# POLITECNICO DI TORINO

**Corso di Laurea Magistrale in Ingegneria Biomedica**


Politecnico
di Torino

**Tesi di Laurea Magistrale**

# Comparazione tramite clustering di estrattori di caratteristiche da tiles istopatologiche tumorali del colon-retto, aggregate con Bag of Visual Words.

**Relatori**

**Prof. Valentina GIANNINI**

**Prof. Samanta ROSATI**

**Candidato**

**Antonino BELLINA**

**Dicembre 2024**

# Sommario

Il cancro del colon-retto (CRC) è il terzo tipo di cancro più diffuso negli esseri umani e la terza causa più comune di morte per cancro sia negli uomini che nelle donne, contribuendo a un significativo problema di salute pubblica a livello mondiale. L'obiettivo del lavoro è analizzare e tentare di estrarre biomarcatori significativi da immagini istopatologiche ottenute da campioni di tumore colonrettale estratti chirurgicamente da due gruppi di soggetti: il gruppo IANG, di età inferiore a 40 anni (9 pazienti), e il gruppo OLD, di età superiore ai 40 anni (8 pazienti).

Tali immagini ad alta risoluzione, di dimensioni ($83456 \times 185600$) pixels, sono state colorate con ematossilina ed eosina (HE) e acquisite con uno scanner Mirax dall'Ospedale Niguarda di Milano. Questa tesi, dunque, trae ispirazione dall'esigenza di estrarre biomarcatori morfologici utili alla medicina di precisione per organizzare un trattamento adeguato e mirato per i pazienti, cercando di evitare di sottoporli a sofferenze e costi inutili.

Per tentare di raggiungere questo obiettivo, è stata condotta un'analisi esplorativa sul raggruppamento delle immagini dei pazienti, basata sull'estrazione di caratteristiche tramite diversi estrattori applicati al tessuto tumorale suddiviso in tasselli (tiles). Le caratteristiche estratte sono state successivamente ridotte in dimensione utilizzando due tecniche di riduzione dimensionale: l'Analisi delle Componenti Principali (PCA) e la Correlazione di Pearson, impiegate anche per combinare i diversi estrattori.

Il raggruppamento delle tiles tumorali è stato effettuato con la tecnica denominata Bag of Visual Words (BoVW), usando una rete SOM per assegnare i fenotipi tumorali, al fine di esplorarli in modo da aggregare i pazienti ed identificare nuovi potenziali biomarcatori. Le motivazioni di tale approccio sono da ricondurre al fatto che alcuni soggetti del gruppo IANG mostrano una risposta alla terapia simile a quella del gruppo OLD.

Tali tipologie di analisi rappresentano una sfida significativa per la medicina

di precisione, per vari motivi, tra cui: le immagini istopatologiche HE sono uno strumento più rapido e meno costoso rispetto all'analisi immunoistochimica (IHC) e alla PCR; tuttavia, l'interpretazione e la preparazione di queste immagini presentano un'intrinseca variabilità che può introdurre bias diagnostici. Inoltre, il numero di anatomopatologi esperti, responsabili della preparazione e analisi manuale dei campioni, è in declino da diversi anni.

Le immagini istopatologiche rappresentano anche una sfida per le tecniche di computer vision e image processing, soprattutto per via dell'enorme risoluzione, che comporta problemi di memoria e RAM durante il processamento.

I risultati dei clustering dei fenotipi estratti, sono stati valutati tramite la metrica della silhouette, che ha permesso di analizzare la coerenza e la separazione dei gruppi generati. Tali risultati sono stati presentati a un oncologo per confrontare i raggruppamenti generati con le caratteristiche cliniche dei pazienti, al fine di individuare potenziali nuovi biomarcatori di interesse clinico.

# Ringraziamenti

Desidero esprimere la mia profonda gratitudine a tutte le persone che mi hanno sostenuto e accompagnato in questo percorso.

Un ringraziamento speciale va a mia madre, che mi ha supportato fin dai primi passi nel mondo degli studi e ha creduto in me e nelle mie capacità, anche nei momenti di maggiore difficoltà. La sua fiducia e il suo incoraggiamento sono stati fondamentali per il raggiungimento di questo traguardo.

A mia sorella Irene, che con il suo amore e il suo affetto mi ha sempre dato la spinta giusta per andare avanti e non mollare, anche quando il percorso sembrava insormontabile.

Al mio amico Paolo, un vero fratello per me, che con la sua presenza sincera e il suo supporto mi ha aiutato ad affrontare tutte le difficoltà, università compresa. Insieme, abbiamo sempre trovato la forza per non demoralizzarci, ricordandoci che oltre alle qualità accademiche, possediamo valori e capacità personali che ci rendono unici, soprattutto nella perseveranza.

A mio padre, che con il suo motto "volli, volli, fortissimamente volli" mi ha insegnato l'importanza della determinazione. Con il suo spirito mi ha accompagnato nel superare le avversità incontrate durante questo percorso.

Infine, alla mia fidanzata Sara, che con lil suo amore e la fiducia in me e nei miei mezzi mi ha dato la forza e la motivazione per portare a termine questo viaggio.

A tutti voi, va il mio più sincero grazie. Questo traguardo è anche vostro.

# Indice

**Elenco delle tabelle** VIII

**Elenco delle figure** IX

**Acronyms** XII

**1 Introduzione** 1
- 1.1 Tumore Colorettale . . . 1
- 1.2 Stadi del tumore al colon retto . . . 5
- 1.3 Eziologia e Instabilità del microsatellite (MSI) . . . 6
- 1.4 Fattori di rischio del tumore colonrettale metastatico. . . . 10
- 1.5 Immagini istopatologiche . . . 11
- 1.6 L'intelligenza artificiale come prossimo passo verso la patologia di precisione . . . 11

**2 Stato dell'arte** 13
- 2.1 Problematiche tecnologiche della Digital Pathology . . . 13
- 2.2 Obiettivi della Digital Pathology . . . 14
- 2.3 Predizione della risposta alla terapia . . . 15

**3 Materiali e metodi** 17
- 3.1 Dataset . . . 17
- 3.2 Pipeline di Analisi . . . 18
  - 3.2.1 Preprocessing e Normalizzazione del Colore . . . 18
  - 3.2.2 Segmentazione delle Tiles Tumorali . . . 21
  - 3.2.3 Estrazione delle Caratteristiche . . . 22
  - 3.2.4 Aggregazione delle Caratteristiche . . . 29
  - 3.2.5 Analisi delle Caratteristiche Estratte . . . 30

**4 Conclusioni** 38
- 4.1 Prospettive Future . . . 38

**Bibliografia** 40

# Elenco delle tabelle

3.1 Risultati della silhouette media per diverse estrazioni e metodi di selezione delle caratteristiche . . . . . 34

# Elenco delle figure


1.1 Tassi di incidenza del tumore colorettale nella popolazione di età inferiore a 50 anni dal 2010 al 2030 [7] . . . . . 3
1.2 Mappa delle regioni in cui è documentato l'aumento di incidenza del tumore colonrettale in età giovanile inferiore ai 50 anni di età [7] . . 4
1.3 Schematizzazione degli stadi del tumore clonrettale . . . . . 6
1.4 Pipeline del processo di campionamento dei tessuti e digitalizzazione in patologia. Dopo che è stata effettuata la biopsia dal paziente (a), viene creato un blocco di tessuto, preceduto dalla fissazione e dall'inclusione in paraffina (b). Dopo il taglio del blocco di tessuto (c), la sezione viene posizionata su un vetrino seguito da una colorazione speciale (d). Successivamente, il vetrino colorato viene inserito in uno scanner specifico per vetrini (e), ottenendo una diapositiva digitale del tessuto (f)[17]. . . . . 12

3.1 Pipeline di analisi per la predizione della risposta alla terapia. I passaggi includono la normalizzazione del colore, la segmentazione delle tiles tumorali, l'estrazione delle caratteristiche, l'aggregazione delle caratteristiche e la classificazione finale. . . . . 19
3.2 Spazio RGB (sinistra) e spazio della densità ottica (destra) . . . . . 20
3.3 Esempio di classificazione dei tessuti di un appartenente al gruppo IANG . . . . . 21
3.4 Esempio di classificazione dei tessuti di un appartenente al gruppo OLD . . . . . 21
3.5 Illustrazione della costruzione del Codebook per il *Bag of Visual Words (BoVW)*. Le caratteristiche locali estratte dalle immagini di allenamento vengono raggruppate in cluster, ciascuno rappresentante una parola visiva (*Visual Word*). Il dizionario visuale così costruito permette di trasformare le caratteristiche locali di una nuova immagine in un istogramma di frequenze delle parole visuali, utilizzabile per l'analisi e il clustering dei pazienti. . . . . 30
3.6 Matrice di correlazione per le caratteristiche estratte da ResNet18 . 31

3.7 Matrice di correlazione per le caratteristiche estratte da CTransPath 32
3.8 Matrice di correlazione per le caratteristiche combinate degli estrattori prime 23 righe GLCM dalla 24-ma alla 536-ma ResNet18 dalla 527-ma all ultima riga Ctranspath . . . . . 32
3.9 Varianza spiegata dalle caratteristiche estratte da ResNet-18 . . . . 33
3.10 Varianza spiegata dalle caratteristiche estratte da CTranspath . . . 33
3.11 valori di silhouette ottenuti con i vari metodi da 1 a 8 IANG e da 9 a 16 OLD . . . . . 35
3.12 valori di silhouette ottenuti con i vari metodi divisione nei due cluster principali . . . . . 35
3.13 Miglior raggruppamento ResNet18 Corr 0.8 . . . . . 36
3.14 Fenotipi discriminanti evidenziati dall aggregazione delle caratteristiche estratte con ResNet18. . . . . 37

# Acronyms

**AI**
artificial intelligence

**DP**
Digital Pathology

**nCRT**
Chemioterapia Neoadiuvante

**CRC**
Colorectal Cancer

**MSI**
Microsatellite Instability

**EO-CRC**
Tumore Colorettale a insorgenza precoce

**CIN**
Instabilità Cromosomica

**CIMP**
fenotipo del metilatore delle isole CpG

**MMR**
Mismatch Repair

**LS**
Lynch Syndrome

**MACS**

Microsattelite and Chromosome Stable

**STRs**

Short Tandem Repeats

**IHC**

Immunoistochimica

**PCR**

Polymerase Chain Reaction

**GLCM**

Gray-Level Co-Occurence Matrix

**WSI**

Whole Slide Image

# Capitolo 1

# Introduzione

In questo capitolo si discuterà del tumore colorettale focalizzandone in particolare l'emergente incidenza nella popolazione giovanile. Inoltre si vuole concentrare l'attenzione sull'eziologia genetica ed epigenetica, oltre che sui fattori di rischio associati a questa malattia nella popolazione giovanile.

## 1.1 Tumore Colorettale

Il cancro colorettale ha la terza più alta morbilità e la seconda più alta mortalità a livello mondiale [1, 2]. Circa un terzo dei tumori è localizzato nel retto e circa il 70% del cancro rettale è cancro rettale localmente avanzato. Il trattamento standard per il cancro rettale localmente avanzato prevede la chemioradioterapia neoadiuvante (nCRT) seguita da chirurgia. Riducendo la stadiazione e la dimensione del tumore, la nCRT aumenta la probabilità di un successivo intervento chirurgico R0 di successo e di un intervento chirurgico di preservazione dello sfintere, e diminuisce la probabilità di recidiva locale. Tuttavia, la nCRT può indebolire il sistema immunitario e causare un intervento chirurgico ritardato per i pazienti che non ne possono trarre beneficio. Pertanto, è necessario identificare biomarcatori per la risposta al trattamento con nCRT per il cancro rettale localmente avanzato, e individuare i pazienti che non ne trarranno beneficio per migliorare la strategia di trattamento e ridurre il dolore e i costi inutili[3].

Al momento della diagnosi, l'età mediana dei pazienti con cancro al colon è di 68 anni per gli uomini e 72 anni per le donne; l'età mediana dei pazienti con cancro al retto è di 63 anni per entrambi i sessi.

Negli ultimi anni, l'incidenza complessiva e la mortalità del CRC negli USA e in Europa è diminuita. Dalla metà degli anni 2000, l'incidenza del CRC negli USA è diminuita del 2-3% all'anno sia negli uomini che nelle donne. Questa riduzione è stata principalmente correlata alla diffusione dei test di screening [test

del sangue occulto nelle feci (FOBT) e colonscopia] che permettono la rilevazione e l'asportazione delle lesioni precancerose, e a una maggiore consapevolezza dei fattori di rischio del CRC tra la popolazione [4]. Dal 2012 in Europa, il tasso di mortalità per CRC è diminuito del 6,7% negli uomini e del 7,5% nelle donne, mentre tra il 2008 e il 2016 l'incidenza del CRC è aumentata del 6% all'anno [5]. La sopravvivenza complessiva (OS) a 5 anni dalla diagnosi è intorno al 60% considerando tutti gli stadi della malattia. Il CRC metastatico, nonostante i progressi terapeutici, mostra una prognosi sfavorevole; allo stato attuale delle conoscenze, solo il 14% dei pazienti è ancora vivo a 5 anni dalla diagnosi [4, 6].

