
Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Biomedica
A.a. 2022/2023
Sessione di Laurea Luglio 2023

# Sviluppo di un sistema di diagnosi automatizzata per il riconoscimento della metaplasia intestinale gastrica

da immagini endoscopiche

Relatori:
Prof. Monica Visintin
Prof. Guido Pagana

Candidato:
Loris Carta

# Indice

# Elenco delle figure

# Elenco delle tabelle

# 1. Sommario

Al giorno d'oggi la medicina si avvale sempre più spesso dell'intelligenza artificiale, in ogni suo ambito. Questa disciplina viene impiegata per garantire maggiore precisione ed efficienza, in particolar modo nelle tecniche di diagnostica per immagini.
In questo lavoro è stata valutata la possibilità di utilizzare l'intelligenza artificiale nella diagnosi della metaplasia intestinale gastrica, un alto fattore di rischio per il carcinoma gastrico, nonché un suo precursore. Esso è il sesto tumore più comune al mondo con un tasso di sopravvivenza a 5 anni inferiore al 40%. La migliore arma di prevenzione è la diagnosi precoce della patologia. A tale scopo è stato sviluppato un sistema di diagnosi automatizzata che, a partire da immagini endoscopiche, riconosca la presenza della metaplasia.
Il campione di immagini utilizzato è costituito da 95 acquisizioni gastroscopiche RGB, provenienti da 25 individui, di sesso maschile e femminile. Le immagini sono state acquisite tramite sonda endoscopica digitale i-scan della società PENTAX Medical e sono state fornite dall'azienda Ospedaliera Ordine Mauriziano di Torino, Ospedale Umberto I.
Le acquisizioni rappresentano porzioni diverse dell'organo gastrico e sono suddivise in immagini provenienti da pazienti sani e da pazienti affetti da metaplasia intestinale gastrica. Quest'ultime sono affiancate da segmentazioni manuali effettuate da un medico, che circoscrivono la regione metaplastica.
Le immagini sono processate in modo da estrarre solo le aree di interesse, successivamente vengono ripulite dagli artefatti indesiderati. Le immagini sono poi convertite in bianco-nero e in seguito vengono enfatizzati texture e pattern attraverso la tecnica di *image sharpening*, con l'utilizzo dei gradienti di Sobel.
A valle del *pre-processing* le immagini sono state suddivise in piccole regioni di interesse (ROI) di dimensioni $50 \times 50$ pixel, con una sovrapposizione bidirezionale del 40%.
Per ogni singola ROI sono state estratte 35 *feature*, ridotte successivamente a 32, suddivise in caratteristiche del primo ordine, GLCM (*gray level co-occurrence matrix*), GLRLM (*gray level run length matrix*) e caratteristiche nel dominio della frequenza.
Un classificatore d'insieme di tipo *random forest* è stato addestrato con un dataset di circa 50.600 osservazioni. Il modello è stato validato con la 5-*fold cross-validation* e testato su un dataset di circa 16.900 campioni.
Le maschere di segmentazione in uscita dal classificatore subiscono un processo di pulizia e affinamento tramite l'algoritmo DBSCAN, rimuovendo parte del rumore.
A valle del *post-processing* il set di test ha ottenuto i seguenti risultati. Un'accuratezza nella classificazione delle ROI pari al 90,5%, con una sensibilità dell'88,8%

e una specificità del 92,3%. Le maschere di segmentazione hanno ottenuto un indice Sørensen-Dice medio sull'intero dataset di 0,612. Il dataset di immagini è stato classificato correttamente in circa il 99% dei casi, sbagliando 1 immagine su 95.
Poiché le funzioni e gli algoritmi implementati potrebbero risultare ostici per un utente non esperto in programmazione e calcolo numerico è stata creata un'interfaccia grafica (GUI), semplice e intuitiva.
Nella GUI è stata implementata una mappa di calore per la visualizzazione delle aree potenzialmente metaplastiche, in scala di colore giallo-rosso. Inoltre è stata introdotta un'indicazione semaforica con lo scopo di fornire al medico un'informazione oggettiva sull'entità della lesione metaplastica, basata sull'estensione dell'anomalia rilevata.

# 2. Introduzione

Lo studio dell'apparato digerente umano ha origini antichissime, già gli antichi egizi si erano interessati ad esso. Nell'antica Grecia al tempo di Ippocrate si producevano farmaci per i disturbi gastrici.
Durante l'Impero Romano, nel II secolo, un medico di nome Claudius Galenus, meglio noto come Galeno di Pergamo, si interessò approfonditamente all'apparato digerente e completò gli insegnamenti della tradizione greca.
Egli riteneva che la digestione consistesse solamente nell'azione meccanica della triturazione del cibo e che lo stomaco fosse un semplice contenitore dove avveniva l'assorbimento degli alimenti. La digestione veniva vista infatti come somma delle quattro facoltà dell'organismo umano: attrazione delle sostanze utili, conservazione, trasformazione ed espulsione delle sostanze estranee. [1].
Questo concetto rimase pressoché immutato fino al XVIII secolo, quando un italiano, Lazzaro Spallanzani iniziò ad ignorare le antiche teorie di Galeno e nel 1780 trovò un'evidenza sperimentale che saliva e succhi gastrici agivano attivamente sugli alimenti [2].
Così alla fine del secolo iniziano ad essere pubblicati i primi trattati sul canale digerente e la digestione, in particolare dal tedesco Johann Georg Zimmermann e dall'austriaco Maximilian Stoll, che scrissero rispettivamente un trattato sulla dissenteria e una descrizione del cancro alla cistifellea [2].
I maggiori passi avanti si ebbero nel XIX secolo quando avvenne il primo tentativo di ispezionare l'interno del corpo umano vivente mediante una sonda, il tedesco Philipp Bozzini nel 1804 inventò infatti il cistoscopio, un primitivo endoscopio con il quale fu capace di ispezionare vescica, uretra, retto e cavità nasali [2].
Nel 1823 William Prout scoprì che i succhi gastrici erano composti da acido cloridrico, ma solo nel 1868 vi fu il primo vero progresso, con l'invenzione del primo rudimentale gastroscopio. Esso consisteva in un tubo rigido che il paziente doveva ingoiare, questo venne poi perfezionato agli inizi del XX secolo ad opera di William Hill [2].
Grazie a questi nuovi progressi tecnologici, Jesse Francis McClendon, nel 1915 riuscì ad effettuare la prima misurazione del pH dello stomaco umano, in situ [2].
Per il primo gastroscopio flessibile occorre invece attendere fino al 1932, quando il fisico tedesco Rudolf Schindler sviluppò e perfezionò il primo vero gastroscopio. Grazie a questa sua invenzione viene ampiamente nominato come il padre della gastroscopia. Egli, infatti, già nel 1940 aveva raccolto oltre 22.000 casi [2].
Nel 1957 venne inventato il primo endoscopio flessibile a fibre ottiche, ad opera di Basil Hirschowitz, che aprì le basi alla moderna gastroenterologia [2].

Già nel XVIII e nel XIX secolo si erano registrate descrizioni post mortem di carcinoma gastrico, ad opera di Giovanni Battista Morgagni nel 1761, Matthew Baillie nel 1793, Carl von Rokitansky nel 1842; nel 1881 venne eseguita la prima gastroenterostomia di successo per carcinoma gastrico ad opera di Anton Wolfer. Ma fino al XX secolo era molto raro diagnosticare un cancro gastrico prima che diventasse terminale [2].
Grazie alle scoperte dei sempre più moderni gastro endoscopi e all'impiego dei raggi X (usati già dal 1914), furono fatti enormi progressi in fase di diagnosi precoce dei carcinomi gastrici che misero le fondamenta per la moderna oncologia [2].

## 2.1 Lo stomaco

Lo stomaco è la porzione del canale digerente in cui si accumulano e sostano gli alimenti temporaneamente, prima dell'assorbimento intestinale. Gli alimenti vengono infatti posti sotto l'azione digestiva tramite succo gastrico e miscelati tramite azione meccanica. Il tempo di stazionamento all'interno dell'organo gastrico dipende invece dalla tipologia degli alimenti [3].
Presenta due orifizi, uno superiore chiamato cardias o orifizio cardiale che lo collega all'esofago, uno inferiore, noto come orifizio pilorico o piloro. Quest'ultimo pone invece in comunicazione lo stomaco con il duodeno.
Lo stomaco può essere anatomicamente diviso in, cardias o parte cardiale, il fondo dello stomaco, il corpo dello stomaco, la parte pilorica, l'antro pilorico, il canale pilorico e il piloro [3].

