

POLITECNICO DI TORINO

Laurea magistrale in Ingegneria Biomedica

Valutazione Automatica e
Quantitativa
dei Linfociti Infiltranti il Tumore
nel Carcinoma Mammario

Relatore e corelatore:

Prof. Filippo Molinari

Ing. Massimo Salvi

Candidato:

Martina Iucci



2019-2020

Abstract

Il tumore della mammella racchiude un gruppo eterogeno di sottotipi che differiscono per una molteplice varietà di fattori e che presentano prognosi molto diverse. Il carcinoma della mammella è un tumore maligno caratterizzato da un'abnorme crescita cellulare, svincolata dai normali meccanismi di controllo dell'organismo che si origina a livello della ghiandola mammaria. Le sue caratteristiche istologiche consentono di individuare, in base allo stato dei recettori ormonali e HER2, almeno 5 sottotipi clinicamente rilevanti. Tra queste categorie ne esistono due, il tumore triplo negativo ed il tumore HER2 positivo, in genere associati a cattiva prognosi a causa dell'alta aggressività del tumore, in cui la componente immunogenica ed in particolare l'entità dell'infiltrazione di linfociti ed altre cellule immunitarie può rappresentare un'importante discriminante per la prognosi. La valutazione dei linfociti infiltranti il tumore (TIL) viene effettuata su campioni di tessuto prelevati tramite biopsia e colorati con Ematossilina ed Eosina (H&E) ed è unicamente di tipo manuale. Il metodo presenta un'alta variabilità intra- ed inter-operatore, per questo motivo sono state recentemente pubblicate delle linee guida per standardizzare il processo, che risulta però ancora poco riproducibile.

Scopo della presente tesi è lo sviluppo di un algoritmo automatico che esegua il conteggio ed il calcolo della percentuale vera dei TIL in modo da aumentare la precisione, accuratezza e ripetibilità di questo marker prognostico e predittivo. Le immagini istopatologiche vengono segmentate per mezzo di una *Convolutional Neural Network* (CNN) che produce delle mappe di probabilità dei pixel di appartenere alla regione interna, esterna o bordo delle cellule; queste mappe sono successivamente sottoposte ad una fase di post-processing per ottenere la segmentazione finale. Infine, i TILS vengono individuati e conteggiati e se ne calcola la percentuale.

L'algoritmo, confrontato con le percentuali calcolate dall'operatore manuale ha fornito risultati soddisfacenti, suggerendo una sua applicazione futura come tool di supporto alla diagnostica nel carcinoma della mammella.

Indice

Abstract	1
Indice	0
INTRODUZIONE	3
Capitolo 1	5
1.1 Il Tumore della mammella	5
1.4 Inquadramento diagnostico	7
1.4.1 Classificazione istologica	8
1.4.2 Classificazione molecolare	8
1.4.3 Stadiazione	10
1.4.4 Fattori prognostici e predittivi	12
1.5 I Linfociti infiltranti il tumore: valenza prognostica	13
1.5.1 Metodologia di valutazione dei TIL	15
Capitolo 2	19
Analisi istologica	19
2.1 Istopatologia	19
2.2 Ematossilina ed Eosina (H&E)	19
2.3 Istopatologia digitale	20
Capitolo 3	21
Stato dell'arte	21
3.1 Tecniche tradizionali per la segmentazione di nuclei	21
3.2 Metodi di segmentazione basati su Machine learning	23
3.3 Estrazione delle feature	25
Capitolo 4 Metodi	26

4.1	Descrizione del dataset	26
4.2	Descrizione dell'algoritmo	28
4.3	Rete Neurale Convolutionale (CNN)	29
4.4	Post processing	35
4.5	Pulizia della maschera binaria	44
4.6	Calcolo TIL%	51
Capitolo 5	54
Risultati	54
Capitolo 6	57
Discussione	57
Conclusioni	58

INTRODUZIONE

Il tumore della mammella è una complessa patologia che raggruppa un insieme vasto ed eterogeneo di sottotipi differenti dal punto di vista istologico, dell'origine cellulare, del potenziale metastatico, mutazioni e progressioni della malattia, nonché della risposta terapeutica e dell'andamento clinico. Esistono infatti una grande varietà di classificazioni a seconda del parametro preso in considerazione. Dal punto di vista molecolare ed in particolare eseguendo un'analisi dello stato dei recettori ormonali estrogenici e progestinici, si possono individuare 5 sottotipi clinicamente rilevanti, tra cui il tumore triplo negativo (TNBC) ed il tumore con iper-espressione della proteina Human Epidermal Growth Factor Receptor 2 (HER2+) costituiscono le tipologie più aggressive. Questi tipi di carcinoma si distinguono dai restanti per un'assenza di sensibilità agli ormoni femminili il che conduce appunto ad una maggior aggressività e soprattutto ad un minor numero di opzioni terapeutiche. Recentemente si stanno sviluppando delle nuove terapie che possano contrastare tali sottotipi; in particolare, trattamenti tramite anticorpi monoclonali ed inibitori tirosin-chinasi anti-HER2 risultano efficaci per il carcinoma HER2+; per il TNBC invece, non ormonosensibile e privo di amplificazione HER2, è più difficile individuare dei trattamenti bersaglio. Il tumore triplo negativo rappresenta circa il 15-20% di tutte le diagnosi di carcinoma mammario, i pazienti affetti da questo tipo di tumore, presentano prognosi peggiore e tasso di recidiva e mortalità più elevato rispetto agli altri tipi [1].

Considerata la carenza di trattamenti mirati per il TNBC, l'immunoterapia si è recentemente rivelata un'attraente opportunità per il trattamento di questa patologia ed ha dimostrato di portare a miglioramenti se utilizzata in combinazione con trattamenti chemioterapici tradizionali [2], [3]. In un processo definito *immunosorveglianza*, le cellule immunitarie possono riconoscere antigeni associati al tumore ed attaccare le cellule tumorali. In genere le cellule tumorali sono in grado di evadere questo sistema tramite l'attuazione di meccanismi adattivi di *immunoresistenza*. L'andamento della patologia tumorale è dunque influenzato da un complesso microambiente costituito da cellule tumorali, cellule stromali e cellule immunitarie. Recentemente è stata riconosciuta una correlazione tra una presenza massiccia di linfociti infiltranti il tumore (Tumor Infiltrating Lymphocytes, TIL) e una miglior prognosi per il carcinoma mammario, in particolare per il TNBC [4], [5]. Più nel dettaglio, la valutazione quantitativa dei TIL è stata validata come fattore prognostico e predittivo nella risposta alla chemioterapia in studi adiuvanti nei tumori TNBC, ma anche HER2+. In seguito alla crescente concordanza nel riconoscere la rilevanza di questo parametro, nel 2014 sono state pubblicate delle linee guida internazionali per la standardizzazione del metodo di valutazione, nel tentativo di aumentarne l'accuratezza e la

ripetibilità [6], [7]. Tali linee guida definiscono il tipo di infiltrato linfocitario da considerare nel conteggio e sono corredate da tabelle che forniscono un'indicazione visiva relativa ad ogni percentuale clinicamente rilevante. La valutazione finale viene eseguita confrontando il proprio vetrino con questi pattern visivi. Le percentuali indicate non sono rigorose e soprattutto ad elevate percentuali non corrispondono alla reale distribuzione di tali cellule sullo stroma, ne risulta che in ogni caso la valutazione sarà influenzata dall'interpretazione individuale del patologo e dalla sua esperienza, conducendo dunque ad un'elevata variabilità inter-operatore che potrebbe intaccare l'affidabilità di tale marker. L'applicazione di metodi di analisi automatica può aiutare a migliorare l'accuratezza della valutazione semplificando l'analisi di distribuzioni spaziali più complesse e provvedere una metrica standardizzata.

Lo scopo del presente lavoro di tesi è lo sviluppo di un algoritmo che possa individuare i linfociti infiltranti il tumore nelle regioni prossime al bordo tumorale per estrarre una percentuale %TIL vera che possa essere correlata all'aggressività del tumore ed utilizzata per estrarre marker prognostici e predittivi del tumore della mammella.

L'elaborato è così organizzato: nel capitolo 1 si introducono e affrontano in modo generale le nozioni base del tumore alla mammella, sua epidemiologia e inquadramento diagnostico; in particolare si discuteranno i vari tipi di classificazione del carcinoma, mettendo in luce l'estesa eterogeneità delle patologie neoplastiche raggruppate sotto il nome di tumore. Successivamente ci si concentra sulla caratterizzazione dei TIL e l'attuale metodologia di valutazione. Nel capitolo 2 si descrive la procedura con la quale vengono ottenute le immagini istopatologiche. Nel capitolo 3 si rassegnano le principali tecniche utilizzate nell'ambito dell'identificazione e segmentazione delle cellule nelle immagini istopatologiche, con riferimento alle cellule cancerose e ai linfociti infiltranti il tumore nel cancro al seno. Nel capitolo 4 si descrive l'approccio seguito per pervenire all'algoritmo di segmentazione e classificazione. Nel capitolo 5 sono esposti i risultati che verranno discussi nel capitolo 6. Infine nel capitolo 7 vengono riassunti i punti di forza e i limiti dell'algoritmo fornendo indicazioni per futuri miglioramenti.

Capitolo 1

1.1 Il Tumore della mammella

Il cancro è una patologia caratterizzata da un'abnorme crescita cellulare, svincolata dai normali meccanismi di controllo dell'organismo; nel cancro al seno, vengono interessate le cellule del tessuto mammario. In realtà con il termine tumore si intendono diverse condizioni patologiche della mammella che possono essere benigne o maligne. I tumori benigni sono detti fibroadenomi, sono frequenti in donne giovani e non rappresentano un pericolo, i tumori maligni sono invece in genere carcinomi e si differenziano a seconda della struttura che colpiscono.

La mammella è costituita da cute, tessuto connettivo, adiposo e ghiandolare. Quest'ultimo è suddiviso in lobi, specializzati nella produzione di latte, a loro volta suddivisi in lobuli, da cui si diparte una rete di dotti che servono invece a condurre il latte fino al capezzolo. Nel primo caso si avranno carcinomi *lobulari*, nel secondo caso invece sono detti *duttali*.

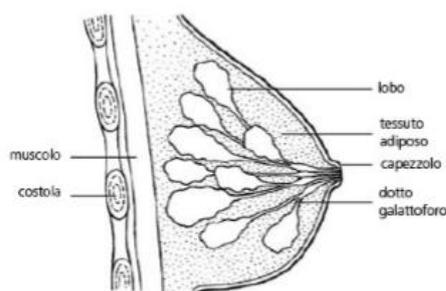


Figura 1.1: Anatomia della mammella [8]

Il processo di trasformazione in cellula neoplastica avviene a causa della mutazione di diverse classi di geni specializzati nel controllo di processi fisiologici quali la proliferazione cellulare. Tali geni sono detti oncogeni e oncosoppressori. I primi controllano la crescita cellulare e se danneggiati conducono ad una attività incontrollata, mentre i secondi regolano processi antagonisti quali la morte cellulare e la riparazione del DNA e se lesionati vengono inattivati. Una mutazione di tali geni, combinata all'assenza di una risposta immunitaria, causa una crescita cellulare abnorme e la progressione della neoplasia.

Il processo che porta alla crescita di un tumore è dunque un processo molto lungo, il danno genetico può verificarsi molto prima della manifestazione della malattia e conferisce alla cellula l'abilità di dividersi, ignorando i segnali di controllo esterni. Tale abilità viene poi trasmessa alle cellule figlie causando un accumulo di mutazioni che accresce l'autonomia proliferativa della

cellula. La formazione di un carcinoma segue quindi vari stadi in cui si passa da una iperplasia, caratterizzata dalla presenza di cellule con spiccata capacità proliferativa e qualche alterazione morfologia (displasia) per poi passare alla formazione del tumore benigno che può accrescersi rapidamente senza travalicare i confini del tessuto; nel momento in cui oltrepassa i confini del tessuto si ha la comparsa del carcinoma *in situ*, assumendo definitivamente le caratteristiche di tumore maligno; successivamente, penetrando nei vasi sanguiferi, svilupperà capacità metastatiche ed invasive, assumendo la forma di carcinoma *invasivo*. Le metastasi si diffondono per infiltrazione diretta della cute sovrastante, del parenchima mammario e della fascia pettorale profonda; per via linfatica possono attaccare i linfonodi ascellari omolaterali, sovraclaveari ed i linfonodi della catena mammaria interna; per via ematica infine possono raggiungere scheletro, polmone, fegato, cervello, rene, ovaio, surrene e tratto gastroenterico.

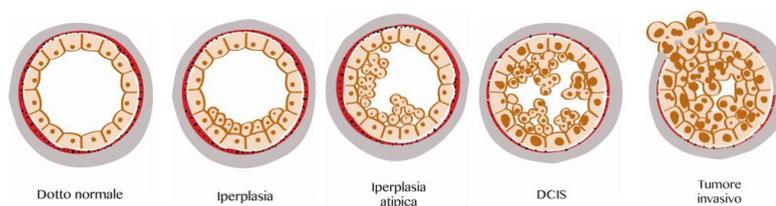


Figura 1.2: Esempio di progressione della malattia nel carcinoma duttale [9]

1.2 Epidemiologia

Il tumore della mammella rappresenta la seconda neoplasia più frequente tra le diagnosi di cancro a livello mondiale, come confermato da una stima del 2018 secondo cui è stato diagnosticato nell'11,6% dei casi di tumore. I numeri crescono se si considera la sola popolazione femminile, per la quale il carcinoma mammario detiene il primato sia per incidenza che per mortalità, pari rispettivamente al 24,2% e 15% [10]. Le stime di incidenza del 2019 su base nazionale mostrano un quadro molto simile, nel 2019 sarebbero stati diagnosticati circa 53 000 nuovi casi di carcinoma della mammella femminile. Escludendo quelli cutanei, anche a livello nazionale tale neoplasia è la più diagnosticata tra le donne (in cui un tumore maligno ogni 3 è un carcinoma mammario) e tra le prime per mortalità (basti pensare che nel 2016 tale patologia è stata la prima causa di morte con 12 000 decessi in Italia [11], [12]). L'introduzione dei programmi di screening e prevenzione ed i progressi terapeutici, hanno modificato il trend di incidenza e di mortalità degli ultimi anni, che hanno subito rispettivamente un aumento di +0.3%/anno e -0.8%/anno [11], [13]. Le stime possono cambiare a seconda della fascia d'età delle pazienti considerate, con un'incidenza più alta nella fascia 0-49 (40%), per poi decrescere leggermente nelle fasce 50-69 e

70+, che presentano rispettivamente un'incidenza del 35% e 22%. La sopravvivenza a 5 e 10 anni delle donne con tumore alla mammella in Italia è pari rispettivamente all'87% ed all'80% e non presenta eterogeneità elevata tra fasce di età. In Italia sono 800 000 le donne a cui è stato diagnosticato un carcinoma mammario, pari al 44% di tutte le donne che convivono con una pregressa diagnosi di tumore e al 24% di tutti i casi prevalenti (uomini e donne) [12].

Vari fattori contribuiscono all'aumentare del rischio di ammalarsi di carcinoma della mammella. Primo fra questi è l'*età*, in particolare al suo aumentare cresce il rischio. Come già detto infatti, il cancro è una malattia a lento sviluppo che si manifesta in seguito all'accumulo di alterazioni epigenetiche e danni al DNA che alterano l'equilibrio di espressione tra oncogeni ed oncosoppressori.

Altri fattori di rischio possono esser riscontrati in:

- Fattori riproduttivi: lunga durata del periodo fertile, nulliparità, prima gravidanza a termine dopo i 30 anni e mancato allattamento al seno;
- Fattori ormonali: terapia endocrina sostitutiva durante la menopausa, uso di contraccettivi orali;
- Fattori dietetici e metabolici: elevato consumo di alcool e di grassi animali e basso consumo di fibre vegetali, obesità, sindrome metabolica;
- Pregressa radioterapia e precedenti displasie o neoplasie mammarie;
- Familiarità ed ereditarietà.

1.4 Inquadramento diagnostico

La diagnosi di neoplasia mammaria viene effettuata in seguito a tre tipi di valutazioni: la valutazione obiettiva, la valutazione di immagine e la valutazione cito-istologica. La valutazione obiettiva è un esame fisico, in cui il medico ispeziona collo, regioni sovraclavari e infraclavari, parete toracica, cavi ascellari ed entrambe le ghiandole mammarie, alla ricerca di eventuali anomalie oppure tumefazioni dominanti e asimmetriche rispetto al seno controlaterale. La valutazione di immagine comprende la mammografia, l'ecografia mammaria ed eventualmente la risonanza magnetica della mammella. La diagnosi definitiva può esser eseguita solo a seguito della valutazione cito-istopatologica che viene ottenuta tramite il prelievo di campioni di tessuto per mezzo di biopsia [14].

1.4.1 Classificazione istologica

Il termine cancro è spesso usato nella patologia neoplastica maligna della mammella come sinonimo di carcinoma, che è il tipo di neoplasia in assoluto più frequente. Il carcinoma mammario origina dalle cellule epiteliali dell'albero ghiandolare e può dar luogo a diversi istotipi. Di questi i più frequenti sono il carcinoma duttale e quello lobulare. Il carcinoma può avere vari gradi di malignità, distinti principalmente attraverso la definizione in situ (o non invasivo) e infiltrante (o invasivo). I carcinomi in situ si hanno nel caso di tumore ancora contenuti entro gli spazi ghiandolari duttali e lobulari. Il carcinoma duttale in situ è una lesione neoplastica iniziale con eccellente prognosi, ma che deve essere asportato completamente. Il carcinoma lobulare in situ è invece una lesione pre-neoplastica microscopica che aumenta il rischio di sviluppare un carcinoma infiltrante. Nel momento in cui il carcinoma diventa invasivo, il tumore ha superato gli spazi ghiandolari e può essersi disseminato e aver prodotto metastasi. La classificazione istologica dei carcinomi infiltranti della mammella prevede almeno 20 istotipi diversi indicati come speciali perché riconoscibili per caratteristiche immunofenotipiche, morfologiche o genomiche. La gran parte (70-80%) dei carcinomi invasivi è però raggruppata sotto un'unica classe, definita come carcinoma invasivo o infiltrante non di tipo speciale (NST); il carcinoma NST, comunemente noto come carcinoma duttale infiltrante, rappresenta una definizione di default che identifica un gruppo eterogeneo di tumori che non presentano caratteristiche sufficienti per poterli classificare come tipi istologici speciali. La componente stromale di questo tipo di tumore può essere popolata da cellule infiammatorie quali linfociti e monociti (TIL) il cui valore prognostico è stato recentemente riconosciuto dall'Organizzazione Mondiale della Sanità nelle nuove linee guida per la classificazione anatomo-patologica del tumore della mammella [15].

1.4.2 Classificazione molecolare

L'eterogeneità della patologia neoplastica mammaria può esser ulteriormente apprezzata considerando l'elevato numero di geni coinvolti nei processi di controllo di crescita, morte e differenziazione cellulare, implicando una diversa progressione della malattia in tumori apparentemente simili per caratteristiche clinico-patologiche. Le indagini sull'analisi di espressione genica mediante "microarray" hanno permesso l'individuazione di diversi sottotipi di carcinoma mammario in base alla "firma intrinseca"; i diversi sottotipi, pur presentando aspetti simili dal punto di vista clinico-patologico, differiscono notevolmente nella prognosi e nei potenziali bersagli terapeutici [16] [17]. La firma intrinseca è definita in base ai gruppi di geni

coinvolti e sono principalmente correlati all'espressione del recettore HER2, all'espressione delle proteine della proliferazione (Luminali) e all'espressione di un fenotipo chiamato basal-like. I sottotipi Luminali sono quelli che presentano prognosi favorevole, gli HER2 ed i basal like sono invece associati a prognosi peggiore [12].

Nella pratica clinica si preferisce utilizzare un'alternativa alla metodica complessa e costosa del microarray, effettuando un'analisi immunohistochimica. Analizzando lo stato dei recettori ormonali (ER e Pgr), del Ki67 (utilizzato per identificare il grado di proliferazione) e di HER2, si possono individuare 4 sottogruppi immunofenotipici che presentano una buona corrispondenza con i 4 gruppi derivati dai profili di espressione genica; tale classificazione consente una differenziazione per quanto riguarda la prognosi, sopravvivenza libera da malattia, sedi di ripresa di malattia e sopravvivenza globale [12].

Tramite immunohistochimica vengono identificati i seguenti tipi di carcinoma:

- *Luminali A*: recettori ormonali positivi, HER2 negativo e bassa attività proliferativa
- *Luminali B/HER2 negativi*: recettori ormonali positivi, HER2 negativo ed alta attività proliferativa;
- *Luminali B/HER2 positivi*: recettori ormonali positivi, HER2 sovraespresso o amplificato, qualsiasi valore di attività proliferativa;
- *HER2 positivi (non luminali)*: HER2 sovraespresso o amplificato ed entrambi i recettori ormonali negativi;
- *Triplo negativi*: assenza di espressione dei recettori ormonali e negatività di HER2.