### Fattori di rischio

La dieta occidentale non mediterranea, l'obesità, la scarsa attività fisica, l'alto consumo di carne rossa e lavorata e il basso apporto di fibre sono i fattori di rischio più rilevanti per lo sviluppo del CRC. Negli Stati Uniti si è recentemente registrato un aumento dell'incidenza dell'obesità, soprattutto tra i giovani pazienti, e, essendo un noto fattore di rischio per il CRC, ciò potrebbe aver contribuito a ridurre l'età di insorgenza del CRC. Il contributo netto di questi fattori non può essere valutato.

Ci sono fattori di rischio per il CRC ben noti che giocano un ruolo più importante tra i giovani individui:

- Le malattie infiammatorie intestinali aumentano di due o tre volte il rischio di CRC rispetto alla popolazione generale, soprattutto quando diagnosticate in giovane età.
- Le sindromi ereditarie predisponenti al cancro o il CRC familiare sono più frequenti tra gli EO-CRC.
- La bassa adesione ai programmi di screening specifici negli individui con sindromi tumorali note o CRC familiare è anche un punto cruciale nei paesi con un sistema sanitario privato o tra le popolazioni con un basso livello socioeconomico.
- Irradiazione addominale precedente (ad esempio, radioterapia per malignità pediatriche curabili): la colonscopia è raccomandata a partire dai 35 anni di età o un decennio dopo un trattamento con radiazioni $> 30$ Gy al bacino.

### Tumore colonrettale nella popolazone giovanile

Nel 2010, il CRC tra i pazienti con meno di 50 anni rappresentava il 4,8% e il 9,5% dei tumori del colon e del retto, rispettivamente. Studi recenti su diversi continenti hanno riportato un aumento dell'incidenza del CRC in questa fascia di età, in particolare tra gli individui con meno di 40 anni.

Secondo il database Surveillance, Epidemiology and End Results Program (SEER) negli Stati Uniti, circa il 5% di tutti i casi di carcinoma colorettale (CRC) viene diagnosticato in pazienti con meno di 45 anni. Il carcinoma del retto viene diagnosticato fino al 18% dei casi in pazienti sotto i 50 anni, sia negli uomini che nelle donne. Il CRC a insorgenza precoce (EO-CRC) è più comunemente diagnosticato tra le minoranze e nelle popolazioni prive di assicurazione sanitaria.

All'inizio del XXI secolo, diversi studi hanno mostrato evidenze di un cambiamento nell'incidenza del CRC in diversi gruppi di età. Negli Stati Uniti, è stato osservato un aumento del numero di diagnosi di CRC nella popolazione sotto i 50 anni, con un incremento particolarmente evidente nei pazienti di età compresa tra 20 e 35 anni. I dati epidemiologici indicano che, dal 1994, l'incidenza di CRC in individui sotto i 55 anni è aumentata di circa il 2% ogni anno.

Uno studio di Bailey et al., basato sui trend di incidenza di CRC tra il 1975 e il 2010 negli Stati Uniti, ha previsto che l'incidenza del CRC nella popolazione sotto i 35 anni sarà quasi raddoppiata entro il 2030, con il maggiore incremento previsto per i tumori del retto-sigma nella fascia di età tra 18 e 35 anni. Entro il 2030 negli Stati Uniti, si prevede che il 10% di tutti i tumori del colon e il 22% di tutti i tumori del retto saranno diagnosticati in pazienti con meno di 50 anni, **Figura 1.1**. Queste previsioni sono significative se confrontate con i dati di 10 anni fa, quando le percentuali erano rispettivamente del 4% e del 9% per i tumori del colon e del retto.

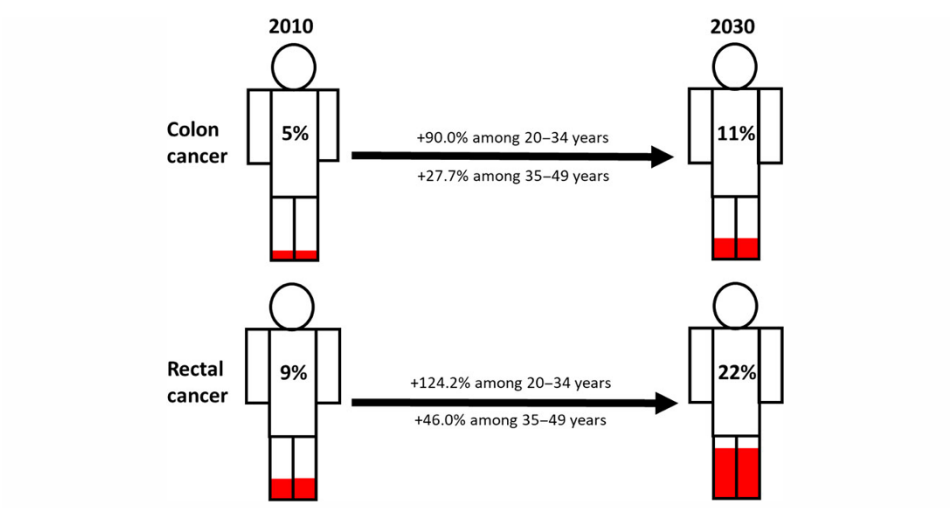


**Figura 1.1:** Tassi di incidenza del tumore colorettale nella popolazione di età inferiore a 50 anni dal 2010 al 2030 [7]

Questi dati indicano che l'EO-CRC rappresenta un problema attuale di sanità pubblica negli Stati Uniti e in altre parti del mondo, **Figura 1.2**[7].

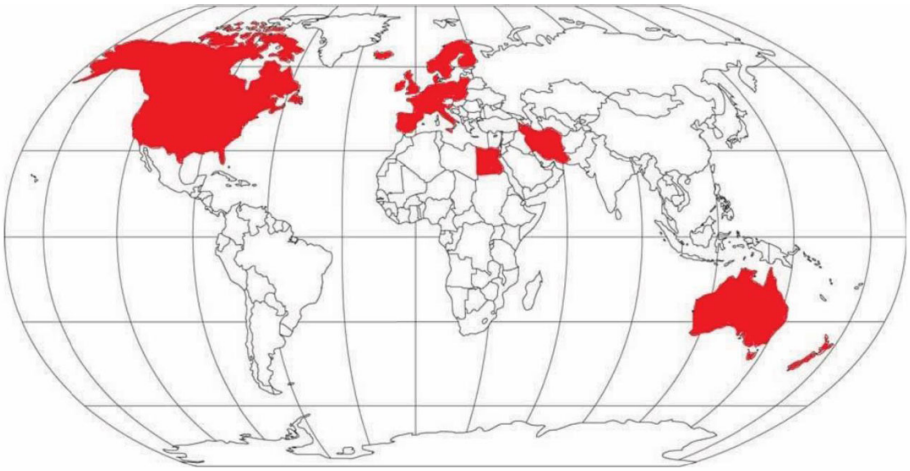

**Figura 1.2:** Mappa delle regioni in cui è documentato l'aumento di incidenza del tumore colonrettale in età giovanile inferiore ai 50 anni di età [7]

### Eziologia del tumore colonrettale nella popolazione giovanile

Nella popolazione generale con carcinoma colorettale (CRC), ci sono tre principali vie di carcinogenesi coinvolte nell'insorgenza e nello sviluppo del CRC: instabilità cromosomica (CIN), instabilità dei microsatelliti (MSI) e il fenotipo metilatore delle isole CpG (CIMP).

Tra i CRC a insorgenza precoce (EO-CRC), così come nella popolazione generale, la CIN è la causa più comune di CRC.In una percentuale variabile di casi, a seconda dello stadio della malattia, il CRC è causato da MSI. Esistono due possibili meccanismi che portano a CRC associato a MSI:

- mutazioni ereditarie germinali nei geni di riparazione dei mismatch (MMR);
- ipermetilazione somatica tumorale del gene MLH1.

In tutti i pazienti con MSI-CRC è raccomandato il rinvio alla consulenza genetica per lo screening della sindrome di Lynch (LS). La sindrome di Lynch predispone all'EO-CRC, e si è osservato che la percentuale di tumori MSI può arrivare fino al 27%, specialmente nei pazienti con meno di 30 anni.

Una piccola percentuale di MSI-CRC può presentare anche CIN, mentre circa il 50% dei CRC MSS non presenta CIN. Questo sottogruppo è stato definito Microsatellite and Chromosome Stable (MACS). Questi tumori sono più frequentemente localizzati a livello rettale o nel colon sinistro, sono caratterizzati da una prognosi sfavorevole e da uno scarso riconoscimento da parte del sistema immunitario.

Una causa importante di EO-CRC è la presenza di una mutazione germinale in un oncogene, che dà origine a una sindrome tumorale ereditaria. La prevalenza delle sindromi ereditarie di CRC tra gli EO-CRC è influenzata dalle diverse fasce di età analizzate negli studi, con una prevalenza maggiore tra i pazienti sotto i 35 anni. Tra gli EO-CRC si possono considerare quattro scenari genetici potenziali:

- sindromi tumorali ereditarie note
- mutazioni germinali de novo di tumori ereditari
- tumori colorettali familiari
- CRC non ereditari e non familiari

## 1.2 Stadi del tumore al colon retto

In questa sezione si vogliono evidenziare le caratteristiche degli stadi del tumore colonrettale.

**stadio 0** è il meno grave, con tutte le lesioni limitate alla mucosa e alla lamina propria. La rimozione locale o la polipectomia semplice con margini chiari è l'opzione di trattamento più comune.

**stadio I** il cancro può essersi sviluppato nella muscolaris mucosae o nella muscolaris propria, ma non si è ancora diffuso più in profondità nella parete muscolare del colon, nei linfonodi vicini o in altri siti distanti. Poiché il CRC in questa fase è ancora localizzato, ha anche un'alta percentuale di guarigione con una resezione chirurgica ampia e anastomosi.

**stadio II** caratterizza il CRC che si è diffuso fino alla sierosa o oltre, e potrebbe essersi sviluppato nei tessuti o organi vicini, ma non nei linfonodi e non ha metastatizzato. La resezione chirurgica è nuovamente il trattamento standard, tuttavia ai pazienti ad alto rischio, come quelli con malattia T4, potrebbe essere offerta la chemioterapia.

**stadio III** è caratterizzato dall'coinvolgimento dei linfonodi e i trattamenti standard sono la resezione chirurgica ampia con anastomosi e la chemioterapia adiuvante.

La malattia allo **stadio IV** è caratterizzata da malattia metastatica.Il trattamento del CRC in questa fase dipende principalmente dai siti di malattia metastatica. Le metastasi al fegato rappresentano circa il 50% dello stadio IV e del CRC ricorrente, e le opzioni di trattamento includono tutte le precedenti, oltre alla radioterapia palliativa, chemioterapia palliativa e terapie mirate. [8]

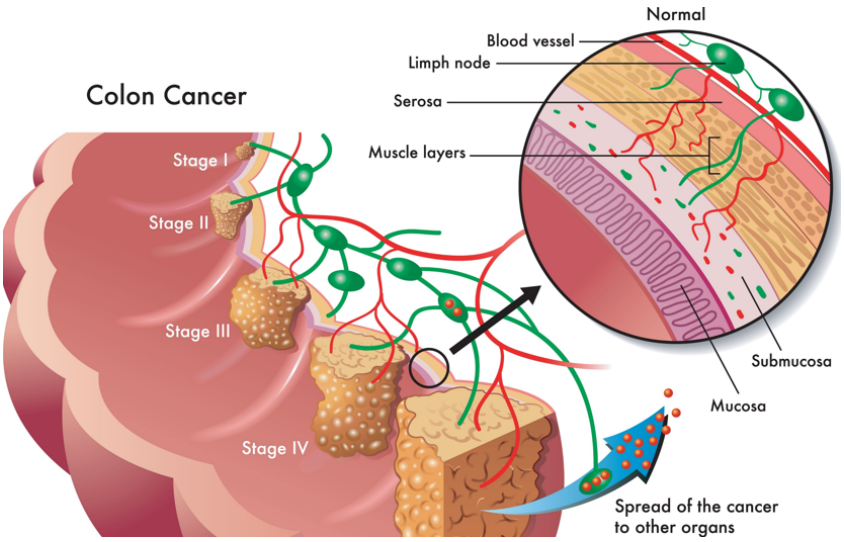


**Figura 1.3:** Schematizzazione degli stadi del tumore clonrettale

## 1.3 Eziologia e Instabilità del microsatellite (MSI)

Il Tumore colnrettale è una malattia eterogenea che è causata dall'interazione tra fattori genetici e ambientali. Le modificazioni molecolari che avvengono nel CRC possono essere divise in 3 categorie:

- **CIN** Instabilità dei cromosomi.
- **MSI** Instabilità del microsatellite.
- **CIMP** CpG island Methylator phenotype.

L'MSI anche conosciuto come Short Tandem Repeats (STRs) sono piccole coppie di basi (1-6) del DNA sparsi in tutto il genoma e rappresentano circa il 3% del genoma dell'essere umano. A causa della loro struttura ripetuta i microsatelliti sono soggetti ad un alto tasso di mutazione. E' un'alterazione unica e un fenotipo ipermutabile che è il risultato di un difetto nel sistema di riparazione del mismatch del DNA e può essere definito come la presenza di sequenze ripetitive di DNA di dimensioni alternate che non sono presenti nel DNA germinale.