- Il **cardias** è una regione molto ristretta e si estende nell'area più prossima all'esofago.
- Il **fondo dello stomaco** è la parte più alta dello stomaco ed è noto come grande tuberosità.
- Il **corpo dello stomaco** è la parte più estesa dello stomaco e congiunge il fondo con il canale pilorico
- La **parte pilorica** ha una forma conica e può essere suddivisa in **antro pilorico** che è la porzione più prossimale, **canale pilorico** che è la parte più distale, e il **piloro**, la sua parte terminale.

In **Figura 2.1** è possibile osservare la caratteristica forma uncinata dello stomaco, vengono inoltre evidenziate le principali regioni dell'organo gastrico.

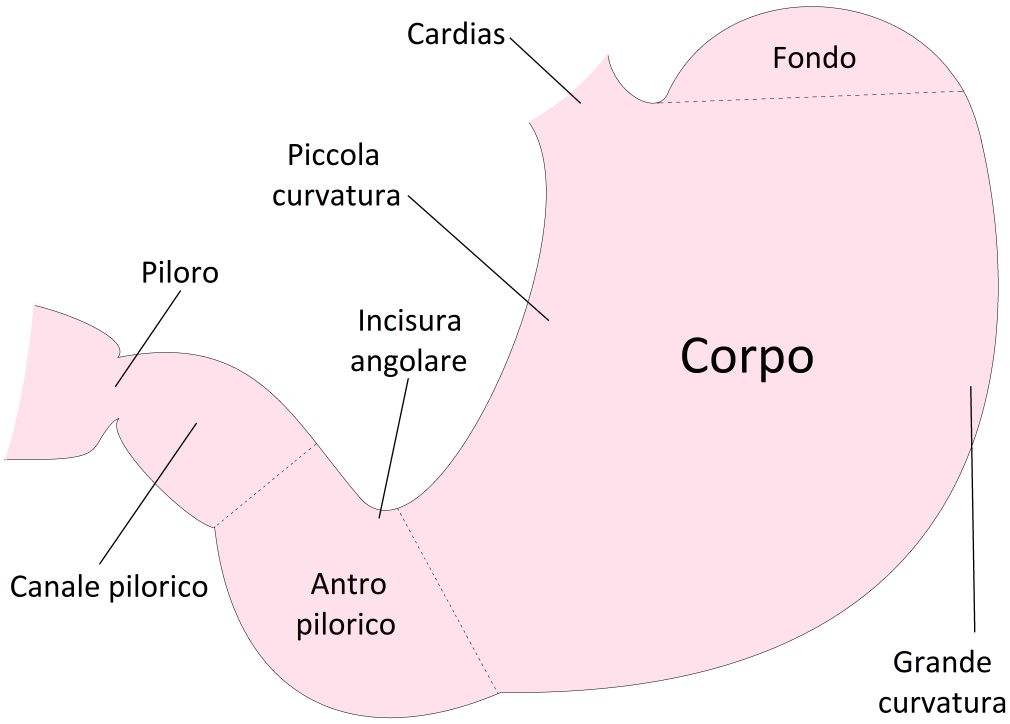



**Figura 2.1: Lo stomaco.** *L'organo gastrico, dalla caratteristica forma uncinata, è suddiviso in quattro regioni principali, parte cardiale, fondo dello stomaco, corpo dello stomaco, parte pilorica.*

I vari strati della parte gastrica sono rappresentati in **Figura 2.2**, si noti la struttura articolata composta dalle aree e dalle fossette gastriche.
La superficie interna dello stomaco si presenta di un colore grigio roseo, ma, durante la fase digestiva, a causa di un maggiore apporto di sangue, tende a diventare più rosso. La superficie interna è completamente segnata da pieghe gastriche determinate dal sollevamento della tonaca mucosa e dalla sottomucosa. Hanno una direzione prevalente lungo l'asse longitudinale e formano una specie di reticolo, essendo anastomizzate fra di loro. Queste pieghe non sono permanenti ma scompaiono con la dilatazione dello stomaco. Sulla piccola curvatura si trova un solco delimitato da due pieghe longitudinali, nota come via gastrica breve [3].
Vi sono invece altri solchi permanenti, visibili quando l'organo è disteso, essi formano un reticolo fine e delimitano piccole aree di 2-4 millimetri di diametro, chiamate areole gastriche. La superficie di questo reticolo è punteggiata da piccoli affossamenti, chiamati appunto fossette gastriche, e al fondo di queste si trovano le ghiandole gastriche. Queste sono molto numerose, in ogni stomaco vi sono circa 15 milioni di ghiandole gastriche e circa 3,5 milioni di fossette gastriche [3].
Le fossette gastriche sono separate da una tonaca mucosa che si solleva in sottili prominenze, dette creste gastriche e possono essere cilindriche o laminari.
Il limite fra tonaca mucosa gastrica e tonaca mucosa esofagea è segnato da un orlo dentellato, anulare, sito in corrispondenza del cardias. Questo è noto come orifizio cardiale, ha un asse verticale e forma ovalare [3].

L'orifizio pilorico è circoscritto da una piega circolare e che sporge in cavità, nota anche come valvola pilorica. Questa è formata da un sollevamento della superficie interna dello stomaco e ha una forma quasi circolare sul piano frontale. La valvola pilorica è formata da tonaca mucosa, tonaca sottomucosa e tonaca muscolare; quest'ultimo è di forma circolare e costituisce il muscolo sfintere pilorico. In via generale il piloro presenta una forma triangolare dove il lato duodenale cade in direzione della superficie intestinale [4].
La parete gastrica ha uno spessore di circa 0,5-0,8 centimetri ed è formata da quattro strati, la tonaca mucosa, la tonaca sottomucosa, la tonaca muscolare e la tonaca sierosa [5].
La tonaca mucosa a sua volta è divisa in epitelio superficiale o di rivestimento, lamina propria e muscolaris mucosae.
L'epitelio superficiale ricopre le creste gastriche ed è costituito da cellule prismatiche dotate di microvilli sulla superficie, il cui nucleo si trova nella parte basale della cellula. Queste secernono un muco ricco di proteoglicani, proteggendo lo stomaco dall'azione degli enzimi e dell'acido. Il secreto neutralizza l'elevata natura acida del contenuto gastrico, innalzando il pH [3].
La lamina propria è costituita da tessuto lasso, ricco di vasi. È suddiviso in due parti, la prima forma le creste gastriche ed è superficiale, la seconda forma diversi tipi di ghiandole gastriche si trova nella parte più profonda. Queste due parti sono ricche di macrofagi, granulociti e plasmacellule e linfociti che spesso formano noduli solitari [3].
Le ghiandole gastriche si trovano sempre nella lamina propria e si dividono in ghiandole cardiali, ghiandole del fondo e del corpo e ghiandole piloriche.
Le ghiandole cardiali occupano area molto ristretta e secernono glicoproteine neutre ma non producono acido cloridrico o enzimi [3].
Le ghiandole del fondo e del corpo sono quelle più abbondanti e occupano la maggior parte della tonaca mucosa. Queste ghiandole invece producono succo gastrico, ricco di enzimi, a pH molto basso. Sono composte da diverse cellule come quelle indifferenziate o staminali che occupano la parte prossimale della ghiandola. Queste provvedono al rimpiazzo delle cellule dell'epitelio superficiale e della ghiandola stessa [3].
Sono presenti poi le cellule del colletto che si localizzano invece nella parte più alta della ghiandola. Sono composte da numerosi granuli di muco nella loro parte apicale mentre presentano il nucleo nella parte basale; producono muco costituito da proteoglicani [3].
Si hanno ancora le cellule principali, dette anche adelomorfe, che predominano nella parte intermedia e nel fondo delle ghiandole. Producono una secrezione sierosa, il pepsinogeno, un precursore della pepsina ma anche la rennina, una proteina che aiuta nella digestione del latte.
Vi sono poi le cellule parietali o delomorfe, note anche come cellule di rivestimento o ancor meglio come ossintiche [3]. Queste cellule sono le più caratteristiche, in quanto producono acido cloridrico, indispensabile per la digestione; si trovano solo nella tonaca mucosa del corpo e del fondo dello stomaco.
Hanno un profilo ovale, sovente sporgono dalla superficie esterna e sono costituite da particolari canalicoli intracellulari [3]. In queste cellule è presente anche il fattore