Va considerato che il gruppo intrinseco basal-like e l'immunofenotipo triplo negativo, presentano una corrispondenza dell'80%, poiché non tutte le forme basal-like sono di tipo triplo negativo.

Dal punto di vista prognostico i quattro sottotipi sono associati a differenze in sopravvivenza libera da malattia, sedi di ripresa di malattia e sopravvivenza globale. Il sottotipo triplo negativo come già detto è quello associato a prognosi peggiore a causa dell'assenza di espressione di tutti e tre i recettori che implicano una sua insensibilità alle terapie standard (ormonoterapia e anti-HER2).

Tabella 1.1: Sottotipi derivanti dai profili di espressione genica.

A Luminali	Marcata espressione di recettori ormonali.
B Luminali	Espressione di recettori ormonali ed elevato indice proliferativo.
HER2	Espressione HER2
Basal like	Assenza di recettori ormonali e di HER2 e aumentata espressione delle citocheratine basali (ad esempio CK5/6 e CK14).

1.4.3 Stadiazione

Il processo decisionale di cura dei pazienti affetti da carcinoma mammario segue un sistema di stadiazione che nasce dall'analisi retrospettiva della storia clinica di diversi casi di pazienti. Il sistema più utilizzato è il Tumor, Node, Metastasis (TNM), che caratterizza l'estensione della malattia sulla base di criteri classificativi anatomici. A partire dal gennaio 2018, l'American Joint Committee on Cancer (AJCC-ottava edizione, [18]) ha aggiornato il sistema di valutazione, affiancando al TNM un sistema di valutazione prognostico del Prognostic Stage Group (PSG) che aggiunge alla classificazione anatomica, una considerazione sui marcatori biologici.

Stadio del tumore su valutazione anatomica: Il sistema TNM si basa sull'estensione del tumore primitivo (T), presenza ed estensione delle metastasi a linfonodi regionali (N) e presenza di eventuali metastasi a distanza (M). A ciascun parametro verrà assegnata una delle seguenti etichette:

Parametro T:

- Tx: tumore non valutabile;
- Tis: tumore in situ;
- T0: nessuna evidenza di tumore;
- T1,2,3,4: in base alle dimensioni ed estensione del tumore.

Parametro N:

- N0: non evidenza metastasi;
- Nx: linfonodi non valutabili;
- N1,2,3: metastasi in numero ed estensione variabili.

Parametro M:

- M0: assenza di metastasi a distanza;
- M1: presenza di metastasi a distanza.

In base alla combinazione di questi parametri, si possono distinguere diversi stadi:

- 1) **Stadio 0:** si tratta dei carcinomi in situ, ovvero forme di cancro nello stadio precoce che non si è ancora diffuso al di fuori della struttura ghiandolare della mammella; costituiscono fattori di rischio per il manifestarsi di forme più aggressive invasive.
- 2) **Stadio I:** rappresenta la fase iniziale, con carcinomi di diametro inferiore ai 2 cm, che non abbiano ancora interessato i linfonodi;
- 3) **Stadio IIA:** carcinomi di dimensioni tra i 2 e i 5 centimetri che non abbiano ancora raggiunto i linfonodi ascellari. Si tratta ancora di una forma iniziale in cui sono inseriti anche tumori in cui le cellule maligne non si trovano nel seno, ma nei linfonodi;
- 4) **Stadio IIB:** carcinomi con dimensioni tra 2 e 5 centimetri che hanno raggiunto i linfonodi, oppure carcinomi di dimensioni superiori a 5 centimetri ma che non hanno ancora raggiunto i linfonodi ascellari.
- 5) **Stadio IIIA:** il tumore può avere dimensioni variabili e si è diffuso ai linfonodi ascellari o ai tessuti vicini al seno;
- 6) **Stadio IIIB:** il tumore ha raggiunto la parete toracica, le costole e la cute;
- 7) **Stadio IIIC:** il tumore ha raggiunto i linfonodi sotto la clavicola, vicini al collo, e potrebbe essersi diffuso anche a quelli nel seno e sotto il braccio e nei tessuti vicini al seno;
- 8) **Stadio IV:** il tumore ha metastatizzato in organi diversi dal seno.

Gli stadi 0-IIA rappresentano gli stadi precoci, con prognosi favorevole e sopravvivenza a 5 anni del 98%; tali tumori sono in genere rimossi chirurgicamente. Gli stadi IIB-III rappresentano la fase localmente avanzata della malattia; a seconda del grado di diffusione questi possono essere operabili o inoperabili e le pazienti ricevono terapia sistemica neoadiuvante preoperatoria. Lo stadio IV rappresenta invece lo stadio avanzato che non è curabile, ma può essere trattato.

Questa stadiazione non è ancora definitiva, poiché può esser modificata dallo stato dei recettori ormonali ed HER2 e dal grado tumorale. Tali modifiche si sono rese necessarie considerati i progressi nella biologia, diagnostica e terapia. Ad esempio, nell'ottava edizione si considera che pazienti con tumori TNBC, indipendentemente dal grado, hanno tassi di sopravvivenza simili a quelli relativi a pazienti di uno stadio superiore rispetto a quelli che esprimono sia HER2 che ER e PR; allo stesso modo si considera che pazienti con grado 3 e tumore di tipo HER2-negativo e ed ER- e PR- positivo hanno tassi di sopravvivenza paragonabili a quelli di pazienti con tumore di uno stadio superiore rispetto a quelli con tumore di grado più basso [19].

Grado tumorale: il grado indica quanto le cellule tumorali differiscano da quelle sane, sia per struttura che per comportamento. La determinazione del grado viene eseguita su sezioni istologiche prelevate tramite biopsia. Il patologo utilizza sistemi di grading standardizzati come quello proposto da Elston ed Ellis che si basa sulla valutazione di: percentuale di formazione di strutture tubulari, numero di mitosi, pleomorfismo nucleare. Sommando i punteggi attribuiti a ciascuno dei parametri (su una scala da 1 a 3) si estrae uno score globale che definisce il grado istologico:

- grado 1, carcinoma ben differenziato (score da 3 a 5);
- grado 2, carcinoma moderatamente differenziato (score 6 o 7);
- grado 3, carcinoma scarsamente differenziato (score 8 o 9).

La prognosi peggiora all'aumentare del grado.

Stato dei recettori ormonali e del recettore HER2: come già detto lo stato dei recettori ormonali estrogenici (ER) e progestinici (PgR) viene effettuato tramite valutazione immunoistochimica, mentre per quanto riguarda il recettore HER2 i metodi più utilizzati sono l'immunoistochimica, che valuta l'eventuale sovraespressione del recettore, e l'ibridazione in situ mediante fluorescenza (FISH) che misura l'amplificazione del gene.

4.4.4 Fattori prognostici e predittivi

Con il termine prognostico ci si riferisce a tutti quei fattori correlati alla prognosi e sopravvivenza del paziente, sono utili nella scelta di del trattamento e possono stimare ad esempio il rischio di recidive in pazienti libere da malattia. I fattori predittivi sono invece quelli utilizzati per predire la risposta alle terapie. Molto spesso alcuni fattori possono essere sia prognostici che predittivi come nel caso di parametri biologici (come recettori ormonali, HER2, firme genomiche), mentre i fattori esclusivamente prognostici sono parametri di tipo clinico -patologico (come dimensione del tumore o numero di linfonodi coinvolti) [14]. Tra i criteri prognostici di maggiore importanza [12] ritroviamo i fattori precedentemente discussi nonché quelli utilizzati per la classificazione, tra cui:

- dimensioni del tumore;
- stato dei linfonodi ascellari;
- il grado istologico;
- attività proliferativa;
- tipo istologico;

- invasione vascolare;
- stato di HER2;
- stato dei recettori ormonali;
- età (<35 anni prognosi peggiore);
- profili di espressione genica;
- multifocalità;
- linfociti intratumorali.

1.5 I Linfociti infiltranti il tumore: valenza prognostica

Il ruolo del sistema immunitario all'interno del processo di trasformazione neoplastica è definito dalla teoria dell'*immunoediting*, processo che descrive la progressione maligna della malattia tramite l'interazione delle cellule tumorali con il sistema immunitario. Il processo si articola in 3 fasi: eliminazione, equilibrio ed evasione [20]. Durante la fase di eliminazione il sistema immunitario avvia un processo in cui le cellule tumorali vengono riconosciute ed attaccate tramite il reclutamento di cellule immunitarie. Il sistema immunitario viene attivato dall'alterazione della struttura tissutale operata dalle cellule tumorali. Le cellule reclutate si infiltrano nel tumore ed eliminano gran parte delle cellule tumorali oltre ad inibire l'angiogenesi. Questa fase si attiva ancor prima che il tumore diventi clinicamente visibile. Durante la fase dell'equilibrio il sistema immunitario mantiene il tumore in uno stato dormiente; contemporaneamente, le cellule tumorali sopravvissute alla fase di eliminazione possono subire mutazioni genetiche ed epigenetiche acquisendo una maggior eterogeneità che le rende in grado di resistere agli attacchi da parte del sistema immunitario e congiuntamente di sopprimerlo. L'evasione rappresenta la fase finale in cui il sistema immunitario non riesce a contenere la crescita di cellule tumorali che diventano a questo punto clinicamente visibili. Inoltre, in questa fase le cellule tumorali stimolano il processo di angiogenesi.

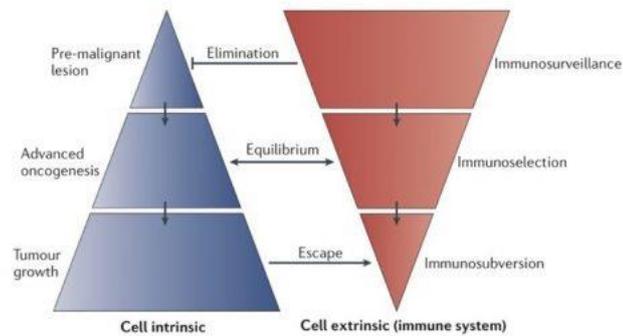


Figura 1.3: Processo di immunoediting. Relazioni tra aspetti intrinseci ed estrinseci delle cellule nella progressione del tumore [21]

L'interazione tra cellule immunitarie e cellule cancerose continua dunque ad avvenire anche a stadi più avanzati del tumore, in particolare la densità, localizzazione ed organizzazione delle cellule immunitarie, definita come “immune contexture” determinata al momento della diagnosi riflette la risposta immunitaria, con prognosi maggiormente favorevoli nel caso di elevato numero di cellule immunitarie [22]. La presenza di una risposta immunologica organizzata infatti potrebbe indicare la generazione di una memoria immunologica con la potenzialità di controllare la malattia minima residua [6].

La composizione delle cellule che costituisce l'infiltrato linfocitario è molto eterogenea e può presentare attività sia anti-tumorogena che pro-tumorogena, ciononostante il solo grado di infiltrazione linfocitaria senza considerare informazioni sulla sua sottopopolazione, valutato tramite analisi di sezioni istologiche colorate con Ematossilina ed Eosina (H&E), ha dimostrato di avere valore prognostico e predittivo nel carcinoma mammario triplo negativo e HER2-positivo [6]. Il primo studio dell'interazione tra l'infiltrato linfocitario e tumore mammario fu pubblicato nel 1990 da Aaltomaa et al. [23]. Alla luce delle loro considerazioni, sono state avviate numerose ricerche successive che investigassero il valore prognostico e predittivo dei TIL sia in terapie adiuvanti che neoadiuvanti. Dati istologici così come dati molecolari indicano che i parametri immunologici, tra cui i TIL stromali sono associati a tassi più elevati di remissione patologica completa (pCR), indipendente da altri fattori prognostici clinico-patologici o il regime chemioterapico [6].

Nella maggioranza degli studi adiuvanti e neoadiuvanti condotti per la valutazione dei linfociti infiltranti il tumore, sono stati considerati sia i TIL stromali che quelli intratumorali.

I TIL intratumorali sono caratterizzati da un contatto cellula-cellula senza stroma intermedio ed interagiscono direttamente con cellule del carcinoma, mentre i TIL stromali si trovano dispersi nello stroma tra le cellule di carcinoma senza entrare direttamente in contatto con esse. Le linee guida pubblicate nel 2014 per standardizzare il processo di valutazione dei TIL raccomandano di valutare solo i TIL stromali; questi vengono preferiti per una questione di maggior riproducibilità, essendo più numerosi, rilevati in un maggior numero di casi e più facili da individuare sui vetrini H&E [6]. Data la rilevanza di questo fattore, un recente studio ha evidenziato come un algoritmo automatico di valutazione possa migliorare la variabilità inter-operatore, in modo tale da fornire supporto per personalizzare la terapia nella gestione clinica del carcinoma mammario [24].

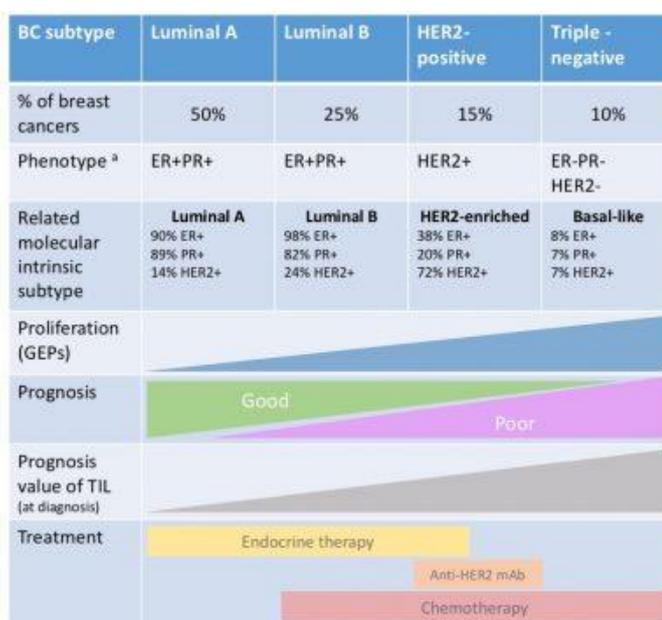


Figura 1.4. Classificazione del carcinoma mammario [25]

1.5.1 Metodologia di valutazione dei TIL

La valutazione dei TIL segue le raccomandazioni del TIL Working Group. Tali linee guida sono state redatte raccogliendo l'esperienza di tutti i patologi partecipanti al progetto. Ai partecipanti fu chiesto di rispondere ad una serie di domande, pubblicata su Annals of Oncology Online, in modo da riuscire a trarre una regola generale. Sulla base dei risultati ottenuti dal quiz si suggerisce di seguire i seguenti passaggi [6]:

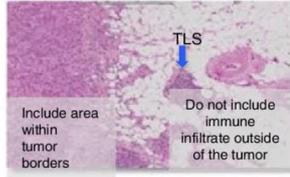
- Step 1: Definizione dell'area per la valutazione dei TIL. Si considerano solo i TIL all'interno del bordo del tumore invasivo il quale deve essere incluso nella valutazione.

Vanno dunque esclusi gli infiltrati immunitari al di fuori del bordo del tumore. Inoltre, vanno escluse aree centrali di necrosi o fibrosi.

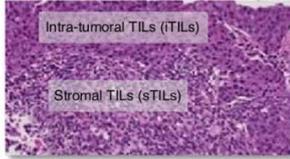
- Step 2: Concentrarsi sui TIL stromali. Lo stroma può presentare aspetto diverso a seconda che sia caratterizzato da predominanza di cellule di tessuto connettivo con poche cellule circolari o caratterizzato da cellule di infiltrato linfocitario tondeggianti con tessuto connettivo difficile da distinguere.
- Step 3: Determinare il tipo di infiltrato infiammatorio, includendo solo infiltrati mononucleari (linfociti e plasmacellule) ed escludendo granulociti nelle aree necrotiche.
- Step 4: Come primo approccio includere il tumore in uno di 3 gruppi considerando ingrandimenti maggiori e valutare la percentuale dei TIL stromali.
- Step 5: Riportare la percentuale di linfociti stromali assegnando la percentuale in base ad un confronto con tabelle riportanti pattern ricorrenti.

Standardized approach for TILs evaluation in solid tumors

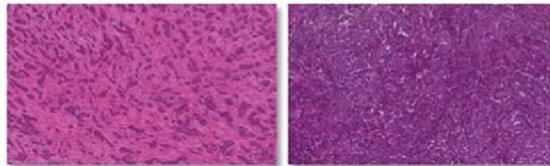
Step 1: Select tumor area



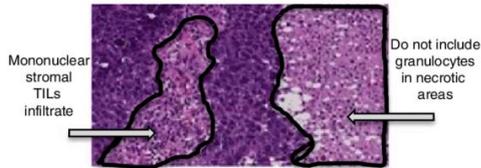
Step 2: Define stromal and intra-tumoral areas



Step 3: Scan at low magnification



Step 4: Determine type of inflammatory infiltrate



Step 5: Assess the percentage TILs

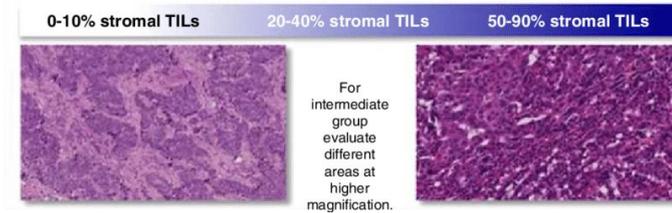


Figura 1.5. Step nella valutazione percentuale dei TIL [26]

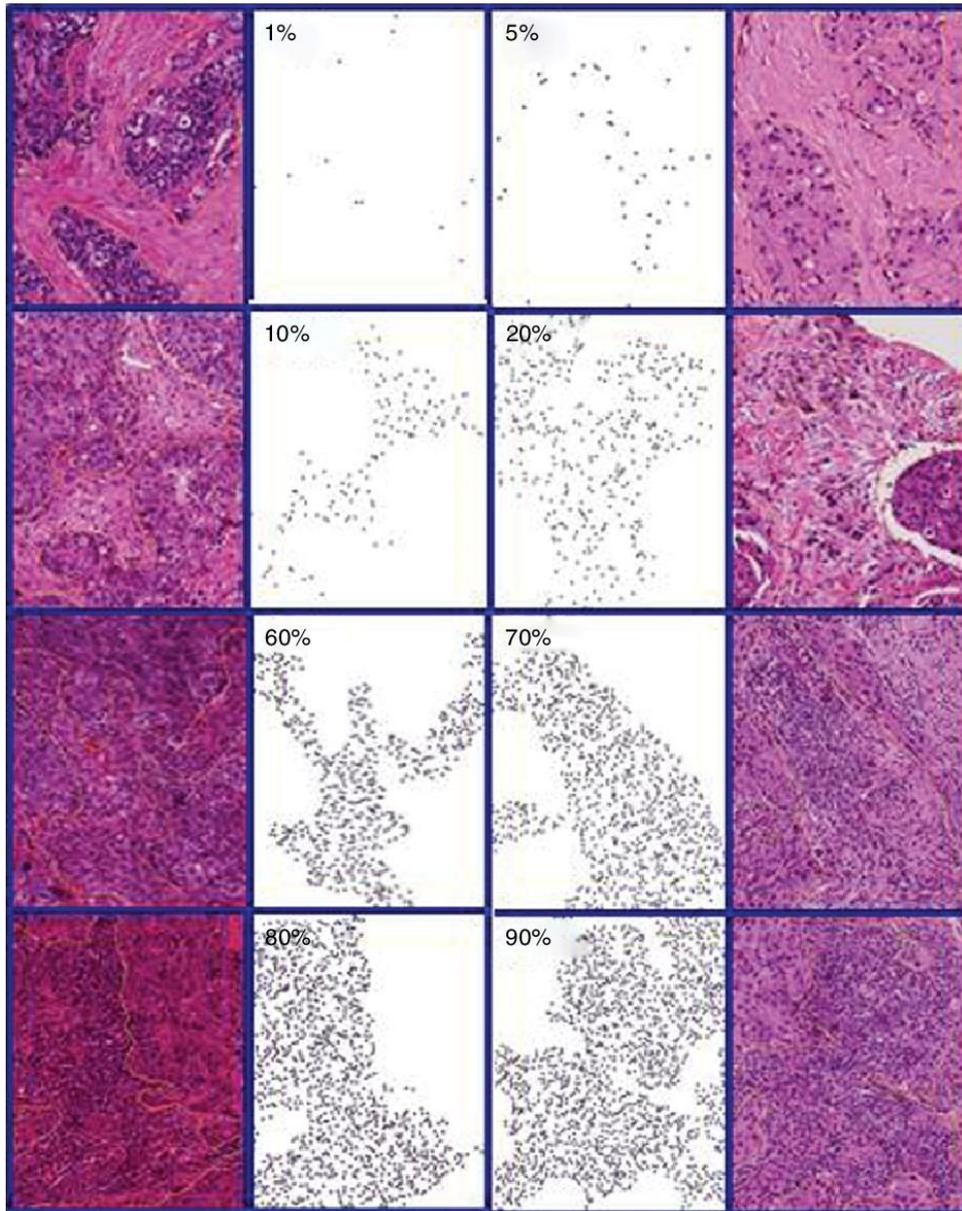


Figura 1.6. Tabella utilizzata nell'assegnazione delle %TIL [6]

Capitolo 2

Analisi istologica

2.1 Istopatologia

L'istologia è un ramo delle discipline biologiche che studia la struttura dei tessuti e delle loro cellule da un punto di vista morfologico, funzionale ed istochimico. L'istopatologia usa le tecniche di indagine proprie dell'istologia per studiare le alterazioni indotte nei tessuti dai diversi processi patologici. L'indagine viene effettuata su campioni di tessuto opportunamente trattati per ottenere il preparato istologico. La procedura di preparazione del campione si articola in più fasi: fissazione, disidratazione e diafanizzazione, inclusione, taglio, colorazione. I campioni di tessuto vengono prima di tutto fissati in modo tale da bloccare l'autolisi delle cellule ed evitare il deterioramento. I metodi di fissazione sono principalmente chimici e prevedono l'impiego di agenti stabilizzanti che instaurano dei legami trasversali tra le proteine. La disidratazione è un processo attraverso il quale l'acqua presente nelle cellule viene rimossa; questo passaggio è necessario per rendere il campione più facilmente visibile al microscopio ed essere successivamente infiltrato dalla paraffina, materiale idrofobo. Il processo viene effettuato tramite immersione del campione in soluzioni alcoliche a crescente concentrazione. Successivamente il campione è immerso in un solvente diafanizzante che lo renderà trasparente. L'inclusione in paraffina è necessaria per conferire al campione la giusta consistenza per il taglio che viene effettuata tramite uno strumento chiamato microtomo. Dal taglio si ottengono delle sezioni di tessuto di spessore compreso tra 5 e 10 μm . Infine, tramite la colorazione permette di evidenziare le strutture rilevanti del tessuto e renderle visibili per l'analisi. Terminata la colorazione, il campione viene sigillato tra 2 vetrini tramite resine sintetiche trasparenti ed è pronto per essere analizzato al microscopio.