La determinazione dello stato di MSI ha implicazioni prognostiche e diagnostiche. I tassi di mutazione nei singoli nucleotidi sono maggiori nei MSI-positivi che non nei CIN-positivi, il 15-20% dei CRC MSI-positivi è causata da un difetto nel sistama di riparazione dei mismatch del DNA.

**Divisione dei tipi di CRC in base all'eziologia e alla genetica della malattia**

- **CRC Sporadico:** E' il più comune e include il 75% dei casi che non mostrano evidenti segni di ereditarietà del disturbo. Tuttavia, non è ancora chiaro quali mutazioni specifiche influenzano la probabilità di sviluppare il CRC.

  Il CRC sporadico è comune in età avanzata probabilmente a causa di fattori ambientali, alimentazione e invecchiamento. I CRC sporadici MSI-H sono spesso (75-90%) causati da alterazioni del gene MLH1 tramite ipermetilazione del promotore somatico.

- **CRC familiare:** Questo è considerato sporadico e non è stato trovato alcun gene associato. Le persone con una storia di tumore colonrettale in un parente di primo grado hanno un rischio aumentato da due a tre volte rispetto alla poplazione normale di contrarre CRC.

- **Poliposi adenomatosa familiare** (tipo ereditario): È la sindrome di poliposi più comune, causata dalla mutazione del gene APC, una proteina multifunzionale che controlla la crescita cellulare e previene lo sviluppo di tumori. L'APC è coinvolta nella regolazione della $\beta$-catenina, una proteina della via di segnalazione Wnt. Mutazioni nel gene APC portano ad un accumulo di $\beta$-catenina; questa mutazione è associata ad altre mutazioni come KRAS, DCC, P53, COX-2, etc., necessarie per lo sviluppo del cancro. Questi pazienti presentano polipi nella parte prossimale del colon e raramente nel retto. L'età media in cui questi pazienti FAP sviluppano polipi è 35 anni e se non prontamente trattati svilupperanno CRC.

  Ci sono quattro FAP causate da differenti linee germinali quali: FAP atenuato, sindrome di Gardner, sindrome di Turcot, adenocarcinoma gastrico e poliposi prossimale.

- **Poliposi associata a MUTYH** (tipo erditario): causata da una mutazione biellelica nel gene MUTYH sul cromosoma 1p34-1, il gene MUTYH codifica un enzima chiamato MYH glicolasi, che è coninvolto in un sistema di riparazione del DNA chiamato BER (Base excision repair). il numero di polipi è minore rispettoa FAP e l'età media per la diagnosi è circa 40-60 anni e se la MAP non viene riconosiuta cè una probabilità dell'80% di contrarre CRC. Solitamente la MAP è MSI stabile.

- **Sindrome di Peutz-Jeghers** (tipo ereditario): presenta la predisposizione a sviluppare polipi benigni nell'intestino tenue sin dall'infanzia. Causa di questa malattia sono le mutazioni germinali nel gene STK11 (serina treonina chinasi 11) nota anche come LKB1 che è un gene soppressore di tumori situato nel

cromosoma 19p13.3. Le mutazioni in questo gene alterano la struttura e/o la funzione della proteina (STK11), compromettendo lasua capacità di limitare la divisione cellulare. L'MSI, la perdita di eterozigosi (LOH) nelle vicinanze del gene APC e le mutazioni KRAS sono state identificate in alcuni tumori.

- **Poliposi serrata** (tipo erditario): causata da mutazioni germinali nei geni coinvolti nella via della senescenza. Questi sono solitamente classificati come MSI-basso o MSS.

- **Sindrome di Lynch** (tipo ereditario): nota come cancro colonrettale non poliposo ereditario, è la sindrome di tumore colonrettale ereditario più comune. Causata da mutazioni germinali dei vari geni coinvolti nella riparazione dei mismatch del DNA, tra cui MSH2 sul cromosoma 2p16, MLH1 sul cromosoma 3p21, MSH6 sul cromosoma 2p16 e PMS2 sul cromosoma 7p22. Le mutaioni MSH2 e MLH1 sono le più comuni. Circa il 2-3% dei casi totali di CRC, il 90% dei tumori colonrettali con LS ha associato un MSI alto.

**Consensous Molecoular Subregions (CMS)**

- **CMS1:** sono (immuni all'instabilità del microsatellite, 14%) , ipermetilati, microsatellite instabili e immunogenici.

- **CMS2:** sono (canonici al 37%) , epiteliali, con attivazione marcata della segnalazione Wnt e MYC, e presntano un'alta sopravvivenza generale.

- **CMS3:** sono (metabolici, 13%), epiteliali e con evidente fenotipo metabolico del cancro.

- **CMS4:** sono (mesenchimali, 23%) , con attivzione prominente del fattore di crescita trasformante-$\beta$, invasione stromale e angiogenesi, e ha la maggiore sopravvivenza.

Il microsatellite instabile è altamente immunogenico, quindi la terapia che attiva il sistema immunitario può avere effetti miracolosi su i tumori instabili. La diagnosi del insatbilità del microsatellite può avvenire con 2 tecniche:

- **Immuno Istochimica:** tramite l'analisi delle protene MMR.

- **Amplificazione di specifiche ripetizioni di microsatellite:** basata su PCR.

### Immuno Istochimica

L'analisi immunoistochimica può determinare la perdita di espressione di una o più proteine MMR (riparazione dei mismatch). In realtà, l'IHC è correlata con l'MSI (instabilità dei microsatelliti), ma non è un test perfetto per la determinazione dell'MSI. Si tratta di un test dell'espressione delle proteine di riparazione dei mismatch nelle cellule. In questo metodo, gli anticorpi contro le proteine MMR come MLH1, MSH2, PMS2 e MSH6 forniscono informazioni sulla funzionalità del sistema MMR. L'analisi IHC con anticorpi PMS2 e MSH6 è in grado di rilevare la maggior parte delle anomalie nei geni corrispondenti, così come mutazioni in MLH1 e MSH2; tuttavia, l'analisi IHC con anticorpi MLH1/MSH2 può rilevare solo una frazione delle anomalie di MLH1 o MSH2, ma non tutte. Pertanto, l'analisi IHC con anticorpi MSH6 e PMS2 ha un potenziale diagnostico maggiore rispetto all'analisi con anticorpi MLH1 e MSH2. La principale rilevanza dell'MSI e dell'IHC è come test di screening per la sindrome di Lynch. L'MSI/IHC universale sui tumori è sempre più utilizzato in tutto il mondo.

### Metodo basato su PCR

Per l'analisi MSI mediante il metodo PCR multiplex fluorescente, abbiamo bisogno di DNA da tessuti tumorali e tessuti normali, una serie di primer, uno dei quali è marcato fluorescentemente (il filamento senso e/o antisenso di ciascun primer), un sequenziatore e un software appropriato. Il principio di questo metodo è misurare la presenza di diverse lunghezze di marcatori di microsatelliti specifici nelle cellule tumorali rispetto alle cellule normali. Nel primo tentativo di diagnosi di MSI nel CRC, una conferenza di consenso ha raccomandato un pannello di marcatori di microsatelliti, includendo tre ripetizioni di dinucleotide (D5S346, D2S123 e D17S250) e 2 ripetizioni di mononucleotide (BAT25 e BAT26). Sono stati descritti tre fenotipi MSI distinti. Se due o più marcatori di microsatelliti sono mutati, il tumore è considerato MSI-alto (MSI-H); se solo uno è mutato, il tumore è definito MSI-basso (MSI-L); e se nessuno dei loci esaminati dimostra instabilità, il tumore sarà considerato Microsatellite Stable (MSS). Questo pannello è noto come pannello di Bethesda (Rodriguez-Bigas et al., 1997).

Pochi anni dopo, si è scoperto che i marcatori di mononucleotide hanno una migliore specificità e sensibilità rispetto ai ripetizioni di dinucleotide (i marcatori di dinucleotide hanno una natura polimorfica), e quindi i criteri delle linee guida di Bethesda sono stati rivisti dal NCI (National Cancer Institute) nella conferenza successiva nel 2004 . Da allora, l'uso di pannelli contenenti più marcatori di mononucleotide è aumentato grazie alla loro maggiore sensibilità e specificità nella diagnosi di MSI nei CRC.

**MSI nei trattamenti.**

L'uso dello stato MSI nella previsione della risposta alla chemioterapia adiuvante è controverso, sebbene sia stato confermato che i tumori colonrettali che mostrano MSI hanno una prognosi migliore rispetto ai tumori MSS. Gli agenti chemioterapici utilizzati nel trattamento del CRC appartengono a tre categorie: antimetaboliti (5-fluorouracile), agenti alchilanti e inibitori della topoisomerasi. Il trattamento chemioterapico è efficace in alcuni pazienti, ma può comunque causare molti effetti avversi. L'MSI-H è uno dei potenziali punti predittivi per l'efficacia del trattamento chemioterapico e per il livello di effetti avversi in un paziente; pertanto, sono stati condotti diversi trial clinici riguardo a questa opinione.

Esistono diverse risposte terapeutiche nei CRC MSI-H a seconda del tipo di chemioterapia adiuvante. Quando un tumore è MSI-alto, per la diagnosi può essere o di tipo Lynch o metilato. Se viene eseguita l'IHC e la proteina non espressa è MSH2, PMS2 o MSH6, allora si tratta di Lynch, e il test germinale è indicato. Se la proteina non espressa è MLH1, potrebbe trattarsi di un tumore CIMP con ipermetilazione del promotore di MLH1, oppure di Lynch. Per distinguere tra le due possibilità, il test di mutazione BRAF o l'analisi di metilazione sul tumore sono utili. Se il tumore è avanzato, è un candidato per la terapia di attivazione immunitaria. Se non è avanzato, ha una buona prognosi e non risponderà alla terapia basata su 5-FU[9].

## 1.4 Fattori di rischio del tumore colonrettale metastatico.

Circa il 20-25% dei pazienti con cancro del colon-retto (CRC) presenta metastasi alla diagnosi iniziale, mentre i pazienti che sembrano privi di cancro dopo l'indagine iniziale sviluppano successivamente recidive locoregionali (18%), recidive a distanza (78%) o entrambe (4%). Le metastasi si verificano quando le cellule tumorali dell'originale tumore riescono a proliferare in tessuti locali, regionali o distanti; linfonodi; o organi attraverso la diffusione linfatica, ematica o addirittura transcoelomica . La recidiva del CRC è definita come recidiva metastatica locale, regionale e distante dopo un periodo di assenza della malattia. La recidiva locale si riferisce alla ricomparsa del CRC nel sito della resezione chirurgica originale, mentre la recidiva regionale si verifica nei linfonodi di drenaggio e/o nei linfonodi pelvici laterali. La recidiva metastatica distante coinvolge il fegato (che rappresenta il 40-50% delle metastasi), i polmoni (che rappresentano il 10-20% delle metastasi), il peritoneo, le ovaie, le ghiandole surrenali, le ossa e il cervello . Si stima che i tassi di sopravvivenza a 5 anni siano intorno al 90%, 70% e 10% per i rispettivi stadi di CRC localizzati, regionali e metastatici a distanza.

Molti studi hanno dimostrato che l'invasione dei vasi sanguigni, che porta alla disseminazione e metastasi delle cellule tumorali, è un forte fattore di rischio per la prognosi della malattia. Una grande proporzione di studi ha investigato l'invasione linfatica e vascolare come fattori di rischio separati, mentre altri degli studi li hanno categorizzati congiuntamente come invasione linfovascolare. Tuttavia, è stato dimostrato che la capacità predittiva dell'invasione linfovascolare è inferiore a quella dell'invasione vascolare.

L'invasione linfatica potrebbe essere un indicatore delle cellule tumorali che metastatizzano ai linfonodi. Questo riscontro è in accordo con tre studi recentemente pubblicati che manifestano che l'invasione linfatica è associata causalmente al rischio di LNM nel CRC. Le linee guida della European Society for Medical Oncology (ESMO), che manifestano che una lesione rettale di meno di 1 cm ha un rischio inferiore di metastasi e pertanto si suggerisce l'escissione locale (TEM)[10].

## 1.5 Immagini istopatologiche

L'importanza della patologia anatomica per diagnosticare e classificare le malattie non può essere sottovalutata. La diagnosi del patologo sui vetrini istologici è al centro della diagnosi, della ricerca clinica e farmaceutica e, soprattutto, delle decisioni su come trattare i pazienti oncologici nella pratica quotidiana. La necessità di precisione nella diagnosi istopatologica del cancro sta aumentando poiché la terapia personalizzata richiede una valutazione accurata dei biomarcatori[11]. Tuttavia, la maggior parte del mondo sta affrontando una carenza urgente di patologi[12].

Diversi studi hanno dimostrato una bassa riproducibilità tra laboratori, tra osservatori e intra-osservatori nella valutazione dei biomarcatori[13, 14]. Questa variabilità sta ostacolando sia il processo di scoperta di nuovi biomarcatori sia la loro utilizzazione nella pratica clinica. L'analisi computazionale delle immagini in patologia esiste da molti anni[15, 16]. Tuttavia, la sua applicazione nella patologia di routine è stata limitata a causa delle capacità limitate di digitalizzazione dei vetrini, dell'hardware dei computer, del tempo di elaborazione e dei metodi di analisi delle immagini, così come dell'archiviazione dei dati[17].

