantitativa della percentuale di cellule positive sul totale. Deve essere in grado di iniziare l'elaborazione delle immagini partendo direttamente dal formato in cui queste sono ottenute tramite il microscopio, procedere con le analisi e calcolare, in un lasso di tempo confrontabile con quello che impiega un operatore esperto, un risultato stabile nel tempo, oggettivo e il più possibile preciso.

2.MATERIALI E METODI

2.1 Ambiente si sviluppo

L'intero progetto è stato portato avanti su un PC portatile modello ASUS F552M Series con processore Intel® Pentium® CPU N3540 @2.16 GHz dotato di 8.00 GB di memoria RAM e scheda grafica NVIDIA GeForce 920M con memoria DDR3 da 2 GB.

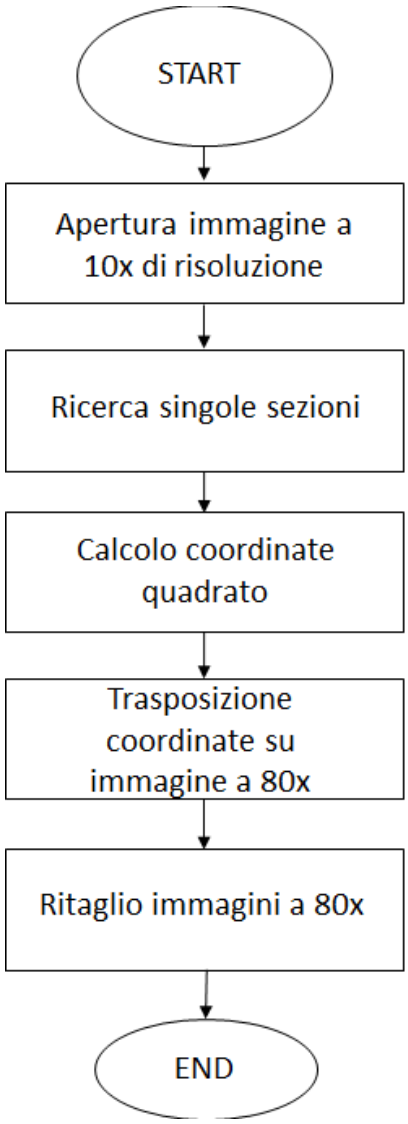
L'algoritmo è stato sviluppato e testato su Matlab R2018a ed inoltre è stato utilizzato anche il software NDP.view 2 per la visualizzazione delle immagini facenti parte del dataset iniziale nel formato originale. Tramite quest'ultimo software sono state anche create delle immagini di dimensioni ridotte, ritagliando le originali, per alleggerire dal punto di vista computazionale l'elaborazione delle stesse.

2.2 Pipeline

Di seguito verrà esposta la pipeline dell'algoritmo, per favorire una più semplice consultazione è stata divisa in due pipeline distinte, una che riguarda la prima parte (sarà chiamata pipeline preprocessing), ovvero gli step che permettono di ottenere delle immagini adatte all'elaborazione, e una seconda (pipeline segmentazione) che concerne la segmentazione cellulare automatica.

2.2.1 Pipeline preprocessing

Il workflow della prima parte dell'algoritmo è il seguente:



- Apertura immagini a 10x di risoluzione: Le immagini si presentano nel formato .ndpi, si tratta del formato con cui vengono salvate dal microscopio attraverso il quale vengono acquisite. Il file contiene la stessa immagine a diverse risoluzioni e quindi con dimensioni in pixel diverse, per questa ragione si è reso necessario scegliere quale di queste utilizzare e trovare il modo per estrapolarla direttamente su Matlab considerando il fatto che esistono delle dimensioni massime di un'immagine su cui questo software possa lavorare. La scelta è ricaduta sulla risoluzione da 10x che permette un buon compromesso tra velocità di calcolo, qualità dell'immagine per permettere il riconoscimento delle zone di interesse e capacità di calcolo del

programma. Questo è stato possibile con la funzione “Imread” una volta indagata quale fosse la risoluzione di ogni singola immagine del file.

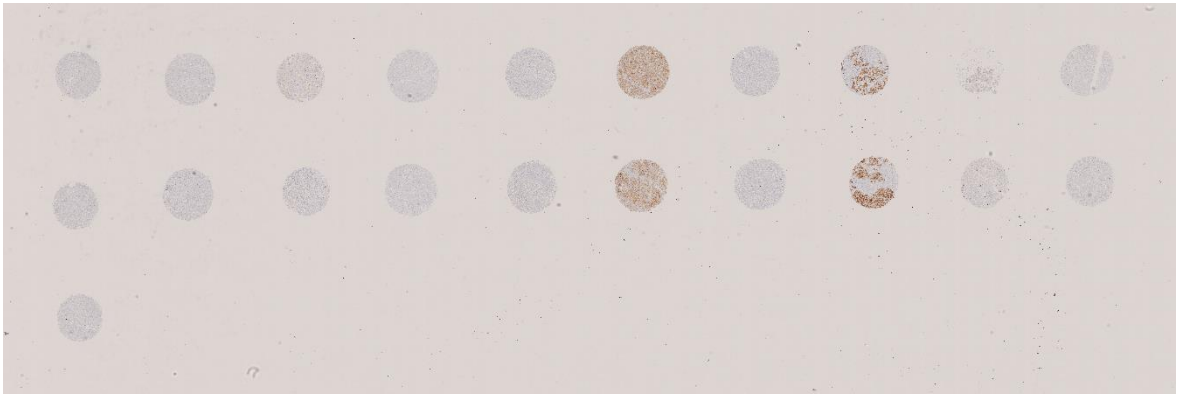


Figura 6: Immagine originale salvata in formato .png

Come è possibile notare nella figura precedente diverse sezioni di biopsia vengono normalmente inserite nella stessa immagine e per questo motivo il passo successivo all’apertura dell’immagine è proprio quello della ricerca di queste singole sezioni.

- Ricerca singole sezioni: per questo step è stata utilizzata la funzione “extract_bg1” che permette, attraverso una soglia sull’intensità del colore e una sulle dimensioni di riconoscere le zone di interesse dallo sfondo dell’immagine. Si tratta di una funzione modificata appositamente per lo scopo del progetto che restituisce una maschera in bianco e nero, con in bianco le sezioni trovate e in nero lo sfondo, e le coordinate di tutti i punti del bordo delle sezioni.

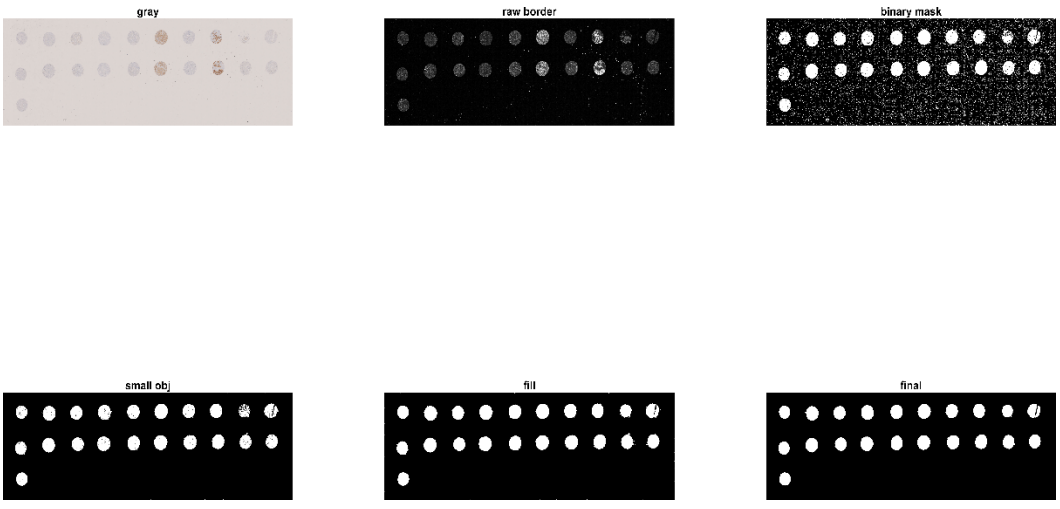


Figura 7 : Step riconoscimento sezioni

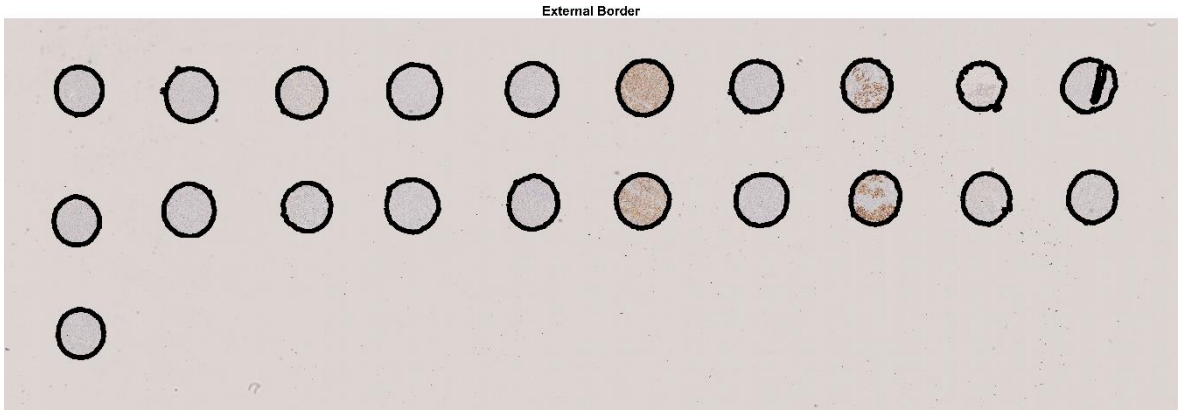


Figura 8: Immagine originale con plot dei bordi delle sezioni individuate

- Calcolo coordinate quadrato: una volta ottenute le coordinate di ogni pixel del bordo viene calcolata un'area che contiene la sezione in esame con un margine di sicurezza di 80 pixel e le sue coordinate come mostrato nell'immagine seguente. È stato utilizzato anche un controllo sulla possibilità che quest'area superi i margini dell'immagine, in quel caso infatti si prende come lato di quest'area proprio il margine dell'immagine.