2.2 Ematossilina ed Eosina (H&E)

Esistono vari tipi di colorazione, ma nell'istopatologia quella con Ematossilina ed Eosina (in inglese Hematoxylin and Eosin, H&E) è la più utilizzata, poiché rende visibili strutture rilevanti come citoplasma e nuclei, affini ad uno dei due coloranti. Questa colorazione si basa sul diverso valore di PH dei componenti della cellula. L'ematossilina è una sostanza basica affine con il nucleo ed i componenti acidi del citoplasma (quali ribosomi e secreti acidi) che risulteranno colorati in viola. I tessuti basici quali tessuto muscolare, connettivo ed osseo, vengono colorati in

rosa tramite una miscela acida di eosina-orange G. In generale quindi, i tessuti colorati con H&E sono caratterizzati da un buon contrasto con i nuclei delle cellule di colore blu ed il citoplasma di colore rosa o rosso.

2.3 Istopatologia digitale

I campioni istologici contengono un numero molto elevato di cellule ed altre strutture circondate da diversi tipi di tessuto; per questo la valutazione di immagini istologiche richiede molta esperienza nonché tempo. L'opportunità di digitalizzare le immagini tramite l'introduzione dei microscopi digitali ha apportato una serie di vantaggi, alleggerendo il lavoro del patologo. Avere a disposizione una versione digitale dei vetrini sul proprio monitor consente ad esempio di inserire note, o misurare le strutture del campione. Più immagini possono essere visualizzate sullo stesso schermo in modo tale da effettuare un confronto tra le diverse porzioni di tessuto o tra i risultati di diverse modalità di analisi del campione. Un altro importante vantaggio risiede nella possibilità di condividere le immagini per un consulto o a scopi didattici, nonché l'opportunità di poter conservare in maniera sicura i campioni. L'applicazione di algoritmi automatici di elaborazione delle immagini per ottenere dei supporti all'analisi offre una serie di vantaggi sia dal punto di vista del tempo di elaborazione, sia per la ripetibilità e accuratezza dei risultati. L'interpretazione e la valutazione di campioni al microscopio, infatti, è non solo molto complicata, ma è anche un processo soggettivo e altamente qualitativo. Nonostante gli sforzi di standardizzare il processo di valutazione, la riproducibilità inter- ed intra-operatore rimane un problema per l'estrazione di dati quantitativi e riproducibili per la predizione del trattamento e la prognosi. Questa necessità di standardizzazione, unita alle crescenti possibilità introdotte dallo sviluppo tecnologico (crescente potere computazionale, alta risoluzione delle fotocamere digitali, introduzione delle *whole-slide-scanner*) hanno costituito una spinta nello sviluppo di sistemi di elaborazione di immagini mediche digitali. Tramite la continua ottimizzazione degli algoritmi e l'introduzione del machine learning, sono stati effettuati notevoli passi avanti. Negli ultimi anni la patologia digitale in particolare ha attirato l'interesse dei ricercatori in seguito all'individuazione di biomarcatori prognostici e predittivi che necessitano di una accurata valutazione quantitativa delle proprietà molecolari ed istologiche del campione.

Capitolo 3

Stato dell'arte

Gli studi relativi alla valutazione dell'infiltrato linfocitario all'interno di immagini istopatologiche H&E sono ad oggi molto pochi nonostante siano stati sviluppati diversi metodi per la caratterizzazione del carcinoma mammario. Tipicamente questi studi si concentrano sulla classificazione di regioni di interesse in benigne o maligne per la caratterizzazione del tumore, oppure nella classificazione del grado istologico tramite metodi di classificazione diretta. Altri approcci prevedono la segmentazione delle cellule, successivamente assegnate alle varie classi in base a caratteristiche quali ad esempio forma, colore o organizzazione del tessuto circostante.

Una delle maggiori complicazioni nella segmentazione dei TIL risiede nell'elevata somiglianza tra i linfociti infiltranti il tumore e le cellule cancerose, nonché nella modificazione del colore del tessuto attorno alle aree tumorali che tende ad assumere un colore molto simile alle cellule stesse. La segmentazione di strutture in immagini istopatologiche di biopsie al seno è stata tentata tramite varie tecniche, le quali a loro volta possono essere divise in metodi non supervisionati o supervisionati, ma in genere seguono tre step: una prima segmentazione per individuare i contorni nell'immagine e separare gli oggetti dallo sfondo; in seguito si identificano tra gli oggetti estratti, quelli di interesse, in questo caso le cellule; infine si estrae un set di features che permettano di discriminare il tipo di cellule segmentate e classificarle considerando parametri riguardanti la forma o la dimensione delle cellule oppure imparando automaticamente le features nel caso di tecniche basate su machine learning.

Al di là della grande varietà di tecniche che possono essere utilizzate per la segmentazione di cellule, si possono individuare 2 principali tipi di approcci: *metodi di segmentazione tradizionali* o *metodi di deep learning*. Di seguito si analizzeranno le principali soluzioni evidenziando eventuali svantaggi.

3.1 Tecniche tradizionali per la segmentazione di nuclei

Tecniche tradizionali per la segmentazione dei nuclei includono vari metodi tra cui *thresholding*, *watershed*, *level set*, approcci basati su *clustering* dei pixel in base al colore e *modelli deformabili*. Queste tecniche in generale seguono tre step: dapprima si esegue una segmentazione grezza per identificare i bordi nell'immagine ed utilizzarli per separare gli oggetti; in seguito si identificano gli oggetti di interesse in base a parametri prestabiliti; infine si estraggono le feature per classificare gli oggetti individuati. Il vantaggio di utilizzare queste tecniche risiede nel fatto che

non c'è bisogno di acquisire un training set come nel caso dei metodi basati sul machine learning. Tuttavia, è richiesta una conoscenza specifica della materia in modo da individuare i parametri necessari per poter estrarre le cellule di interesse. Di seguito si analizzeranno le tecniche più utilizzate.

L'*adaptive thresholding* [27] è una tecnica molto popolare, in genere rifinita tramite operazioni morfologiche. Il *thresholding* è una tecnica con la quale un'immagine viene segmentata estraendo tutti i pixel la cui intensità supera una certa soglia. L'*adaptive thresholding* non usa una soglia unica come nel caso del *global thresholding*, ma la cambia dinamicamente sull'immagine selezionandola di volta in volta per ogni pixel, in base al range dei valori di intensità nel suo vicinato. Essa tuttavia, si basa su differenze di colore, tende a lavorare solo su immagini uniformi e non produce risultati consistenti se c'è variabilità considerevole all'interno dell'immagine.

Gli algoritmi *watershed* vengono largamente impiegati [28] [29] [30]; tuttavia questa tecnica è altamente dipendente dai marcatori iniziali e tende a sovrasegmentare l'immagine se la regione non è uniforme. Il funzionamento di questa tecnica verrà approfondito nelle sezioni successive.

Altre tecniche prevedono di partire da *seed points* trovati tramite *trasformata wavelet* [31]. Questa tecnica è utilizzata nell'analisi dei segnali la cui frequenza varia nel tempo. Si può adattare al caso delle immagini considerando che esse siano caratterizzate da andamenti di lungo periodo alle basse frequenze (come lo sfondo) e andamenti di breve periodo ad alte frequenze (bordi). Attraverso un banco di filtri l'immagine viene trasformata in una versione a bassa risoluzione (utilizzando dei filtri convoluzionali passa-basso) e in una versione più dettagliata (utilizzando dei filtri convoluzionali passa-alto).

Un'altra tecnica ampiamente sfruttata è il *region growing* di cui si discuterà meglio nelle sezioni successive. In [25], viene implementato per individuare cellule linfocitarie, sfruttando il fatto che i loro nuclei sono identificati come regioni circolari, continue, di bassa intensità, circoscritte da contorni ben definiti. Per tener conto della variabilità nelle dimensioni dei linfociti, l'immagine viene filtrata applicando dei kernel di diverse dimensioni, in modo da ottenere una rappresentazione multiscala. A questo punto i pixel più scuri vengono individuati in base a differenze locali di luminanza. Questi pixel definiscono un set di *seed points* che rappresentano i centri dei candidati linfociti. Ognuno di essi viene fatto crescere usando una procedura a 4 step. Dato che il *region growing* viene ripetuto con *seed points* provenienti da una varietà di scale, il problema delle regioni sovrapposte è risolto rigettando le regioni che non soddisfano determinati criteri. L'individuazione iniziale dei linfociti è poi rifinita incorporando informazioni riguardo la dimensione dei linfociti, luminanza e prossimità spaziale.

I *contorni attivi* usano informazioni di gradiente per isolare automaticamente i nuclei. Queste tecniche potrebbero presentare dei limiti a causa della variabilità visiva delle cellule tumorali, inoltre spesso non sono in grado di distinguere strutture sovrapposte e possono essere sensibili all'inizializzazione, per cui bisogna fornire una specifica congiuntamente ad algoritmi per individuare e separare le strutture sovrapposte [32].

Un problema spesso ricorrente in immagini di questo tipo è dovuto alla presenza di cellule sovrapposte. Per risolvere il problema della sovrapposizione delle cellule, sono stati esposti diverse soluzioni [30], [31], [28], molto spesso si utilizzano algoritmi watershed basati su marcatori derivanti da mappe di distanza euclidea o dall'individuazione di oggetti a simmetria radiale. Altri studi hanno proposto dei criteri morfologici per individuare e separare le cellule, partendo dal presupposto che le cellule sono oggetti convessi e quindi oggetti sovrapposti possono essere individuati attraverso un controllo sulla loro concavità. Ad esempio, in [26] si individuano i punti del contorno ad alta concavità che verranno utilizzati come punto di partenza e punto finale di una linea tracciata per separare le cellule. Queste tecniche falliscono nel caso in cui ci siano agglomerati di cellule sovrapposte costituiti da più di due elementi, cosa molto frequente all'interno di queste immagini.

3.2 Metodi di segmentazione basati su Machine learning

Sotto il termine di machine learning si raggruppano una varietà di algoritmi, che vengono addestrati utilizzando degli ampi dataset di annotazioni manuali. Queste tecniche possono essere divise in due classi, in un caso si usano delle feature specifiche, definite dall'utente in base alla propria conoscenza sulle caratteristiche discriminanti dei nuclei; nel secondo caso, invece si utilizzano modelli di deep learning, quali le *Convolutional Neural Networks* (CNN) che imparano a riconoscere automaticamente gli oggetti in base a feature imparate durante la fase di addestramento.

Uno dei problemi che si affrontano con le tecniche tradizionali di segmentazione infatti risiede nella loro sensibilità a variazioni biologiche o introdotte da rumore all'interno delle immagini, rendendo difficile la generalizzazione dei risultati in tessuti con alta densità di nuclei e cromatina sparsa. Per questo motivo recentemente l'attenzione si è spostata su questi approcci basati su machine learning o su tecniche ibride. Queste ultime incorporano modelli probabilistici in cui si modella l'intensità dell'immagine, considerata come variabile random, campionata da una o più distribuzioni che si adattino bene ai dati che modellano. Ad esempio, in [33] si effettua un clustering dei pixel tramite una *Support Vector Machine (SVM)*, per effettuare una prima segmentazione delle regioni contenenti nuclei e utilizzarle come inizializzazione per contorni

attivi. In [34] sono stati realizzati dei modelli dei fenotipi nucleari utilizzando delle “*Gaussian Mixture Model*” (GMM) costruiti sulla base di set di regioni annotate. Dall’immagine viene estratta una maschera binaria ed eventuali nuclei sovrapposti vengono divisi sulla base di considerazioni geometriche. Essendo questa una tecnica basata su un modello, riesce a catturare le variazioni intrinseche nella densità e intensità dei nuclei. Un altro metodo di segmentazione viene proposto in [35], in cui è stato usato un *region growing* combinato con metodi probabilistici a partire da segmentazioni manuali nel training set.

Recentemente si sta sviluppando un interesse verso i metodi di *deep learning*. In particolare, c’è una vasta diffusione nell’uso delle CNN applicate all’individuazione e segmentazione di nuclei o ghiandole. Le CNN sono state largamente utilizzate nel campo della classificazione delle immagini e se ne parlerà più approfonditamente nella sezione successiva. Considerando l’individuazione dei nuclei la CNN può essere utilizzata per assegnare una probabilità ai pixel dell’immagine che sarà tanto più alta quanto più il pixel è vicino al centroide di un nucleo [36]. Un approccio simile viene utilizzato in [37], in cui la CNN viene addestrata per generare le posizioni dei nuclei e il loro corrispondente livello di confidenza in una data patch.

Per quanto riguarda invece la segmentazione nucleare, i modelli delle CNN possono essere addestrati per segmentazione di regioni o di nuclei. In [38] la CNN viene utilizzata per assegnare una prima classificazione del grado tumore a partire da una iniziale segmentazione effettuata tramite *level set*. La classificazione finale viene eseguita da una SVM sulla base dell’integrazione delle feature ottenute dalla CNN con altre che descrivano la texture ed il livello di organizzazione delle cellule (ad esempio la relazione dei nuclei all’interno e tra le diverse colonie di cellule). Tuttavia, il modello utilizzato per il level set assume che i nuclei siano ben isolati ed abbiano un’alta quantità di cromatina, cosa che non si verifica all’interno di immagini relative a carcinoma mammario. In [39] la segmentazione dei nuclei è stata eseguita tramite conversione dell’immagine in toni di grigio, filtraggio e applicazione della CNN per dividere i pixel in oggetto e sfondo per poi rifinire la segmentazione tramite operazioni morfologiche. Un simile approccio è stato eseguito in [40], in cui la CNN genera mappe di probabilità iniziali su cui viene eseguito un modello deformabile che separi nuclei sovrapposti. Come si può notare, in genere le CNN vengono utilizzate per segmentare regioni nucleari e regioni non nucleari, recentemente però si è dimostrato che inserire una terza classe può aumentare le performance per quanto riguarda la separazione di cellule sovrapposte. Ad esempio, in [41], la CNN fornisce in output anche una mappa di probabilità dei pixel di appartenere al bordo dei nuclei. Altri approcci recentemente seguiti prevedono l’uso di tre diverse reti specializzate sul riconoscimento di regione

intranucleare, regione extra nucleare e bordi i cui output vengono successivamente fusi per fornire la segmentazione finale [42].

3.3 Estrazione delle feature

L'estrazione delle feature relative alla morfologia ed all'arrangiamento delle strutture è un passo essenziale per la corretta classificazione. Le caratteristiche maggiormente utilizzate per discriminare i vari tipi all'interno della popolazione cellulare si basano per lo più sulla caratterizzazione del colore e della *texture*. Con il termine texture si intende una disposizione geometrica e/o ripetitiva dei livelli di grigio di una immagine, descritta attraverso un insieme di statistiche locali o altre proprietà definite in base alla distribuzione spaziale dei livelli di intensità nel vicinato di ogni pixel. Tramite questo set di parametri si possono descrivere caratteristiche quali ad esempio la levigatezza, la grossolanità o l'omogeneità della trama. Nell'ottica dell'individuazione dei linfociti, molto spesso si preferisce utilizzare un set più piccolo di feature introducendo informazioni anche sulla forma. Gli articoli in letteratura in realtà, si concentrano sull'individuazione dei piccoli linfociti piuttosto che su tutta la popolazione dell'infiltrato linfocitario. I piccoli linfociti presentano caratteristiche in termini di omogeneità del colore e di grandezza molto diverse da tutte le altre cellule presenti. Sono facilmente individuabili dunque, tramite un controllo sulla deviazione standard e sulla loro area nucleare. È interessante considerare anche l'arrangiamento spaziale dei linfociti all'interno del tessuto; in questo caso si potrà definire, più che una valutazione quantitativa, il grado di infiltrazione in maniera qualitativa. Per farlo si ricorre alla descrizione di caratteristiche architettoniche che possono essere descritte tramite l'uso di grafi (*voronoi*, *delaunay triangulation*, *minimum spanning tree*) [43]. Lo stesso metodo è stato applicato in [29] per caratterizzare l'arrangiamento strutturale e l'apparenza morfologica dei linfociti. In questo articolo è stato dimostrato che semplici caratteristiche quali la deviazione standard e la radice dell'area nucleare erano sufficienti per descrivere la morfologia delle cellule. Le altre feature basate sulla descrizione della trama degli oggetti sono state scartate poiché si sono dimostrate incompatibili per applicazioni CADx nel cancro al seno che si basano su informazioni spaziali [43]. I nuclei dei linfociti così trovati sono a questo punto vengono analizzati considerandone l'arrangiamento spaziale oltre che tramite statistiche che ne descrivano il numero e la densità.

Capitolo 4

Metodi

4.1 Descrizione del dataset

Dataset per la valutazione della percentuale dei TIL

La valutazione della percentuale di Linfociti infiltranti il tumore è stata eseguita su 100 immagini ottenute a partire da vetrini di campioni chirurgici di carcinoma mammario HER2+ e TNBC, colorati con ematossilina ed eosina e digitalizzati tramite Aperio. Le immagini, acquisite con un ingrandimento 20x, sono state selezionate in modo da garantire eterogeneità al dataset, costituito da immagini rappresentative di diverse porzioni del bordo del tumore nei vari campioni provenienti da diversi pazienti. L'apparenza delle sezioni di tessuto infatti differisce molto in base al tipo e alla densità cellulare, o alla presenza di accumuli adiposi. Una bassa densità cellulare fa sì che le cellule siano più isolate l'una dall'altra, disperse all'interno di tessuto connettivo. Nel caso di alta densità cellulare, invece esse appariranno ammassate fra loro mentre il tessuto connettivo sarà più difficile da riconoscere. Per questo motivo si è cercato di garantire una buona proporzione fra le diverse caratteristiche delle immagini inserendone di molto diverse fra loro.

La percentuale dei Linfociti infiltranti il tumore è stata valutata da un patologo esperto seguendo le indicazioni del TILs Working Group. Oltre alle valutazioni manuali, sono state fornite anche le segmentazioni manuali delle zone tumorali utili per definire lo stroma sul quale eseguire il calcolo.