## 1.6 L'intelligenza artificiale come prossimo passo verso la patologia di precisione

L'apparizione dell'analisi delle immagini digitali promette di migliorare sia il volume che la precisione della valutazione istomorfologica. Recentemente, l'apprendimento automatico, e in particolare il deep learning, ha permesso rapidi progressi nella patologia computazionale. L'integrazione dell'apprendimento automatico nella cura di routine sarà una pietra miliare per il settore sanitario nel prossimo decennio,

e l'istopatologia è al centro di questa rivoluzione. Storicamente, la patologia diagnostica è stata eseguita mediante valutazione microscopica di sezioni di tessuto o biopsie su vetrini. La digitalizzazione delle immagini microscopiche consente l'analisi quantitativa delle immagini basata su macchine[17] **Figura1.4**.

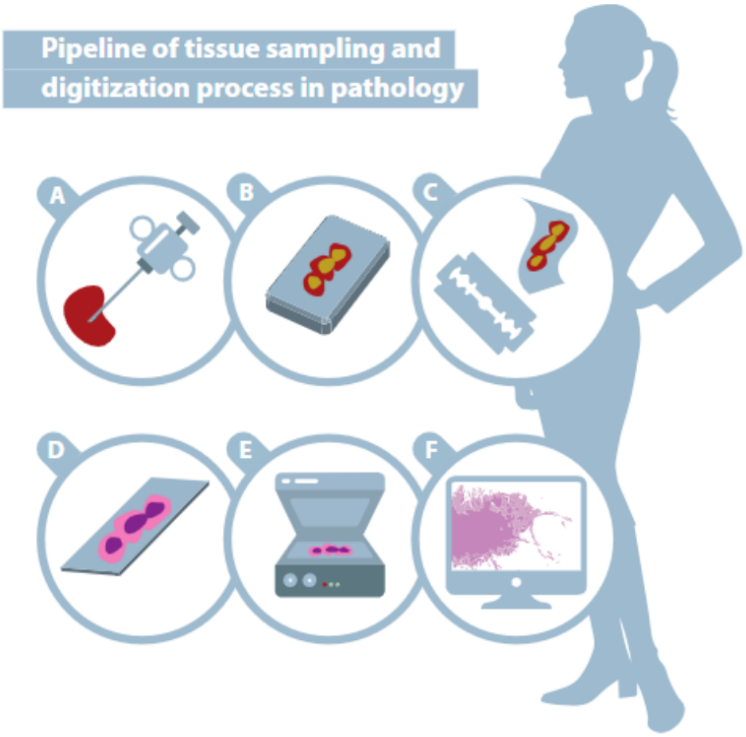


**Figura 1.4:** Pipeline del processo di campionamento dei tessuti e digitalizzazione in patologia. Dopo che è stata effettuata la biopsia dal paziente (a), viene creato un blocco di tessuto, preceduto dalla fissazione e dall'inclusione in paraffina (b). Dopo il taglio del blocco di tessuto (c), la sezione viene posizionata su un vetrino seguito da una colorazione speciale (d). Successivamente, il vetrino colorato viene inserito in uno scanner specifico per vetrini (e), ottenendo una diapositiva digitale del tessuto (f)[17].

# Capitolo 2

# Stato dell'arte

In questo capitolo vengono presentati i lavori sulla predizione alla terapia che hanno ispirato le tecniche utilizzate in questo lavoro di tesi, oltre che una panoramica sui compiti generali della Digital Parthology, che dunque, riguardano anche la ricerca sui CAD che lavorano su immagini patologiche, appartenenti anche ad altri tipi di tumore.

## 2.1 Problematiche tecnologiche della Digital Pathology

L'analisi di immagini istopatologiche digitalizzate rappresenta una rivoluzione nel campo della patologia, offrendo supporto agli anatomopatologi e oncologi nella diagnosi e prognosi dei tumori. Tuttavia, nonostante i progressi, questa tecnologia presenta ancora diverse sfide tecnologiche che ne limitano l'adozione su larga scala.

Uno dei principali ostacoli è rappresentato dalla dimensione e complessità delle immagini istopatologiche, che richiedono ingenti risorse computazionali per essere elaborate. Le immagini intere dei vetrini (Whole Slide Images, WSI) sono di dimensioni molto elevate, spesso superiori a diversi gigabyte per singolo file, e ciò pone problemi sia di archiviazione sia di elaborazione. Inoltre, gli algoritmi di Intelligenza Artificiale (AI) necessitano di grandi dataset per il training e la validazione, aumentando ulteriormente la richiesta di dati annotati.

Un'altra problematica critica riguarda la necessità di etichette accurate e dettagliate fornite da esperti anatomopatologi per addestrare modelli AI efficaci. L'annotazione manuale è un processo estremamente laborioso e dispendioso, soprattutto considerando che ogni immagine può contenere milioni di pixel e richiedere segmentazioni precise. Per mitigare questa difficoltà, negli ultimi anni sono emersi approcci innovativi come la supervisione debole, che permette di utilizzare etichette meno dettagliate a livello di immagine o caso, riducendo il carico di lavoro per gli

esperti. Tuttavia, tali approcci richiedono generalmente dataset più ampi rispetto alle tecniche di supervisione forte e pongono sfide nella generalizzabilità dei modelli.

Infine, un'altra sfida è rappresentata dall'eterogeneità dei campioni biologici e dalla variabilità introdotta durante il processo di preparazione dei vetrini, che può influenzare significativamente le prestazioni degli algoritmi AI.

## 2.2 Obiettivi della Digital Pathology

Negli ultimi anni, la Digital Pathology (DP) ha visto un rapido incremento delle applicazioni basate sull'AI e sulla computer vision. L'obiettivo principale è sviluppare strumenti rapidi, riproducibili ed efficaci per migliorare i processi diagnostici e prognostici dei tumori più comuni. Le applicazioni più promettenti si suddividono in diverse macroaree:

- **Segmentazione:** La segmentazione dei tessuti rappresenta un compito fondamentale in istopatologia digitale ed è spesso il primo passo per affrontare altri obiettivi diagnostici. Attraverso l'uso di reti neurali convoluzionali (CNN), è possibile identificare automaticamente regioni di interesse, come tessuti tumorali, linfociti o altre strutture cellulari. Un esempio è il lavoro che utilizza una rete VGG19 per segmentare immagini istopatologiche colorettali in classi come tessuto tumorale, normale o linfocitario.

- **Stadiazione:** La stadiazione del tumore è cruciale per stabilire l'estensione della malattia e pianificare il trattamento. Modelli AI sono stati sviluppati per integrare informazioni provenienti dalle immagini con i dati clinici per una stratificazione accurata dei pazienti.

- **Predizione della risposta alla terapia:** Gli algoritmi di deep learning possono prevedere la risposta del tumore a specifici trattamenti, offrendo uno strumento prezioso per la medicina personalizzata. Studi recenti hanno dimostrato la capacità di AI di correlare caratteristiche morfologiche con outcome terapeutici.

- **Predizione della sopravvivenza:** Modelli AI avanzati utilizzano immagini istopatologiche combinate con dati molecolari per stimare la sopravvivenza dei pazienti, fornendo un supporto decisionale nella gestione del paziente.

- **Predizione dello stato di mutazione dei geni:** È stato dimostrato che algoritmi AI possono prevedere mutazioni genetiche direttamente da immagini H&E standard. Un esempio significativo è la predizione dello stato di instabilità microsatellitare (MSI) nel cancro colorettale, un biomarcatore chiave per la scelta della terapia immunologica.

- **Predizione della malignità:** Modelli AI stanno contribuendo a migliorare la classificazione dei tumori in base al grado di aggressività, integrando dati istopatologici con caratteristiche molecolari e cliniche.

Questi ambiti applicativi mostrano il potenziale dell'AI nel migliorare l'accuratezza diagnostica, ridurre i bias e offrire nuove intuizioni sui processi patologici che potrebbero essere difficilmente rilevabili da revisione umana [18].

## 2.3 Predizione della risposta alla terapia

In questo lavoro si è tentato di approcciare alle possibili sluzioni per predire la risposta alla terapia dall'analisi delle immagini istopatologiche di tumore colorettale.

Tutti i lavori riguardanti la predizione della risposta alla terapia hanno in comune alcuni passaggi che qui vengono vengono sintetizzati:

- **Normalizzazione del colore:** questo passaggio di preprocessing si rende necessario a causa della bassa riproducibilità delle immagini istopatologiche, dovuta alla variabilità sia nella preparazione del campione che alla diversità di scanner digitali utilizzati. Le tecniche di normalizzazione maggiormente utilizzate sono suddivisibili in due categorie:
  - **Utilizzano un immagine target:**
    - **Rehinhard:** metodo per una forma più generale di correzione del colore che prende in prestito le caratteristiche cromatiche di un'immagine da un'altra[19].
    - **Deep Learning guidate:** tecniche che utilizzano reti come le GAN [20] allenate su immagini target per effettuare la normalizzazione del colore.
  - **Senza immagine target:**
    - **Macenko:** questa trasformazione, invece, si occupa di rappresentare le intensità nello spazio della densità ottica operando una deconvoluzione del colore che rende le colorazioni di ematossilina ed eosina linearmente separabili[21].
- **Estrazione tiles tumorali:** questa fase è approcciata in maniera differente in base alla disponibilità di immagini con annotazioni sulla posizione del tumore all'interno dell'immagine, altrimenti vengono utilizzati algoritmi di segmentazione che evidenziano la zona tumorale e la separano dagl altri tessuti come nel presente lavoro.

- **Estrazione delle Caratteristiche dalle tiles tumorali:** Questa parte è affrontata nei vari lavori con tecniche deterministiche precedenti alle tecniche di AI [3] o con classificatori allenati con tecniche generalmente scarsamente supervisionate [22] o auto-supervisionate [23], anche se non mancano lavori con classificatori supervisionati con dati riguardanti lo stato molecolare com nello studio di Sirinukunwattana[24].
- **Aggregazione delle tiles tumorali:** Viene svolta con diverse tecniche che puntano a classificare l' insieme di tiles apparteneti ad un paziente, e quindi di fornire un'etichetta globale alle tiles di un paziente in base alle caratteristiche precedentemente estratte, e dunque fornire la predizione alla terapia vera e prorpria.

# Capitolo 3

# Materiali e metodi

## 3.1 Dataset

### Origine del Dataset

Il dataset utilizzato in questo studio è stato fornito dall'Ospedale Niguarda di Milano. Esso è composto da campioni istopatologici ottenuti da 17 pazienti affetti da tumore colorettale, suddivisi in due gruppi principali:

- **Gruppo OLD**: 8 pazienti anziani, di cui uno presentava metastasi.
- **Gruppo IANG**: 9 pazienti giovani.

### Preparazione dei Campioni

I campioni sono stati colorati utilizzando la colorazione standard *ematossilina ed eosina (H&E)*, che consente di evidenziare le strutture cellulari e tissutali attraverso una doppia colorazione:

- **Ematossilina**: colora i nuclei cellulari di blu-viola.
- **Eosina**: colora il citoplasma e la matrice extracellulare di tonalità rosa.

### Acquisizione Digitale

Le sezioni istologiche sono state digitalizzate utilizzando uno scanner *Mirax*, che consente di ottenere immagini ad alta risoluzione. Questi dati rappresentano immagini *Whole Slide Images (WSI)*, che coprono intere sezioni di tessuto, garantendo una rappresentazione completa del campione istopatologico.

#### Composizione del Dataset

Le immagini del dataset sono state analizzate per la predizione della risposta alla terapia, considerando le caratteristiche tissutali e cellulari rilevabili tramite le WSI. La suddivisione in sottogruppi (**OLD** e **IANG**) è stata utilizzata per esplorare eventuali correlazioni tra l'età del paziente, la presenza di metastasi e le risposte ai trattamenti terapeutici.

#### Etichettatura dei Dati

Non sono state fornite annotazioni manuali sulle posizioni precise delle regioni tumorali all'interno delle immagini. Per ovviare a questa mancanza, sono stati applicati metodi di segmentazione automatica per identificare e isolare le regioni tumorali dalle altre componenti tissutali, della proceduramper la segmentazione automatica si tratterà più avanti.

#### Considerazioni Etiche

Il dataset è stato utilizzato rispettando tutte le normative vigenti in materia di protezione dei dati personali e delle informazioni sensibili. L'anonimizzazione dei dati dei pazienti è stata garantita dall'Ospedale Niguarda di Milano prima della loro consegna per l'analisi.

## 3.2 Pipeline di Analisi

La pipeline di analisi sviluppata in questo studio è stata progettata per prevedere la risposta alla terapia a partire dalle immagini istopatologiche dei pazienti affetti da tumore colorettale. Di seguito vengono descritti i passaggi principali del processo, riassunti nella Figura 3.1.

### 3.2.1 Preprocessing e Normalizzazione del Colore

Per garantire la coerenza cromatica tra le immagini, è stata effettuata una normalizzazione del colore. Questa fase è cruciale per ridurre la variabilità introdotta dai diversi metodi di colorazione e scannerizzazione. Nel nostro caso, è stato utilizzato il metodo di *Macenko*, che consente di ottenere un'adeguata standardizzazione delle immagini senza necessità di un'immagine target.