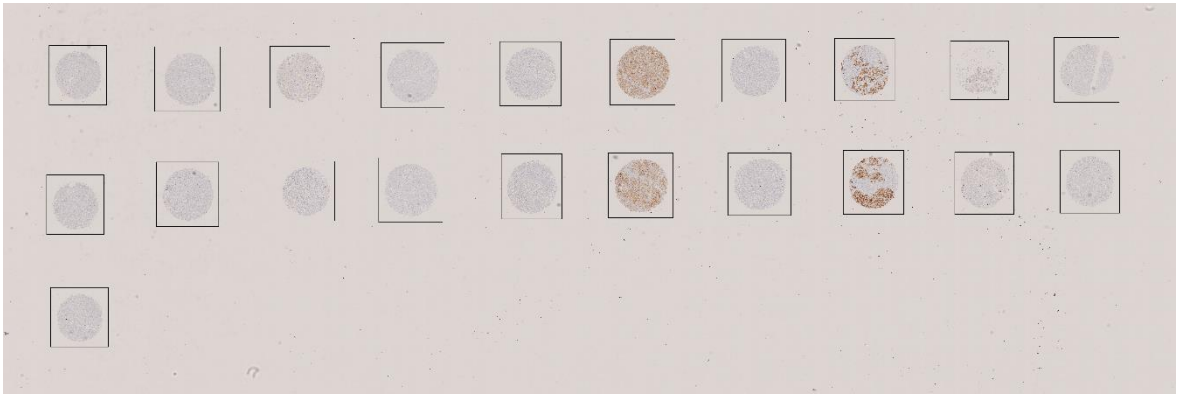


Figura 9: Immagine originale con plot del perimetro delle aree che saranno ritagliate

- Trasposizione coordinate su immagine a 80x: le coordinate calcolate nel punto precedente fanno riferimento all'immagine a 10x e per ottenere le corrispettive nell'immagine a 80x è sufficiente moltiplicarle per 5 avendo sempre un controllo sui margini dell'immagine completa come al punto precedente.

- Ritaglio immagini a 80x: questo è l'ultimo step di questa prima pipeline ed è stato eseguito con la funzione "bfopen1", una modifica della funzione "bfopen" facente parte della libreria open source "bfmatlab"[7]. Con questo comando è possibile creare una nuova immagine in cui sarà presente solo un'area selezionata di una delle immagini che compongono il file originale nel formato "ndpi". È stata scelta la risoluzione massima ed alcuni dei risultati saranno esposti di seguito.

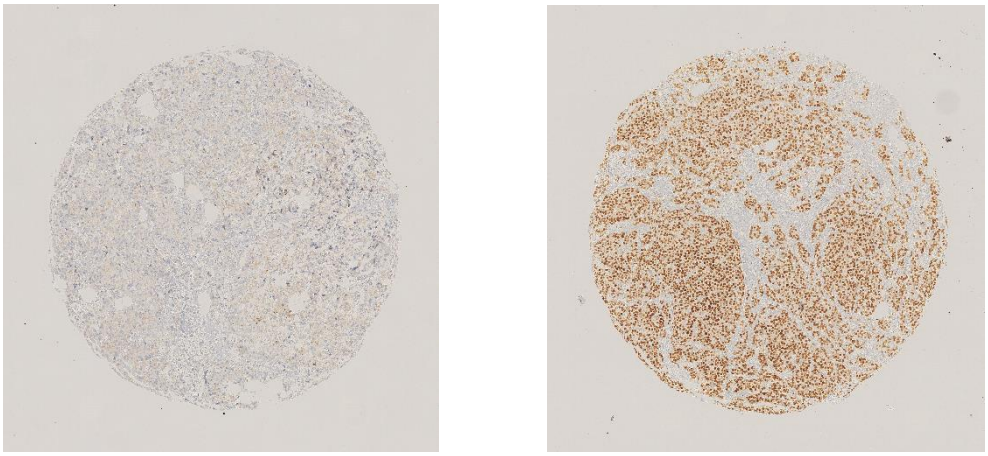


Figura 10: ritaglio immagini ad alta definizione

2.2.2 Pipeline segmentazione

Per la fase di progettazione e di verifica di questa sezione sono state utilizzate delle immagini ritagliate alla risoluzione massima direttamente dalle immagini originali attraverso il software NDP.view2. Questo accorgimento è stato fondamentale per poter snellire dal punto di vista computazionale entrambe le fasi, i parametri utilizzati possono essere gli stessi anche per le immagini ottenute come output dalla pipeline precedente data la risoluzione identica di entrambe.

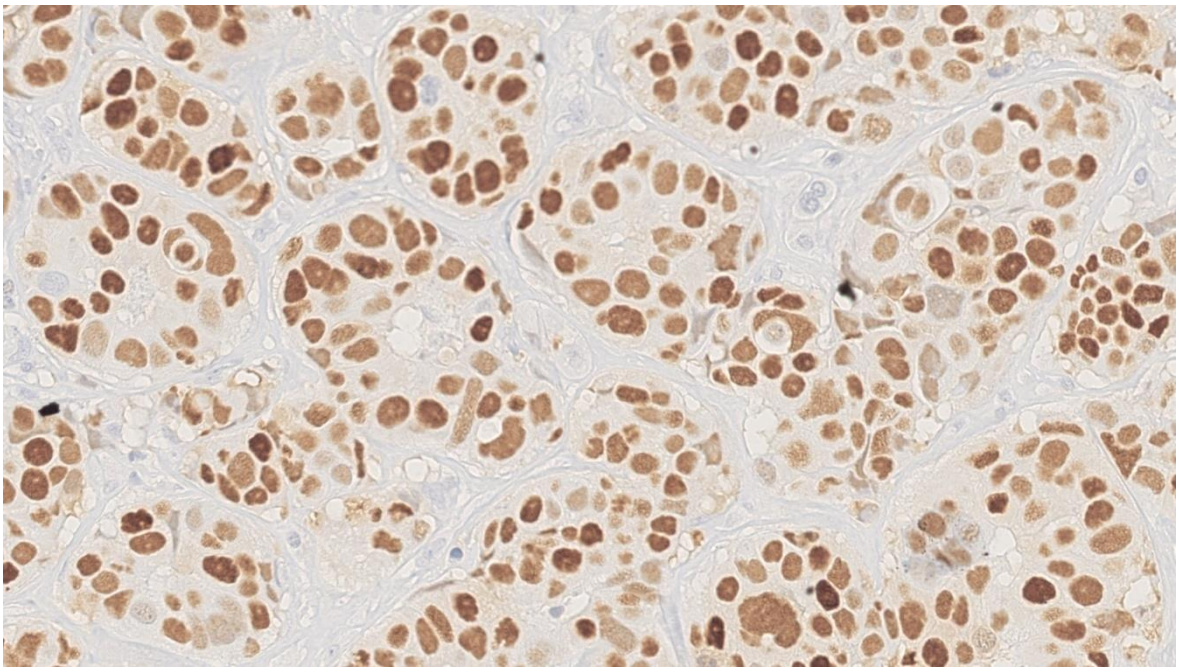
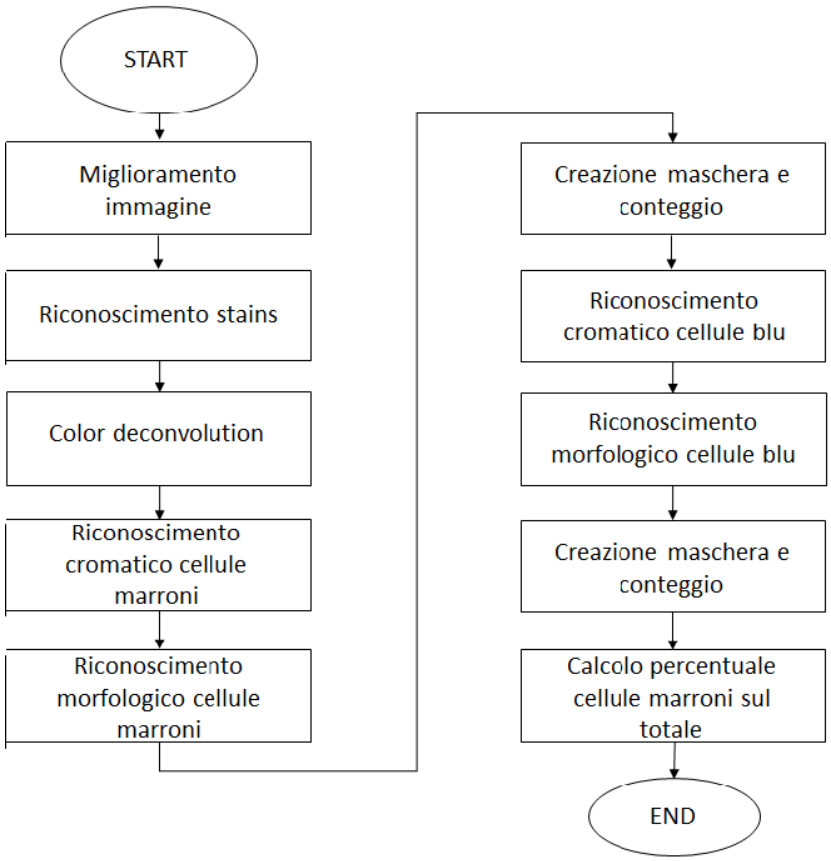


Figura 11 : Immagine esemplificativa del dataset creato

Il seguente workflow sintetizza i vari step della segmentazione.



- Miglioramento immagine: si tratta di una fase che precede la vera e propria segmentazione ma è stata fondamentale per l'ottenimento di buoni risultati. Le immagini fornite come dataset presentavano notevoli diversità dal punto di vista dell'intensità dei colori e questa si poteva riscontrare anche all'interno della stessa immagine, dato che ai fini del progetto l'interesse non era focalizzato sull'intensità della colorazione ma solo sull'eventuale presenza o meno, un miglioramento generale della qualità dell'immagine non ha inficiato in alcun modo i risultati. I miglioramenti sono stati portati seguendo tre vie, la prima e più influente è l'aumento di contrasto effettuato direttamente sull'immagine a colori con la funzione "adapthisteq" di Matlab. Le altre due vie sono rispettivamente l'eliminazione dei pixel bianchi e dei pixel neri che potrebbero causare un'errata stima dei colori nella fase successiva.