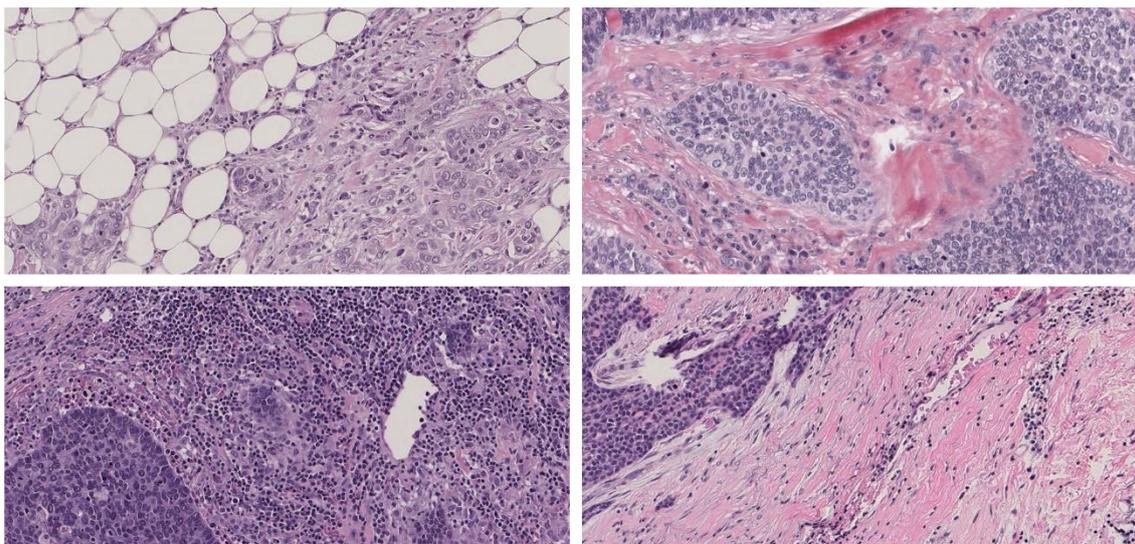


Figura 4.1: Esempio di immagini presenti nel dataset

Training e Testing Dataset

Un secondo dataset, utilizzato per l'addestramento della rete neurale, comprende un totale di 92 immagini H&E di carcinoma mammario, provenienti da "The Cancer Genome Atlas" (TCGA), da uno studio non pubblicato sul carcinoma mammario triplo negativo e da una coorte Scandinava di campioni di carcinoma mammario, acquisite con ingrandimento 40x per i primi due dataset e 20x nell'ultimo. Le annotazioni, di cui si riporta un esempio in figura 1, sono state eseguite da patologi esperti e rese pubblicamente accessibili da studi precedenti [44], [45], [41], [46]. Le segmentazioni manuali evidenziano un totale di 8096 cellule tra piccoli linfociti, plasmaciti, fibroblasti, cellule epiteliali, endoteliali, mitotiche, adipose e cancerose. Le annotazioni sono state effettuate in modo tale che cellule adiacenti non si sovrapponevano.

Le immagini utilizzate provengono da diversi pazienti e da diverse aree del tessuto e rappresentano anche in questo caso l'eterogeneità dei campioni in termini di densità cellulare e prevalenza di tessuto adiposo e stroma. Rappresentano inoltre una certa variabilità in termini di colorazione, essendo stato il tessuto trattato in laboratori diversi. Le immagini originali presentano dimensioni diverse, la maggior parte di esse è 512x512. Le annotazioni necessarie per l'addestramento della rete prevedevano sia la definizione dei nuclei, sia i loro contorni; per ottenere questi ultimi dagli oggetti sulla maschera binaria fornita dalle annotazioni presenti nei dataset sono stati estratti i contorni, i quali sono stati successivamente dilatati.

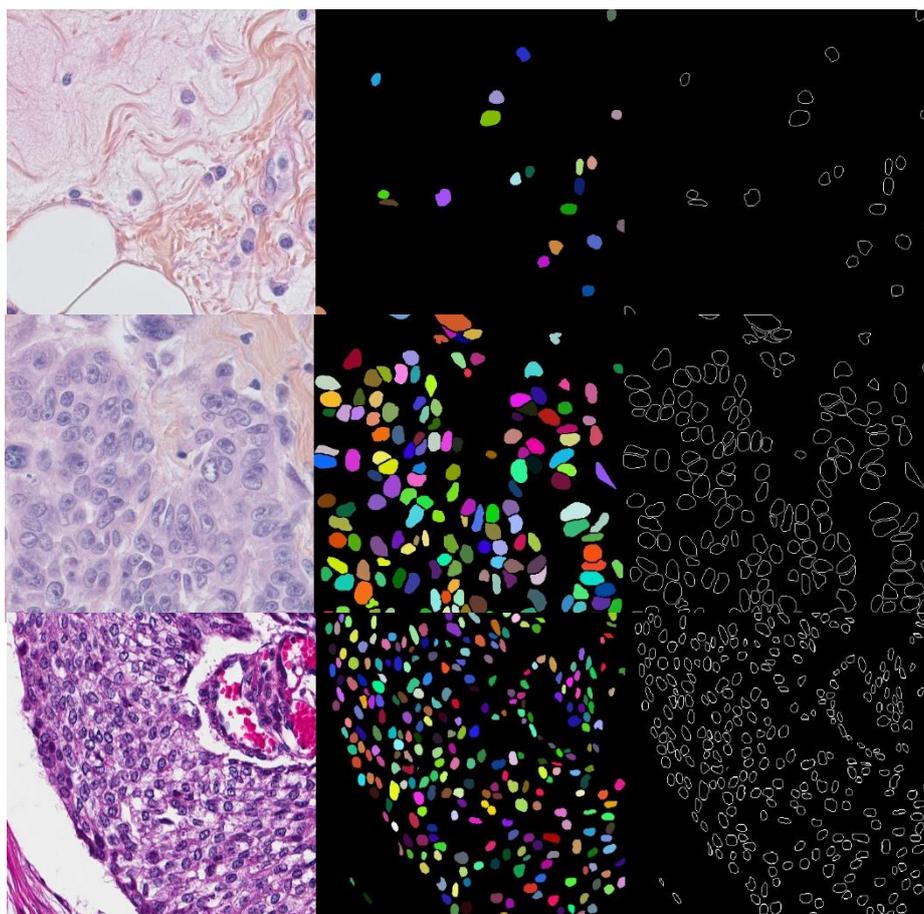


Figura 4.2: Esempio di immagini utilizzate nel training set

4.2 Descrizione dell'algoritmo

L'algoritmo si articola in 4 step principali: estrazione delle regioni e dei contorni nucleari, segmentazione grezza delle cellule, rifinitura della segmentazione ed individuazione e calcolo della percentuale di TIL. L'immagine ricevuta in input viene processata da una CNN che genera 3 mappe di probabilità dei pixel di appartenere alla regione cellulare, extracellulare oppure contorno delle cellule. Le *softmax* in uscita dalla rete vengono a questo punto sottoposte ad un algoritmo di *region growing* i cui *seed* iniziali vengono individuati imponendo una serie decrescente di soglie. La segmentazione grezza ottenuta, che definisce i bordi interni e bordi esterni della cellula, viene ulteriormente rifinita per separare eventuali oggetti ancora sovrapposti e per adattarne meglio i contorni tramite una *watershed*, eseguita utilizzando come *marker* le regioni racchiuse dal bordo interno della cellula ed utilizzando come gradiente ancora la softmax corrispondente alle regioni di bordo delle cellule. Eventuali falsi positivi e falsi negativi vengono individuati tramite confronto con la *color deconvolution* ed infine i TIL vengono discriminati

sulla base della loro intensità media e deviazione standard. La percentuale viene infine calcolata rapportando l'area dei TIL nello stroma peritumorale all'area di quest'ultimo, definito come le cellule ed il citoplasma in una zona di circa 200 μm dal bordo del tumore. I risultati quantitativi sono stati confrontati con le valutazioni fornite dal patologo.

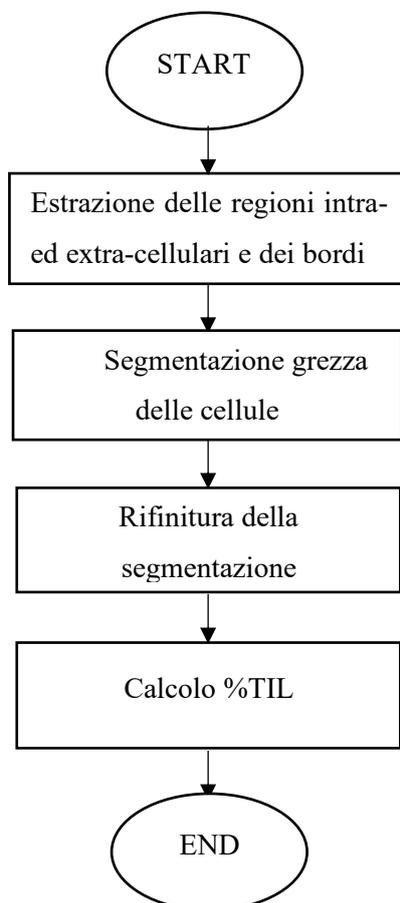


Figura 4.3: Flusso logico dell' algoritmo

4.3 Rete Neurale Convolutionale (CNN)

Le CNN rappresentano uno degli algoritmi più comuni di deep learning, un tipo di machine learning che trova campo di applicazione principalmente nella Computer Vision, riconoscimento vocale ed elaborazione di immagini. Questo tipo di algoritmo è particolarmente utile nel campo dell'elaborazione delle immagini. Le CNN sono in grado di riconoscere autonomamente dei pattern ricorrenti all'interno di un dataset ed apprendere automaticamente le feature necessarie per classificare, eliminando così lo step di estrazione manuale. L'utilizzo delle CNN per il deep learning sta guadagnando sempre più interesse grazie anche ai vantaggi in termini di costo computazionale. Con l'introduzione delle GPU, infatti, si è stati in grado di velocizzare

enormemente il tempo di elaborazione per l'addestramento di un modello che elabori centinaia di immagini. Le CNN sono un tipo di rete neurale artificiale profonda di tipo *feed forward*. Il funzionamento di base di queste reti non differisce da quello di una normale rete neurale, essendo costituita da un layer di input, una serie di layer nascosti ed infine un layer di output. Ogni layer è costituito da neuroni composti da *pesi* e *bias*, parametri che vengono aggiornati durante l'addestramento attraverso la retropropagazione dell'errore (*backpropagation*) e discesa stocastica del gradiente (*gradient descent*). Ad ogni iterazione si computa la *funzione di perdita*, che indica la distanza tra il valore della variabile predetto ed il valore attuale, l'obiettivo è la minimizzazione di tale funzione. La discesa stocastica del gradiente permette alla rete di individuare la direzione che il modello dovrebbe seguire per ridurre l'errore. Tramite la retropropagazione dell'errore, i parametri vengono aggiornati ad ogni iterazione in maniera proporzionale alla derivata parziale della funzione di perdita rispetto al parametro stesso, fino ad arrivare ad un punto di convergenza in cui il modello ha ottimizzato i parametri per minimizzare la funzione di perdita. La proporzionalità tra l'aggiornamento dei pesi ed il gradiente è definita dal *learning rate* che controlla quanto velocemente il modello viene adattato al problema.

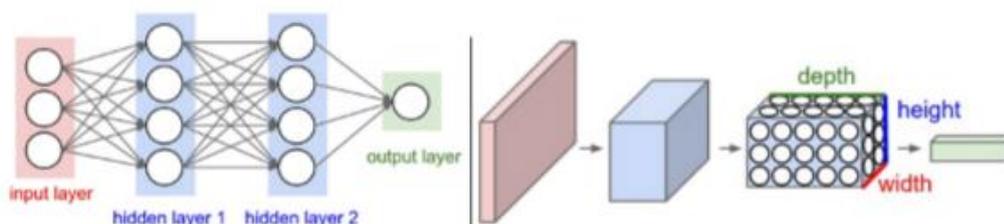


Figura 4.4 Differenze tra i layer di una rete neurale classica (a sinistra) ed una rete convoluzionale (a destra) [47].

Come indicato dal nome, queste reti si basano sull'operazione matematica della *convoluzione* tra l'immagine ed un filtro di dimensioni ridotte ma stesso numero di canali, indicato come kernel, che scorre su di essa. Il risultato finale sarà un'immagine costituita dalla somma dei prodotti degli elementi dell'immagine selezionati dal kernel che scorre su di essa. Ogni filtro rappresenta una feature, lasciando che tutti i valori del kernel siano parametri e lasciando che la rete li impari automaticamente dai dati, essa può individuare automaticamente le feature. La convoluzione rende più agevole l'applicazione delle reti neurali all'elaborazione delle immagini, le quali essendo costituite da migliaia di pixel, richiederebbero una rete con tante unità quanti sono i pixel dell'immagine e tante connessioni quanti sono i pixel del kernel; si otterrebbe in tal caso una matrice di pesi molto grande con un elevato numero di parametri. Le CNN risolvono questo problema perché i neuroni in ogni layer sono organizzati in tre dimensioni e connessi ad un

numero ristretto di neuroni del layer precedente. La regione di input considerata di volta in volta dal kernel è detta “*receptive field*”.

Ogni layer applica i filtri in diverse risoluzioni e l’output di ogni immagine convoluta, chiamato “*feature map*”, viene utilizzato come input del layer successivo. I filtri possono essere inizialmente molto semplici per poi diventare via via più specifici generando così una serie di feature di livello via via più alto, a partire da feature di basso livello come linee fino ad arrivare a comprendere interi oggetti nell’ultimo strato. Nella definizione di un layer convoluzionale, altri *iperparametri* da tenere in considerazione sono:

- 1) *Profondità*: corrisponde al numero di filtri utilizzati per ogni layer convoluzionale e definisce il numero di neuroni collegati alla stessa regione di input;
- 2) *Stride*: indica il passo tra i pixel sui quali si applicano convoluzioni successive, determina dunque la dimensione dell’output;
- 3) *Padding*: aggiunge una cornice di pixel, tipicamente di 0, all’immagine in modo tale da riuscire a regolare la dimensione dell’output ed evitare che i pixel negli angoli o ai bordi siano selezionati dal kernel un numero di volte inferiore rispetto ai pixel centrali, rischiando di perdere informazioni ai bordi.

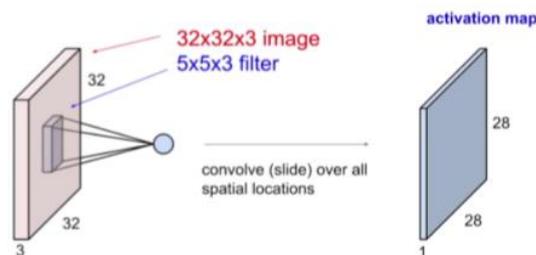


Figura 4.5. Esempio di convoluzione di un filtro su un volume di input. Ogni neurone è collegato ad un campo recettivo del volume di input [48].

Altri layer comuni sono:

- *layer di attivazione*: applica all’input una funzione di attivazione, tipicamente l’*unità lineare rettificata* (ReLU), che permette un addestramento più rapido ed efficace, mappando tutti i valori negativi a 0 e mantenendo valori positivi. Questo layer è necessario per introdurre un fattore di non linearità.
- *layer di pooling*: serve a semplificare l’output, riducendo il numero di parametri ed il costo computazionale della rete. In genere l’operazione più utilizzata è il *max-pooling* che seleziona il massimo tra i valori assunti dalla feature map all’interno di una regione selezionata da un filtro. Il risultato è una feature map sottocampionata che evidenzia la feature maggiormente attivata in ogni regione di interesse selezionata dal filtro.

- *fully connected layer*: rappresenta l'ultimo layer, in cui ogni neurone è connesso a tutti i neuroni attivati del layer. La dimensione dell'output corrisponde al numero di classi che si vuole riconoscere. Le connessioni sono pesate tramite parametri che vengono aggiornati nella fase di addestramento della rete tramite un algoritmo di retropropagazione dell'errore (*backpropagation*) e discesa stocastica del gradiente (*gradient descent*).

Uno dei problemi maggiori delle reti neurali convoluzionali è l'*overfitting* dei dati, una condizione in cui il modello si adatta a caratteristiche specifiche del training set, ma non riesce a generalizzare su nuove immagini. Per prevenire questa condizione esistono diversi metodi in letteratura:

- Divisione del dataset in un *Training set*, *Validation set* e *Test set*. Nel training set sono presenti gli esempi utilizzati dalla rete per l'aggiornamento dei parametri, mentre il validation e test set conterranno immagini non visionate dalla rete durante l'addestramento e servono per la sua validazione. La proporzione delle immagini nei vari subset dipende dal numero di immagini che si ha a disposizione, ma quella più utilizzata è del tipo 60-20-20;
- Aumento dei dati di training: l'ideale sarebbe avere a disposizione un gran numero di immagini così che il modello non riesca ad adattarsi ad un subset così ampio. Quando non sia possibile reperire più esempi, si possono applicare delle trasformazioni dell'input quali rotazioni, deformazioni, ingrandimenti o rotazioni;
- Ridurre la complessità del modello;
- Terminazione anticipata dell'allenamento: nel caso in cui non si registrino progressi per un numero di iterazioni fissato, l'allenamento viene interrotto;
- Regolarizzazione: modifica la funzione d'errore aggiungendo un fattore dipendente dai pesi e da un parametro λ , detto *tasso di regolarizzazione*.
- Regolarizzazione con dropout: si disattivano in maniera casuale i neuroni o le connessioni della rete durante l'allenamento. In questo modo la rete non può far affidamento sugli stessi neuroni per estrarre feature.

Descrizione del modello utilizzato

L'architettura della rete utilizzata è stata scelta in modo da considerare anche l'individuazione dei bordi tra cellule adiacenti, introducendo una terza classe alla consueta classificazione binaria dei pixel in oggetto e sfondo, adottando la struttura suggerita da Kumar et al. [49]. Questo tipo di

architettura è stato utilizzato, nell'articolo, per estrarre i nuclei di tessuti provenienti da organi diversi ed ha dimostrato buone performance; per questo motivo si è pensato potesse adattarsi bene al dataset a disposizione, considerata la sua spiccata eterogeneità.

Per il training si è utilizzato il dataset annotato, costituito da 92 immagini. Il dataset è stato assemblato da più dataset open source per cui presentava caratteristiche diverse anche per quanto riguarda la risoluzione. Per questo motivo le immagini sono state ricampionate in modo da avere una risoluzione 40x. La rete è stata costruita in ambiente MATLAB[®] 2019a. Per generare le annotazioni per la terza classe si sono estratti e dilatati i bordi dalle cellule del dataset annotato. Per le due classi rimanenti, *regione intracellulare* e *regione extracellulare*, sono state utilizzate le annotazioni fornite dai dataset.

L'architettura della rete è sintetizzata nella Tabella 2. La rete riceve in input una patch di dimensioni 51x51 centrata attorno al pixel del quale si vuole fornire la predizione. La dimensione della patch è tale da riuscire a coprire la maggior parte dei nuclei ed allo stesso tempo essere più piccola possibile. L'input viene processato attraverso 3 *layer convoluzionali* con filtri di dimensione crescente. All'output di ognuno di questi layer viene applicata un'operazione di *max-pooling* riducendo di volta in volta la dimensione della *feature map*. Seguono 2 *fully connected layers* (FC) ed infine un *softmax layer* che dà in output un vettore contenente le probabilità di quel pixel di appartenere ad una delle tre classi, tale che la loro somma sia pari ad 1. In ogni layer si è utilizzata un'unità di non linearità *ReLU* in modo da velocizzare il training evitando il problema della scomparsa del gradiente (in inglese *vanishing gradients*); tale problema si verifica nel caso di reti profonde che utilizzino funzioni di attivazione non lineari, in cui un grande cambiamento nell'input induce un piccolo cambiamento nell'output; per questo motivo il gradiente della funzione di perdita risulta troppo piccolo, rendendo più difficile l'addestramento della rete. Il *learning rate* è stato impostato inizialmente ad un valore pari a 0.01 e poi diminuito di un fattore 10 quando l'errore di validazione risultava praticamente costante. Per evitare l'overfitting, la rete è stata regolarizzata tramite dropout con diversi dropout rate, crescenti in layer successivi.

L'algoritmo di ottimizzazione scelto è quello della discesa del gradiente stocastico (SGDM) con momento pari a 0.9. I parametri della rete vengono aggiornati tramite le seguenti equazioni:

$$W_{l+1} = W_l - \alpha \nabla L(W_l) + \gamma(W_l - W_{l-1}) \quad (4.1)$$

$$b_{l+1} = b_l - \alpha \nabla L(b_l) + \gamma(b_l - b_{l-1}) \quad (4.2)$$

Con l numero dell'iterazione, α learning rate, L è la funzione d'errore e γ è il momento. Il momento introduce un contributo alla variazione sui pesi da effettuare che deriva dal passo

precedente, imponendo una sorta di inerzia al sistema. La funzione di perdita utilizzata è una “*cross entropy*” ed è definita secondo:

$$loss = - \sum_{i=1}^N \sum_{j=1}^K t_{ij} \log(y_{i,j}) \quad (4.3)$$

Con N numero dei campioni, K numero delle classi, t_{ij} è un indicatore binario uguale a 1 se la classe predetta j corrisponde alla corretta classificazione per l’osservazione i, log indica il logaritmo naturale e y è la probabilità predetta che l’osservazione i appartenga alla classe j.

Tabella 4.1. Architettura della CNN adottata

Layer	Dimensione filtri	Attivazione	Output	Dropout rate
Input	-	-	51x51x3	-
Conv 1	4x4	ReLU	48x48x25	0.1
Pool 1	2x2	Max	24x24x25	-
Conv 2	5x5	ReLU	20x20x50	0.2
Pool 2	2x2	Max	10x10x50	-
Conv 3	6x6	ReLU	5x5x80	0.25
Pool 3	2x2	Max	3x3x80	-
FC 1	-	ReLU	1024	0.5
FC 2	-	ReLU	1024	0.5
Output	-	SoftMax	3	-

Dopo aver allenato la rete, si è proceduto a fornire in ingresso le immagini del secondo dataset per ottenerne una prima segmentazione. Il risultato è ancora grezzo, avendo la rete un’accuratezza non troppo alta. Essendo infatti che le immagini utilizzate nel dataset erano poche e provenivano da una distribuzione diversa rispetto a quelle utilizzate per il testing, si è preferito concentrarsi su un algoritmo di post processing che a partire dai risultati non ottimali della rete riuscisse a segmentare tutte le cellule.