#### Conversione nello Spazio della Densità Ottica

Le immagini colorate con Ematossilina ed Eosina (H&E) vengono inizialmente convertite dal dominio RGB al dominio della densità ottica. La relazione tra le

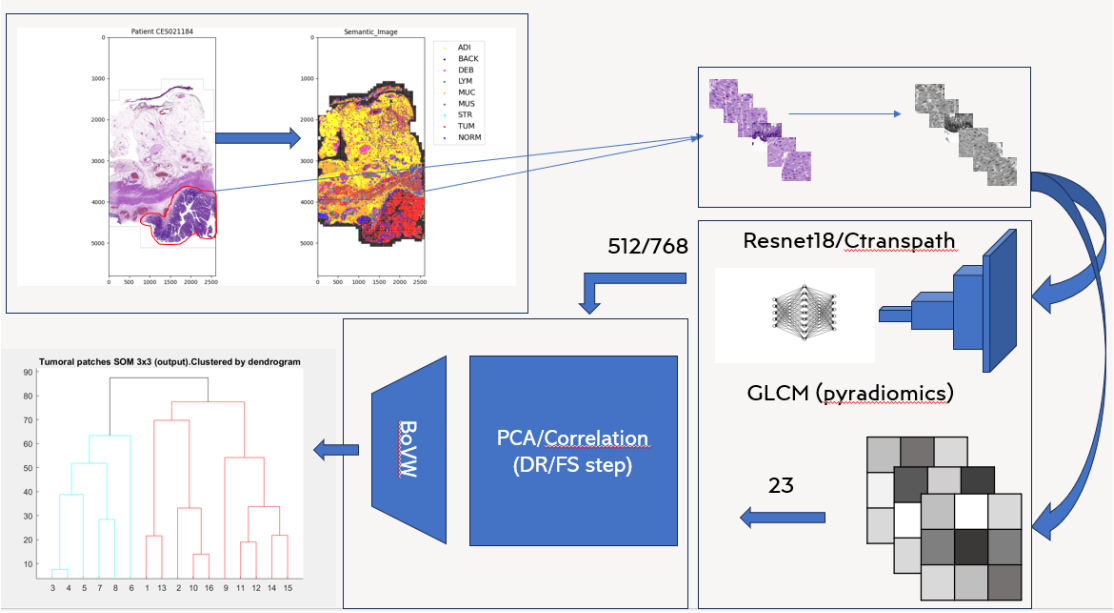


**Figura 3.1:** Pipeline di analisi per la predizione della risposta alla terapia. I passaggi includono la normalizzazione del colore, la segmentazione delle tiles tumorali, l'estrazione delle caratteristiche, l'aggregazione delle caratteristiche e la classificazione finale.

intensità RGB ($I$) e la densità ottica ($OD$) è data da:

$$OD = -\log\left(\frac{I}{I_0}\right)$$

dove $I$ rappresenta i valori di intensità dei pixel e $I_0$ è l'intensità massima (tipicamente 255 per immagini a 8 bit).

### Decomposizione del Colore

La deconvoluzione del colore separa le componenti corrispondenti ai coloranti (ematossilina ed eosina). Si rappresenta la densità ottica di un pixel come combinazione lineare delle componenti cromatiche:

$$OD = C \cdot S$$

dove:

- $C$ è una matrice $3 \times 2$ contenente i vettori direzionali per le componenti cromatiche di H&E.
- $S$ è una matrice $2 \times N$, dove $N$ è il numero di pixel, che contiene le quantità dei due coloranti per ciascun pixel.

### Trasformazione Lineare e Normalizzazione

Una volta ottenuta la rappresentazione nel dominio del colore, i vettori direzionali $C$ vengono standardizzati rispetto a un riferimento cromatico globale, calcolato utilizzando statistiche derivate da una popolazione di immagini. Questo processo garantisce che i colori delle immagini sorgente siano coerenti con il riferimento cromatico standard. La trasformazione viene effettuata tramite una mappatura lineare dei vettori direzionali della sorgente sul riferimento globale:

$$C' = C \cdot T$$

dove $T$ è una matrice di trasformazione calcolata utilizzando i *percentili* della distribuzione cromatica globale. Questo approccio evita la necessità di un'immagine target specifica, rendendo il metodo robusto alla variabilità inter-dataset.

### Ricostruzione dell'Immagine Normalizzata

Infine, l'immagine normalizzata viene ricostruita convertendo nuovamente dallo spazio della densità ottica al dominio RGB inverso:

$$I' = I_0 \cdot \exp(-OD')$$

dove $OD'$ è la densità ottica trasformata dopo la normalizzazione[21].

Questo processo garantisce la standardizzazione cromatica tra immagini provenienti da diverse fonti, minimizzando l'impatto della variabilità introdotta durante la preparazione e la digitalizzazione del campione. Nella **Figura 3.2** sottostante viene riportata la trasformazione delle coordinate dei colori di ematossilina e eosina dallo spazio RGB allo spazio della densità ottica.

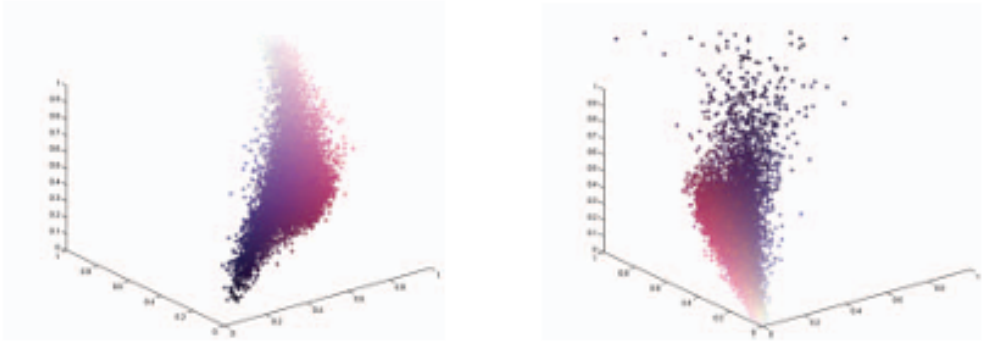

**Figura 3.2:** Spazio RGB (sinistra) e spazio della densità ottica (destra)

### 3.2.2 Segmentazione delle Tiles Tumorali

La segmentazione dei tessuti di un immagine istopatologica è il prmo step per poter approcciare agl' altri compiti elencati. Un lavoro molto interessante è stato prodotto dal Kather Lab, nel quale è stato costruito un classificatore tramite l allenamento di una VGG19 che classifica le tiles di 224x224 pixels delle immagini istopatologiche di tumore colorettale in 9 classi tra cui: tessuto adiposo, mucosa, debris, tessuto muscolare, normale, tumorale, linfociti e background[25]. Nel presente lavoro di tesi si è utilizzato tale classificatore in cui di seguito viene mostrato un esempio di classificazione per un paziente IANG **Figura 3.3** e uno OLD **Figura 3.4**.

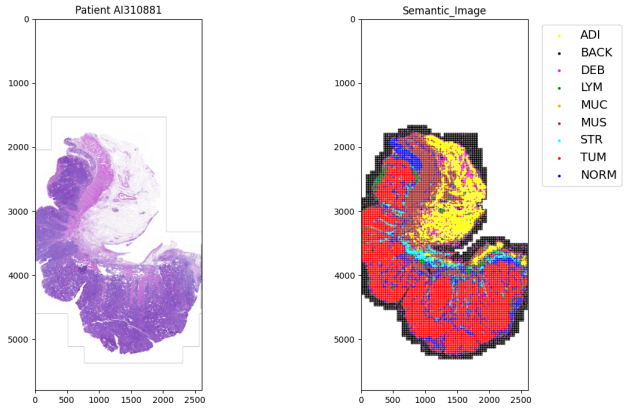


**Figura 3.3:** Esempio di classificazione dei tessuti di un appartenente al gruppo IANG

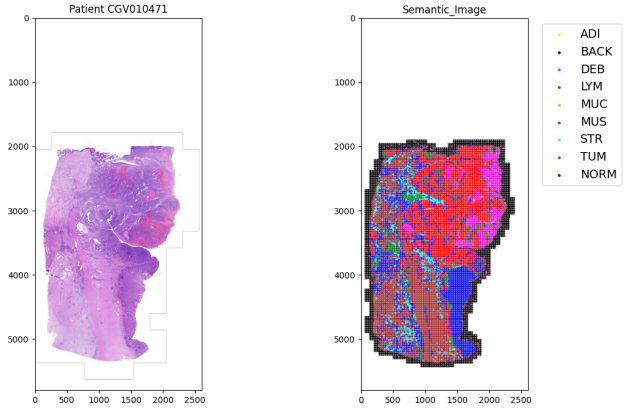


**Figura 3.4:** Esempio di classificazione dei tessuti di un appartenente al gruppo OLD

.

### 3.2.3 Estrazione delle Caratteristiche

Dalle tiles tumorali sono state estratte caratteristiche tramite reti neurali convoluzionali (*Convolutional Neural Networks, CNN*) pre-addestrate o attraverso il calcolo di caratteristiche quantitative. Queste reti sono state utilizzate come estrattori di caratteristiche, senza ulteriori fasi di addestramento. Le caratteristiche estratte rappresentano informazioni quantitative relative a tessuti, cellule e strutture presenti nelle immagini.

#### Caratteristiche Testurali: Gray Level Co-occurrence Matrix (GLCM) e Caratteristiche Derivate

Il primo metodo utilizzato è basato sulle Matrici di Co-Occorrenza a Livelli di Grigio (*Gray-Level Co-occurrence Matrices*, GLCM), calcolate attraverso il framework `PyRadiomics` [26]. Questo approccio si concentra sull'estrazione di 23 caratteristiche testurali per ciascuna *tile*, che includono metriche come l'omogeneità, l'entropia, la correlazione e il contrasto. Tali caratteristiche sono comunemente impiegate per catturare la struttura testurale delle immagini istopatologiche e la loro distribuzione cromatica.

La Gray Level Co-occurrence Matrix (GLCM) è una matrice $N_g \times N_g$ che rappresenta la funzione di probabilità congiunta di secondo ordine di una regione dell'immagine. L'elemento $(i, j)$ della matrice rappresenta il numero di volte in cui i livelli di grigio $i$ e $j$ si trovano in due pixel separati da una distanza $\delta$ lungo un angolo $\theta$. La distanza $\delta$ viene definita secondo la norma infinita. Per esempio, in 2D con $\delta = 1$ e $\theta = 0°$, si considerano i pixel alla sinistra e alla destra del pixel centrale.

Le probabilità normalizzate della GLCM sono calcolate dividendo ogni elemento della matrice per la somma totale degli elementi. Inoltre, PyRadiomics utilizza di default una GLCM simmetrica, che rende la matrice invariante rispetto alla direzione dei pixel considerati.

Nel presente lavoro, per l'estrazione delle caratteristiche tramite GLCM sono stati utilizzati i valori di default di PyRadiomics:

- Distanza tra i pixel ($\delta$) pari a 1,
- Larghezza dei bin (`binWidth`) pari a 25.

Con questi parametri, i livelli di grigio dell'immagine, originariamente distribuiti nell'intervallo $[0, 255]$, sono stati quantizzati in $N_g = 11$ bin. Di conseguenza, la GLCM calcolata ha dimensioni $11 \times 11$. Di seguito sono riportate le principali caratteristiche statistiche derivate dalla GLCM.