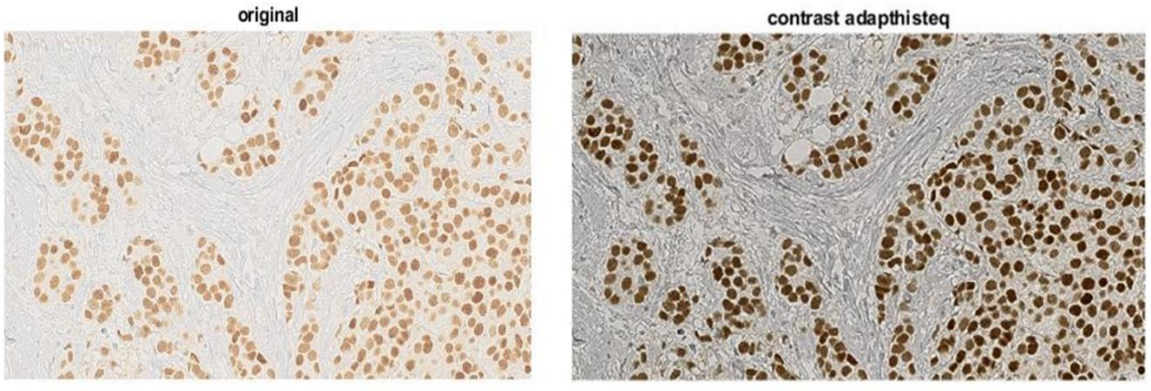


Figura 12: Immagine pre e post applicazione del miglioramento del contrasto

- Riconoscimento stains: questo step è deputato al riconoscimento dei colori presenti nell'immagine. Sono state provate due funzioni diverse per questo scopo, la prima è stata "EstUsingMacenko" che sfrutta le diverse lunghezze d'onda assorbite dai vari colori[8], mentre la seconda funzione è stata "EstUsingHSD" creata sulla base della "hue saturation intensity"[9]. Quest'ultima ha condotto a risultati decisamente migliori ed è stata quindi scelta in via definitiva. L'immagine viene fornita in ingresso alla funzione non come matrice, come viene vista in realtà da Matlab, ma come vettore ed è proprio in

questo momento che intervengono i due meccanismi che sono stati menzionati al punto precedente per l'eliminazione dei pixel bianchi e di quelli neri in modo tale che questa stima non sia perturbata. L'output di questa funzione è un vettore con la stima dei colori presenti nell'immagine.

- Color deconvolution: sfruttando il vettore in uscita al punto precedente vengono ricostruite due immagini, una per ogni colore individuato. Nel caso in cui nell'immagine fosse presente un solo colore è stato previsto un controllo che annulla la ricostruzione di un'immagine. Le funzioni utilizzate in questo caso sono due, "Deconvolve" per creare un'immagine formata da tre layer in scala di grigi in cui ogni layer mette in evidenza uno dei colori trovati e "PseudoColourStains" che ricostruisce effettivamente due immagini distinte avendo come input sia il vettore degli stains che l'output di "Deconvolve"

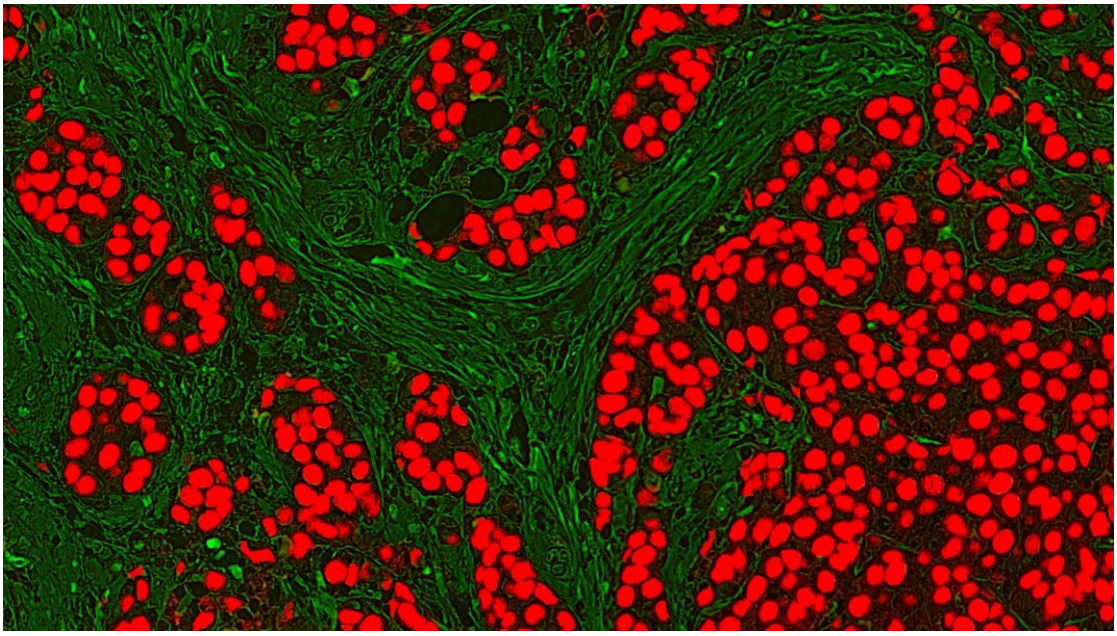


Figura 13: Immagine in uscita dalla funzione deconvolve

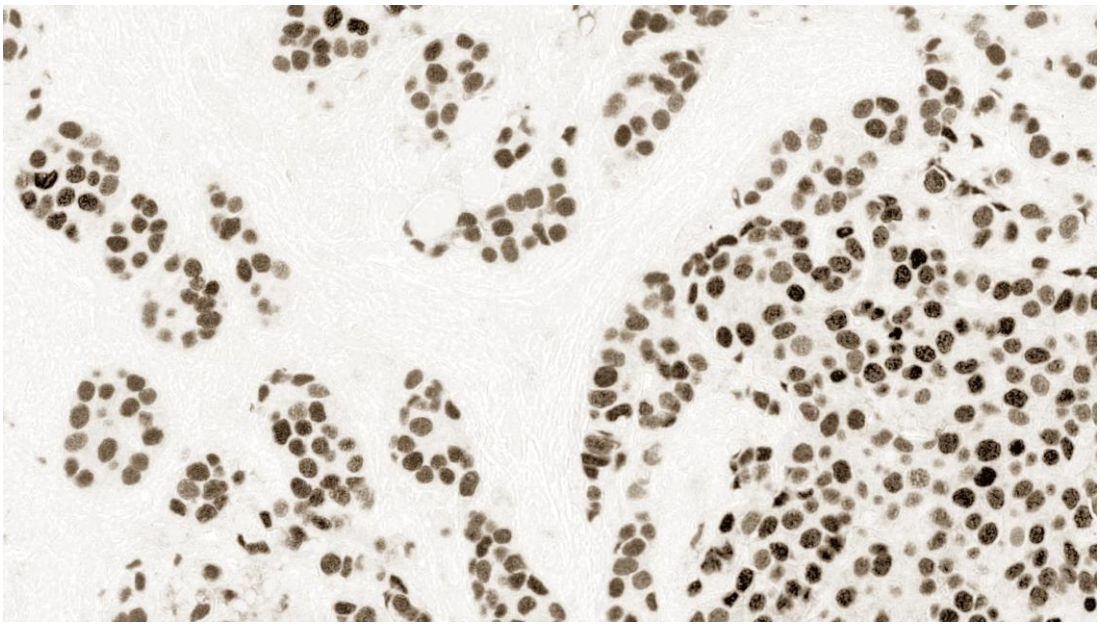


Figura 14: Prima immagine in uscita dalla funzione PseudoColourStains

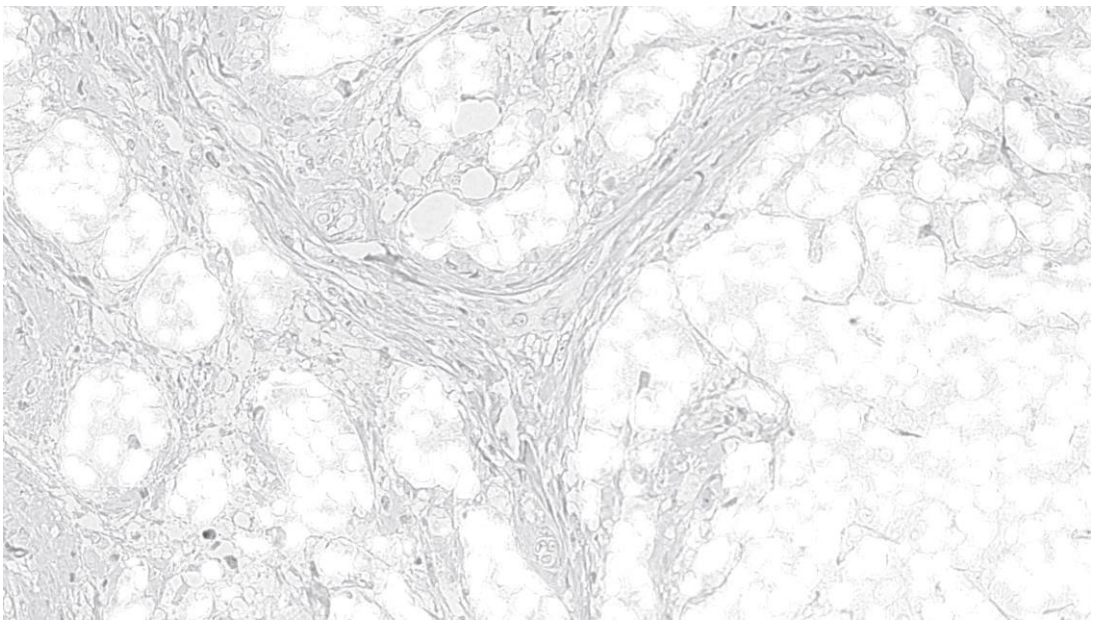


Figura 15: Seconda immagine in uscita dalla funzione PseudoColourStains

- Riconoscimento cromatico cellule marroni: da questo punto inizia il riconoscimento delle sole cellule marroni, quelle positive, per farlo si utilizza il primo layer dell'immagine in uscita dalla funzione "Deconvolve"