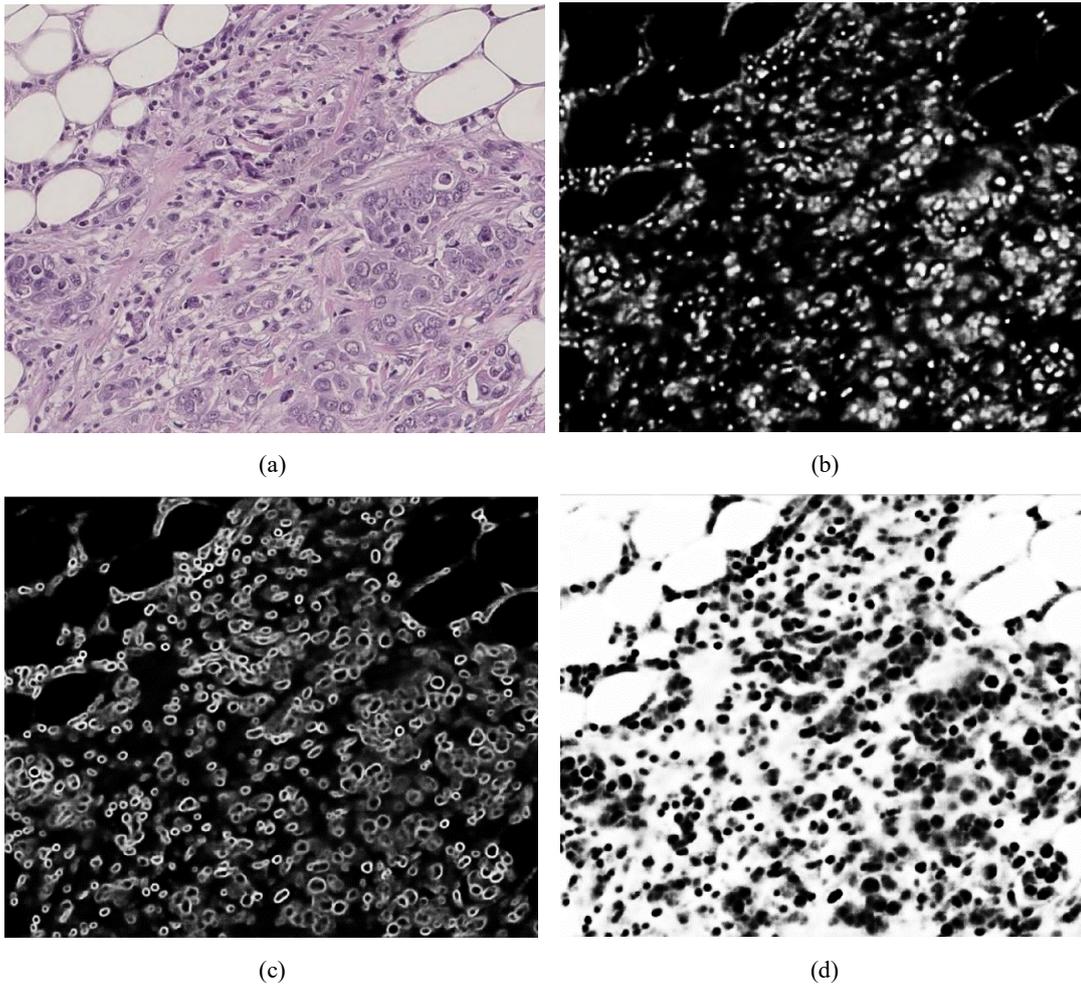


Figura 4.6: (a) immagine originale; (b) mappa delle regioni intracellulari; (c) mappa dei bordi; (d) mappa delle regioni extracellulari.

4.4 Post processing

L'output della rete è costituito da 3 immagini le cui intensità dei pixel rappresentano la probabilità che esso appartenga alle tre classi, in modo tale che la somma delle intensità del pixel sui 3 canali sia sempre uguale a 1. L'algoritmo si pone come obiettivo l'estrazione di una maschera binaria delle cellule a partire dall'output della rete. A questo scopo le mappe di probabilità vengono sottoposte ad una fase di post-processing per rifinire la segmentazione.

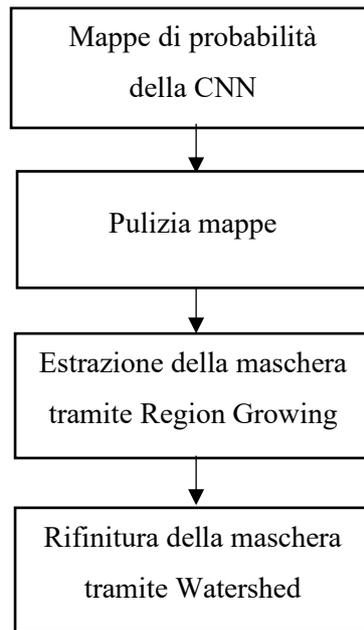


Figura 4.7: Diagramma a blocchi dell'algorithmo di Post-processing

Pulizia delle mappe di probabilità

Prima di tutto si è eseguita un'operazione di pulizia sulle maschere cercando di eliminare i pixel appartenenti allo sfondo dalla mappa relativa alle regioni intracellulari. Si sono individuati quei pixel che avessero un'alta probabilità di appartenere alla regione intracellulare congiuntamente ad un'alta probabilità di non appartenere alla regione extracellulare sommando la prima al complemento della seconda ed applicando una soglia di 0.9. I pixel che superavano la soglia sono stati mantenuti, mentre gli altri scartati. Infine, alla mappa pulita è stata sottratta la mappa dei bordi, moltiplicata per un fattore 0.2 in modo da cercare di isolare i centri dei nuclei. Sulla mappa dei bordi invece si è deciso di non eseguire nessuna pulizia, dato che la delineazione della maschera binaria viene effettuata su di essa.

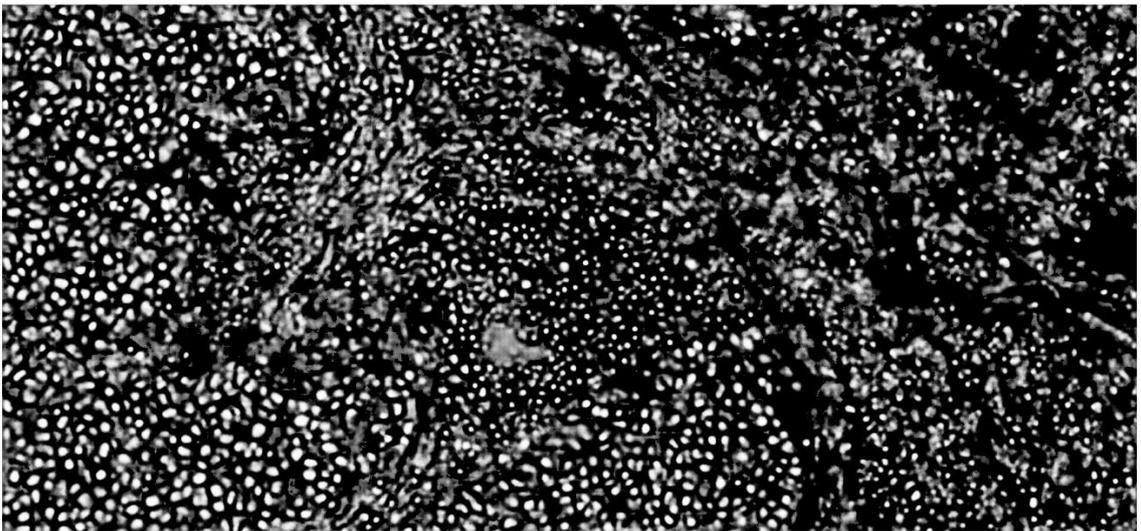
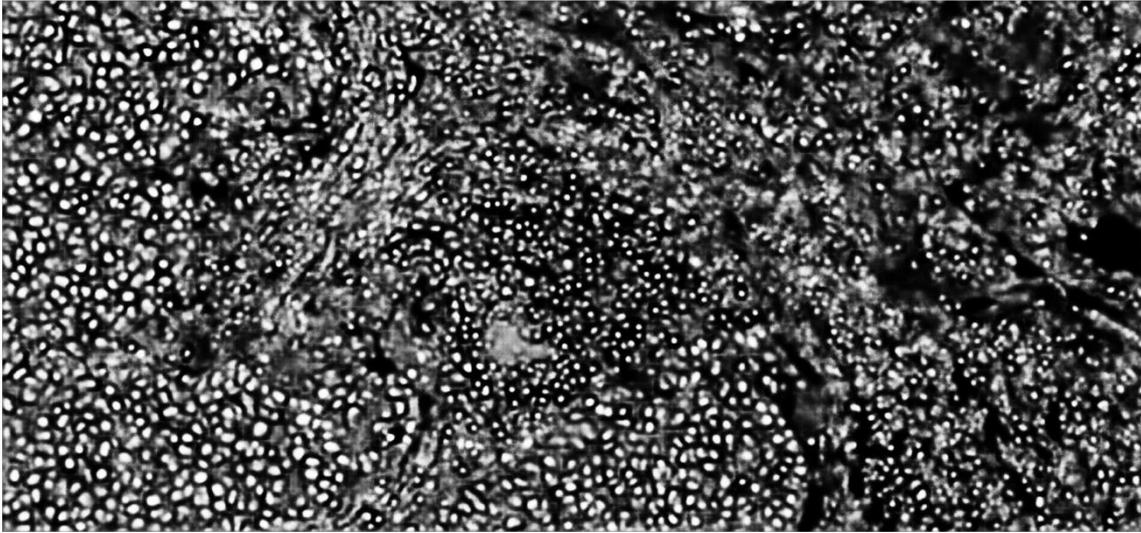


Figura 4.8: In alto mappa delle regioni intra cellulari originale. In basso mappa processata

Region Growing

L'output della rete è costituito da 3 immagini le cui intensità dei pixel rappresentano la probabilità che esso appartenga alle tre classi, in modo tale che la somma delle intensità del pixel sui 3 canali sia sempre uguale a 1. Ogni bordo sulla mappa non è costituito da un unico perimetro, ma da più corone concentriche con intensità variabile; si può individuare dunque un perimetro interno ed uno esterno. L'algoritmo applicato si pone come obiettivo una prima creazione di una maschera binaria delle cellule tramite la delineazione del perimetro interno ed esterno dei bordi individuati sulla relativa mappa di probabilità in uscita dalla rete.

Si effettuano delle operazioni di dilatazione e un algoritmo di region growing. La tecnica di region growing consente di individuare selettivamente una regione, mantenendo una coerenza

morfologica. L'algoritmo raggruppa regioni di pixel caratterizzati da intensità simili; in particolare a partire da gruppi di pixel che ricadano all'interno della regione da riconoscere chiamati semi (in inglese *seed*), l'algoritmo accresce tale insieme fino a ricoprire l'intera regione. Questi semi possono essere ricavati in modi diversi, ma il criterio più comune è l'apposizione di una soglia sull'intensità. I pixel vengono aggiunti uno alla volta e selezionati all'interno di una regione di esplorazione definita attraverso un elemento strutturale. Tale elemento è formato da *adiacenze* e *contorni* del seme considerato; le adiacenze sono i punti direttamente connessi sulla colonna o sulla riga del pixel di interesse, mentre i contorni sono i punti connessi che si trovano sulla diagonale. Nel caso in cui si considerino solo le adiacenze si avrà un insieme 4-connesso, mentre considerando anche le connessioni si avrà un insieme 8-connesso. Il criterio di selezione o esclusione dei pixel da aggiungere alla regione si basa sull'applicazione di una seconda soglia, inferiore alla prima nel caso di oggetti chiari su sfondo scuro, in modo tale che quando la maschera cresce e si avvicina al bordo i pixel assumono intensità sempre minore, fino ad avere un valore inferiore alla soglia; in questo caso il pixel verrà attribuito alla maschera dello sfondo.

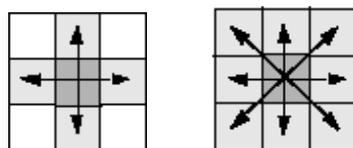


Figura 4.9. Esempi di insieme 4-connesso (sinistra) e 8-connesso (destra). [50]

Le mappe in uscita dalla fase di pulizia contengono ancora regioni appartenenti a cellule diverse connessi fra di loro. La rete ha fornito in output regioni cellulari caratterizzate da diverse intensità dei pixel; in particolare cellule piccole ed isolate presentavano una bassa intensità, molto inferiore rispetto a quella assegnata a cellule o aggregati di cellule più grandi. Risultava particolarmente difficile individuare un'unica soglia per definire dei semi che ricadessero su tutte le cellule ed allo stesso tempo garantire un seme per ogni cellula risultava impossibile ed i risultati non erano soddisfacenti. Per questo motivo individuare i semi in maniera iterativa si è rivelata una strategia migliore; In particolare si è eseguito un primo thresholding dei pixel ad elevata probabilità, per applicare di volta in volta soglie di valore inferiore nell'intervallo [0.9 0.7 0.5 0.4 0.3]. Le cellule individuate ad ogni step vengono eliminate dalla maschera utilizzata per individuare i semi prima di applicare la soglia successiva.

Per garantire una migliore precisione, non viene imposta una soglia nella condizione di stop del region growing ma si utilizza un criterio adattato ad ogni singola cellula per individuare il bordo della regione interna ed esterna. L'algoritmo si basa sulle seguenti assunzioni:

- 1) Non è possibile definire una soglia unica per individuare i centri di tutte le cellule: cellule piccole ed isolate presentano un'intensità molto inferiore rispetto a cellule più grandi o presenti in regioni ad alta densità cellulare;
- 2) La mappa delle regioni interne assume valori maggiori nell'intorno del centro della cellula per diventare via via inferiori man mano che ci si allontana dal centro;
- 3) La mappa dei bordi presenta pixel più scuri all'interno e più chiari in corrispondenza del bordo fino a raggiungere un picco di intensità in corrispondenza del bordo vero, per poi assumere di nuovo dei valori via via inferiori man mano che ci si allontana dal bordo. Nel caso in cui questa assunzione non sia rispettata, non ci si trova in corrispondenza di un bordo vero, ma di rumore.
- 4) Per lo stesso motivo del punto 2, bordi di cellule diverse possono avere valori medi di intensità diversi, per cui non è possibile definire una soglia univoca per individuare tutti i bordi delle cellule.
- 5) La differenza di intensità media tra i pixel che compongono il perimetro interno o il perimetro esterno dei bordi individuati sulla mappa di probabilità è molto maggiore rispetto alla differenza tra le intensità medie dei pixel che costituiscono corone concentriche del bordo.

Ogni seme viene fatto crescere nella corrispondente regione utilizzando la procedura descritta di seguito. Si estrae una maschera dei primi semi imponendo la prima soglia. Il contorno di ogni componente connessa individuata viene dilatato, applicando un elemento strutturale di forma circolare. Questo passaggio serve nel caso di cellule molto grandi, in modo tale da permettere al region growing di individuare il bordo eseguendo un numero inferiore di iterazioni. La dilatazione viene eseguita per un numero di iterazioni massimo pari a 4. Ad ogni iterazione si calcola la differenza tra l'intensità media del bordo attuale e quella del bordo precedente; infine si individua l'iterazione nella quale questa differenza raggiunge il primo picco e si sceglie il bordo corrispondente all'iterazione precedente, che raggrupperà pixel appartenenti alla regione interna. Tale scelta si è resa necessaria considerando che il processo di dilatazione utilizza un elemento strutturale di forma circolare, che nella maggior parte dei casi non corrisponde alla forma vera della cellula; ciò implica che questa forma circolare si adatterà solo parzialmente a cellule più allungate. Per questo motivo questo primo step serve unicamente a raggruppare pixel appartenenti alla regione interna del bordo da

utilizzare come semi per un region growing e non a definire effettivamente il contorno della cellula. A questo punto si esaminano i pixel presenti nel vicinato della regione individuata e si aggiunge di volta in volta il pixel con intensità minima tra quelli nel vicinato. L'obiettivo di questo step è individuare il perimetro interno del bordo della cellula. Per farlo si utilizza una procedura simile a quella proposta in [29].

Si definiscano le seguenti regioni:

- CR rappresenta la regione contenente tutti i pixel interni al bordo;
- IB rappresenta il perimetro della regione interna CR;
- CB rappresenta i pixel 8-connessi a IB.

L'algoritmo esegue i seguenti passaggi:

- 1) Si individua il pixel con minore intensità tra i pixel appartenenti al perimetro CB;
- 2) Il pixel viene rimosso da CB ed inserito nella regione CR;
- 3) CB viene aggiornato con i pixel 8-connessi al pixel appena rimosso e non presenti in CR;
- 4) Il perimetro interno IB viene riaggiornato sostituendolo con il contorno della nuova regione interna CR;
- 5) Si esegue la differenza tra l'intensità media dei pixel di CB e l'intensità media dei pixel di IB.

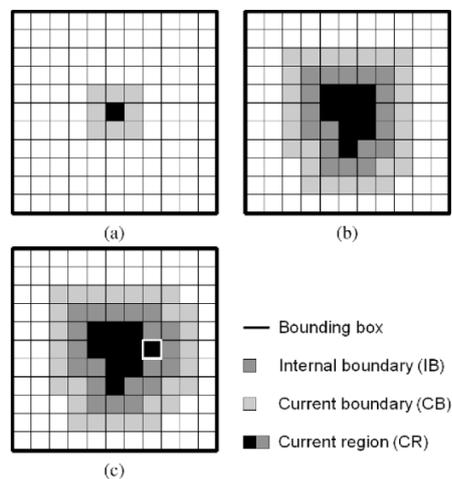


Figura 4.10: Illustrazione dell'algoritmo di region growing [29]

Il procedimento viene ripetuto finché la differenza non diventa negativa o per un massimo di 1000 iterazioni. Il valore delle differenze di intensità fra i bordi di ogni iterazione viene inserito in un vettore, all'interno del quale si va a ricercare il massimo fra i primi 3 picchi. Il valore trovato identificherà il perimetro della regione interna al bordo sulla mappa dei contorni delle cellule.

Calcolando ad ogni iterazione la differenza tra le intensità media dei pixel costituenti contorni successivi, si avrà che queste rimangono molto basse in regioni costituite da pixel con intensità media simile (quali la regione interna o i pixel appartenenti alla regione di bordo sulla mappa di probabilità) mentre saranno molto alte nel passaggio da regione interna a regione di bordo o da regione di bordo a regione esterna.

Individuata la regione interna al bordo sulla mappa dei bordi fornita dalla rete, bisogna ora individuare il perimetro esterno. Si esegue una serie di dilatazioni del perimetro interno individuato allo step precedente, per poi scegliere quello corrispondente al picco minimo delle differenze di intensità tra dilatazioni successive. In questo caso infatti l'intensità media in iterazioni successive crescerà fino ad arrivare ad un massimo in corrispondenza del centro del bordo sulla mappa dei contorni delle cellule, dopo di che decrescerà allontanandosi da esso. Queste differenze saranno basse finché il contorno dilatato si trova sulla regione di bordo, ma si avrà una differenza in valore assoluto massima nel passaggio dalla zona chiara del bordo alla zona scura della regione extra-cellulare.

A questo punto il bordo vero della cellula viene individuato come quello presente ad una distanza intermedia tra il bordo esterno e quello interno.

I perimetri esterni delle cellule individuati vengono inoltre dilatati e utilizzati per aggiornare una maschera binaria che serve per cancellare le cellule appena individuate, dalla mappa utilizzata per estrarre i semi. Imponendo infatti una soglia minore, si andrebbero ad riconsiderare anche cellule processate precedentemente, aumentando inutilmente i costi computazionali.