intrinseco, esso si lega alla vitamina B12 degli alimenti e ne aiuta l'assorbimento nell'intestino [3].
Sono presenti inoltre cellule endocrine, site nella parte intermedia dei tubuli ghiandolari e che secernono diversi ormoni.
Le ghiandole piloriche nella omonima parte anatomica dello stomaco e si presentano con fossette gastriche particolarmente profonde e creste alte e sottili. Queste ghiandole secernono una sostanza che protegge l'antro pilorico e sono costituite da cellule che secernono gastrina, che stimola la produzione di acido cloridrico [3].
La muscolaris mucosae della tonaca mucosa è formata da cellule muscolari lisce e la loro contrazione favorisce l'espulsione del secreto.
La tonaca sottomucosa è costituita da tessuto connettivo di tipo lasso e cellule adipose, e aderisce sia alla tonaca mucosa che a quella muscolare [3].
La tonaca muscolare si presenta come uno strato particolarmente spesso. Internamente è costituito da uno strato circolare mentre esternamente da uno strato longitudinale a cui si aggiunge uno strato di fibre oblique, necessario all'intenso sforzo peristaltico. Nella regione pilorica è particolarmente sviluppato lo strato circolare interno che forma il muscolo sfintere pilorico.
La tonaca sierosa è costituita dal peritoneo e da uno strato sottomesentiale ed è unito alla tonaca muscolare dalla tela sottosierosa [3].

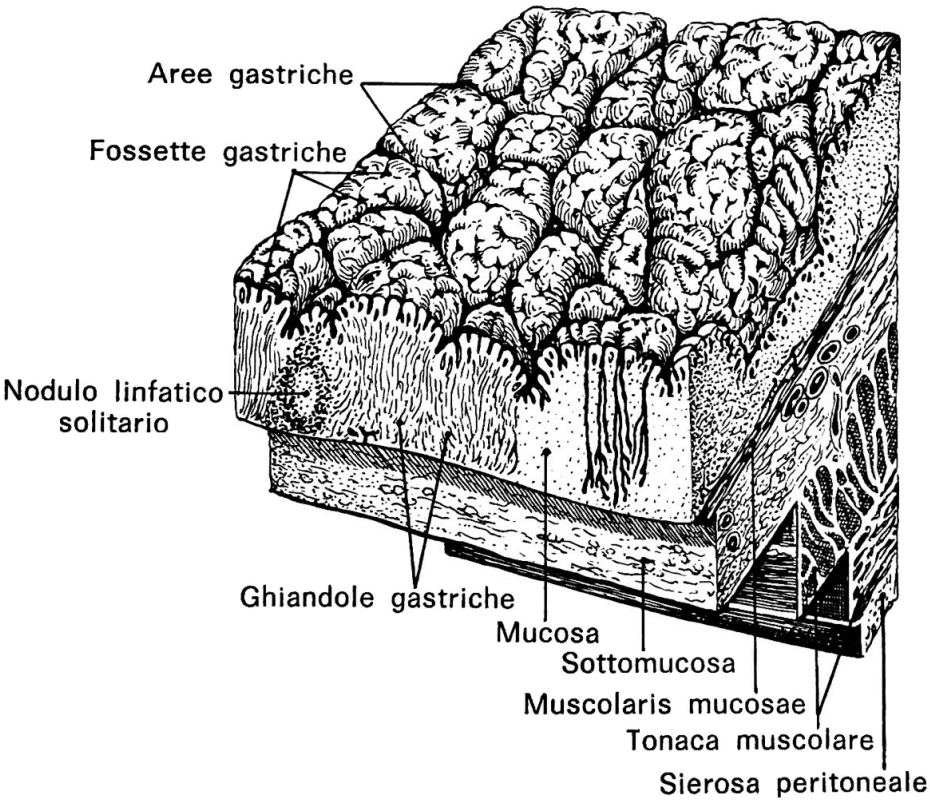



**Figura 2.2: Il tessuto gastrico.** *La parete interna dell'organo gastrico è molto complessa, nella parte più superficiale, la mucosa, sono presenti le ghiandole gastriche, mentre più in profondità si trovano la sottomucosa, la tonaca muscolare e la sierosa peritoneale, immagine tratta da [4].*

## 2.2 I tumori e il carcinoma gastrico

In ambito mondiale, vi è circa il 20% di possibilità di contrarre il cancro durante la propria vita, per un'età inferiore ai 75 anni, con un rischio di morire per cancro che arriva al 10% [6].
Circa un quarto dei casi di cancro stimati nel mondo è a carico del continente europeo, nonostante questo rappresenti meno del 9% della popolazione mondiale. Gli organi che maggiormente sono colpiti da questa patologia sono il colon e i polmoni, entrambi rappresentano una percentuale superiore al 12%, sul totale dei tumori. Il tumore allo stomaco, per entrambi i sessi, rappresenta il 3,4% del totale dei casi. Nonostante possa sembrare un valore esiguo, esso rappresenta il sesto organo più colpito, preceduto solo da vescica (5%), melanoma della pelle (3,7%) e reni (3,5%) [7].
A livello mondiale il cancro allo stomaco colpisce 1,1 milioni di persone ogni anno, con una mortalità globale di 800.000 persone. Il continente più colpito è l'Asia,

principalmente in Cina, dove l'aspettativa di vita, a cinque anni dalla diagnosi, è inferiore al 20% [8].
La comparsa della patologia cresce con l'aumentare dell'età e risulta veramente raro per individui giovani, con un'età inferiore ai 45 anni [8].
Nei paesi mediorientali il tumore allo stomaco è il cancro più comunemente diagnosticato, nonché la prima causa di morte legata a patologie tumorali in Medioriente e gran parte dell'America meridionale. Il tasso di mortalità in Europa, Nord America e Africa risulta invece essere inferiore [8].
Nel 2018, solo in Europa, sono stati registrati oltre 133.000 nuovi casi, con un tasso di mortalità molto elevato, circa il 77% dei casi. Il cancro allo stomaco colpisce in prevalenza il genere maschile, con un rapporto di circa 60%-40% [7].
Si stima che in Italia vengano invece diagnosticati circa 14.500 nuovi casi all'anno, di cui oltre 8.000 sono di genere maschile [9].
La **Tabella 2.1** presenta le stime di GLOBOCAN 2020 sull'incidenza annuale dei tumori in Italia e sono mostrati, in ordine di occorrenza, i primi 15 organi colpiti. I dati sono disponibili sul sito ufficiale del *Global Cancer Observatory* (GCO) dell'*International Agency of Research on Cancer* (IARC) [9]. Il tumore allo stomaco risulta essere il sesto per numero di casi registrati e per numero di morti.

| Cancer | Incidence | Mortality | Mortality % |
|---|---|---|---|
| Breast | 55133 | 12633 | 23% |
| Lung | 41953 | 33602 | 80% |
| Prostate | 39317 | 6902 | 18% |
| Colon | 33957 | 16629 | 49% |
| Bladder | 28336 | 7108 | 25% |
| **Stomach** | **14372** | **8853** | **62%** |
| Pancreas | 14155 | 12917 | 91% |
| Non-Hodgkin lymphoma | 14032 | 5175 | 37% |
| Rectum | 13326 | 4812 | 36% |
| Melanoma of skin | 12515 | 2224 | 18% |
| Kidney | 12306 | 4280 | 35% |
| Thyroid | 12288 | 526 | 4% |
| Liver | 11739 | 9798 | 83% |
| Corpus uteri | 10013 | 2152 | 21% |
| Leukaemia | 9352 | 6324 | 68% |

**Tabella 2.1: Incidenza e mortalità annuale dei tumori in Italia.** *I dati in tabella sono stime effettuate da GLOBOCAN 2020 sull'incidenza dei tumori in Italia. Il tumore allo stomaco è sesto per incidenza e numero di morti. Ha però un'alta percentuale di mortalità (62%), preceduto solo da pancreas, fegato, polmoni e dalla leucemia. I dati sono stati prelevati dal sito ufficiale del GCO dell'IARC [9].*

Solo in Italia, nel 2020, GLOBOCAN 2020 ha stimato quasi 383.000 nuovi casi di tumore, considerando entrambi i sessi. La maggior parte interessa il cancro al seno (oltre il 14% dei casi) seguiti da polmoni e prostata, attorno al 10-11%. In **Figura 2.3** sono rappresentati i primi sette organi interessati, per ordine di incidenza, in Italia.