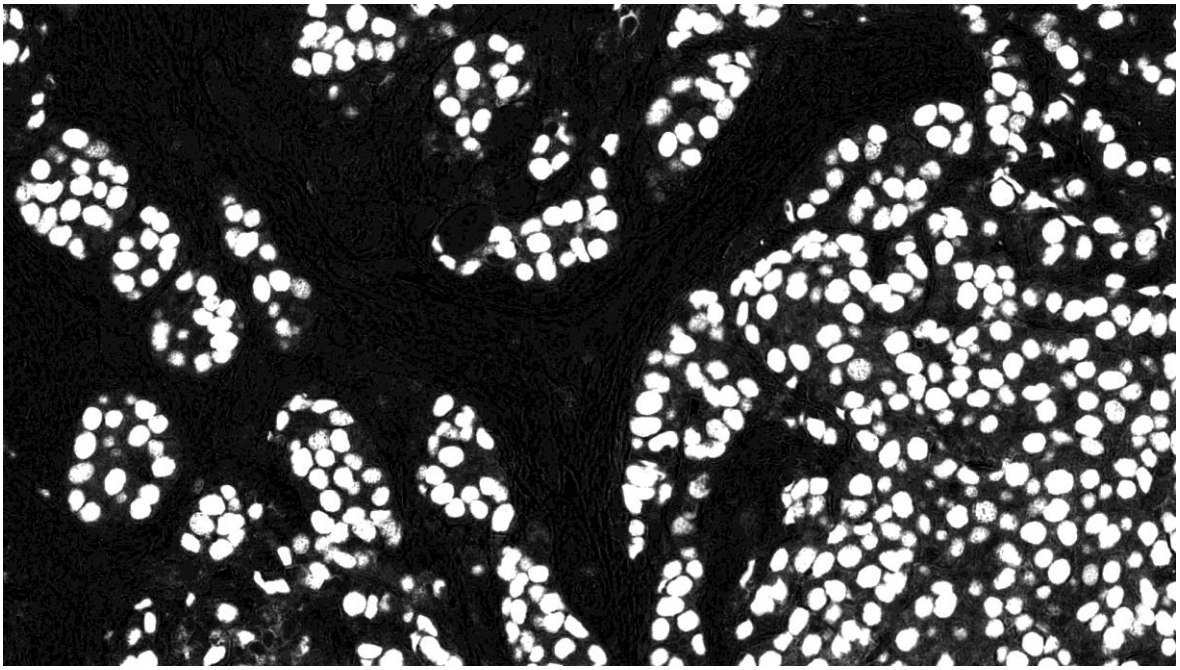


Figura 16: Primo layer dell'immagine in uscita dalla funzione deconvolve

e si applica una soglia per effettuare thresholding ed ottenere un'immagine in bianco e nero con in bianco tutto ciò che fino a questo punto è stato identificato come cellula positiva.

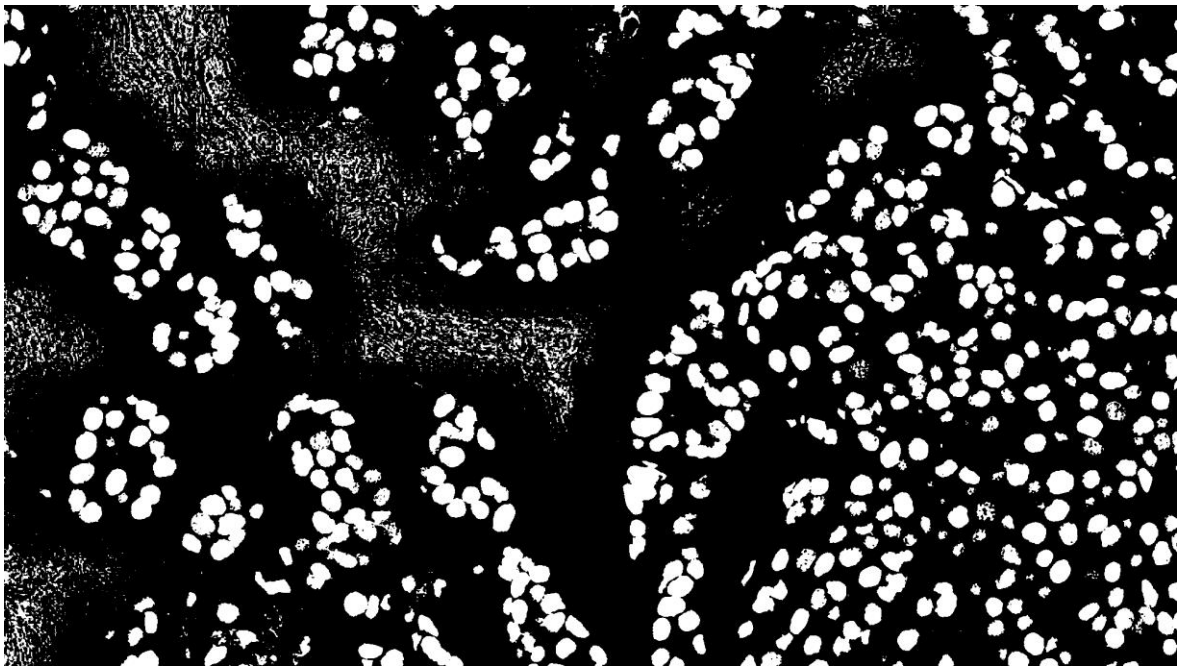


Figura 17: Uscita dal blocco di thresholding

- Riconoscimento morfologico cellule marroni: tutte le aree che hanno superato il primo controllo di tipo cromatico vengono sottoposte ad un controllo di tipo morfologico. I criteri scelti sono due, la dimensione e la forma. Per ciò che concerne la dimensione ne è stata scelta una minima quantificata in pixel che deve essere superata per poter considerare quell'area una vera e propria cellula.

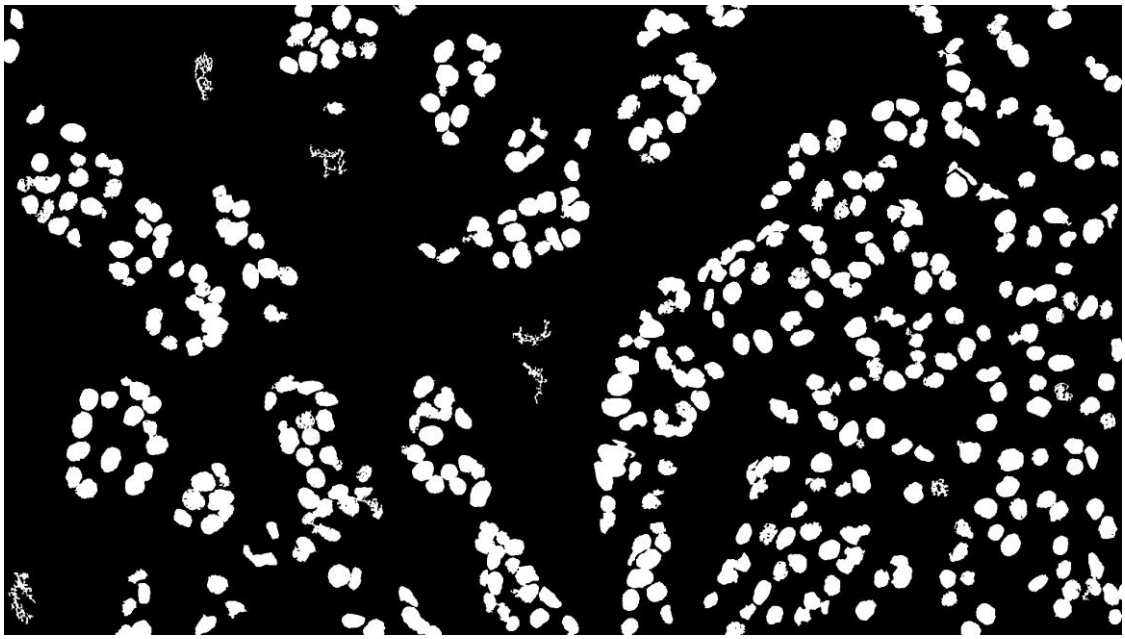


Figura 18: Eliminazione oggetti piccoli

A questo punto è stata inserita una parte di codice per fare in modo che l'algoritmo riempia automaticamente le aree bianche che però hanno all'interno qualche pixel bianco nero.

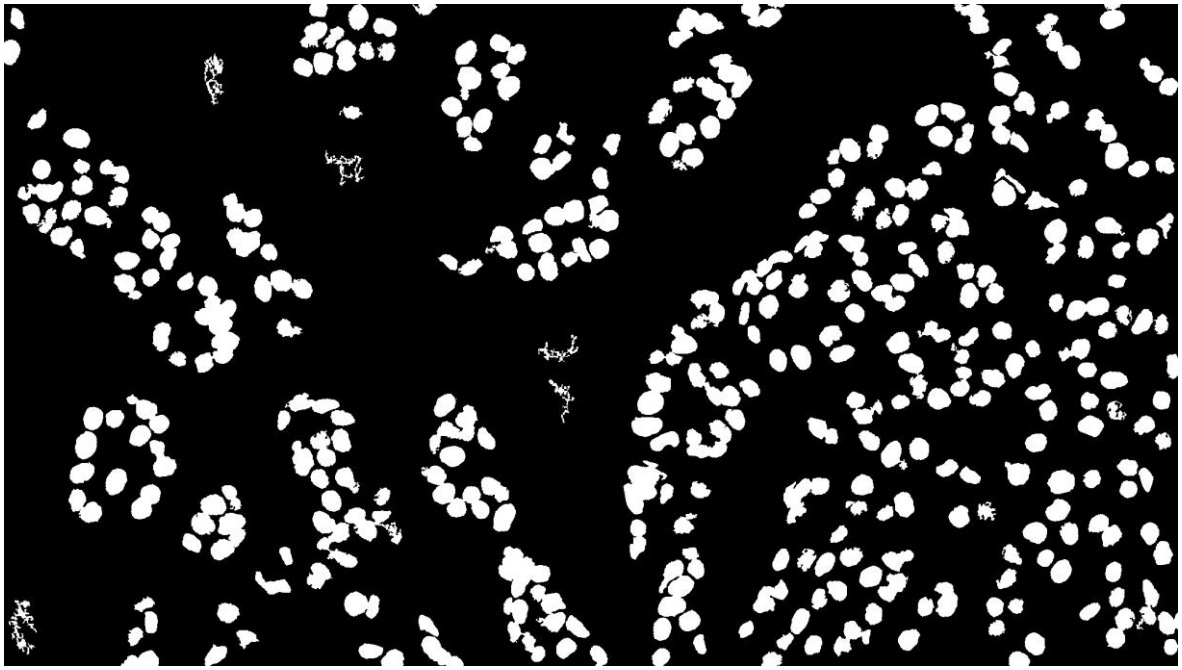


Figura 19: Riempimento oggetti

Per ciò che invece concerne la forma si è deciso di calcolare la circolarità di ogni area come $C = \frac{4\pi A}{P^2}$ con C=circolarità, P=perimetro, A=area, e un valore di tale parametro che deve essere superato per classificare definitivamente quell'elemento come cellula positiva.