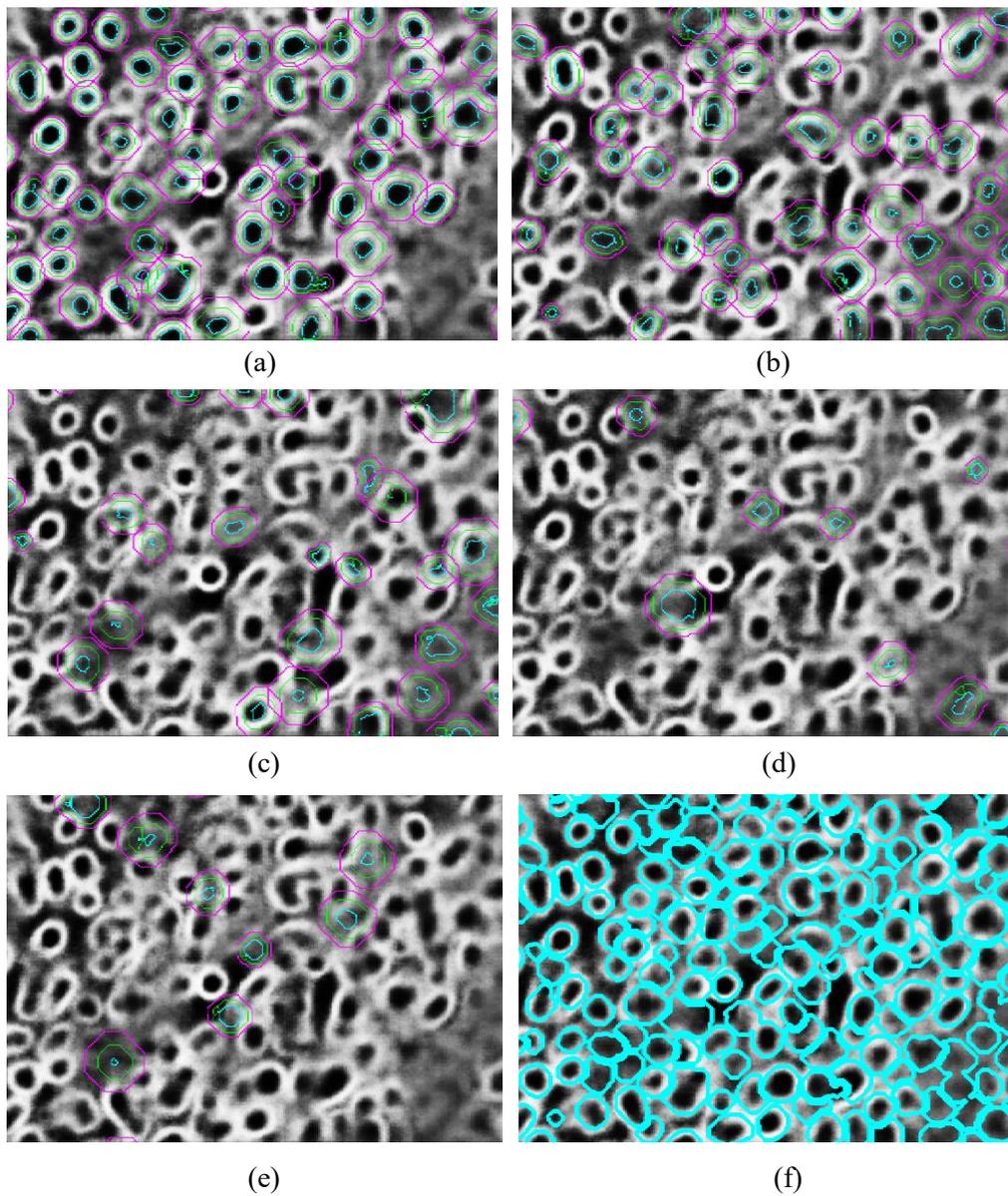


Figura 4.11: Esempio di perimetri interno ed esterno individuati a partire dai semi individuati tramite le varie soglie e risultato finale. (a) soglia 0,9; (b) soglia 0,7; (c) soglia 0,5; (d) soglia 0,4; (e) soglia 0,3; (f) risultato finale.

Trasformata Watershed

La maschera binaria ottenuta nello step precedente viene a questo punto rifinita utilizzando un algoritmo di watershed basato su marker. Gli oggetti estratti sono infatti ancora grezzi ed in alcuni casi sovrapposti. L'algoritmo consente inoltre di adattare meglio i contorni anche delle cellule che erano state isolate nel punto precedente, permettendo di stringere meglio la maschera sul contorno vero della cellula. La trasformata watershed si basa sull'interpretazione di un'immagine a toni di grigio come un rilievo topografico, le cui altitudini corrispondono ai livelli di grigio dei

pixel. Considerando questa analogia, un corso d'acqua scivolerà attorno ai rilievi topografici secondo un percorso che lo conduca ad un bacino di raccolta, individuato in un minimo locale. I percorsi seguiti da questi corsi separano i bacini tra di loro. Allo stesso modo tornando a considerare l'immagine a toni di grigio, tali percorsi corrisponderanno a linee di separazione fra gli oggetti. Le segmentazioni watershed di solito si basano sulla *distance transform* che sostituisce al valore di ogni pixel l'entità della sua distanza dal pixel nero più vicino. Applicando la *distance transform* sul complemento della maschera binaria si ottengono le distanze dai pixel non nulli della maschera iniziale. Applicando la watershed sul complemento di questa immagine si ottiene la segmentazione. In molti casi si verifica il fenomeno della sovrasedgmentazione. La watershed basata su marker permette di evitare questo problema imponendo i minimi in corrispondenza dei marker. Bisogna identificare due tipi di marker:

- piccole componenti connesse posizionate su ogni oggetto che si vuole separare: di solito vengono individuate eseguendo delle operazioni morfologiche atte a separare gli oggetti dallo sfondo;
- linee di separazione che ricadano nelle zone di sfondo;

L'algoritmo viene applicato in genere sul gradiente dell'immagine, che presenta valori di intensità più alti in corrispondenza dei pixel costituenti i contorni degli oggetti, su cui sono stati imposti punti di minimo in corrispondenza dei marker precedentemente individuati.

In questo lavoro si è preferito utilizzare come funzione di segmentazione la mappa dei bordi in uscita dalla CNN e come marker degli oggetti le regioni interne individuate nello step precedente; i marker dello sfondo corrispondono alla *distance transform* applicata sul complemento della maschera binaria ricavata nello step precedente.

La mappa dei bordi viene quindi aggiornata imponendo i minimi in corrispondenza dei marker interni ed esterni. Gli oggetti sovrapposti vengono individuati imponendo una soglia di 0.875 sulla *solidity*. La *solidity* è definita come il rapporto tra l'area dell'oggetto e l'area del suo involucro convesso (in inglese *convex hull*), un poligono convesso che racchiuda tutti i punti dell'oggetto. La soglia è stata determinata in maniera empirica e presentava il giusto compromesso tra il numero di oggetti correttamente individuati come sovrapposti e il numero di oggetti che invece corrispondevano già a cellule isolate. Le cellule isolate verranno comunque sottoposte alla segmentazione con la watershed, ma processate separatamente rispetto agli oggetti sovrapposti per diminuire il costo computazionale; nel caso degli oggetti sovrapposti, infatti, la watershed viene eseguita su una componente connessa alla volta poiché questa accortezza permetteva di ottenere dei risultati migliori.

Durante questo processo inoltre si riescono ad individuare molti falsi positivi sfuggiti dallo step precedente, poiché questi oggetti vengono ridotti a punti con area molto inferiore agli altri ed individuabili quindi con un semplice controllo sul numero di pixel che li compongono.

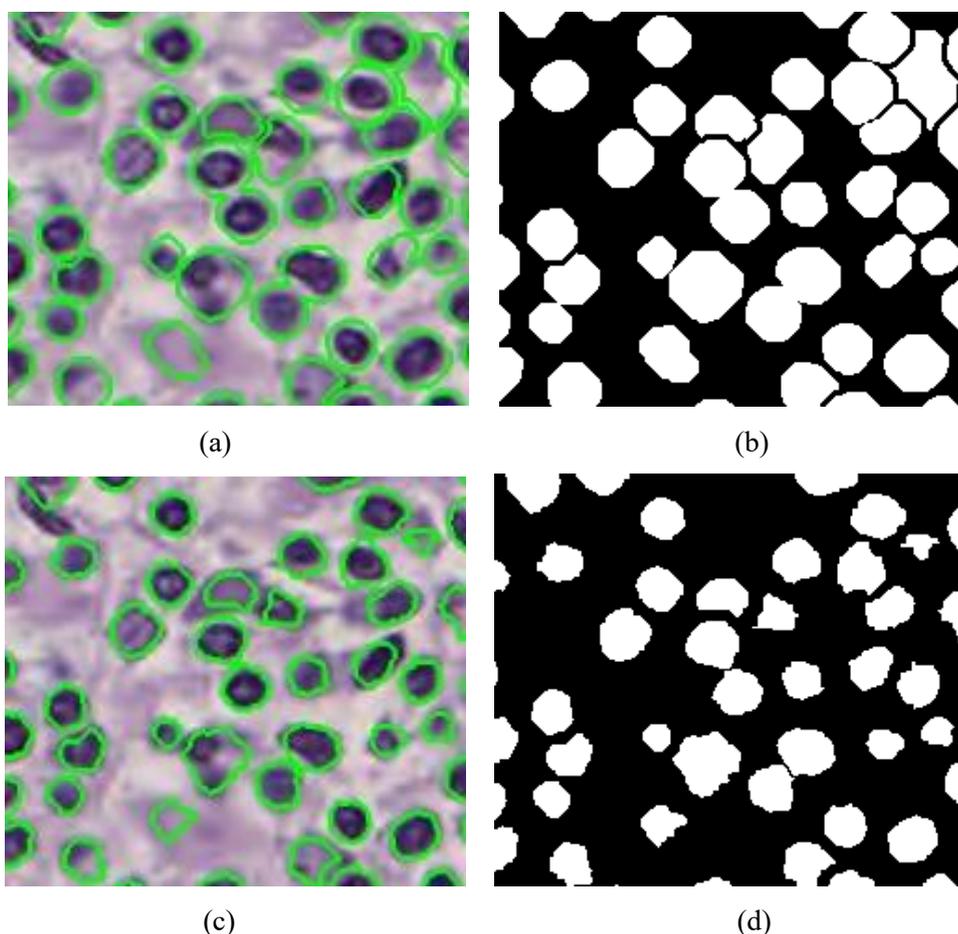


Figura 4.12: Risultato della trasformata watershed. (a)segmentazione da region growing; (b) maschera binaria da region growing; (c) risultato tramite watershed; (d) maschera binaria della watershed.

4.5 Pulizia della maschera binaria

Le maschere ottenute presentano ancora un alto numero di falsi positivi, ovvero oggetti erroneamente identificati dalla rete come cellule. Questi oggetti vengono individuati tramite un confronto con la *color deconvolution*. La color deconvolution è un tipo di processamento dell'immagine effettuata per separare i contributi delle tinte utilizzate durante il processo di colorazione dei campioni di tessuto sui vetrini. Le sostanze coloranti che si legano alle strutture biologiche, assorbendo la luce emanata dal microscopio a lunghezze d'onda differenti, assegnerà alle strutture rilevanti del campione una colorazione univoca. In questo lavoro si è utilizzata la

Sparse Nonnegative Matrix Factorization (SNMF) per stimare le colorazioni principali. Come già detto nel caso di immagini istologiche per la valutazione dei TIL si utilizza la colorazione Ematossilina-Eosina che assegna ai nuclei una colorazione viola e al citoplasma e allo stroma una colorazione rosa. Il primo passo consiste quindi in una trasformazione dell'immagine dallo spazio colore RGB in una *mappa di densità* o *assorbimento* per le due colorazioni. Per preservare le strutture biologiche durante l'operazione, si utilizzano dei modelli per le mappe di densità del colore tenendo conto delle seguenti proprietà:

- *Non-negatività*: il risultato della trasformazione è una mappa di densità ottica e di densità colore, per cui tenendo conto di ciò, ottenere dei valori negativi assumerebbe fisicamente il significato di emissione di luce. Nel nostro caso invece si può avere assenza o presenza di assorbimento della luce, per cui gli unici valori possibili sono nulli o positivi.
- *Sparsity*: convenzionalmente si dice che l'ematossilina (H) si leghi ai nuclei e l'eosina (E) al citoplasma. In realtà ogni struttura si lega in proporzioni diverse ai due coloranti. In particolare, si può definire una colorazione effettiva ed una colorazione pura. Le colorazioni pure sono quelle associabili ai due coloranti separatamente, la colorazione effettiva è quella assunta dalla struttura biologica in esame la quale presenterà una miscela con diverse proporzioni tra le due colorazioni pure; nel caso ad esempio di un nucleo, la colorazione effettiva sarà costituita da una grande quantità di colorazione pura H e una piccola quantità di colorazione pure E. La colorazione che si va a stimare è dunque quella effettiva.
- *Soft-classification*: nel caso di pixel di bordo tra due diversi materiali biologici, si permette la presenza di entrambe le colorazioni.

La trasformazione dell'immagine in una mappa della densità di colore viene eseguita applicando la legge di *Lambert-Beer*, la quale definisce la dipendenza dell'intensità della luce trasmessa attraverso un campione con la quantità di colorante per il proprio specifico fattore di assorbimento. Sia $I \in R^{m \times n}$ la matrice di intensità dell'immagine, con $m=3$ per i canali RGB ed n dato dal numero di pixel, e sia I_0 l'intensità della luce incidente sul campione. Definiamo con $W \in R^{m \times r}$ la matrice dei vettori colorazione, le cui colonne rappresentano la base colore di ciascuna colorazione e r il numero di colorazioni; indichiamo con $H \in R^{r \times n}$ la mappa di densità del colore, le cui righe rappresentano la concentrazione di ciascuna colorazione. La matrice delle intensità I può esser scritta come:

$$I = I_0 \exp(-WH) \quad (4.4)$$

Si nota che la trasmissione della luce è dunque i livelli di grigio di ogni canale sull'immagine, dipendono dalla concentrazione di colorazione in una maniera non lineare per cui i valori di

intensità dell'immagine non possono essere usati direttamente per la separazione e la misurazione di ognuna delle colorazioni. Tuttavia, definendo con V la densità ottica relativa:

$$V = \log \frac{I_0}{I} \quad (4.5)$$

si può scrivere:

$$V = WH \quad (4.6)$$

La densità ottica per ogni canale è lineare con la concentrazione del materiale assorbente e può quindi esser facilmente utilizzata per separare il contributo di più tinte all'interno del campione.

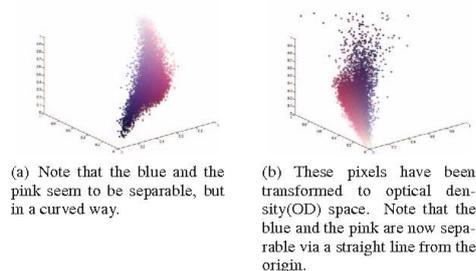


Figura 4.13: Ogni punto nello spazio 3D corrisponde ad un pixel selezionato in maniera casuale dall'immagine e colorato col colore originale in Ematossilina Eosina [51]

Dalla matrice di densità ottica, si ricava la densità di colorazione, contenente le componenti della densità ottica per ogni tinta sui tre canali, e la mappa di densità colore, che contiene le componenti della miscela di colorazioni per ogni pixel. Il metodo di separazione dei contributi relativi a ciascun colorante utilizzato in questo lavoro è la *SMNF* che si basa sulla non *negative matrix factorization* (NMF), ma aggiunge un elemento che considera il fatto che ogni struttura biologica, come ad esempio un nucleo, ha un'estensione spaziale finita, caratterizzata dal suo spettro di assorbimento o dalla sua colorazione effettiva. L'introduzione di questo termine permette di simulare la specificità del target di ognuna delle tinte nel processo di colorazione, permettendo di preservare la struttura biologica.

Una volta ottenuta la color deconvolution si procede alla pulizia della maschera. Imponendo una soglia sulla media e la deviazione standard dei pixel costituenti i singoli oggetti sulla color deconvolution è possibile individuare i falsi positivi. Dato che l'apparenza delle cellule è molto diversa tra loro, quelle caratterizzate da una cromatina densa e scura sono facilmente individuabili imponendo una intensità media alta ed una bassa deviazione standard; invece le cellule caratterizzate da cromatina poco densa e sparsa (come nel caso di cellule tumorali) o con nucleoli molto piccoli e molto citoplasma (come nel caso di altre cellule immunitarie mononucleari) appaiono invece con una bassa intensità media e alta deviazione standard. Anche i falsi positivi si presentavano diversi all'aspetto: in alcuni casi in cui lo stroma appariva omogeneo,

presentavano bassa intensità media e bassa deviazione standard, in altri invece in cui lo stroma risultava esser caratterizzato da un'alta presenza di fibre e tessuto connettivo, appariva caratterizzato dalla presenza di striature più o meno lunghe sulla color deconvolution; tali caratteristiche facevano sì che falsi positivi che fossero ricaduti su quelle zone presentassero un'alta deviazione standard e potessero così esser confusi con TIL (diversi da piccoli linfociti) o con cellule tumorali. Per questo motivo si è deciso di utilizzare anche una seconda maschera ottenuta dopo aver applicato operazioni morfologiche sull'immagine a toni di grigio relativa alla color deconvolution, atte a filtrare le strutture striate; si è imposto un elemento strutturale con raggio minimo pari alla metà di un piccolo linfocita, tale è la misura dei nucleoli dei plasmaciti.

Anche in questo caso si è voluto tener conto della diversità dell'aspetto della color deconvolution ottenuta per immagini con densità cellulari diverse. Si sono individuati due sottocasi:

- 1) Nel caso in cui la densità cellulare sia bassa, le cellule appaiono spesso isolate e scure, il che si traduce in un minor contrasto sulla color deconvolution;
- 2) Nel caso in cui la densità cellulare sia alta, le cellule risultano ammassate fra loro e presentano una intensità maggiore.

I due casi sono stati discriminati eseguendo un *thresholding di Otsu* sull'immagine della color deconvolution precedentemente processata tramite operatori morfologici, come spiegato sopra. Se la soglia individuata dall'algoritmo di Otsu supera un valore determinato empiricamente e pari a 0.3, l'immagine è riconosciuta come avente contrasto maggiore e le cellule vengono sottoposte ad un criterio di pulizia differente.

Si procederà ora a definire il criterio utilizzato per immagini con bassa densità cellulare (caso 1). Si è applicata congiuntamente una soglia sull'intensità media e sulla deviazione standard. In primo luogo si sono eliminati unicamente gli oggetti troppo scuri, utilizzando una soglia pari a 0.1. La soglia scelta è molto bassa ma è un compromesso per evitare di perdere cellule vere ed allo stesso tempo eliminare più falsi positivi possibile; per questo motivo si sono imposti altri parametri in modo da migliorare i risultati di questo step. Il passo successivo è stato individuare altri oggetti da eliminare alzando leggermente la soglia sull'intensità media ed aggiungendone una sulla deviazione standard (soglia su media=0.15, soglia su deviazione standard=0.1).

Nel caso 2 di immagini a più alta densità cellulare, si sono considerati anche dei criteri morfologici. Piuttosto che apporre una soglia sulla media per individuare i piccoli linfociti si è preferito individuare le strutture da rielaborare tramite un criterio di sovrapposizione tra la maschera ottenuta da *thresholding* e gli oggetti individuati tramite *watershed*. Gli oggetti appartenenti alla maschera binaria in uscita dalla *watershed* che si sovrapponevano per più del

30% alla maschera ottenuta tramite thresholding sulla color deconvolution sono stati riconosciuti come cellule effettive, mentre gli altri sono stati sottoposti a processi di soglia e criteri morfologici. Tra gli oggetti che non hanno superato questo primo criterio, infatti ci potrebbero essere dei TIL mononucleari (come i plasmaciti) caratterizzati come già detto da un piccolo nucleolo molto scuro e abbondante citoplasma più chiaro; il nucleolo viene rilevato sulla maschera binaria proveniente dalla color deconvolution, mentre il citoplasma no. Questi oggetti presentano in ogni caso un'area molto maggiore dei piccoli linfociti, per cui vengono recuperati applicando i seguenti controlli:

- Oggetti con un nucleolo piccolo e convesso, individuato sulla maschera binaria della color deconvolution, e citoplasma più scuro ed omogeneo sulla color deconvolution (intensità media pari a 0.1 e deviazione standard pari a 0.15);
- Oggetti molto grandi ed leggermente più scuri (media superiore a 0.2).

Si ottiene così una maschera depurata dalla maggior parte dei falsi positivi. Eventuali altri falsi positivi vengono eliminati nello step successivo.

Durante queste operazioni di pulizia, si potrebbero verificare casi in cui il perimetro dell'oggetto individuato non si sovrappone al bordo vero della cellula; ciò si verifica soprattutto in corrispondenza di piccoli linfociti. Tale risultato si ottiene a causa di un errore nell'individuazione dei bordi ad opera della rete. Per recuperare questi piccoli linfociti ed altri sfuggiti alla rete, si aggiungono cellule individuate dalla color deconvolution che non si sovrappongono agli oggetti già presenti sulla maschera. Le cellule da aggiungere vengono individuate tra quelle che presentano una circolarità e una dimensione comparabile a quella di un piccolo linfocita su due maschere provenienti dalla color deconvolution. Una prima maschera ottenuta tramite thresholding sulla color deconvolution ed una seconda ottenuta imponendo una soglia pari a 0.4 sulla color deconvolution processata tramite operazioni morfologiche. In questo modo dalla maschera ottenuta tramite operazioni morfologiche si potevano riuscire a separare oggetti debolmente connessi ed allo stesso tempo, con il controllo sulla seconda maschera, evitare che venissero cancellati oggetti di interesse.