**Definizione dei Parametri utili al calcolo delle caratteristiche con GLCM**

- $\epsilon$: Un valore positivo arbitrariamente piccolo (es. $\approx 2.2 \times 10^{-16}$).
- $P(i,j)$: Matrice di co-occorrenza per un valore arbitrario di $\delta$ (distanza spaziale) e $\theta$ (angolo).
- $p(i,j)$: Matrice di co-occorrenza normalizzata, calcolata come:

$$p(i,j) = \frac{P(i,j)}{\sum P(i,j)}.$$

- $N_g$: Numero di livelli di intensità discreti nell'immagine.
- $p_x(i)$: Probabilità marginale per riga, calcolata come:

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j).$$

- $p_y(j)$: Probabilità marginale per colonna, calcolata come:

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j).$$

- $\mu_x$: Media dei livelli di intensità per $p_x$, definita come:

$$\mu_x = \sum_{i=1}^{N_g} p_x(i) \cdot i.$$

- $\mu_y$: Media dei livelli di intensità per $p_y$, definita come:

$$\mu_y = \sum_{j=1}^{N_g} p_y(j) \cdot j.$$

- $\sigma_x$: Deviazione standard di $p_x$.
- $\sigma_y$: Deviazione standard di $p_y$.
- $p_{x+y}(k)$: Probabilità della somma degli indici $i + j = k$, definita come:

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \quad \text{dove } i + j = k,\ k = 2, 3, \ldots, 2N_g.$$

- $p_{x-y}(k)$: Probabilità della differenza assoluta degli indici $|i-j|=k$, definita come:

$$p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j), \quad \text{dove } |i-j|=k,\ k=0,1,\ldots,N_g-1.$$

- $H_X$: Entropia di $p_x$, calcolata come:

$$H_X = -\sum_{i=1}^{N_g} p_x(i)\log_2(p_x(i)+\epsilon).$$

- $H_Y$: Entropia di $p_y$, calcolata come:

$$H_Y = -\sum_{j=1}^{N_g} p_y(j)\log_2(p_y(j)+\epsilon).$$

- $H_{XY}$: Entropia della matrice $p(i,j)$, calcolata come:

$$H_{XY} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2(p(i,j)+\epsilon).$$

- $H_{XY1}$: Entropia calcolata considerando $p_x(i)p_y(j)$ come distribuzione di riferimento:

$$H_{XY1} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2(p_x(i)p_y(j)+\epsilon).$$

- $H_{XY2}$: Entropia calcolata considerando $p_x(i)p_y(j)$ come distribuzione effettiva:

$$H_{XY2} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j)\log_2(p_x(i)p_y(j)+\epsilon).$$

**Formule delle Caratteristiche**

- **Autocorrelation**

$$\text{Autocorrelation} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} i\cdot j\cdot p(i,j)$$

- **Joint Average**

$$\text{Joint Average} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot i$$

- **Cluster Prominence**

$$\text{Cluster Prominence} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^4 \cdot p(i,j)$$

- **Cluster Shade**

$$\text{Cluster Shade} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^3 \cdot p(i,j)$$

- **Cluster Tendency**

$$\text{Cluster Tendency} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^2 \cdot p(i,j)$$

- **Contrast**

$$\text{Contrast} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}|i-j|^2 \cdot p(i,j)$$

- **Correlation**

$$\text{Correlation} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x\sigma_y}$$

- **Difference Average**

$$\text{Difference Average} = \sum_{k=0}^{N_g-1} k \cdot p_{x-y}(k)$$

- **Difference Entropy**

$$\text{Difference Entropy} = -\sum_{k=0}^{N_g-1} p_{x-y}(k) \cdot \log_2(p_{x-y}(k)+\epsilon)$$

- **Difference Variance**

$$\text{Difference Variance} = \sum_{k=0}^{N_g-1} (k-\mu_{x-y})^2 \cdot p_{x-y}(k)$$

- **Joint Energy**

$$\text{Joint Energy} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$$

- **Joint Entropy**

$$\text{Joint Entropy} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \cdot \log_2(p(i,j) + \epsilon)$$

- **IMC1 (Informational Measure of Correlation 1)**

$$\text{IMC1} = \frac{H_{XY} - H_{XY1}}{\max(H_X, H_Y)}$$

- **IMC2 (Informational Measure of Correlation 2)**

$$\text{IMC2} = \sqrt{1 - e^{-2(H_{XY2} - H_{XY})}}$$

- **IDM (Inverse Difference Moment)**

$$\text{IDM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1 + (i-j)^2}$$

- **IDMN (Normalized IDM)**

$$\text{IDMN} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1 + \frac{(i-j)^2}{N_g^2}}$$

- **ID (Inverse Difference)**

$$\text{ID} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1 + |i-j|}$$

- **IDN (Normalized ID)**

$$\text{IDN} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1 + \frac{|i-j|}{N_g}}$$

- **Inverse Variance**

$$\text{Inverse Variance} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{|i-j|^2}, \quad i \neq j$$

- **Maximum Probability**

$$\text{Maximum Probability} = \max(p(i,j))$$

- **Sum Entropy**

$$\text{Sum Entropy} = -\sum_{k=2}^{2N_g} p_{x+y}(k) \cdot \log_2(p_{x+y}(k) + \epsilon)$$

- **Sum Squares**

$$\text{Sum Squares} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i - \mu_x)^2 \cdot p(i,j)$$

Le definizioni sono tratte dalla documentazione ufficiale di PyRadiomics: `https://pyradiomics.readthedocs.io/en/latest/features.html#module-radiomics.glcm`.

### Caratteristiche Derivate da ResNet18

Nel presente lavoro, è stato utilizzato un modello pre-allenato di rete neurale convoluzionale ResNet-18, adottato da un approccio di Multiple Instance Learning (MIL) per l'estrazione di caratteristiche dalle tiles tumorali. ResNet-18, una rete profonda composta da 18 layer, è nota per la sua capacità di evitare il problema del *vanishing gradient* grazie all'uso di *skip connections* (connessioni di salto), che permettono una migliore propagazione del gradiente attraverso la rete[27].

Nel contesto dell'analisi istopatologica, la rete è stata utilizzata per estrarre un vettore di caratteristiche da ciascuna tile tumorale, rimuovendo il layer finale di classificazione. In questo modo, la rete funge da estrattore di caratteristiche generando un vettore di 512 dimensioni, che rappresenta informazioni strutturali e morfologiche delle tiles.

La rete è stata allenata inizialmente per il compito di classificazione dello stato MSI/MSS in immagini istopatologiche, e successivamente adattata per l'estrazione di caratteristiche. A ciascuna tile tumorale veniva assegnata l'etichetta dell'intera immagine WSI (Whole Slide Image) nel contesto di un approccio MIL. Sebbene la rete non fosse utilizzata per la classificazione diretta, le caratteristiche estratte sono state utilizzate come input per classificatori successivi, con l'obiettivo di predire la risposta alla terapia. Questo approccio ha consentito di estrarre caratteristiche rilevanti dalle immagini istopatologiche, migliorando la qualità dei dati utilizzati nei modelli predittivi.

Per il secondo metodo, è stata utilizzata una rete neurale convoluzionale pre-allenata, la ResNet18, ottimizzata per il task di predizione dello stato di instabilità

microsatellitare (MSI/MSS) su immagini istopatologiche tumorali. Il layer di output della rete è stato rimosso, e l'output del penultimo layer è stato utilizzato come rappresentazione delle *tile*, generando 512 caratteristiche per ogni immagine. Questo approccio sfrutta la capacità della rete di apprendere rappresentazioni gerarchiche direttamente dai dati visivi.

### CTransPath: Estrattore di caratteristiche istopatologiche basato su apprendimento auto-supervisionato

Un dataset di grandi dimensioni e ben annotato è un fattore chiave per il successo del deep learning nell'analisi delle immagini mediche. Tuttavia, costruire annotazioni su larga scala è molto impegnativo, soprattutto per le immagini istopatologiche, che presentano caratteristiche uniche come dimensioni delle immagini in gigapixel, molteplici tipi di cancro e ampie variazioni di colorazione. Per mitigare questo problema, è stato proposto l'utilizzo dell'apprendimento auto-supervisionato (self-supervised learning, SSL), che si basa esclusivamente su dati non annotati per generare rappresentazioni informative e si adatta bene a vari compiti downstream anche con annotazioni limitate.

Nel contesto di questo lavoro, è stato adottato come estrattore di caratteristiche il modello CTransPath, un modello ibrido progettato integrando una rete neurale convoluzionale (CNN) con un'architettura Swin Transformer multi-scala. Per migliorare ulteriormente l'efficacia del modello, è stata implementata una strategia innovativa di apprendimento auto-supervisionato denominata Semantically-Relevant Contrastive Learning (SRCL). Questo approccio confronta la rilevanza tra istanze per individuare un numero maggiore di coppie positive, aumentando così la diversità delle coppie allineate e generando rappresentazioni più informative rispetto al contrastive learning tradizionale, che si limita a creare due viste della stessa istanza.

CTransPath è stato pre-addestrato su un ampio insieme di immagini istopatologiche non annotate, servendo come estrattore collaborativo di caratteristiche locali e globali. Questo pre-addestramento consente al modello di apprendere rappresentazioni universali delle caratteristiche che risultano particolarmente adatte ai compiti istopatologici.

L'efficacia di CTransPath pre-addestrato con SRCL è stata dimostrata su cinque diversi compiti downstream, tra cui:

- ricerca di patch (patch retrieval)
- classificazione di patch (patch classification)
- classificazione debolmente supervisionata di immagini di intere sezioni (weakly-supervised whole-slide image classification)
- rilevazione della mitosi (mitosis detection)

- segmentazione delle ghiandole di adenocarcinoma colorettale (colorectal adenocarcinoma gland segmentation)

coprendo un totale di nove dataset pubblici. I risultati ottenuti mostrano che le rappresentazioni visive apprese tramite SRCL non solo raggiungono prestazioni all'avanguardia per ciascun dataset, ma risultano anche più robuste e trasferibili rispetto ad altri metodi di apprendimento auto-supervisionato e al pre-addestramento su ImageNet (sia supervisionato che auto-supervisionato)[28].

Si è fatto uso CTransPath per estrarre vettori di caratteristiche rappresentativi delle immagini tumorali, utilizzando il modello pre-addestrato con SRCL. Questa configurazione si è dimostrata particolarmente utile per le analisi basate su raggruppamenti non supervisionati. Per ulteriori dettagli tecnici e per il codice sorgente, si rimanda alla documentazione ufficiale del progetto.

### 3.2.4 Aggregazione delle Caratteristiche

Le caratteristiche estratte da tutte le tiles di un singolo paziente sono state aggregate utilizzando il metodo del *Bag of Visual Words (BoVW)*. Questo approccio consente di ottenere una rappresentazione globale del paziente attraverso l'analisi dell'istogramma delle occorrenze di specifici fenotipi. I fenotipi sono stati individuati utilizzando una rete auto-organizzante (*Self-Organizing Map*, SOM) 3x3, che ha categorizzato le tiles in 9 cluster distinti sulla base della similarità euclidea tra le caratteristiche estratte. Ogni paziente è stato quindi descritto da un vettore di frequenze relative, ovvero un istogramma delle occorrenze dei fenotipi nelle sue immagini tumorali. Infine, gli istogrammi di tutti i pazienti sono stati utilizzati per un'analisi di clustering gerarchico mediante dendrogramma, basato sempre sulla distanza euclidea tra i vettori di frequenza.

#### Costruzione del Codebook

Un passaggio cruciale del metodo BoVW è la costruzione del **dizionario visuale** (*Codebook*), che consiste in un insieme rappresentativo di "parole visive" (*Visual Words*). Il Codebook è costruito attraverso un algoritmo di clustering o di *vector quantization*, applicato a un set di descrittori locali estratti da una collezione di immagini di allenamento.

Nel dettaglio:

1. **Estrazione dei descrittori locali**: le caratteristiche calcolate su ciascuna patch vengono raccolte, indipendentemente dall'immagine sorgente, in un unico dataset che rappresenta l'intera collezione di immagini.

2. **Clustering dei descrittori**: questi descrittori sono raggruppati in un numero predeterminato di cluster utilizzando tecniche di clustering (ad esempio, *k-means* o SOM stessa).

3. **Definizione delle Visual Words**: ogni cluster rappresenta una "parola visiva", ovvero un elemento del dizionario che cattura informazioni significative della collezione. Il risultato è un Codebook che consente di quantizzare le caratteristiche locali di nuove immagini, trasformandole in sequenze di occorrenze delle parole visuali[29].

La costruzione del Codebook è fondamentale per garantire che il modello possa catturare in modo robusto le caratteristiche globali del dataset, indipendentemente dalle variazioni nelle singole immagini.

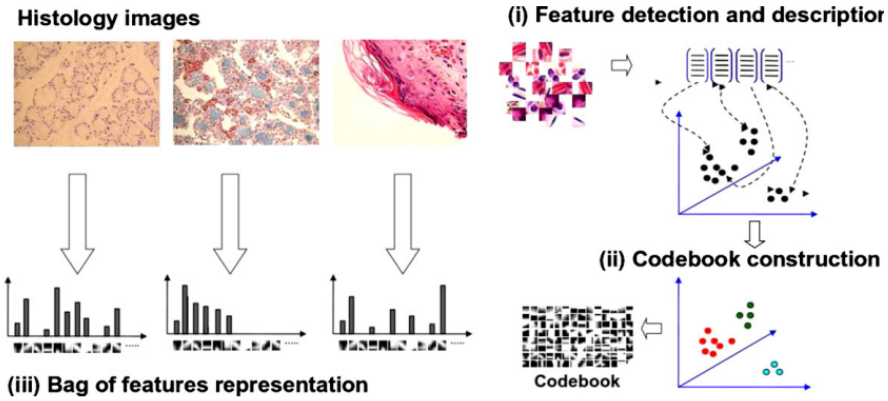


**Figura 3.5:** Illustrazione della costruzione del Codebook per il *Bag of Visual Words (BoVW)*. Le caratteristiche locali estratte dalle immagini di allenamento vengono raggruppate in cluster, ciascuno rappresentante una parola visiva (*Visual Word*). Il dizionario visuale così costruito permette di trasformare le caratteristiche locali di una nuova immagine in un istogramma di frequenze delle parole visuali, utilizzabile per l'analisi e il clustering dei pazienti.

### 3.2.5 Analisi delle Caratteristiche Estratte

Le caratteristiche utilizzate nel presente studio sono state estratte attraverso tre metodi distinti: GLCM, ResNet18 e CTransPath. Tutte le combinazioni di estrazione e riduzione/selezione delle caratteristiche sono state valutate tramite un approccio Bag of Visual Words (BoVW), poiché in questo studio non erano presenti etichette dei dati. L'unica informazione disponibile riguardava l'osservazione, fornita dall'oncologo, secondo cui alcuni pazienti giovani si raggruppavano con quelli anziani.

### Prove iniziali senza riduzione

Inizialmente, sono state eseguite prove senza alcuna riduzione della dimensionalità sui dataset derivanti dalle estrazioni tramite GLCM, ResNet18 e CTransPath. Queste analisi hanno fornito un punto di riferimento per valutare l'impatto delle successive tecniche di riduzione/selezione delle caratteristiche.