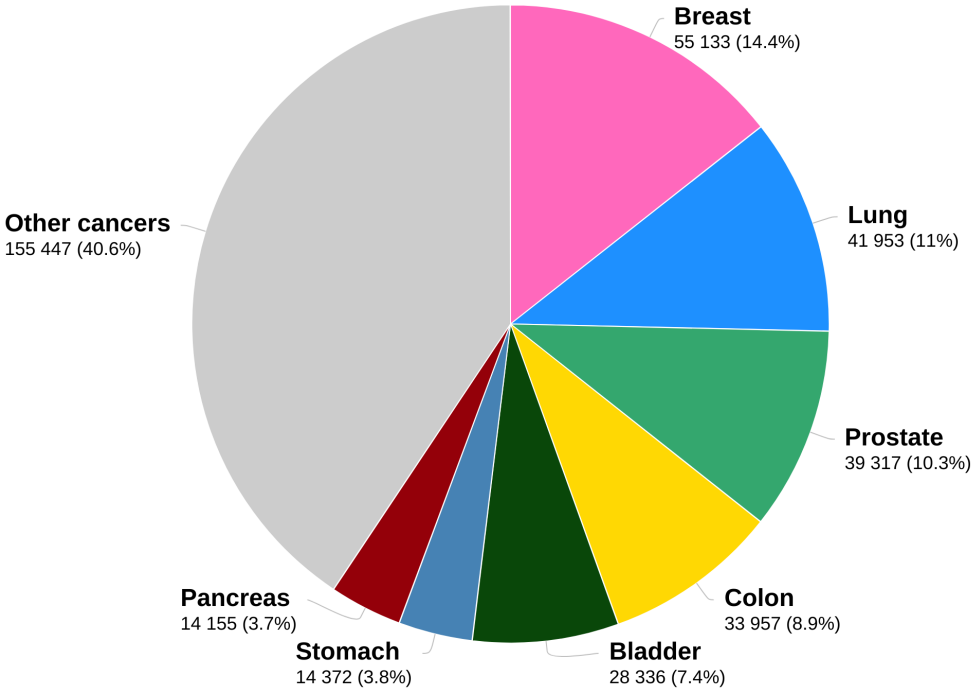



**Figura 2.3: Percentuale di incidenza dei tumori in Italia.** *In Italia, il cancro con maggiore incidenza è quello al seno, con oltre 55.000 casi stimati nel 2020. Considerando entrambi i sessi, il tumore allo stomaco è il sesto per incidenza, ricade infatti nel 3.8% dei casi. Grafico generato dal dal sito ufficiale del GCO dell'IARC [9], con dati stimati da GLOBOCAN 2020.*

Nonostante si registri una leggera flessione a livello mondiale della mortalità, il cancro allo stomaco è tuttora una vasta fetta del carico globale di cancro e molti fattori di rischio non sono stati completamente compresi.
Il tumore gastrico è una patologia multifattoriale e quindi comprende sia i fattori di rischio ambientali che lo stile di vita. Tra i vari fattori si possono citare, quelli dovuti alla dieta e alimentazione, un basso status economico e sociale, elevata assunzione di cibi molto salati, assunzione di cibi affumicati e che contengono nitriti e nitrati, radiazioni, reflusso gastroesofageo, scarsa o assente attività fisica, un basso consumo di frutta, verdura e di vegetali in genere [8]. Ma anche elevata assunzione di fibre, di amidi, cibi grassi o conservati sott'olio, di alcol, la pratica del fumo e quindi consumo di tabacco. Quest'ultimo risulta la seconda causa di cancerogenesi e da solo aumenta del 40%-60% il rischio di contrarre il tumore allo stomaco [10].
Un ruolo importante nella cancerogenesi lo ricopre l'infezione da *Helicobacter pylori*, è un batterio che sopravvive ai pH estremamente acidi dell'ambiente gastrico [11]. Esso rappresenta il primo fattore di rischio per patologia, aumentando la probabilità di contrarre il tumore di oltre 2.5 volte rispetto ad un soggetto sano [10].
A seguito di eradicazione del batterio tramite trattamento antibiotico, la neoplasia recede nell'oltre il 50% dei casi. Infatti, il batterio non agisce direttamente sulle cellule, ma stimola i linfociti T a produrre citochine, fondamentali per la diffusione delle cellule cancerose [5].
Un altro fattore di rischio importante è avere la predisposizione ereditaria. Infatti, alterazioni di alcuni geni come p53 e APC, possono portare all'insorgenza della

patologia, in questo caso si parlerebbe di sindrome di Lynch di tipo II [12]. Altri importanti fattori di rischio sono legati ad altre patologie presenti quali, l'AIDS, il diabete mellito, ulcera peptidica, l'anemia perniciosa, gastrite cronica atrofica, poliposi gastrica e metaplasia intestinale.
Nonostante la conoscenza dei sopra citati fattori di rischio, l'eziologia del tumore gastrico non è stata ancora totalmente chiarita e risulta essere di alto interesse scientifico, soprattutto nell'ambito della ricerca [8].
La diffusione della malattia può avvenire con differenti meccanismi, direttamente o indirettamente: dall'esofago o dal peritoneo, quindi in maniera diretta, per via linfatica ai linfonodi o per via ematica, quindi in maniera indiretta. In quest'ultimo caso può portare alla metastasi del fegato, delle ovaie, dei polmoni o delle ossa [13].
Circa il 90% di tutti i tumori dello stomaco sono adenocarcinomi, mentre altri tipi come linfoma, sarcoma, tumori neuroendocrini, risultano essere molto più rari [8]. Per adenocarcinoma si intende un tumore maligno dell'epitelio ghiandolare [14], dove la loro citoarchitettura ricorda il tessuto ghiandolare di origine. Questi si differenziano dagli adenomi, che sono invece tumori benigni dell'epitelio ghiandolare [14].
Il tumore allo stomaco può essere classificato in base a differenti criteri. Può essere suddiviso in base al sito interessato, alla posizione anatomica, al tipo di stadio e in base al tipo istologico. L'adenocarcinoma gastrico viene classificato in due diversi siti: cancro dello stomaco cardiale, che deriva dalla parte superiore dello stomaco, appunto il cardias, e cancro dello stomaco non cardiale, che invece deriva dalle altre parti dello stomaco [8].
Il cancro allo stomaco può essere anatomicamente differenziato in distale o prossimale. Nel primo caso la neoplasia si trova in posizione anatomica vicina al piloro o al duodeno, nel secondo caso più vicina all'esofago. Fra tutti, il tumore della regione denominata antro è il più comune e rappresenta oltre il 50% di tutti i tumori dello stomaco [13].
La malattia può essere suddivisa in due stadi, iniziale o avanzato. Il cancro gastrico iniziale ha più ampie probabilità di guarigione, con una sopravvivenza a cinque anni che arriva oltre il 90%. A questo stadio la patologia non coinvolge vasi o linfonodi e non forma ulcere, possiede un'ampiezza massima di 2 centimetri di diametro; inoltre presenta un aspetto ben differenziato al microscopio [15].
Proprio per questo motivo la ricerca in tale ambito è in continua crescita, una diagnosi precoce della patologia è fondamentale e aiuterebbe a ridurre la progressione del tumore, una risposta tempestiva sarebbe più efficace per contrastare la malattia.
Gli adenocarcinomi allo stomaco sono istologicamente suddivisi in due tipi: adenocarcinomi diffusi e adenocarcinomi intestinali, questi due tipi di tessuto differiscono per peculiarità epidemiologiche, come il rapporto tra i sessi e l'età alla diagnosi [13].
Il tumore dello stomaco di tipo diffuso ha una frequenza leggermente inferiore. Considerata un età media di 45 anni, questo tipo di neoplasia colpisce in maniera indifferente donne e uomini. Si genera dalla normale mucosa gastrica priva di metaplasia e penetra in maniera molto profonda negli strati tissutali delle pareti dello stomaco. Successivamente questa si espande in direzione laterale generando sovente ulcere. Queste cellule hanno una morfologia molto particolare e vengono spesso chiamate

"ad anello con castone", visto che la loro principale caratteristica istologica è la presenza di cellule simili a un anello con una gemma incastonata [13].
Il tumore dello stomaco più frequente è quello di tipo intestinale; esso è associato alla trasformazione dell'epitelio gastrico in epitelio intestinale, condizione meglio nota come metaplasia intestinale. Questo tumore colpisce in prevalenza il sesso maschile e di solito, si presenta come diffusa infiltrazione della parete del viscere oppure con formazioni simil-polipoidi [13].