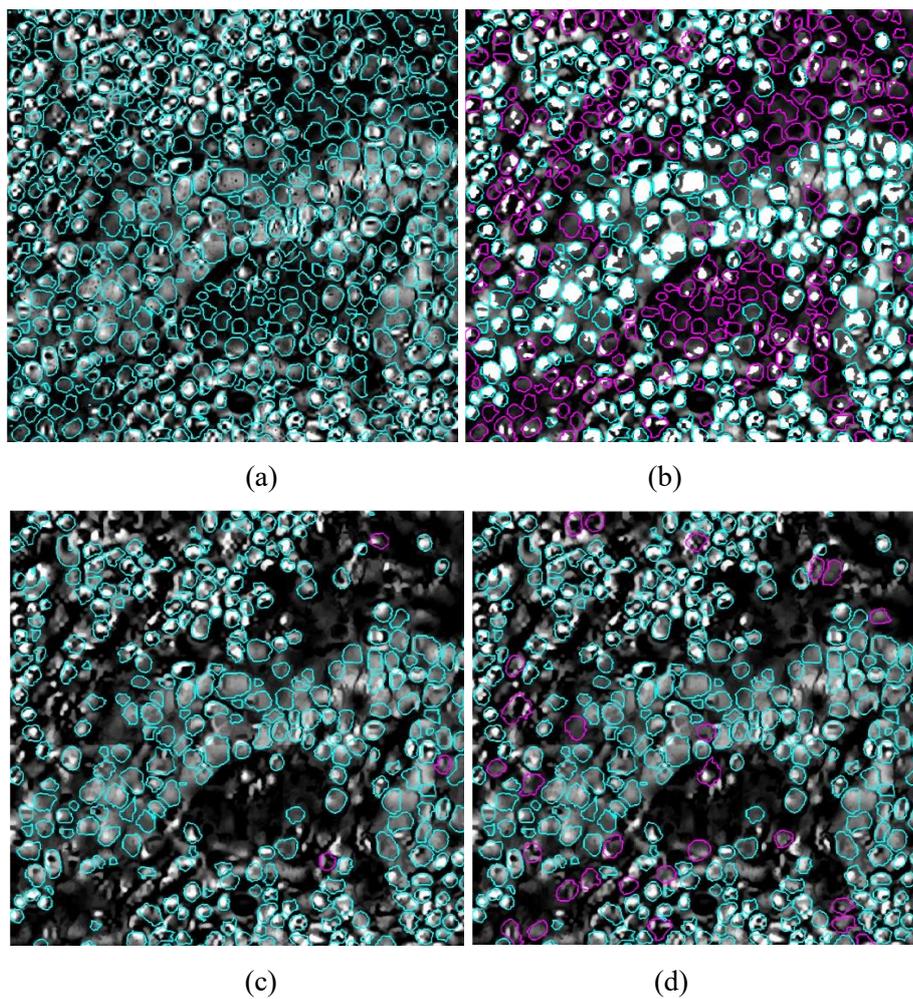
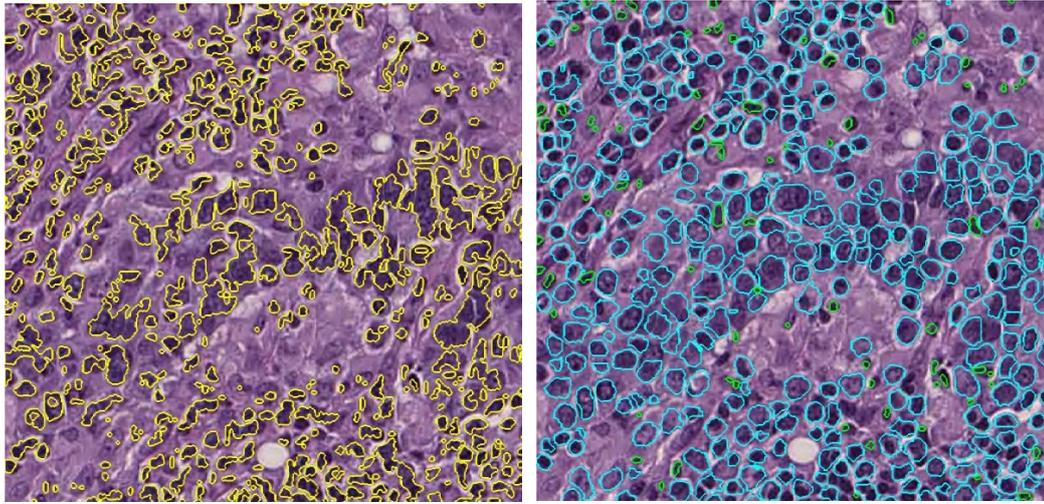
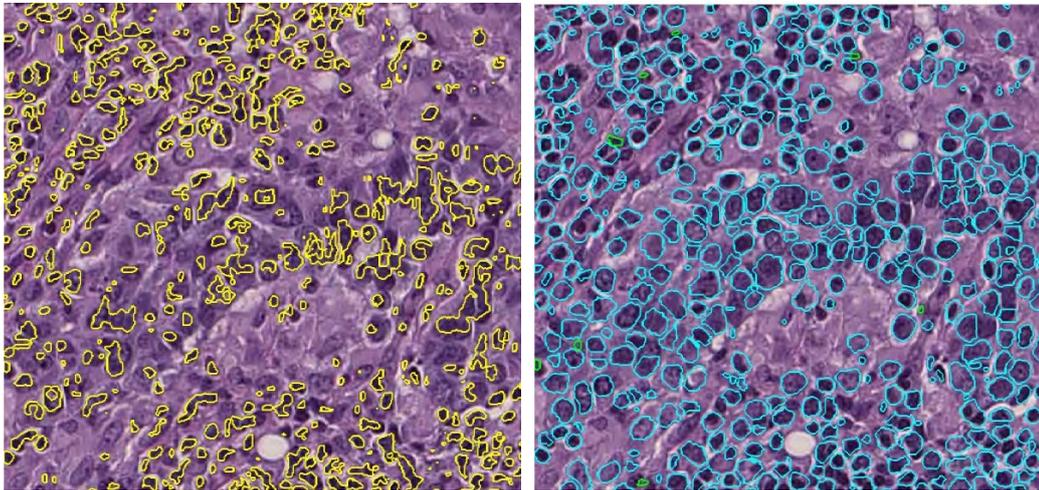


Figura 4.14: Esempio di immagine relativa al caso 2: (a) contorni iniziali dalla watershed; (b) individuazione degli oggetti con sovrapposizione minore del 30%;(c) individuazione di oggetti con nucleolo piccolo e convesso; (d) individuazione di oggetti grandi con media superiore a 0,2;



(a)

(b)



(c)

(d)

Figura 4.15: Esempio di immagine relativa al caso 2: (a) maschera da thresholding sulla color deconvolution processata con operatori morfologici; (b) cellule aggiunte; (c) maschera da thresholding sulla color deconvolution; (d) cellule aggiunte dalla seconda maschera.

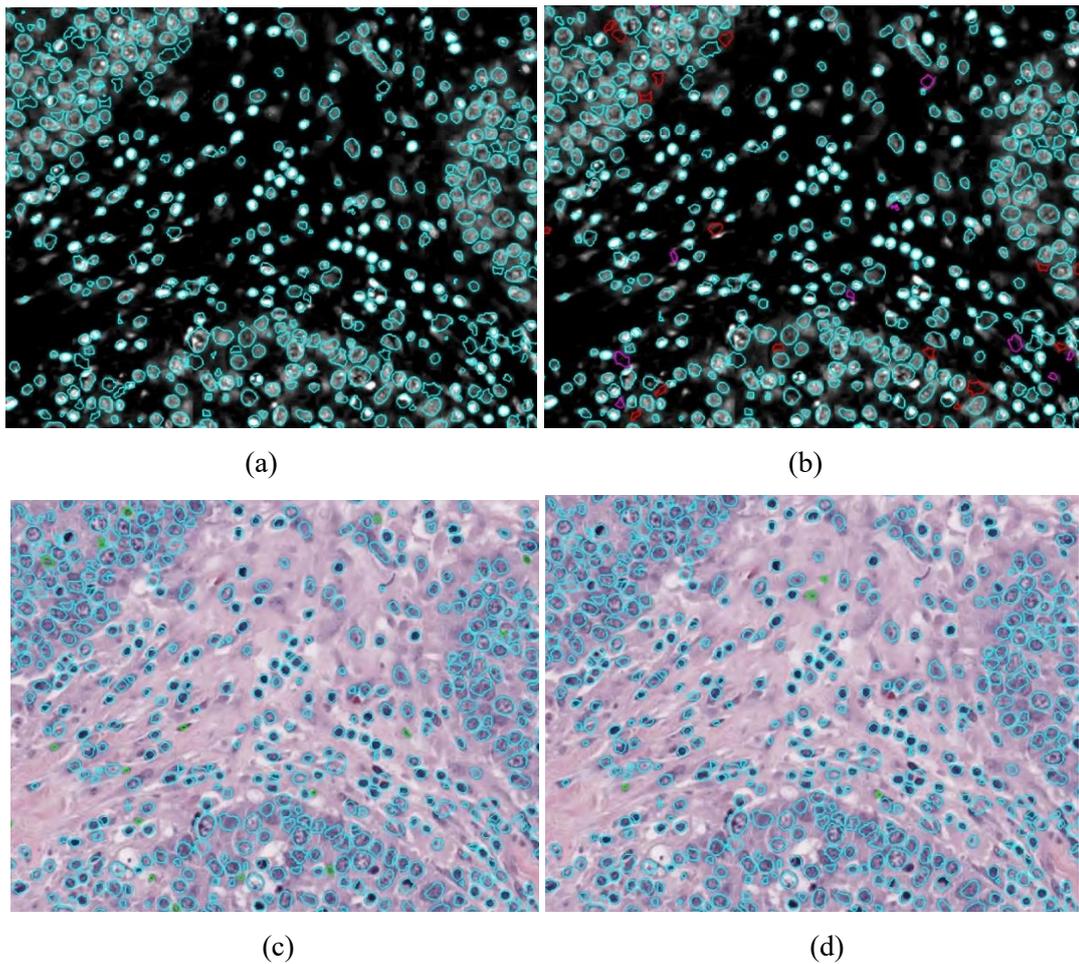


Figura 4.16: Esempio di immagine del caso 1. (a) contorni iniziali dalla watershed; (b) in rosso oggetti con intensità media minore di 0,1, in magenta oggetti con intensità media minore di 0,15 e deviazione standard 0,1; (c) risultato dopo aver aggiunto oggetti dalla maschera ottenuta dalla color deconvolution sottoposta ad operazioni morfologiche; (d) cellule aggiunte dalla maschera dalla color deconvolution ottenuta tramite thresholding.

4.6 Calcolo TIL%

La percentuale di linfociti infiltranti il tumore è data dal rapporto tra l'area dei TIL e l'area dello stroma in modo che teoricamente un 100% di TIL corrisponda ad uno stroma interamente costituito da queste cellule. Prima ancora di individuare i TIL è dunque necessario definire lo stroma. Le aree tumorali vengono individuate a partire dalle maschere di segmentazione del tumore fornite dal patologo. Lo stroma interessato dalla risposta infiammatoria dei TIL è generalmente quello vicino al bordo del tumore, per cui è stato definito come quell'area entro una distanza di circa 200 μm dal bordo del tumore, depurata da accumuli di adipe e zone caratterizzate da assenza di citoplasma, cellule o tessuto connettivo con un semplice individuazione dei pixel bianchi sull'immagine a toni di grigio.

Non essendo disponibili dei dataset annotati per la classificazione dei TIL, ci si è basati su un altro criterio per la loro definizione. Dato che ad eccezione dei fibroblasti, la gran parte delle cellule è TIL, è bastato considerare un thresholding sull'immagine. Una semplice clusterizzazione tramite k-means, utilizzando come feature media e deviazione standard, è risultata essere in grado di individuare gran parte dei veri positivi, tuttavia molte delle cellule individuate come non TIL si sono dimostrate in realtà dei falsi negativi. Per questo motivo, gli oggetti inseriti nei cluster relativi alle cellule non TIL sono state sottoposte ad un thresholding previa trasformazione nello spazio colore *Hue-Saturation-Value* (HSV). Tale spazio colore risulta più vicino alla percezione umana in termini di tinta, sfumatura ed intensità e risulta più accurato ed efficiente nel distinguere i colori degli oggetti [52]. La tonalità H viene misurata attraverso l'angolo rispetto alla posizione del rosso, che corrisponde a 0 gradi. La saturazione misura la tendenza al bianco; infine l'intensità del colore viene misurata attraverso la brillantezza.

I valori utilizzati per effettuare il thresholding sono stati individuati tramite un Tool Matlab che permettesse la visualizzazione 3D della distribuzione di tali colori.

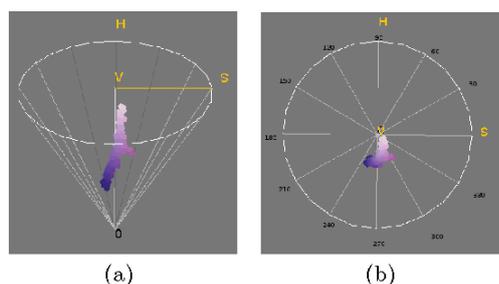


Figura 4.17: Rappresentazione dell'intensità dei pixel nello spazio colore HSV [52]

Infine, la percentuale dei TIL stromali viene calcolata rapportando l'area dei TIL individuati all'area dello stroma.

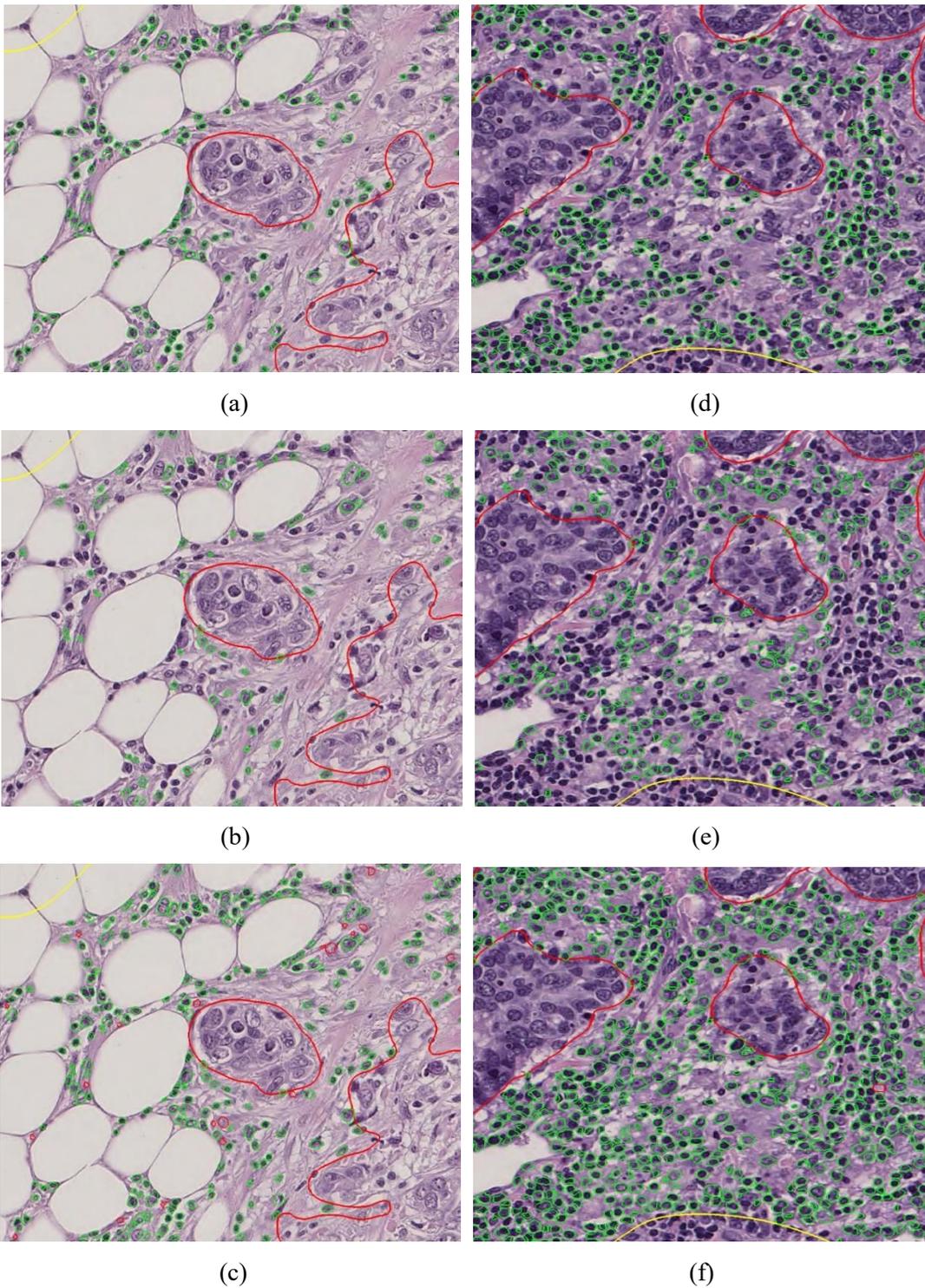


Figura 4.18: Procedura per l'individuazione dei TIL nello stroma in due diverse immagini. L'area delimitata in rosso rappresenta il bordo tumorale, l'area in giallo rappresenta lo stroma peritumorale. (a),(d) cellule identificate come veri positivi tramite k-means; (b),(e) cellule individuate tramite apposizione di soglia; (c),(f) risultato finale, con TIL in verde e cellule scartate in rosso.

Capitolo 5

Risultati

La CNN allenata su 92 immagini per riconoscere le zone intra-cellulari, extracellulari e dei bordi raggiunge un'accuratezza di corretta classificazione sul validation set del 75,60%. L'allenamento si ferma al raggiungimento massimo di iterazioni pari a 2200.

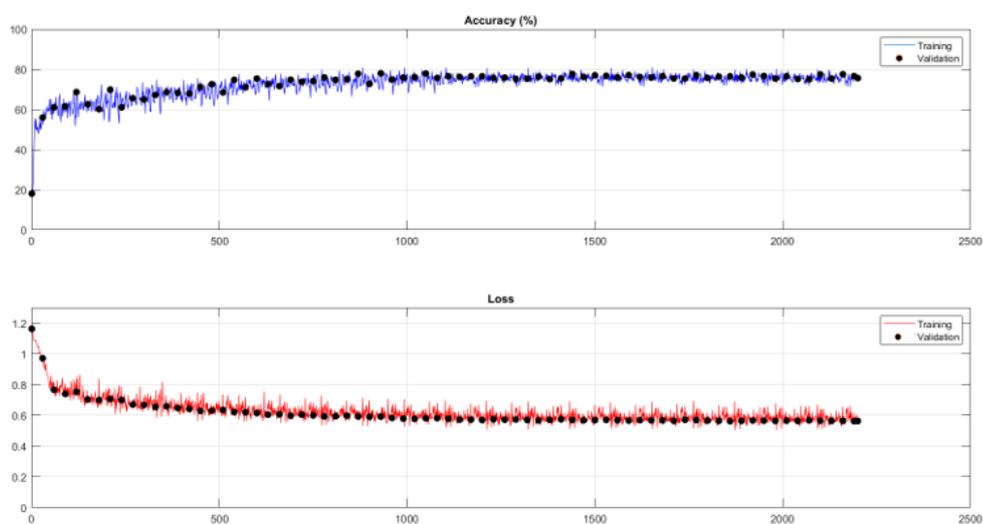


Figura 5.1: Allenamento della rete

Considerato l'alto numero di cellule si è deciso di effettuare una valutazione quantitativa sulla segmentazione dei TIL eseguita da un patologo esperto, che ha fornito un punteggio su una scala da 0 a 5, tale che 5 corrispondesse ad una segmentazione ottima e 0 ad una segmentazione pessima. La valutazione è stata eseguita sui nuclei segmentati nello stroma peritumorale su cui si è calcolata la %TIL, i risultati sono riportati in figura 5.2.

La maggior parte delle segmentazioni (82%) ha ricevuto un punteggio pari o superiore a 4; in particolare, il 51% delle segmentazioni ha raggiunto il punteggio massimo, mentre il 31% ha raggiunto un punteggio pari a 4. Il punteggio minimo assegnato è stato pari a 2 e corrisponde al 3,4% delle immagini sottoposte a valutazione. Le segmentazioni hanno raggiunto un punteggio medio di 4.3 punti;

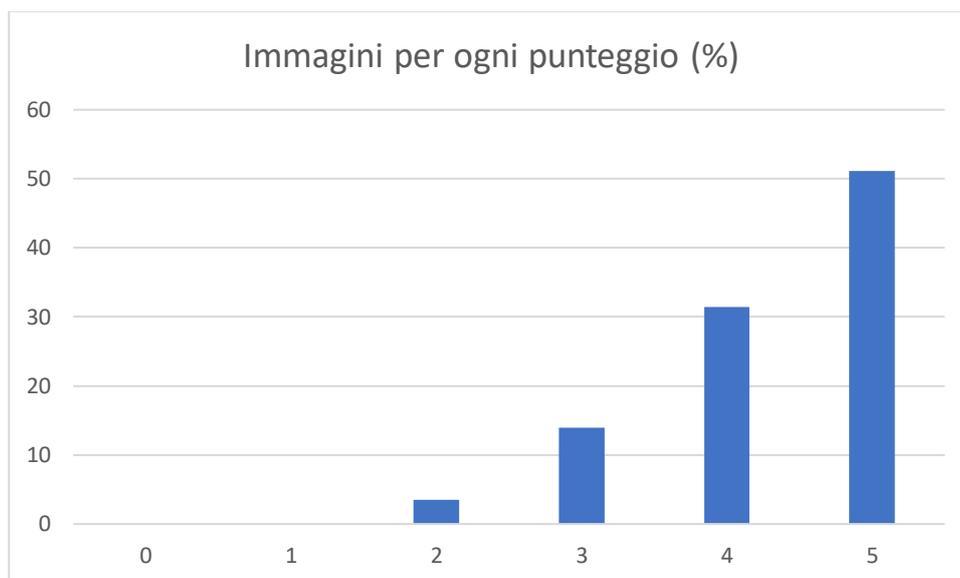


Figura 5.2: Valutazione della segmentazione da parte del patologo. Si esprime la percentuale di immagini a cui sono stati attribuiti i diversi punteggi

Le percentuali dei TIL estratte automaticamente dall'algoritmo sono state poi confrontate con quelle fornite dal patologo. Si nota come ci sia una certa concordanza per le immagini con bassa percentuale di TIL, mentre ci si discosta sempre più all'aumentare della percentuale assegnata.