### Riduzione della dimensionalità tramite soglia di correlazione

Per le estrazioni da ResNet18 e CTransPath, è stata applicata una soglia di correlazione di Pearson per ridurre la dimensionalità delle caratteristiche. Sono stati testati diversi valori di soglia, selezionando 0.8 per ResNet18 e 0.5 per CTransPath. Questi valori hanno permesso di ridurre sufficientemente la dimensionalità mantenendo informazioni rilevanti per l'analisi.

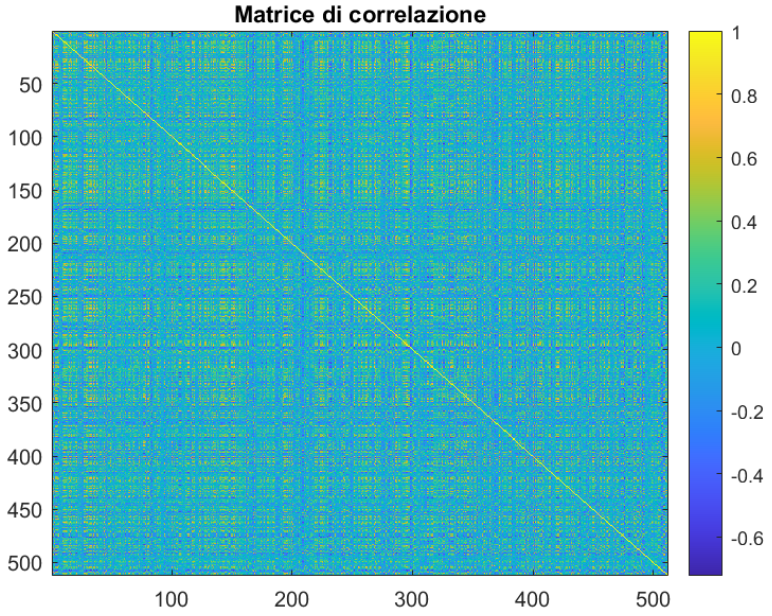


**Figura 3.6:** Matrice di correlazione per le caratteristiche estratte da ResNet18

### Riduzione della dimensionalità tramite PCA

La riduzione della dimensionalità è stata valutata anche tramite analisi delle componenti principali (PCA). Sono state condotte prove variando la *explained variance* al 85%, 90% e 95%, riscontrando variazioni marginali nei risultati tra le

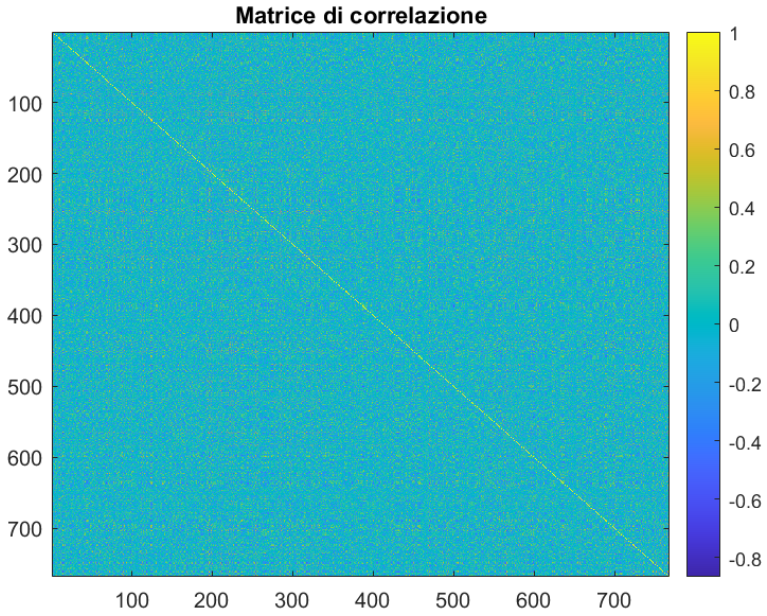


**Figura 3.7:** Matrice di correlazione per le caratteristiche estratte da CTransPath

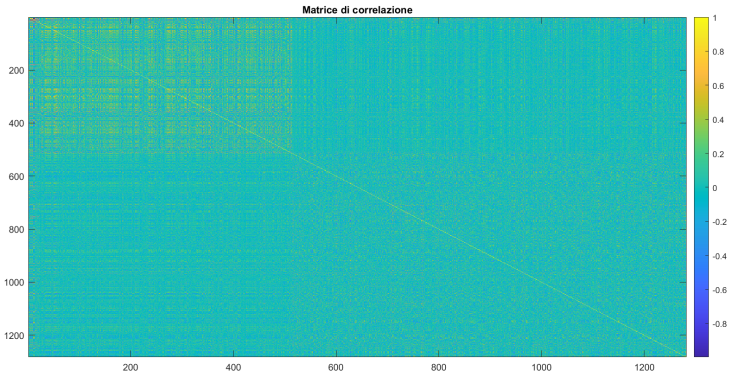


**Figura 3.8:** Matrice di correlazione per le caratteristiche combinate degli estrattori prime 23 righe GLCM dalla 24-ma alla 536-ma ResNet18 dalla 527-ma all ultima riga Ctranspath

diverse configurazioni. Tuttavia, l'applicazione della PCA ha mostrato risultati meno soddisfacenti nel contesto dell'analisi BoVW.

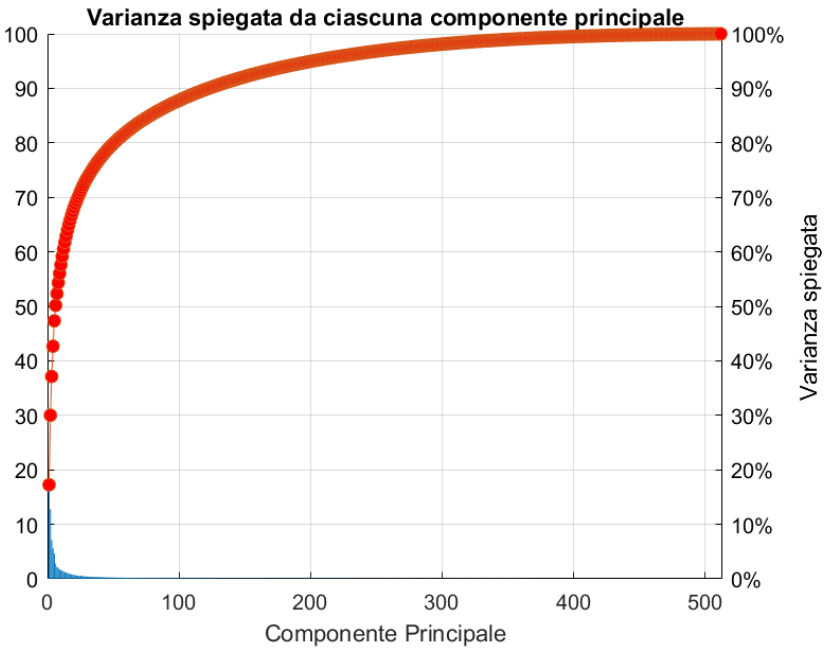


**Figura 3.9:** Varianza spiegata dalle caratteristiche estratte da ResNet-18

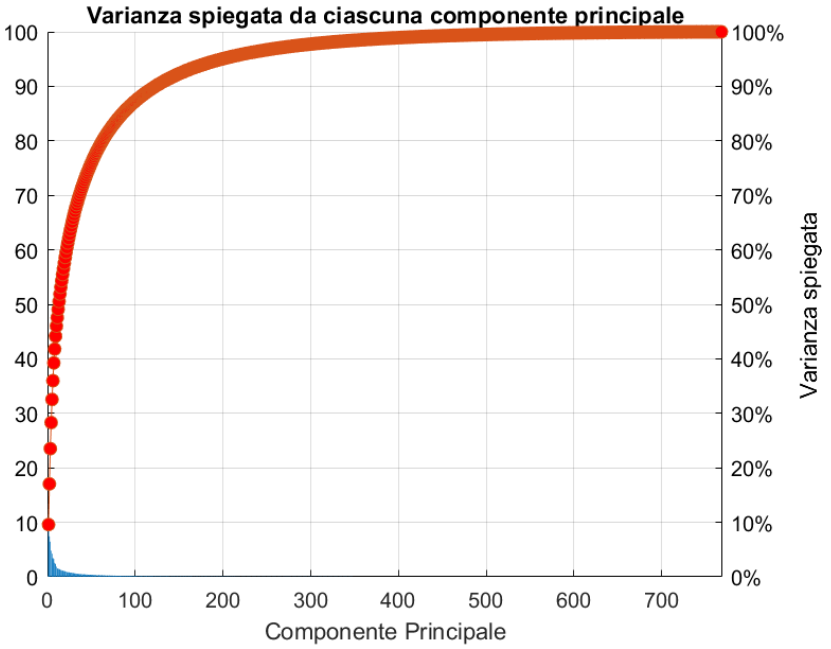


**Figura 3.10:** Varianza spiegata dalle caratteristiche estratte da CTranspath

### Analisi combinata e Bag of Visual Words

Tutte le combinazioni delle caratteristiche estratte e ridotte sono state analizzate con un approccio Bag of Visual Words (BoVW), utilizzando una Self-Organizing Map (SOM) di dimensioni $3 \times 3$ e un dendrogramma per costruire l'istogramma dei fenotipi dei pazienti. Questa metodologia è stata applicata per evidenziare pattern di raggruppamento tra i pazienti, in particolare tra giovani e anziani, sulla base delle osservazioni fornite dall'oncologo.

Sono state eseguite le seguenti combinazioni:

- Le sole caratteristiche estratte tramite GLCM, ResNet18 e CTransPath senza riduzione,
- Le caratteristiche di ResNet18 e CTransPath ridotte tramite soglia di correlazione (0.8 per ResNet18 e 0.5 per CTransPath),
- La combinazione delle caratteristiche GLCM, ResNet18 e CTransPath con riduzione tramite soglia di correlazione,
- La combinazione sopra citata con l'aggiunta della PCA applicata a valle.

### Risultati e selezione del miglior modello

Dall'analisi dei risultati, è emerso che la migliore configurazione è stata ottenuta utilizzando le caratteristiche estratte da ResNet18 con una soglia di correlazione di 0.8. Questa configurazione si è dimostrata particolarmente efficace nel raggruppare alcuni pazienti giovani insieme a quelli anziani, suggerendo un potenziale legame tra le caratteristiche estratte e i fenotipi clinici.

**Tabella 3.1:** Risultati della silhouette media per diverse estrazioni e metodi di selezione delle caratteristiche

| **Configurazione** | **Silhouette Media** |
|---|---|
| Comb FE std GLCM RES Corr0.8 Ctran Corr0.5 | 0.3486 |
| Comb FE std GLCM RES Corr0.8 Ctran Corr0.5 Corr0.8 | 0.1465 |
| Comb FE std GLCM RES Corr0.8 Ctran Corr0.5 PCA85 | 0.3468 |
| Ctranspath std Corr0.5 | 0.1077 |
| Ctranspath std Corr0.5 PCA85 | 0.3468 |
| Resnet18 std Corr0.8 | 0.2464 |
| Resnet18 std Corr0.8 PCA85 | 0.2452 |
| GLCM | 0.4718 |