## 2.3 Le trasformazioni metaplastiche

Con in termine metaplasia si intende una trasformazione da un tipo cellulare differenziato ad un altro tipo cellulare differenziato, nella vita post natale [16].
Per differenziazione si intende la capacità di una cellula di formare diversi tipi di citotipi, che andranno poi a costituire i vari tessuti [14]. Pur rimanendo a DNA costante, ogni citotipo acquisisce differenti caratteristiche funzionali e morfologiche. Si potrebbe interpretare come la funzionale repressione di alcuni geni in favore dell'attivazione di altri, a seguito di determinati fattori di trascrizione [14].
Sebbene il carattere adattivo dei cambiamenti sia ben noto, il meccanismo con il quale avviene la trasformazione non è stato del tutto chiarito. Il passaggio da un tipo cellulare all'altro può essere parte di un normale processo di maturazione della cellula oppure causato da uno stimolo esterno, anomalo. Se quest'ultimo viene rimosso o cessa, i tessuti tornano al loro schema di differenziazione normale. La metaplasia non è considerata un vero e proprio cancro, non è pertanto sinonimo di neoplasia [17]. È doveroso anche distinguerla dalla transdifferenziazione, quest'ultima è infatti la trasformazione di un tipo cellulare differenziato in un altro tipo cellulare completamente diverso, presente nel tessuto [17].
Di solito la metaplasia tende a manifestarsi nei tessuti costantemente esposti agli agenti ambientali, quindi sistema polmoni e trachea e il tratto gastrointestinale, a causa dei loro contatti rispettivamente con l'aria e il cibo, spesso di per sé dannosi per natura [17].
Le metaplasie possono essere omeotiche, ossia quando un tessuto appropriato in una determinata posizione anatomica si trasforma in un altro tessuto che risulta invece appropriato per un'altra posizione anatomica [16]. In questo caso il nuovo tessuto è, a livello ultrastrutturale, immunologico e istochimico, identico alla sua controparte normale in altre parti del corpo [16]. Metaplasie di questo genere possono essere, ad esempio, tessuto intestinale nello stomaco (metaplasia intestinale), tessuto gastrico nel digiuno o nel duodeno, tessuto intestinale nella vescica urinaria, tessuto intestinale o endocervicale o uroteliale nelle ovaie o ancora, tessuto endocervicale nell'utero [16], [17].
Vi sono poi le metaplasie squamose in cui l'epitelio monostratificato si trasforma in epitelio multistratificato, come per esempio accade nelle vie aeree e nei polmoni, nella cervice, nelle ghiandole sebacee e mammarie, nella pelle [17].
Quelle più frequenti si verificano nell'epitelio alveolare o nell'epitelio delle vie aeree. Il primo prevede la sostituzione delle cellule alveolari con epitelio squamoso, mentre il secondo la sostituzione dell'epitelio bronchiolare o bronchiale. La metaplasia squamosa può generare vari gradi di infiammazione, fibrosi e necrosi [17].

Un altro caso avviene nella cervice uterina, l'endocervice è composto da epitelio ghiandolare mentre l'ectocervice da epitelio squamoso stratificato. Si crea così una giunzione tissutale eterogenea che può innescare una risposta metaplastica, inizialmente avviene in modo molto irregolare ma successivamente l'epitelio squamoso sostituisce in toto quello ghiandolare. Più raramente occorrono metaplasie in altre parti del corpo [17].
Esistono inoltre metaplasie che si trovano in parti diverse del corpo o ancora metaplasie che non hanno una apparente controparte normale [16]. Metaplasie di quest'ultimo genere sono ad esempio: tessuto simil-intestinale nell'esofago, stomaco o nella colecisti oppure il tessuto pseudopilorico nel corpo dello stomaco o nell'intestino [16], [17]. Vi è poi il caso isolato delle metaplasie acino-duttale, note anche come ADM, che avvengono nel pancreas [17].
Un tessuto può essere sostituito da un altro tramite due differenti meccanismi. Il primo è per colonizzazione, con cellule di origine diversa, il secondo è quando è presente un cambiamento durante lo sviluppo cellulare. Solo quest'ultimo caso rappresenta la vera metaplasia, essa avviene nei tessuti epiteliali di rinnovamento con la rigenerazione cronica causata da stimolazioni esterne anomale, ormonali, infezioni o traumi [16]. Vista la loro natura di continuo rinnovamento e proliferazione si ritiene che la trasformazione avvenga a livello di cellule staminali.
Poiché la metaplasia si manifesta anche in piccoli focolai, si pensa che essa abbia un'origine monoclonale. Visto che poi assumerà dimensioni macroscopiche dovrà competere con il tessuto normale in rigenerazione, ed è proprio qui che entra in gioco il carattere adattivo del cambiamento cellulare [16]. La trasformazione cellulare da tessuto ghiandolare a tessuto squamoso conferisce una maggiore resistenza chimica e meccanica.
La natura del cambiamento cellulare evidenzia che la trasformazione avviene a livello molto profondo, non è semplicemente un'espressione genica errata, ma piuttosto un cambiamento evolutivo della cellula staminale, da un tipo tissutale ad un altro [16].
Le metaplasie effettuano cambiamenti in un solo passaggio, e questo suggerisce che avvengano solo fra tessuti similmente contigui. Infatti spesso le coppie di tessuti coinvolti appartengono ad organi fisicamente vicini.
Le metaplasie non causano problemi in sé, possono al più generare secrezioni inattese o anomale, come nel caso dei tessuti gastrici eterotropici; nella maggior parte dei casi sono infatti innocue [16].
Il problema più importante è come queste lesioni possano causare predisposizione al cancro, sono infatti associate a danno tissutale cronico [16].
Le metaplasie creerebbero molti nuovi confini con elevata discontinuità tissutale, aumentando così la probabilità di generare un nuovo cancro. proprio per questo motivo, spesso le neoplasie occorrono in prossimità di giunzioni tissutali, come nel caso della giunzione gastroduodenale o la giunzione ecto-endocervicale [16]. Nel caso in cui la neoplasia si sviluppi a partire da una metaplasia, probabilmente questa assomiglierà più alla metaplasia piuttosto che al tessuto genitore [16].
Fra tutte, le metaplasia più diffuse e studiate sono quelle intestinali. Esse si dividono principalmente in metaplasia intestinale esofagea, meglio noto come esofago di Barret, e metaplasia intestinale gastrica.

L'esofago di Barret si manifesta quando la mucosa squamosa che normalmente riveste l'esofago distale viene sostituita da un epitelio colonnare simile a quello dello stomaco o dell'intestino [18]. Esso può essere sede di ulcera, stenosi e adenocarcinoma esofageo, quest'ultimo in particolare è aumentato in frequenza di oltre sei volte negli ultimi decenni [19]. Questa patologia è comune in circa il 6%-12% dei pazienti che accusano reflusso gastroesofageo e ne risulta direttamente collegato [20].
La metaplasia intestinale gastrica verrà trattata separatamente, essendo oggetto della tesi.