- Creazione maschera e conteggio: dopo l'ultimo controllo al punto precedente si costruisce l'ultima maschera sulla quale poi verranno contate tutte le aree chiuse e il numero ottenuto determina il numero di cellule positive nel campione.

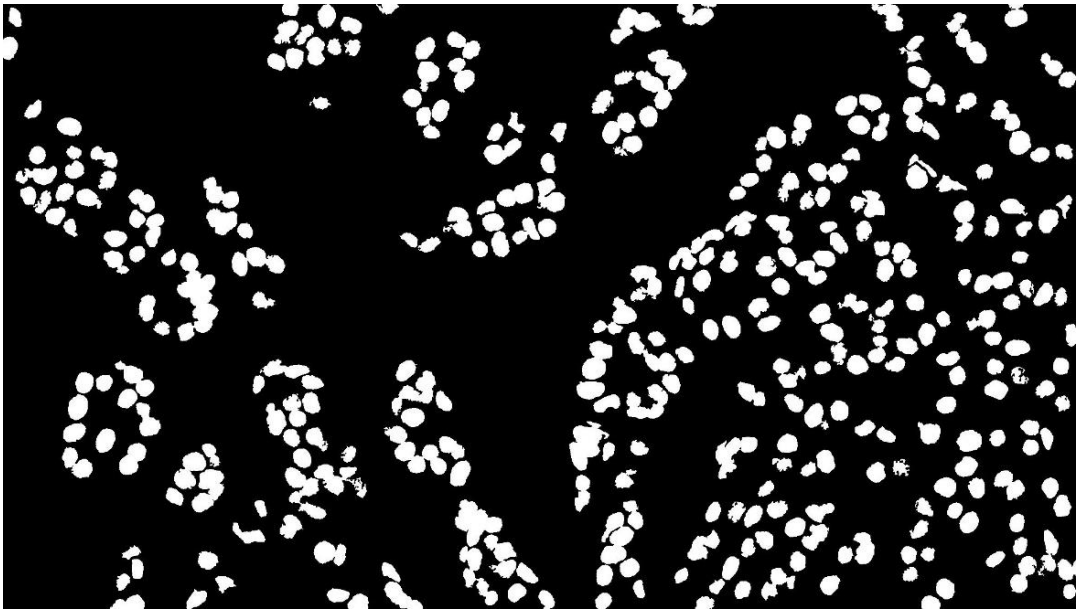


Figura 20: Maschera finale

- Riconoscimento cromatico cellule blu: per il riconoscimento delle cellule blu il procedimento è concettualmente identico rispetto a quello delle cellule marroni, in questo caso però si utilizza il secondo layer dell'immagine in uscita dalla funzione "Deconvolve".

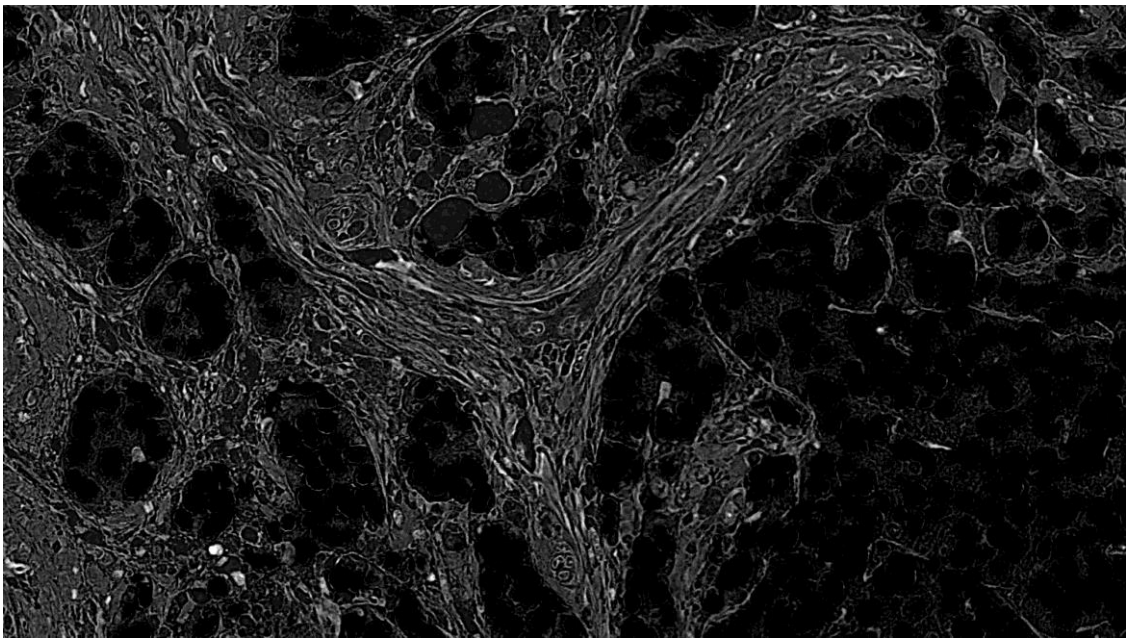


Figura 21:Secondo layer immagine in uscita dalla funzione deconvolve

Si applica un thresholding con una soglia più bassa per creare una prima maschera binaria e si ottiene la seguente immagine

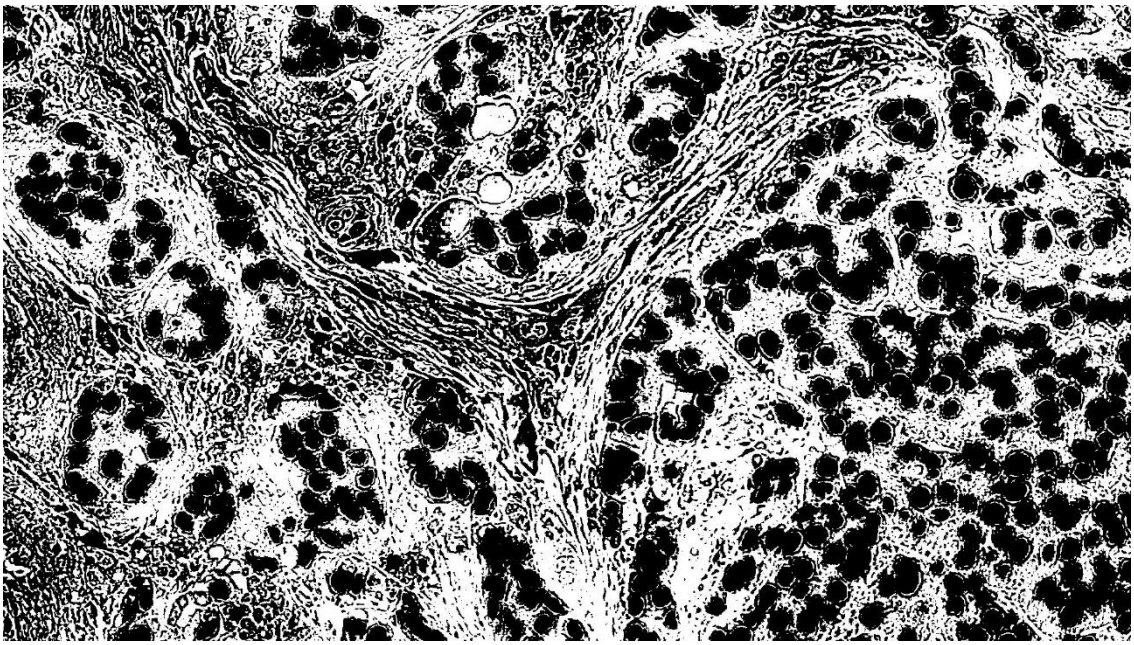


Figura 22: Immagine in uscita dal blocco di thresholding

- Riconoscimento morfologico cellule blu: anche in questo caso gli step sono gli stessi rispetto al caso delle cellule marroni ma vengono cambiati i valori delle

soglie, infatti si è deciso di diminuire la dimensione minima dell'area di una cellula e al contempo di scegliere un valore minimo di circolarità più stringente, questo una volta notato che le cellule negative tendono ad avere una forma più tondeggiante e ad essere mediamente più piccole.