Nei casi in cui il patologo ha assegnato la percentuale massima pari a 90%, l'algoritmo automatico ha fornito un valore compreso tra il 35 ed il 40%. La percentuale minima assegnata dal patologo è stata pari al 5%, corrispondente ad un valore dell'algoritmo automatico compreso tra il 3 ed il 5%. Le restanti percentuali assegnate coprono tutte le sottoclassi possibili proposte dalle tabelle raccomandate dal TIL Working Group per la valutazione dei TIL.

Interpolando i dati con un polinomio di secondo grado si può ottenere una corrispondenza tra le valutazioni manuali e le valutazioni automatiche; si nota come sia possibile ricavare una regola per il comportamento dei due set di dati.

Il polinomio interpolante, di cui si riporta l'equazione di seguito, presenta un valore R^2 pari a 0.88.

$$y = 0,0251x^2 + 1,3929x \quad (5.1)$$

In questo modo nel caso in cui si volesse fornire una valutazione più vicina a quella eseguita dal patologo, si potrebbe utilizzare questa curva per risalire alla percentuale manuale.

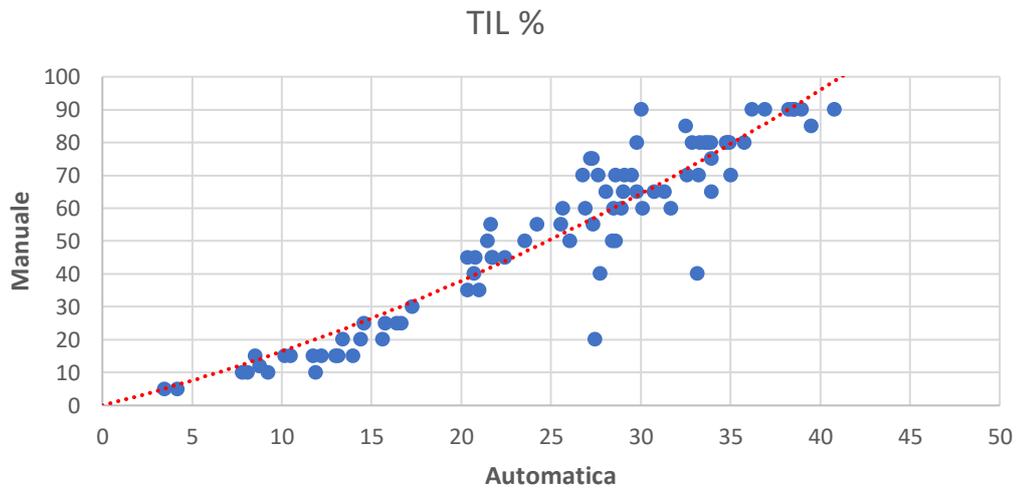


Figura 5.3: Grafico delle percentuali manuali rispetto alle percentuali automatiche con polinomio interpolante

Capitolo 6

Discussione

I risultati mostrano come nonostante l'accuratezza della rete non sia molto alta, la segmentazione finale ottenuta tramite gli step successivi dell'algoritmo proposto, riesca a raggiungere risultati soddisfacenti per lo scopo preposto. La segmentazione è stata eseguita considerando tutte le cellule presenti nell'immagine; infatti non avendo un dataset annotato per la classificazione dei TIL, si è preferito arrivare allo step finale con tutta la popolazione cellulare individuata negli step precedenti per poi applicare un criterio finale per poterli discriminare. La qualità della segmentazione valutata dal patologo ha fornito risultati soddisfacenti, ciò indica che l'algoritmo individua e si adatta bene alle cellule utili al conteggio dei TIL.

Le percentuali estratte dall'algoritmo automatico si discostano da quelle fornite dal patologo per alti valori, mentre per bassi valori le due valutazioni sono paragonabili. Considerando che la percentuale è estratta tramite un semplice rapporto tra l'area dei TIL individuati e l'area dello stroma e che le segmentazioni dei TIL sono state valutate positivamente, ne deriva che il discostamento tra le due misure è dovuto alla percezione visiva, essendo la valutazione manuale effettuata tramite confronto visivo sulle tabelle; il fatto che ci siano scostamenti maggiori per livelli più alti delle percentuali calcolate, è dovuto anche al fatto che le soglie consigliate dalle tabelle di riferimento non hanno una corrispondenza lineare con la densità vera delle cellule.

Considerato che i pattern suggeriti per l'assegnazione delle percentuali non riflettono la vera distribuzione dei TIL sullo stroma, è stata eseguita un'analisi di correlazione tra la percentuale automatica e quella manuale visiva fornita dal patologo. Si è notato che esiste una regola che correla le due valutazioni; in particolare per piccoli valori si raggiungono valori molto simili mentre per percentuali elevate si discostano di un fattore circa pari a 2. L'algoritmo considera come soglia limite una percentuale pari al 40% (dovuta al fatto che tra i TIL ci sarà sempre dello stroma), mentre nel caso delle linee guida la soglia massima arriva al 100%. Il fatto che ci sia una correlazione tra i dati, considerando che il patologo ha valutato positivamente le segmentazioni dei TIL, indica che ci sia comunque una buona concordanza tra le percentuali automatiche e manuali.

I dati sono stati interpolati tramite un polinomio di secondo grado dimostrando che l'algoritmo nonostante si discosti dalle percentuali manuali, segue una regola. Utilizzando la curva, si può infatti risalire ad un valore più simile a quello fornito dal patologo.

Capitolo 7

Conclusioni

La seguente tesi propone un algoritmo automatico per la segmentazione e la valutazione percentuale dei linfociti infiltranti il tumore in immagini istopatologiche di carcinoma mammario. L'algoritmo fornisce una percentuale quantitativa che possa essere usata dal patologo come supporto per le proprie valutazioni. Considerata la crescente rilevanza assunta da questo tipo di cellule come marker prognostico e predittivo nella valutazione del carcinoma mammario, si è reso necessario fornire un metodo accurato e ripetibile. Gli sforzi di standardizzazione operati dai gruppi di ricerca per fornire una regola comune non sono sufficienti a risolvere il problema della variabilità fra gli operatori. Nonostante le raccomandazioni fornite siano esaustive, la valutazione finale della percentuale dei TIL si basa in ogni caso sulla percezione visiva del patologo, inevitabilmente influenzata dalla sua individuale esperienza. Fornire al patologo uno strumento di supporto che calcoli delle percentuali ripetibili, può essere utile nella sua valutazione finale.

Si è dimostrato che le segmentazioni tramite le quali l'algoritmo calcola le percentuali hanno raggiunto prestazioni soddisfacenti, ottenendo un punteggio medio pari a 4.3. Le percentuali calcolate, confrontate con quelle fornite dal patologo, hanno dimostrato di avere una relazione con queste ultime. Tramite l'interpolazione dei dati, inoltre è possibile risalire a percentuali più vicine a quelle fornite dal medico, in modo da rendere più facilmente interpretabile il risultato.

Il tempo impiegato per il processing di un'immagine varia a seconda del numero di cellule presenti, e assume un valore medio pari a 4 ore. Un costo computazionale così elevato si è reso necessario per ottenere dei risultati il più precisi possibile, non avendo infatti a disposizione un dataset annotato, si è preferito concentrarsi sulla qualità dei risultati. Avere a disposizione un dataset maggiore anche per l'addestramento della rete, potrebbe portare a risultati migliori, attribuendo alla rete una maggior capacità di generalizzazione.

Uno sviluppo futuro potrebbe vedere un'estensione dell'algoritmo all'intera biopsia, in modo tale da ottenere un'analisi più approfondita dell'intero stroma peritumorale, estraendo informazioni anche riguardo il loro arrangiamento spaziale.

Inoltre, considerato che la percentuale massima calcolata dall'algoritmo è pari al 40%, sarebbe di interesse fissare tale valore come soglia massima ed esplorare meglio le sottoclassi, utilizzando delle valutazioni maggiormente corrispondenti alla distribuzione reale delle cellule nel campione.

Bibliografia

- [1] M. De Laurentiis, D. Cianniello, R. Caputo, B. Stanzione, G. Arpino, S. Cinieri, V. Lorusso e S. De Placido, «Treatment of triple negative breast cancer (TNBC): current options and future perspectives,» *Cancer treatment Reviews*, vol. 36, n. Supplement 3, pp. S80-S86, 2010.
- [2] C. K. Anders, V. Abramson, T. Tan e R. Dent, «The Evolution of Triple-Negative Breast Cancer: From Biology to Novel Therapeutics,» *American Society of Clinical Oncology Educational Book*, vol. 36, pp. 34-42, 2018.
- [3] S. A. H. S. R. A. S. C. H. B. H. I. V. D. R. H. S.-A. I. G. S. W. V. H. Peter Schmid, «Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer,» *The New England Journal of Medicine*, vol. 379, n. 22, pp. 2108-2121, 2018.
- [4] G. J. Bates, S. B. Fox, C. Han, R. D. Leek, J. F. Garcia, A. L. Harris e A. H. Bonham, «Quantification of Regulatory T Cells Enables the Identification of High-Risk Breast Cancer Patients and Those at Risk of Late Relapse,» *Journal of Clinical Oncology*, vol. 24, n. 34, pp. 5373-5380, 2006.
- [5] S. Loi, N. Sirtaine, F. Piette, R. Salgado, G. Viale, F. Van Eenoo, G. Rouas, P. Francis, J. P. Crown, E. Hitre, E. de Azambuja, E. Quinaux, A. Di Leo, S. Michiels, M. J. Piccart e C. Sotiriou, «Prognostic and Predictive Value of Tumor-Infiltrating Lymphocytes in a Phase III Randomized Adjuvant Breast Cancer Trial in Node-Positive Breast Cancer Comparing the Addition of Docetaxel to Doxorubicin With Doxorubicin-Based Chemotherapy: BIG 02-98,» *Journal of Clinical Oncology*, vol. 31, n. 7, pp. 860-867, 2013.
- [6] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. Baehner, F. Penault-Llorca, E. A. Perez, E. A. Thompson, W. F. Symmans, A. Richardson, J. Brock, C. Criscitiello, H. Bailey, M. Ignatiadis, G. Floris, J. Sparano, Z. Kos, T. Nielsen, D. L. Rimm, K. H. Allison, J. Reis-Filho, S. Loibl, C. Sotiriou, G. Viale, S. Badve, S. Adams, K. Willard-Gallo e S. Loi, «The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014,» *Annals of Oncology*, vol. 26, pp. 259-271, 2015.

- [7] M. V. Dieci, N. Radosevic-Robin, S. Fineberg, G. van den Eyndeng, N. Ternes, F. Penault-Llorcac, G. Pruneri, T. M. D'Alfonso, S. Demaria, C. Castaneda, J. Sanchez, S. Badve, S. Michiels, V. Bossuyt, F. Rojo, B. Singh, T. Nielsen, G. Viale, S.-R. Kim, S. Hewitt, S. Wienert, S. Loibl, D. Rimm, F. Symmans, C. Denkert, S. Adams, S. Loi e R. Salgado, «Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: A report of the International ImmunoOncology Biomarker Working Group on Breas,» *Seminars in Cancer Biology*, vol. 52, pp. 16-25, 2018.
- [8] aimac. [Online]. Available: <https://www.aimac.it/schede-tumori/tumore-seno>.
- [9] [Online]. Available: <https://www.infoseno.org/il-tumore-del-seno/diagnosi/tipi-di-tumore>.
- [10] F. J. S. I. S. R. T. L. J. A. Bray F, «Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,» *CA Cancer J Clin.*, vol. 68, n. 6, pp. 394-424, Nov 2018.
- [11] AIOM, AIRTUM e Fondazione AIOM, *I numeri del cancro in Italia 2019*, 2019.
- [12] A. I. d. O. M. (AIOM), «Linee Guida Neoplasie della Mammella. Edizione 2019,» 2019. [Online]. Available: https://www.aiom.it/wp-content/uploads/2019/10/2019_LG_AIOM_Mammella.pdf.
- [13] AIOM, AIRTUM e Fondazione AIOM, *I numeri del cancro in Italia 2018*, 2018.
- [14] M. Airolid, C. Angiolini, L. Annaratone e e. al., 2018: il carcinoma della mammella, Roma: Il Pensiero Scientifico Editore, 2018, pp. 43-46.
- [15] W. World Health Organization, 2019. [Online]. Available: tumourclassification.iarc.who.int.
- [16] T. Sørlie, C. M. Perou e R. e. a. Tibshirani, «Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,» *Proceedings of the National Academy of Sciences of the United of America*, vol. 98, n. 19, pp. 10869-10874, 2001.
- [17] T. Sørlie, R. Tibshirani e J. e. a. Parker, «Repeated observation of breast tumor subtypes in independent gene expression data sets.,» *Proc Natl Acad Sci U S A*, vol. 100, n. 14, pp. 8418-8423, 2003.

- [18] E. S. G. F. e. a. Amin M B, AJCC (American Joint Committee on Cancer) Cancer Staging Manual; 8th edition, Chicago: Springer, 2018.
- [19] A. E. Giuliano, J. L. Connolly, S. B. Edge e et al., «Breast Cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual,» *CA: A Cancer Journal for Clinicians*, vol. 67, n. 4, pp. 290-303, 2017.
- [20] D. Mittal, M. Gubin, R. D. Schreiber e M. J. Smyth, «New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape,» *Current opinion in immunology*, vol. 27, pp. 16-25, 2014.
- [21] L. A. T. a. G. K. Zitvogel, «Cancer despite Immunosurveillance: Immunoselection and Immunosubversion,» *Nature Reviews Immunology* , vol. 6, n. 10, p. 715–27, 2006.
- [22] W. H. Fridman , J. Galon e F. Pagès, «Prognostic and Predictive Impact of Intra- and Peritumoral Immune Infiltrates,» *Cancer Research*, vol. 71, n. 17, pp. 5601-5, 2011.
- [23] S. P. L. M. E. V. M. K. S. M. E. A. a. K. S. Aaltomaa, «Lymphocyte Infiltrates as a Prognostic Variable in Female Breast Cancer,» *European Journal of Cancer*, vol. 28, n. 4-5, pp. 859-864, 1992.
- [24] C. Denkert, S. Wienert, A. Poterie e e. al., «Standardized Evaluation of Tumor-Infiltrating Lymphocytes in Breast Cancer: Results of the Ring Studies of the International Immuno-Oncology Biomarker Working Group,» *Modern Pathology*, vol. 29, n. 10, 2016.
- [25] [Online]. Available: <https://www.tilsinbreastcancer.org/what-are-tils/>.
- [26] S. Hendry, R. Salgado, T. Gavaert e et al., «Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group,» *Advances in anatomic pathology*, vol. 24, n. 5, pp. 235-251, 2017 .
- [27] G. F. H. M. K. C. T. A. Petushi S, «Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer,» *BMC Med Imaging*, vol. 6, n. 14, 2006.
- [28] P. J. v. D. R. K. A. H. M. A. V. J. P. W. P. Mitko Veta, «Automatic Nuclei Segmentation in H&E Stained Breast Automatic Nuclei Segmentation in H&E Stained Breast,» *PLoS ONE*, vol. 8, n. 7, 2013.

- [29] A. N. Basavanhally, S. Ganesan, S. Agner e et al., «Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology,» *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 57, n. 3, 2010.
- [30] B. S. B. G. R. V. J. R. S. O. G. M. Dundar MM, «Computerized Classification of Intraductal Breast Lesions Using Histopathological Images,» *IEEE Trans Biomed Eng.*, vol. 58, n. 7, pp. 1977-84, 2011.
- [31] X. I. H. Y. m. L. X. Z. Pin Wang, «Automatic cell nuclei segmentation and classification of breast cancer histopathology images,» *Signal Processing*, vol. 122, pp. 1-13, 2016.
- [32] H. Fatakdawala, J. Xu, A. Basavanhally e e. al., «Expectation–Maximization-Driven Geodesic Active Contour With Overlap Resolution (EMaGACOR): Application to Lymphocyte Segmentation on Breast Cancer Histopathology,» in *Ninth IEEE International Conference on Bioinformatics and Bioengineering*, Taichung, Taiwan, 2009.
- [33] P. S. D. A. C. P. R. P. D. G. C. N. Dimitris Glotsos, «Automated segmentation of routinely hematoxylin-eosin-stained microscopic images by combining support vector machine clustering and active contour models,» *Analytical and quantitative cytology and histology / the International Academy of Cytology [and] American Society of Cytology*, vol. 26, n. 6, pp. 331-40, 2004.
- [34] J. H. A. B. L. L. J. W. G. P. T. S. a. B. P. H. Chang, «“Invariant Delineation of Nuclear Architecture in Glioblastoma Multiforme for Clinical and Molecular Association,» *IEEE Transactions on Medical Imaging*, vol. 32, n. 4, pp. 670-682, 2013.
- [35] W. CW, «Robust Automated Tumour Segmentation on Histological and Immunohistochemical Tissue Images,» *PLOS ONE* , vol. 6, n. 4, 2011.
- [36] S. E. A. R. Y. W. T. D. R. J. S. I. A. C. a. N. M. R. K. Sirinukunwattana, «Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images,» *IEEE Transactions on Medical Imaging*, vol. 35, n. 5, pp. 1196-1206, 2016.
- [37] X. K. F. X. F. L. H. S. a. L. Y. Y. Xie, «Deep Voting: A Robust Approach Toward Nucleus Localization in Microscopy Images,» in *Medical Image Computing and*

Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, 2015.

- [38] J. C. J. C. a. Z. Q. T. Wan, «Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features,» *Neurocomputing*, vol. 229, pp. 34-44, 2017.
- [39] L. L. H. Y. Z. L. J. Y. L. Z. a. Y. F. X. Pan, «Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks,» *Neurocomputing*, vol. 229, pp. 88-99, 2017.
- [40] Y. X. a. L. Y. F. Xing, «An Automatic Learning-Based Framework for Robust Nucleus Segmentation,» *IEEE Transactions on Medical Imaging*, vol. 35, n. 2, pp. 550-566, 2016.
- [41] R. V. S. S. S. B. A. V. a. A. S. N. Kumar, «A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology,» *IEEE Transactions on Medical Imaging*, vol. 36, n. 7, pp. 1550-1560, Luglio 2017.
- [42] B. P. M. Khoshdeli, «Deep Learning Models Delineates Multiple Nuclear Phenotypes in H&E Stained Histology Sections,» 2018.
- [43] S. Doyle, S. Agner, A. Madabhushi, M. Feldman e J. Tomaszewski, «Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features,» in *2008 5th IEEE international symposium on biomedical imaging: From nano to macro*, 2008.
- [44] F. R. M. L. T. W. P. Naylor, «A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology,» *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1-1, 21 Agosto 2018.
- [45] G. B. P. M. Khoshdeli, «Fusion of encoder-decoder deep networks improves delineation of multiple nuclear phenotypes,» *BMC Bioinformaticx*, vol. 19, n. 1, pp. 1-11, 2018.
- [46] P. Naylor, M. Laé, F. Reyat e T. Walter, «Nuclei segmentation in histopathology images using deep neural networks,» in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, Melbourne, VIC, Australia , 2018.
- [47] «Convolutional Neural Networks,» [Online]. Available: <http://cs231n.github.io/convolutional-networks/>.

- [48] «Convolutional Neural Networks,» [Online]. Available: <https://medium.com/@udemudofia01/basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17>.
- [49] R. V. S. S. B. A. V. a. A. S. Neeraj Kumar, «A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology,» *IEEE TRANSACTIONS ON MEDICAL IMAGING*, vol. 36, n. 7, July 2017.
- [50] MathWorks, «MathWorks Documentation,» [Online]. Available: <https://it.mathworks.com/help/images/ref/bwconncomp.html>.
- [51] M. Macenko, M. Niethammer e N. E. Thomas, «A method for normalizing histology slides for quantitative analysis,» *International Symposium on Biomedical Imaging: from Nano to Macro*, 2009.
- [52] M. Kuse, T. Sharma e S. Gupta, «A Classification Scheme for Lymphocyte Segmentation in H&E Stained Histology Images,» in *Patterns in Signals, Speech, Images and Videos. ICPR 2010. Lecture Notes in Computer Science, vol 6388. Springer, Berlin, Heidelberg, 2010*.