## 2.4 La metaplasia intestinale gastrica

La metaplasia intestinale rappresenta la trasformazione metaplastica più frequente degli epiteli gastrici [5]. Sia le ghiandole ossintiche che le ghiandole della mucosa antrale possono assumere il fenotipo delle ghiandole intestinali, caratterizzato da vacuolo mucosecretivo sovranucleare, ossia epiteli caliciformi mucipari simili a quelli dell'intestino [5]. Le metaplasie intestinali gastriche si dividono in base al tipo di secreto che queste producono. Se il secreto è a bassa acidità le ghiandole sono costituite da sialomucine e in questo caso si dice che la metaplasia è matura, o di tipo tipo intestino tenue o più semplicemente di tipo I [5]. Se il secreto è ad alta acidità le ghiandole sono costituite da sulfomucine, si dice che la metaplasia è immatura, o di tipo tipo intestino crasso o più semplicemente di tipo II o III [5].
In generale, più cresce l'acidità del secreto e più aumenta il rischio neoplastico, quindi il tipo III rappresenta la metaplasia a più alto rischio neoplastico [5].
I tessuti metaplastici hanno epiteli con nuclei atipici pseudostratificati e allungati e le ghiandole presentano cellule caliciformi mucipare [5].
Vi sono due differenti teorie sulla genesi della metaplasia:

- Cellule metaplastiche derivanti dalla SPEM
- Cellule metaplastiche derivanti direttamente dall'istmo gastrico

La prima teoria collega lo sviluppo della neoplasia all'infezione da *Helicobacter pylori* , e SPEM è l'acronimo inglese di *Spasmolytic Polypeptide-Expressing Metaplasia* [17]. La fase iniziale a seguito di questa infezione è la risposta infiammatoria alla lesione, che comporta il reclutamento di tessuto linfoide e neutrofili nella mucosa gastrica. Successivamente si verifica l'apoptosi e la proliferazione cellulare, ma il tasso di perdita cellulare è maggiore della proliferazione, e ciò causa un assottigliamento della mucosa che porta alla gastrite atrofica. Ulteriori modifiche sull'architettura cellulare e modifiche genetiche portano alla progressione verso stadi più avanzati e alla fine al cancro [21].
A seguito di infezione da *Helicobacter pylori* infatti, avviene la perdita di cellule parietali e quindi infiammazione cronica. Questo comporta iperplasia foveolare e SPEM. L'epitelio gastrico è costituito da diversi tipi di cellule, partendo dall'esterno si trovano le cellule superficiali, poi più internamente le cellule staminali (o cellule progenitrici), poi le cellule parietali che producono l'acido e infine le cellule principali [17].

Nell'antro gastrico sono presenti cellule $LGR5^{+}$ con proprietà di auto rinnovamento, sebbene pare che queste non diano origine a SPEM o metaplasia intestinale, è stato scoperto che danno origine a un cancro gastrico precoce [17].
La seconda teoria sostiene che l'origine della metaplasia intestinale gastrica risieda nell'istmo gastrico, direttamente dalle cellule staminali o dalle cellule principali [17]. Questo accade per esempio nelle cellule staminali $MIST1$ nell'istmo gastrico che mostrano espansione clonale nell'ambiente della mutazione del gene $Kras^{G12D}$ [17]. Esse hanno come conseguenza la formazione di metaplasia intestinale gastrica [17]. L'asportazione delle cellule staminali blocca la proliferazione della metaplasia, inoltre, poiché le cellule metaplastiche sono molto longeve, la riprogrammazione avviene probabilmente nelle cellule staminali longeve e non comporta transdifferenziazione [17].
Sia SPEM che percorso diretto sono precursori della displasia, ossia uno sviluppo cellulare anomalo che comporta la sostituzione delle cellule mature con delle cellule immature[22] e successivamente dell'adenocarcinoma [17].
Nella **Figura 2.4** sono rappresentati i due differenti percorsi possibili della genesi della metaplasia intestinale gastrica: SPEM e percorso diretto.

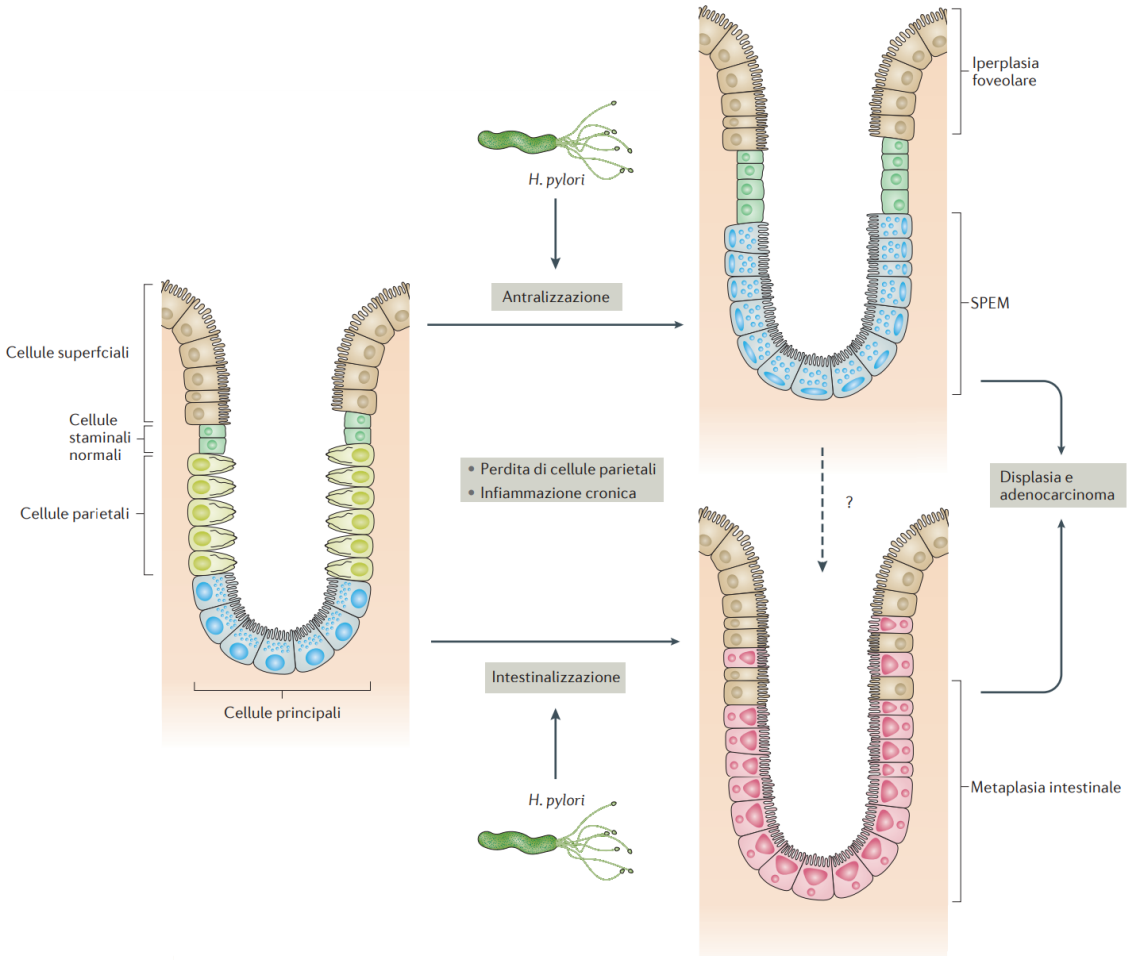



**Figura 2.4: La SPEM e il percorso diretto.** *Sono stati studiati due differenti percorsi possibili per la genesi della metaplasia gastrica intestinale, il primo è la SPEM (Spasmolytic Polypeptide-Expressing Metaplasia), in cui l'Helicobacter pylori gioca un ruolo chiave creando un'infiammazione cronica, la seconda è invece un percorso diretto, dove la sostituzione cellulare avviene a seguito di mutazione genetica. Immagine tratta da [17].*

A livello anatomico, la metaplasia intestinale gastrica può essere suddivisa in metaplasia del cardias o metaplasia distale, non del cardias.
La mucosa cardiale è prossimale alla mucosa ossintica ed è situata in posizione distale alla giunzione squamo colonnare, ha una morfologia simile a quella dell'antro e in superficie è costituita da epitelio colonnare e ghiandole mucose [23].
La metaplasia intestinale del cardias e l'esofago di Barret differiscono nel loro rischio di trasformazione maligna, per questo è fondamentale distinguere i due casi tramite biopsia della giunzione gastro esofagea [23].
Per esempio l'epitelio squamoso sovrastante, i dotti ghiandolari esofagei e le ghiandole ibride si osservano esclusivamente nell'esofago di Barret, così come la metaplasia intestinale di tipo incompleto e l'epitelio multistrato. Distinguerle risulta molto complesso e spesso la diagnosi è determinata solamente dalla posizione endoscopica della biopsia [23].