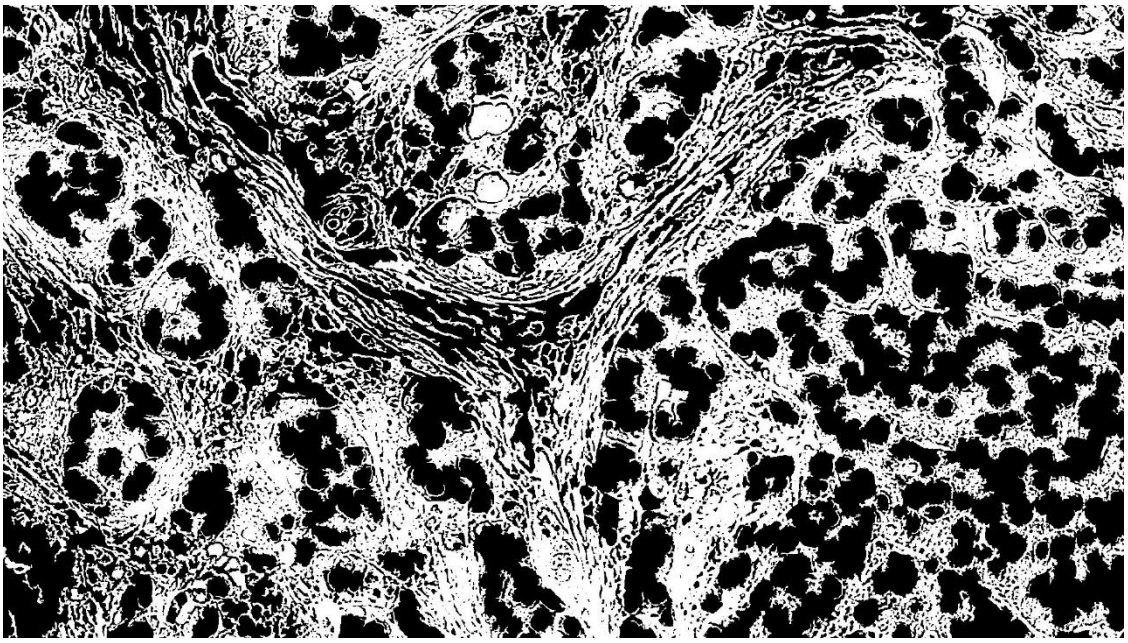


Figura 23: Rimozione oggetti piccoli

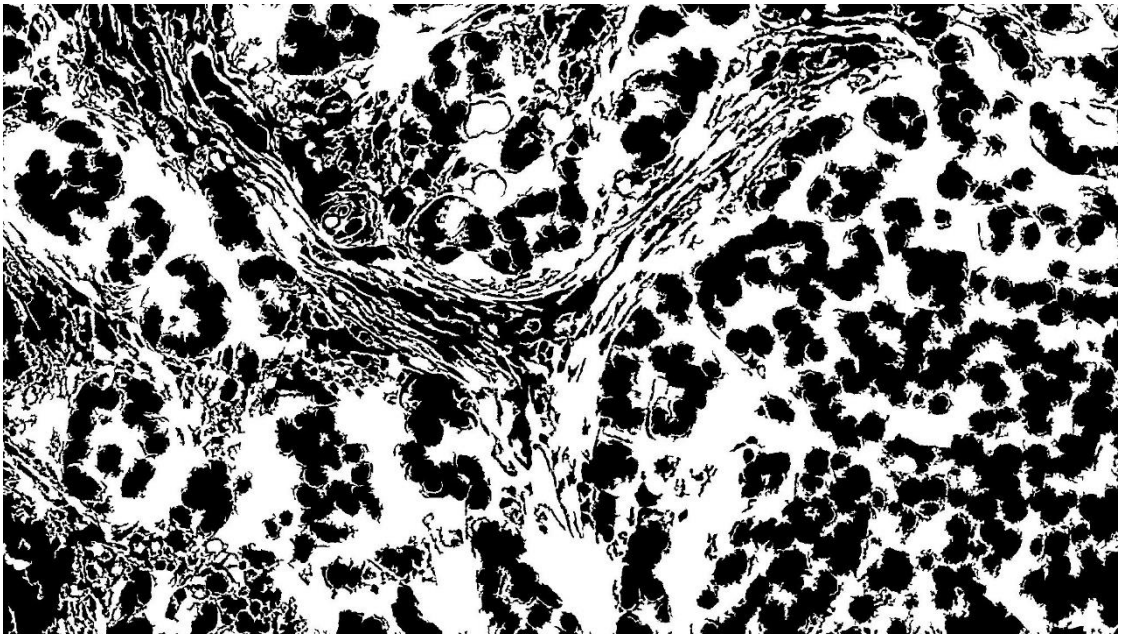


Figura 24: Riempimento oggetti

- Creazione maschera e conteggio: una volta eliminate le forme con un valore di circolarità sufficiente si procede con la creazione dell'ultima maschera e al conteggio delle cellule blu come si era fatto per quelle marroni.

Figura 25: Maschera finale

- Calcolo percentuale cellule marroni sul totale: una volta effettuate tutte le elaborazioni sopra descritte l'algoritmo è pronto a restituire in output un risultato numerico finale calcolando la somma tra le cellule positive e quelle negative che sono state trovate e calcolando a quanto ammonta la percentuale di queste è stata decretata positiva al test.

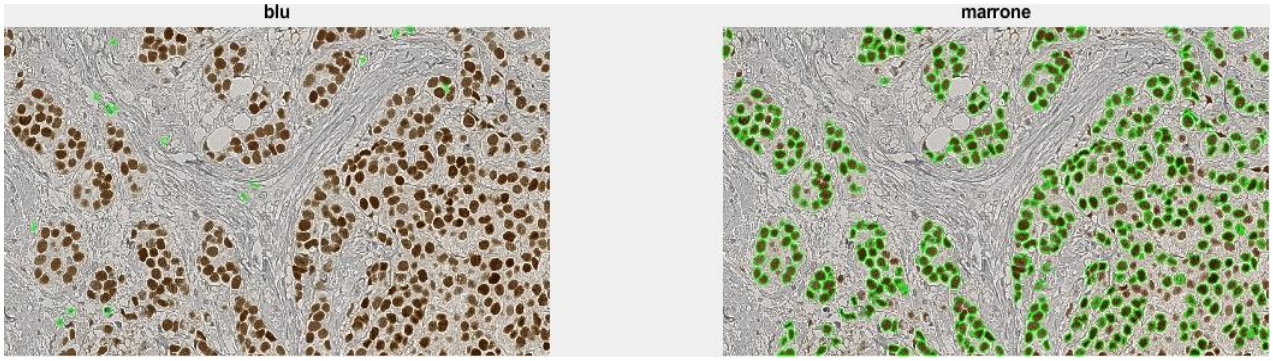


Figura 26: Immagine finale con plot dei bordi di tutte le cellule individuate sull'immagine originale

3. RISULTATI

Vengono riportate di seguito alcune immagini rappresentative dei risultati ottenuti tramite l'algoritmo di segmentazione automatica. Come è possibile evincere anche solo da una prima analisi visiva delle stesse, la segmentazione delle cellule marroni ha prodotto ottimi risultati riuscendo ad individuarne una grossa percentuale e non scambiando mai cellule negative per positive. La segmentazione delle cellule blu invece presenta dei risultati meno buoni ma anche in questo caso non capita mai che vengano individuate come cellule negative cellule che in realtà risultano essere positive.