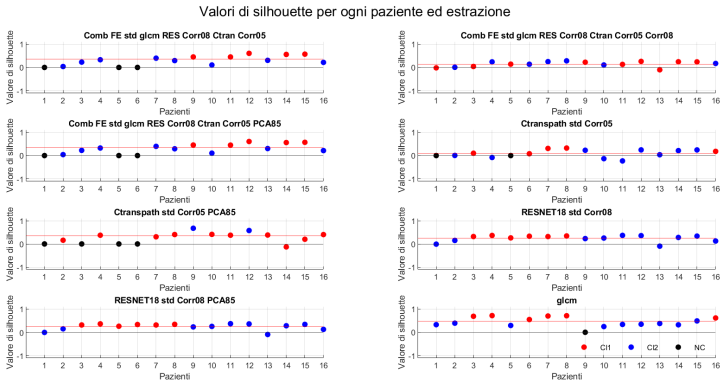


**Figura 3.11:** valori di silhouette ottenuti con i vari metodi da 1 a 8 IANG e da 9 a 16 OLD

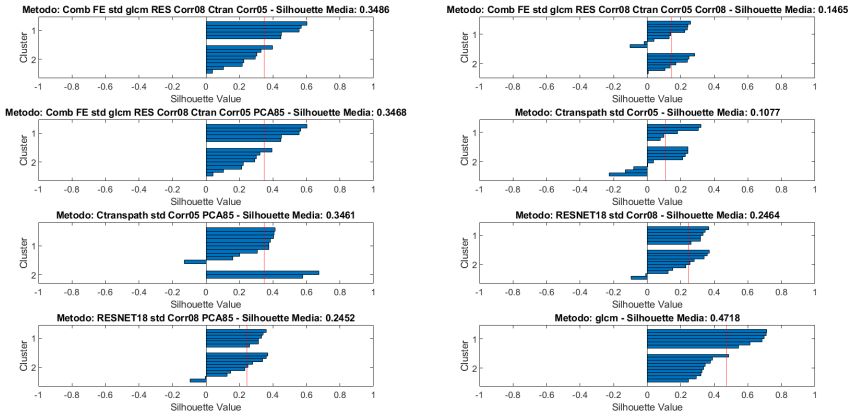


**Figura 3.12:** valori di silhouette ottenuti con i vari metodi divisione nei due cluster principali

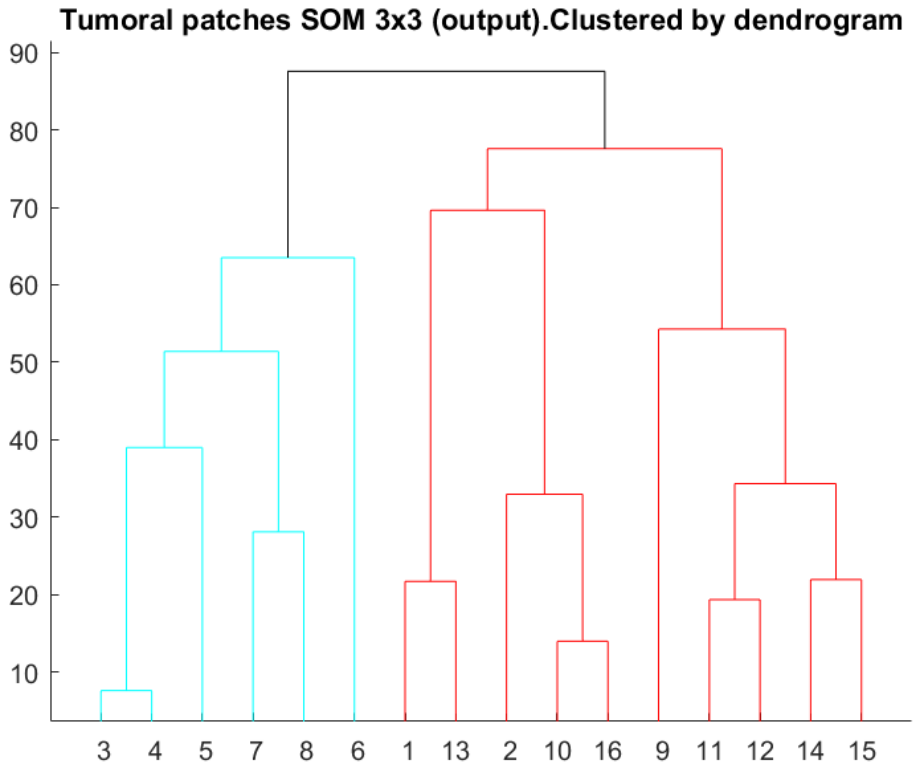


**Figura 3.13:** Miglior raggruppamento ResNet18 Corr 0.8

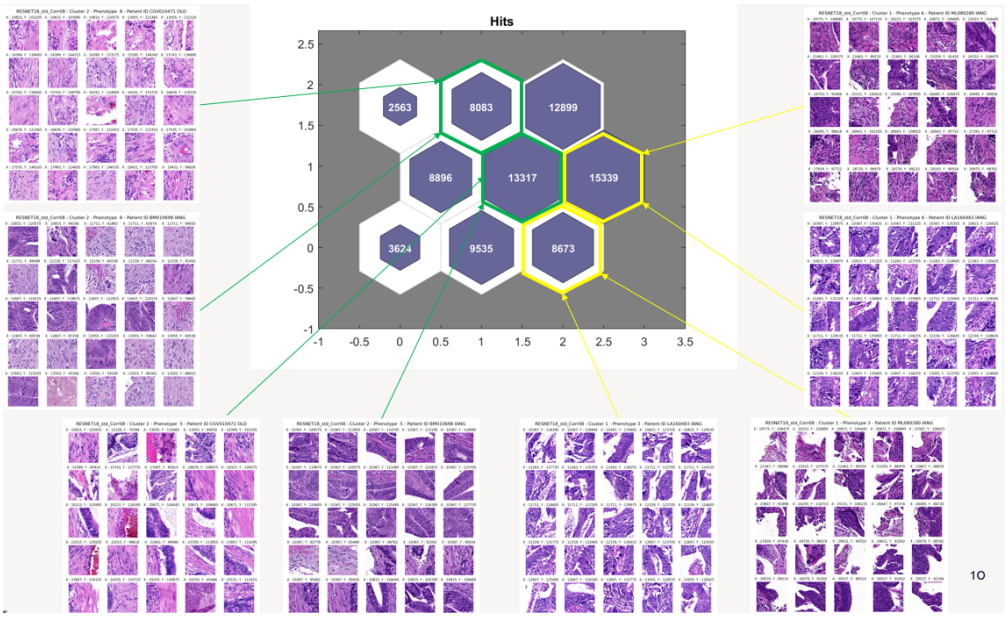


**Figura 3.14:** Fenotipi discriminanti evidenziati dall aggregazione delle caratteristiche estratte con ResNet18.

# Capitolo 4

# Conclusioni

In questo studio, abbiamo esplorato diverse strategie di estrazione e riduzione delle caratteristiche per analizzare immagini istopatologiche colorate con ematossilina ed eosina. L'obiettivo era identificare pattern comuni tra i pazienti, con particolare attenzione alla possibilità di raggruppare alcuni giovani con gli anziani, come suggerito dall'oncologo.

Tra le varie configurazioni analizzate, l'estrazione delle caratteristiche tramite **ResNet-18** con una soglia di correlazione di **0.8** è risultata la più promettente nel raggruppare i pazienti secondo il pattern previsto. Questo risultato suggerisce che le caratteristiche apprese dalla ResNet-18, dopo una riduzione basata sulla correlazione, potrebbero catturare informazioni rilevanti per la discriminazione fenotipica.

Tuttavia, i risultati devono essere interpretati con cautela per diverse ragioni:

1. **Mancanza di Confronto Clinico**: Non è stato possibile validare i raggruppamenti identificati confrontandoli con le caratteristiche cliniche dei pazienti o con ulteriori indicazioni fornite dall'oncologo.

2. **Limiti della Validazione Interna**: Sebbene l'indice di silhouette e altre metriche abbiano fornito indicazioni utili sulla coerenza dei cluster, la validità biologica e clinica di questi raggruppamenti rimane incerta.

3. **Performance delle Altre Configurazioni**: Configurazioni alternative, come la combinazione di più estrazioni o l'uso della PCA, non hanno prodotto miglioramenti significativi rispetto all'approccio con ResNet-18 e correlazione a 0.8.

## 4.1 Prospettive Future

Per rafforzare le conclusioni di questo studio e incrementarne il valore clinico:

- **Confronto con Dati Clinici**: È fondamentale effettuare un confronto diretto tra i gruppi identificati e le caratteristiche cliniche dei pazienti, in collaborazione con oncologi esperti.
- **Valutazione su Dataset Più Ampi**: Ripetere l'analisi su dataset di maggiore dimensione potrebbe confermare la generalizzabilità dei risultati.
- **Integrazione con Dati Multi-Omici**: L'integrazione di dati genetici o molecolari potrebbe fornire ulteriori informazioni per validare i raggruppamenti osservati.

# Bibliografia


[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre e Ahmedin Jemal. «Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424 (cit. a p. 1).

[2] Al B Benson et al. «NCCN guidelines insights: colon cancer, version 2.2018». In: *Journal of the National Comprehensive Cancer Network* 16.4 (2018), pp. 359–369 (cit. a p. 1).

[3] Fang Zhang et al. «Predicting treatment response to neoadjuvant chemoradiotherapy in local advanced rectal cancer by biopsy digital pathology image features». In: *Clinical and translational medicine* 10.2 (2020) (cit. alle pp. 1, 16).

[4] Andrew MD Wolf et al. «Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society». In: *CA: a cancer journal for clinicians* 68.4 (2018), pp. 250–281 (cit. a p. 2).

[5] M Malvezzi, P Bertuccio, F Levi, C La Vecchia e E Negri. «European cancer mortality predictions for the year 2014». In: *Annals of oncology* 25.8 (2014), pp. 1650–1656 (cit. a p. 2).

[6] Gianluca Mauri, Andrea Sartore-Bianchi, Antonio-Giampiero Russo, Silvia Marsoni, Alberto Bardelli e Salvatore Siena. «Early-onset colorectal cancer in young individuals». In: *Molecular oncology* 13.2 (2019), pp. 109–131 (cit. a p. 2).

[7] AG Russo, A Andreano, A Sartore-Bianchi, G Mauri, A Decarli e S Siena. «Increased incidence of colon cancer among individuals younger than 50 years: a 17 years analysis from the cancer registry of the municipality of Milan, Italy». In: *Cancer epidemiology* 60 (2019), pp. 134–140 (cit. alle pp. 3, 4).

[8] Xingzhi Yue, Neofytos Dimitriou e Ognjen Arandjelovic. «Colorectal cancer outcome prediction from H&E whole slide images using machine learning and automatically inferred phenotype profiles». In: *arXiv preprint arXiv:1902.03582* (2019) (cit. a p. 5).

[9] Jafar Nouri Nojadeh, Shahin Behrouz Sharif e Ebrahim Sakhinia. «Microsatellite instability in colorectal cancer». In: *EXCLI journal* 17 (2018), p. 159 (cit. a p. 10).

[10] Wei Xu, Yazhou He, Yuming Wang, Xue Li, Jane Young, John PA Ioannidis, Malcolm G Dunlop e Evropi Theodoratou. «Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies». In: *BMC medicine* 18 (2020), pp. 1–19 (cit. a p. 11).

[11] Dongfeng Tan e Henry T Lynch. *Principles of molecular diagnostics and personalized cancer medicine*. Lippincott Williams & Wilkins, 2012 (cit. a p. 11).

[12] Stephanie Robertson, Hossein Azizpour, Kevin Smith e Johan Hartman. «Digital image analysis in breast pathology—from image processing techniques to artificial intelligence». In: *Translational Research* 194 (2018), pp. 19–35 (cit. a p. 11).

[13] Zsuzsanna Varga et al. «How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast-and Gynecopathologists». In: *PloS one* 7.5 (2012), e37379 (cit. a p. 11).

[14] Mei-Yin C Polley et al. «An international Ki67 reproducibility study». In: *Journal of the National Cancer Institute* 105.24 (2013), pp. 1897–1906 (cit. a p. 11).

[15] Ewert Bengtsson. «The measuring of cell features.» In: *Analytical and quantitative cytology and histology* 9.3 (1987), pp. 212–217 (cit. a p. 11).

[16] SH Ong, XC Jin, R Sinniah et al. «Image analysis of tissue sections». In: *Computers in biology and medicine* 26.3 (1996), pp. 269–279 (cit. a p. 11).

[17] Balázs Acs, Mattias Rantalainen e Johan Hartman. «Artificial intelligence as the next step towards precision pathology». In: *Journal of internal medicine* 288.1 (2020), pp. 62–81 (cit. alle pp. 11, 12).

[18] David F Steiner, Po-Hsuan Cameron Chen e Craig H Mermel. «Closing the translation gap: AI applications in digital pathology». In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1875.1 (2021), p. 188452 (cit. a p. 15).

[19] Erik Reinhard, Michael Adhikhmin, Bruce Gooch e Peter Shirley. «Color transfer between images». In: *IEEE Computer graphics and applications* 21.5 (2001), pp. 34–41 (cit. a p. 15).

[20] Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen AWM van der Laak e Peter HN de With. «Stain normalization of histopathology images using generative adversarial networks». In: *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 573–577 (cit. a p. 15).

[21] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt e Nancy E Thomas. «A method for normalizing histology slides for quantitative analysis». In: *2009 IEEE international symposium on biomedical imaging: from nano to macro*. IEEE. 2009, pp. 1107–1110 (cit. alle pp. 15, 20).

[22] Wei-Wen Hsu, Jing-Ming Guo, Linmin Pei, Ling-An Chiang, Yao-Feng Li, Jui-Chien Hsiao, Rivka Colen e Peizhong Liu. «A weakly supervised deep learning-based method for glioma subtype classification using WSI and mpMRIs». In: *Scientific Reports* 12.1 (2022), p. 6111 (cit. a p. 16).

[23] Charlie Saillard, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoit Schmauch e Simon Jegou. «Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers». In: *arXiv preprint arXiv:2109.05819* (2021) (cit. a p. 16).

[24] Korsuk Sirinukunwattana et al. «Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning». In: *Gut* 70.3 (2021), pp. 544–554 (cit. a p. 16).

[25] Jakob Nikolas Kather et al. «Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study». In: *PLoS medicine* 16.1 (2019), e1002730 (cit. a p. 21).

[26] Joost JM Van Griethuysen et al. «Computational radiomics system to decode the radiographic phenotype». In: *Cancer research* 77.21 (2017), e104–e107 (cit. a p. 22).

[27] Rui Cao et al. «Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer». In: *Theranostics* 10.24 (2020), p. 11080 (cit. a p. 27).

[28] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang e Xiao Han. «Transformer-based unsupervised contrastive learning for histopathological image classification». In: *Medical image analysis* 81 (2022), p. 102559 (cit. a p. 29).

[29] Angel Cruz-Roa, Juan C Caicedo e Fabio A González. «Visual pattern mining in histology image collections using bag of features». In: *Artificial intelligence in medicine* 52.2 (2011), pp. 91–106 (cit. a p. 30).