La metaplasia intestinale distale è spesso diagnosticata nelle popolazioni ad alto rischio di cancro gastrico. Lo sviluppo dell'adenocarcinoma consiste nelle seguenti fasi precancerose: gastrite non atrofica, gastrite atrofica multifocale, metaplasia intestinale e infine displasia [23].
I focolai metaplastici solitamente compaiono prima alla giunzione antro-corpo, in modo particolare nell'incisura angularis. Successivamente questi si allargano e si uniscono, estendendosi sia alla mucosa dell'antro che a quella del corpo. In queste macchie metaplastiche possono poi comparire le prime displasie [23].
In generale, a livello istopatologico, la metaplasia intestinale è facilmente riconoscibile tramite ematossilina ed eosina, ma anche, nei casi di metaplasia completa, tramite presenza di enzimi digestivi dell'intestino tenue.
La metaplasia intestinale deriva dalle cellule staminali gastriche che vengono deviate dalla proliferazione in cellule specifiche dello stomaco verso quelle dell'intestino tenue, le cellule di Paneth, le cellule assorbenti e le cellule caliciformi. Ciò è solitamente innescato da un'irritazione persistente della mucosa gastrica [24].
Per discriminare se una metaplasia è di tipo completo o di tipo incompleto, dai patologi viene ampiamente studiate le morfologie delle sezioni colorate con ematossilina ed eosina[23].
La metaplasia completa viene diagnosticata quando l'epitelio assomiglia al fenotipo dell'intestino tenue, con cellule caliciformi ben formate e enterociti eosinofili che mostrano un bordo a spazzola ben definito [24].
La metaplasia incompleta invece assomiglia a un fenotipo dell'epitelio del colon con assenza di un bordo a spazzola e goccioline di mucina multiple, variabili e irregolari [23]. Un altro tipo di classificazione sperimentale è la discriminazione tramite pH. Infatti dal punto di vista istochimico, le normali mucine gastriche hanno un pH neutro. Con la metaplasia intestinale le mucine gastriche originali sono sostituite da mucine acide [23].
La metaplasia completa, detta di tipo I, esprime solo sialomucine, quella di tipo III, o incompleta esprime invece sulfomucine. La metaplasia di tipo II è sempre di genere incompleto ma è una forma ibrida che esprime una commistione di mucine gastriche e intestinali [23]. Le mucine neutre presenti nella mucosa normale diminuiscono in maniera graduale durante lo sviluppo iniziale della metaplasia, mentre le scialomucine compaiono e diventano quelle predominanti; nelle fasi più avanzate invece, predominano le sulfomucine [24].
La metaplasia intestinale di tipo I è associata a un basso rischio di cancro gastrico. La metaplasia intestinale incompleta ha la maggiore associazione con il cancro, con un rischio quattro volte maggiore di sviluppare il cancro rispetto a quelli di tipo completo [24].
La metaplasia intestinale gastrica nella piccola curvatura (dal cardias al piloro) è associata a un rischio maggiore di cancro gastrico rispetto alla metaplasia intestinale antrale [24].Tuttavia, non tutti i pazienti con metaplasia intestinale progrediranno verso il cancro gastrico [24].

## 2.5 La gastroscopia moderna

I disturbi del tratto gastrointestinale superiore sono malattie globali e comuni, ampiamente diffuse in tutto il mondo [25].
L'avanzamento del cancro gastrico è preceduto da gastrite cronica, atrofia gastrica, metaplasia intestinale gastrica e displasia. Ciò mette in evidenza che una diagnosi precoce e l'identificazione tempestiva dei soggetti a rischio risulta essere fondamentale.
Le caratteristiche macroscopiche che vengono comunemente ricercate in un esame endoscopico sono chiazze bianche grigiastre e in rilievo, circondate da mucosa gastrica pallida o di colore normale, oppure eritema a chiazze [21]. Si ricercano anche goccioline lipidiche denominate "sostanza opaca bianca"; esse sono un classico marker endoscopico di metaplasia intestinale e tumori epiteliali [21].
Come indicatori invece, per la diagnosi della metaplasia intestinale nella gastroscopia convenzionale, vengono utilizzati i seguenti criteri endoscopici: mucosa biancastra, superficie della mucosa ruvida o irregolare, aspetto villoso, arrossamento a chiazze, venule collettori atipiche con morfologia anomala e distribuzione irregolare [26]. I vari tipi di tessuti metaplastici sopra elencati sono rappresentati in **Figura 2.5**.

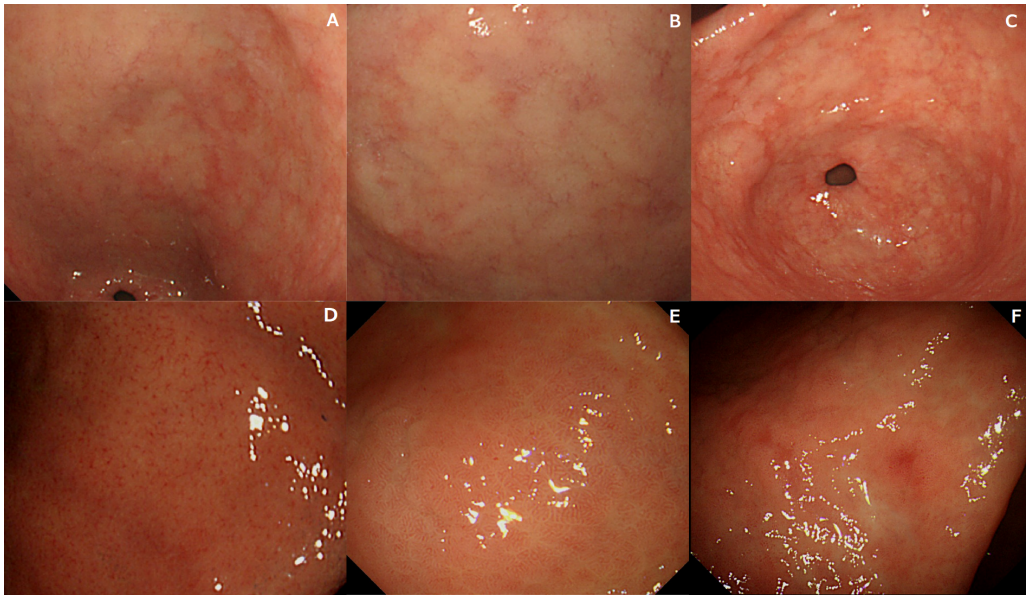



**Figura 2.5: I criteri endoscopici per diagnosticare la metaplasia intestinale.** *In figura i principali criteri con la quale viene diagnosticata la metaplasia intestinale gastrica. (A) Mucosa biancastra, a grappolo irregolare. (B) Mucosa a superficie ruvida. (C) Mucosa a superficie irregolare biancastra, granuli sparsi di diverse dimensioni e forme. (D) Mucosa con venula collettore atipica, di forma e distribuzione irregolare. (E) Mucosa con aspetto villoso. (F) Mucosa con rossore a chiazze. Immagini tratte da [26].*