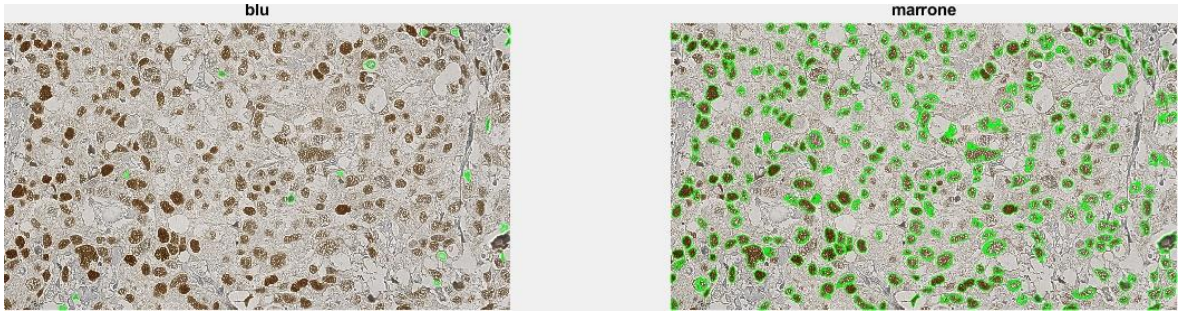


Figura 27: Immagine finale elaborazione

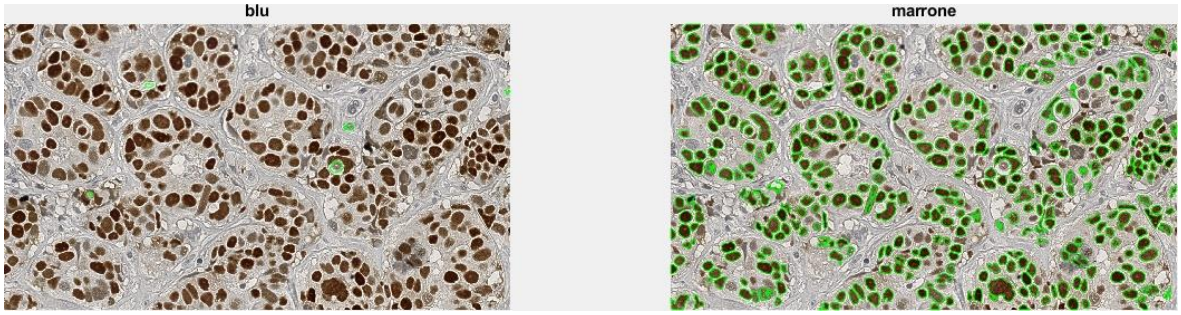


Figura 28: Immagine finale elaborazione

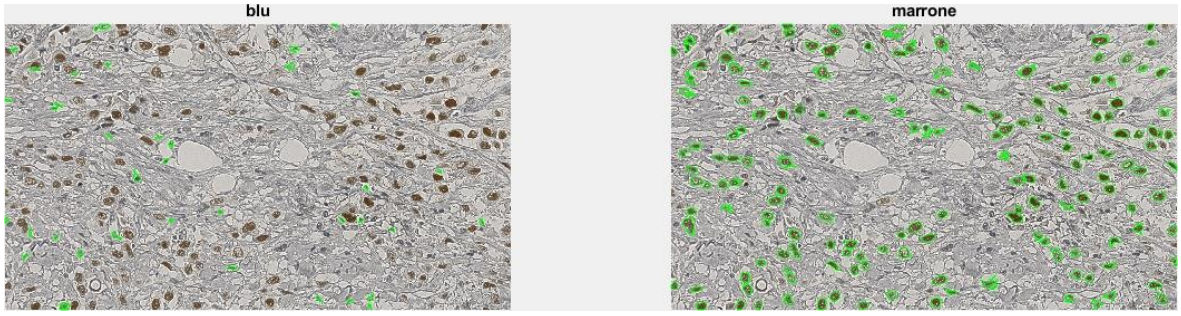


Figura 29: Immagine finale elaborazione

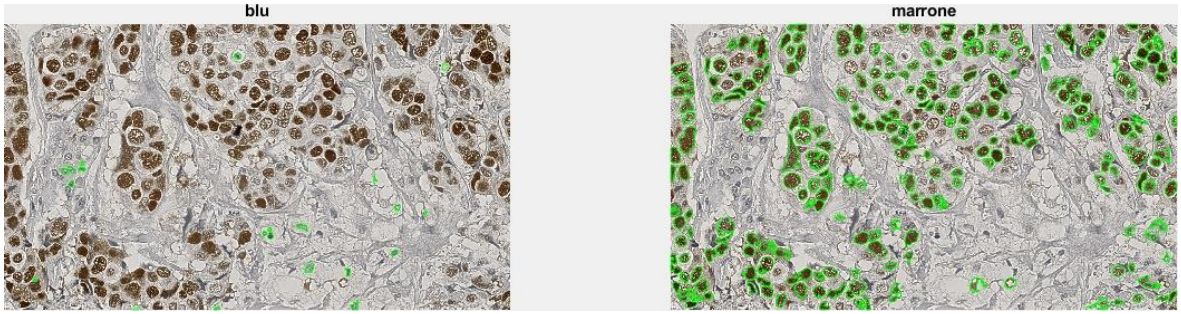


Figura 30: Immagine finale elaborazione

3.1 Risultati numerici

Per poter ottenere dei risultati quantitativi analizzabili con precisione si è deciso di confrontare quelli ottenuti attraverso la segmentazione automatica con i risultati derivanti dalla segmentazione manuale delle stesse immagini. Tale segmentazione è stata effettuata sul software Matlab attraverso un'interfaccia grafica che permetteva di tracciare con il mouse collegato al pc il contorno di qualsiasi area di interesse sulle immagini del dateset, calcolando automaticamente delle maschere in bianco e nero con evidenziate in bianco le aree selezionate manualmente e creando anche un'immagine con i bordi delle suddette aree plottate sulle immagini originali. Sono state così create due maschere distinte per ogni immagine, una per la segmentazione manuale delle cellule marroni e una per la segmentazione manuale delle cellule blu.

I dati ottenuti vengono riportati nelle tabelle seguenti e saranno analizzati nel dettaglio.

Nome	Precisione	Recall	F1 score	N° cell auto	N° cell manual
img_1.jpg	0,436065657	0,735974	0,547649	102	90
img_2.jpg	0,861730881	0,729135	0,789907	306	513
img_3.jpg	0,77216666	0,803211	0,787383	284	307
img_4.jpg	0,863978244	0,834893	0,849187	218	304
img_5.jpg	0,656706886	0,898209	0,758703	155	164
img_6.jpg	0,888742014	0,698524	0,782235	340	467

img_7.jpg	0,831042472	0,79121	0,810637	357	607
img_8.jpg	-	-	-	0	0
img_9.jpg	0,597747499	0,865116	0,706998	138	184
img_10.jpg	0,822313242	0,762061	0,791041	206	245
img_11.jpg	0	-	-	1	0
img_12.jpg	-	-	-	0	0
img_14.jpg	0,773858681	0,796083	0,784814	142	192
img_15.jpg	-	-	-	0	0
img_16.jpg	-	-	-	0	0
img_17.jpg	0,67757116	0,706612	0,691787	181	321
img_18.jpg	-	-	-	0	0
img_19.jpg	-	-	-	0	0
img_20.jpg	0,236359043	0,920129	0,376106	88	36
img_21.jpg	0	0	-	4	1
img_22.jpg	-	-	-	0	0
img_23.jpg	-	-	-	0	0
img_24.jpg	0,524193505	0,896001	0,661427	148	147
img_25.jpg	0	-	-	15	0
img_26.jpg	-	-	-	0	0
img_27.jpg	0,688489749	0,854132	0,762418	188	335
img_28.jpg	0	-	-	8	0
img_29.jpg	0,532751092	0,997664	0,694591	1	1
img_55.jpg	0,866523835	0,796063	0,8298	183	273

img_56.jpg	0,539145839	0,855829	0,661541	119	98

Figura 31 Tabella confronto risultati automatici e manuali segmentazione cellule marroni

La precedente tabella si riferisce alla segmentazione delle sole cellule marroni e presenta nella prima colonna il nome dell'immagine in esame, nella seconda colonna il valore di precisione, nella terza il valore di recall, nella quarta l'f1 score, nella quinta il numero di cellule positive calcolate dall'algoritmo automatico e nella sesta ed ultima colonna il numero di cellule positive calcolate attraverso la segmentazione manuale.

I valori di precisione, recall ed f1 score sono stati calcolati pixel per pixel con le seguenti formule:

$$Precisione = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad F1\ Score = 2 * \frac{P*R}{P+R}$$

Considerando:

- TP= veri positivi; pixel considerati marroni sia nella segmentazione manuale che in quella automatica.
- FP= falsi positivi; pixel considerati marroni nella segmentazione automatica ma non in quella manuale.
- FN= falsi negativi; pixel considerati marroni nella segmentazione manuale ma non in quella automatica.
- P=precisione
- R= recall

I valori di precisione spesso non molto alti anche quando il numero di cellule è molto simile possono essere spiegati facilmente andando a vedere quelli relativi all'immagine 29, qui troviamo un numero identico di cellule marroni ma la precisione è sotto il 60%, questo può trovare una spiegazione nel fatto che, per quanto si sia cercato di essere il più precisi possibile nel seguire il bordo della cellula nella segmentazione manuale, può essere capitato che un certo numero di pixel esterni non sia stato considerato e che quindi sia stata trovata correttamente in entrambi i casi la cellula positiva ma l'area di interesse sia maggiore nel caso della segmentazione automatica che sicuramente ha una sensibilità pixel per pixel molto maggiore di quanto non la possa avere l'occhio umano in particolar modo in un'immagine con una risoluzione tanto alta. I casi in cui invece il numero di cellule calcolato nei due modi differisce sensibilmente ma si hanno valori alti di precisione, come nell'immagine 6, descrive i casi in cui l'algoritmo ha identificato come un'unica cellula una grossa area

marrone che però nella segmentazione manuale, grazie all'esperienza dell'operatore, in realtà è stata riconosciuta come formata da più cellule molto vicine, magari poco distinguibili ma comunque riconosciute. Come in alcuni casi visti in precedenza l'operatore può aver segmentato il bordo di alcune cellule rimanendo più interno rispetto alla realtà, allo stesso modo possono essere stati tracciati dei contorni più ampi rispetto alla forma vera e propria della cellula il ché può giustificare alcuni valori bassi di recall che dipende proprio dai falsi negativi.

Hanno prodotto risultati eccellenti invece le immagini nelle quali non erano presenti cellule marroni riuscendo, nella maggior parte dei casi in esame, a riconoscerle e a restituire un risultato pari a 0 o comunque molto piccolo.

Nome	Precisione	Recall	F1 score	N° cell auto	N° cell manual
img_1.jpg	0,478707224	0,464004914	0,471241422	18	21
img_2.jpg	0,172406132	0,434490482	0,246858597	15	6
img_3.jpg	0,196122416	0,375114925	0,257575758	17	7
img_4.jpg	0,045485951	0,12425291	0,06659361	14	7
img_5.jpg	0,068775678	0,052621652	0,059623878	37	46
img_6.jpg	0,126021004	0,026922598	0,044366848	5	22
img_7.jpg	0,159901599	0,119120342	0,136530719	9	7
img_8.jpg	0,336312623	0,152608006	0,209948183	120	100
img_9.jpg	0,115775149	0,157437722	0,133429847	28	19
img_10.jpg	0,326994189	0,180214277	0,23236608	21	28
img_11.jpg	0,359910556	0,162309417	0,223725157	37	85
img_12.jpg	0,47236017	0,300927423	0,367641043	198	241
img_14.jpg	0,239587431	0,503351391	0,324647637	60	25
img_15.jpg	0,150697032	0,129188754	0,139116474	70	47

img_16.jpg	0,562092897	0,254673302	0,35052884	142	205
img_17.jpg	0,074415517	0,278575373	0,117455329	31	10
img_18.jpg	0,368524333	0,108714109	0,167898438	62	90
img_19.jpg	0,358593206	0,116091907	0,175399514	51	99
img_20.jpg	0,231690899	0,234841938	0,233255777	64	51
img_21.jpg	0,64432013	0,143482891	0,234700585	28	107
img_22.jpg	0,316796851	0,140324983	0,194497438	113	144
img_23.jpg	0,424260921	0,351868321	0,384688451	121	142
img_24.jpg	0,171959083	0,226384989	0,195453919	52	27
img_25.jpg	0,396433264	0,295930672	0,338887559	119	140
img_26.jpg	0,293979844	0,030814064	0,055781303	45	163
img_27.jpg	0,164095662	0,324329855	0,217929328	49	34
img_28.jpg	0,348499275	0,317837391	0,33246287	167	138
img_29.jpg	0,398709677	0,146268233	0,214021739	88	132
img_55.jpg	0,052916184	0,032184808	0,04002532	19	23
img_56.jpg	0,211619174	0,108807996	0,143719761	30	59

Figura 32: Tabella confronto risultati automatici e manuali segmentazione cellule blu

Da questa tabella si evince un trend comune a tutte le immagini, infatti anche quando il numero di cellule è molto simile precisione e recall sono sempre molto basse, questo dimostra la capacità dell'algoritmo di riconoscere zone negative a bassa densità cellulare ma alla stesso modo di non riuscire perfettamente a distinguere la zona cellulare da quella extracellulare. Anche le segmentazioni manuali in questo caso sono state molto più complesse e legate in maniera più significativa all'esperienza maturata. Il comportamento peggiore è stato osservato nelle immagini contenenti poche cellule marroni e tante blu, in questo caso infatti spesso l'algoritmo non è stato in grado di individuare la membrana

cellulare e, anche quando nei primi step di elaborazione, veniva individuato il nucleo non riusciva poi a ricostruire a dovere la morfologia della cellula. Un aspetto positivo, come già detto, è la capacità della segmentazione automatica di riconoscere zone a bassa densità cellulare e quindi di dare risultati in termini percentuali molto simili ad una segmentazione manuale.

Nome	Percentuale segmentazione automatica	Percentuale segmentazione manuale
img_1.jpg	0,85	0,810810811
img_2.jpg	0,953271028	0,988439306
img_3.jpg	0,943521595	0,977707006
img_4.jpg	0,939655172	0,977491961
img_5.jpg	0,807291667	0,780952381
img_6.jpg	0,985507246	0,955010225
img_7.jpg	0,975409836	0,988599349
img_8.jpg	0	0
img_9.jpg	0,831325301	0,906403941
img_10.jpg	0,907488987	0,897435897
img_11.jpg	0,026315789	0
img_12.jpg	0	0
img_14.jpg	0,702970297	0,884792627
img_15.jpg	0	0
img_16.jpg	0	0
img_17.jpg	0,853773585	0,96978852
img_18.jpg	0	0

img_19.jpg	0	0
img_20.jpg	0,578947368	0,413793103
img_21.jpg	0,125	0,009259259
img_22.jpg	0	0
img_23.jpg	0	0
img_24.jpg	0,74	0,844827586
img_25.jpg	0,111940299	0
img_26.jpg	0	0
img_27.jpg	0,793248945	0,907859079
img_28.jpg	0,045714286	0
img_29.jpg	0,011235955	0,007518797
img_55.jpg	0,905940594	0,922297297
img_56.jpg	0,798657718	0,624203822

Figura 33: Tabella valori percentuali delle cellule positive rispetto al totale calcolati in maniera automatica e manuale

Nella tabella precedente vengono mostrati i risultati calcolati come numero di cellule positive fratto il numero totale di cellule presenti nel campione sia attraverso l'algoritmo automatico, sia attraverso la segmentazione manuale. Si tratta del risultato finale dell'elaborazione, quello che può essere utilizzato come supporto decisionale da un anatomopatologo. I risultati dell'elaborazione automatica sono molto simili a quelli ottenuti manualmente e descrivono, sotto questo punto di vista, una performance ottima dell'algoritmo sul dataset elaborato.

3.2 Risultati statistici

Per poter analizzare a fondo i risultati ottenuti è fondamentale utilizzare anche degli indicatori statistici, per questo progetto è stato calcolato l'errore assoluto, come differenza in valore assoluto tra la percentuale di cellule positive sul totale calcolata manualmente meno quella calcolata in maniera automatica per ogni immagine; l'errore con segno, come nel caso precedente ma mantenendo il segno del risultato; media e deviazione standard degli errori assoluti e media e deviazione standard degli errori con segno.

Nome	Errore con segno	Errore assoluto
img_1.jpg	0,039189189	0,039189189
img_2.jpg	-0,035168278	0,035168278
img_3.jpg	-0,034185412	0,034185412
img_4.jpg	-0,037836789	0,037836789
img_5.jpg	0,026339286	0,026339286
img_6.jpg	0,030497021	0,030497021
img_7.jpg	-0,013189512	0,013189512
img_8.jpg	0	0
img_9.jpg	-0,07507864	0,07507864
img_10.jpg	0,010053089	0,010053089
img_11.jpg	0,026315789	0,026315789
img_12.jpg	0	0
img_14.jpg	-0,18182233	0,18182233
img_15.jpg	0	0
img_16.jpg	0	0

img_17.jpg	-0,116014935	0,116014935
img_18.jpg	0	0
img_19.jpg	0	0
img_20.jpg	0,165154265	0,165154265
img_21.jpg	0,115740741	0,115740741
img_22.jpg	0	0
img_23.jpg	0	0
img_24.jpg	-0,104827586	0,104827586
img_25.jpg	0,111940299	0,111940299
img_26.jpg	0	0
img_27.jpg	-0,114610133	0,114610133
img_28.jpg	0,045714286	0,045714286
img_29.jpg	0,003717158	0,003717158
img_55.jpg	-0,016356703	0,016356703
img_56.jpg	0,174453896	0,174453896

Figura 34: Tabella con errori assoluti e con segno

Media E con segno	Media E assoluto	Dev.st E con segno	Dev.st E assoluto
0,00066749	0,049274	0,076617861	0,057958

Figura 35: Tabella con errore medio assoluto e con segno e loro deviazioni standard

Come si evince in particolar modo dall'ultima tabella presentata la media dell'errore con segno è molto bassa mentre quella dell'errore assoluto è leggermente più alta, intorno a 5%; le deviazioni standard vanno dal 5.7% per l'errore assoluto al 7.6% per l'errore con segno.

Tali valori, maggiori in entrambi i casi rispetto alla loro media di riferimento, trovano una giustificazione nel fatto che sono pesantemente influenzati dagli errori calcolati su un numero ristretto di immagini rispetto alla totalità di campioni analizzati i quali hanno valori percentuali di cellule positive o molto alti o molto bassi, ovvero immagini dove comunque l'algoritmo ha dato risultati assolutamente paragonabili alla realtà.

4.CONCLUSIONI E SVILUPPI FUTURI

Grazie all'analisi dei risultati del capitolo precedente è possibile trarre delle conclusioni sull'effettivo raggiungimento o meno degli obiettivi prefissati all'inizio del progetto e soprattutto anticipare quelli che potrebbero essere gli sviluppi futuri dello stesso. Si è riusciti a realizzare un algoritmo completamente automatico in grado di iniziare l'analisi direttamente dall'immagine ottenuta attraverso un microscopio ottico; l'output è una misura quantitativa e oggettiva della percentuale di cellule risultate positive rispetto al totale; i tempi di elaborazione sono contenuti e ragionevoli rispetto a quelli di un'analisi qualitativa effettuata da personale qualificato; la precisione del risultato finale è buona.

Riguardo gli ultimi due punti appena citati si può fare un discorso più ampio. I tempi di elaborazione sono diversi per la parte di preprocessing e per la parte di segmentazione, la prima è quella più onerosa da questo punto di vista data la dimensione delle immagini e il fatto che per ogni immagine ne vengono create e soprattutto salvate una per ogni step di elaborazione, il salvataggio delle suddette immagini è stato fondamentale nella fase di progettazione e verifica dell'algoritmo ma può essere assolutamente omesso durante un suo utilizzo standard il che accorcerebbe notevolmente i tempi. La parte di segmentazione ha durata molto più breve, si riesce ad elaborare un'immagine ogni 30'', ovvero più di 2800 immagini ogni 24 ore. Anche qui però per ogni step viene creata un'immagine che viene successivamente salvata sempre per scopi di progettazione e verifica e anche in questo caso è un passaggio omettibile. Va anche sottolineato che le immagini utilizzate in questa parte sono state create selezionando una porzione delle immagini ottenute come output del preprocessing e hanno dimensioni ridotte rispetto a quest'ultime la cui analisi può richiedere tempi più lunghi, accorciabili utilizzando una macchina con potenza di calcolo e prestazioni migliori rispetto a quella a disposizione.

Per quanto riguarda la precisione del risultato finale è importante ricordare il contesto in cui si inserisce il progetto di tesi. Come detto in precedenza si tratta di immagini attualmente analizzate in modo qualitativo ed è tramite il confronto con questa metodica che vanno valutati i risultati ottenuti. Nonostante ciò, quindi confrontandoli con segmentazioni manuali che non vengono mai fatte nella quotidianità del lavoro di un anatomopatologo, i risultati sono soddisfacenti e si ritiene che possano essere utilizzati come supporto decisionale al personale medico senza ovviamente poterne sostituire l'esperienza.

Nel futuro, volendo usare come base questo progetto, sarà importante sviluppare alcuni punti ottenendo quindi un algoritmo con prestazioni migliorate. La segmentazione delle cellule blu è sicuramente uno dei più rilevanti soprattutto nelle immagini in cui non sono presenti cellule marroni o se ne individua un numero molto limitato, si potrebbe infatti utilizzare un approccio diverso per questo tipo di immagini andando a ricercare solamente i nuclei piuttosto che l'intera cellula valutando un miglioramento o meno del riconoscimento delle cellule negative. Sarà fondamentale anche ampliare il numero di segmentazioni manuali a disposizione e migliorarne l'accuratezza facendo crescere quindi anche il dataset ottenendo di conseguenza una validità statistica maggiore dei risultati ottenuti.

Si ritiene possa anche essere utile avere a disposizione delle analisi qualitative effettuate da personale medico esperto in modo da poter fare un doppio confronto, ovvero sia tra queste analisi e la segmentazione manuale, sia tra la segmentazione manuale e quella automatica calcolando l'errore nei due casi. In questo modo sarebbe possibile mettere in evidenza l'impatto che un progetto di questo tipo possa avere nella routine di diagnosi e prognosi di una patologia che come visto ha un'incidenza importante a livello nazionale.

5.BIBLIOGRAFIA

[1] "tumore-del-seno @ www.airc.it." .

[2] D. Basile, L. Gerratana, G. Pelizzari, and F. Puglisi, *Trattamento primario del carcinoma mammario operabile e del carcinoma mammario localmente avanzato non operabile*. 2018.

[3] "mammografia-ed-altri-esami @ www.andosonlusnazionale.it." .

[4] "agoaspirato-o-agobiopsia @ www.humanitas.it." .

[5] "c09d3cf2d826031d0c072d35f7ef531e982b7192 @ www.istologia.unige.it." .

[6] H. Liu, "Application of immunohistochemistry in breast pathology: A review and update," *Arch. Pathol. Lab. Med.*, vol. 138, no. 12, pp. 1629–1642, 2014, doi: 10.5858/arpa.2014-0094-RA.

[7] "index @ docs.openmicroscopy.org." .

[8] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," *Proc. - 2009 IEEE Int. Symp. Biomed. Imaging From Nano to Macro, ISBI 2009*, pp. 1107–1110, 2009, doi: 10.1109/ISBI.2009.5193250.

[9] J. A. W. M. Van Der Laak, M. M. M. Pahlplatz, A. G. J. M. Hanselaar, and P. C. M. De Wilde, "Hue-saturation-density (HSD) model for stain recognition in digital images from transmitted light microscopy," *Cytometry*, vol. 39, no. 4, pp. 275–284, 2000, doi: 10.1002/(SICI)1097-0320(20000401)39:4<275::AID-CYTO5>3.0.CO;2-8.