Recentemente, sono state introdotte nuove tecnologie come la cromoendoscopia combinata con l'endoscopia con ingrandimento e la microscopia con laser confocale. Tuttavia, queste tecniche avanzate di endoscopia forniscono solo immagini della su-

perficie della mucosa e l'accuratezza diagnostica dipende ancora dalle operazioni standardizzate di endoscopisti esperti [27].
Negli ultimi anni sono state introdotte nuove piattaforme online per migliorare l'apprendimento di gastroenterologi e questo permette loro di padroneggiare con le proprie capacità diagnostiche. Al fine di individuare precocemente, tramite rilevamento endoscopico, possibili malattie del tratto gastrointestinale, consentendo esiti più favorevoli [25]. Di contro, risulta molto difficile attuare uno screening di massa, e questo spesso porta ad esiti poco soddisfacenti. Inoltre si è visto che circa il 10% dei tumori del tratto gastrointestinale non viene riconosciuto durante gli esami endoscopici eseguiti tre anni prima della diagnosi [28].
Le capacità diagnostiche, e quindi terapeutiche, a seguito di gastroscopia, sono fortemente condizionate dall'esperienza, e dalle capacità tecniche e decisionali del medico. Per questo motivo sono stati introdotti per esempio endoscopi robotici guidati dall'intelligenza artificiale, che calcolano la giusta forza che dovrebbe essere applicata alla parete gastrica, oppure la corretta posizione tridimensionale dei movimenti [28]. Tutte queste tecnologia tentano di ridurre la probabilità che accadano eventi avversi. Per esempio, nel caso della dissezione sottomucosa endoscopica è possibile rimuovere i tumori gastrici precoci, e altre lesioni sottomucose maligne, con interventi minimamente invasivi [28].
Queste tecnologie cercano di diminuire in questo modo la probabilità che accadano eventi avversi, come nel caso della dissezione sottomucosa endoscopica, che consente la rimozione minimamente invasiva dei tumori gastrici precoci e di altre lesioni sottomucose maligne
Un'altra tecnologia futuribile è quella che utilizza capsule ingeribili contenenti microtelecamere, wireless e con possibilità di essere controllate da remoto, al fine di creare una ricostruzione 3D dell'intero apparato gastrico.
Attualmente l'approccio migliore per rilevare e classificare accuratamente i pazienti a rischio risulta essere l'endoscopia con immagini migliorate, combinata con il campionamento tramite biopsia, per analisi istopatologica [25]. Tutte le lesioni identificate dovrebbero essere corredate da indicatori di qualità misurabili, come la registrazione del tempo, la documentazione fotografica e la descrizione delle lesioni utilizzando la terminologia standard.
Per tale scopo esistono numerosi gastroscopi convenzionali con diverse caratteristiche tecnologiche. I più moderni hanno un diametro ridotto, inferiore ai 10 millimetri e ampio angolo di curvatura che arriva fino a 210 gradi; esistono persino gastroscopi monouso [25].
Un primo tentativo di migliorare la precisione nella diagnosi è stato associare le sonde endoscopiche convenzionali all'utilizzo di mezzi di contrasto, come l'indigotina, che consente la visualizzazione delle minuscole irregolarità della superficie della mucosa e il pattern dell'area gastrica senza l'utilizzo di speciali apparecchiature endoscopiche [26].
Con l'utilizzo del mezzo di contrasto è possibile classificare le lesioni in quattro tipologie: $P_0$, $P_1$, $P_2$ e $P_3$ [26]. Man mano che si avanza dal tipo $P_0$ al tipo $P_3$, i solchi diventano più ampi ed evidenti, inoltre le aree gastriche sono classificate in base al grado di irregolarità della larghezza del solco, alle dimensioni e alla forma delle aree gastriche [26]. Nel tipo $P_0$ le aree gastriche sono fini, fittamente disposte

e divise da solchi molto stretti. Nel tipo $P_1$ le aree gastriche sono disposte a griglia o di forma quasi circolare, con scanalature strette e fittamente disposte. Nei tipi $P_2$ e $P_3$ le aree gastriche sono di forma e dimensioni irregolari. In particolare nel tipo $P_3$, dove i solchi sono più marcati e le aree gastriche sono di forma e dimensioni irregolari. Queste ultime sono separate da solchi di larghezza irregolare e hanno un aspetto villoso [26]. Nella **Figura 2.6** sono rappresentate le quattro tipologie di tessuto gastrico sopra elencate.

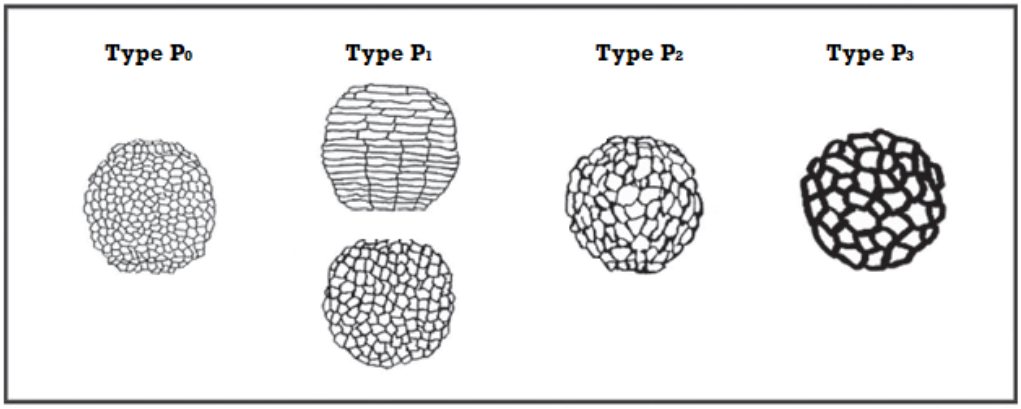



**Figura 2.6: Le quattro tipologie di metaplasia con utilizzo di indigotina.** *Con utilizzo di indigotina come mezzo di contrasto si possono distinguere quattro tipologie di tessuto. Progredendo da tipo $P_0$ verso tipo $P_3$ l'intensità della lesione metaplastica aumenta. Immagine tratta da [26].*

I nuovi strumenti endoscopici consentono funzioni ottiche ad alta potenza, ad esempio alta definizione delle immagini, luce bianca ad alta potenza, imaging a banda stretta, imaging con autofluorescenza, modalità che permettono miglioramento del colore, miglioramento dello spettro e persino intelligenza artificiale [25].
Tutte queste migliorie in ambito tecnologico mirano ad eliminare la componente soggettiva del processo di diagnosi, cercando di stabilire una serie di indicatori chiave verificabili. In particolare mirano a ottimizzare la diagnosi di neoplasie precoci e condizioni premaligne, al fine di alterare positivamente il decorso verso neoplasie maligne, riducendo il bias associato al fattore umano [25].
L'utilizzo della cromoendoscopia virtuale con imaging a banda stretta, meglio conosciuta come NBI, acronimo di *narrow-band imaging* [29] è ampiamente impiegata nella moderna gastroscopia. Essa è infatti una metodologia di endoscopia con immagini migliorate che ottengono tassi di accuratezza superiori all'85%-90% per la diagnosi di metaplasia e displasia intestinale. Viene anche associata ad endoscopie ad alta risoluzione per aumentarne le prestazioni [30]. Infatti l'uso dell'endoscopia standard a luce bianca mostra spesso una scarsa correlazione con l'istologia nel rilevare le trasformazioni metaplastiche [31]. Per questo motivo vengono spesso preferiti gli endoscopi che utilizzano l'imaging a banda stretta. Con questa tecnologia i tessuti affetti da metaplasia appaiono come riflessi a macchia di colore bianco-blu e sono situati sui margini epiteliali; vengono comunemente chiamati "creste azzurre" [21].
I migliori risultati in termini di sensibilità e specificità si ottengono quando viene combinato l'imaging a banda stretta con l'endoscopia standard a luce bianca ad

alta risoluzione. Con questa tecnologia è stata implementata una nuova scala per definire il grado di avanzamento della metaplasia intestinale gastrica. Essa è nota come EGGIM, ossia *endoscopic grading of gastric intestinal metaplasia*, e utilizza una scala da 0 a 10 con ordine di avanzamento della patologia crescente [32]. Il risultato è ottenuto dalla somma di valutazioni eseguite su cinque aree diverse, ad ognuna delle quali può essere assegnato valore 0, 1 o 2. Un risultato totale di 0 significa che la metaplasia è assente, valori da 1 a 4 indicano che la metaplasia è focale o moderata, valori da 5 a 10 indicano una metaplasia estesa [32].
In **Figura 2.7** i miglioramenti ottenuti con l'utilizzo della tecnologia dell'imaging a banda stretta.

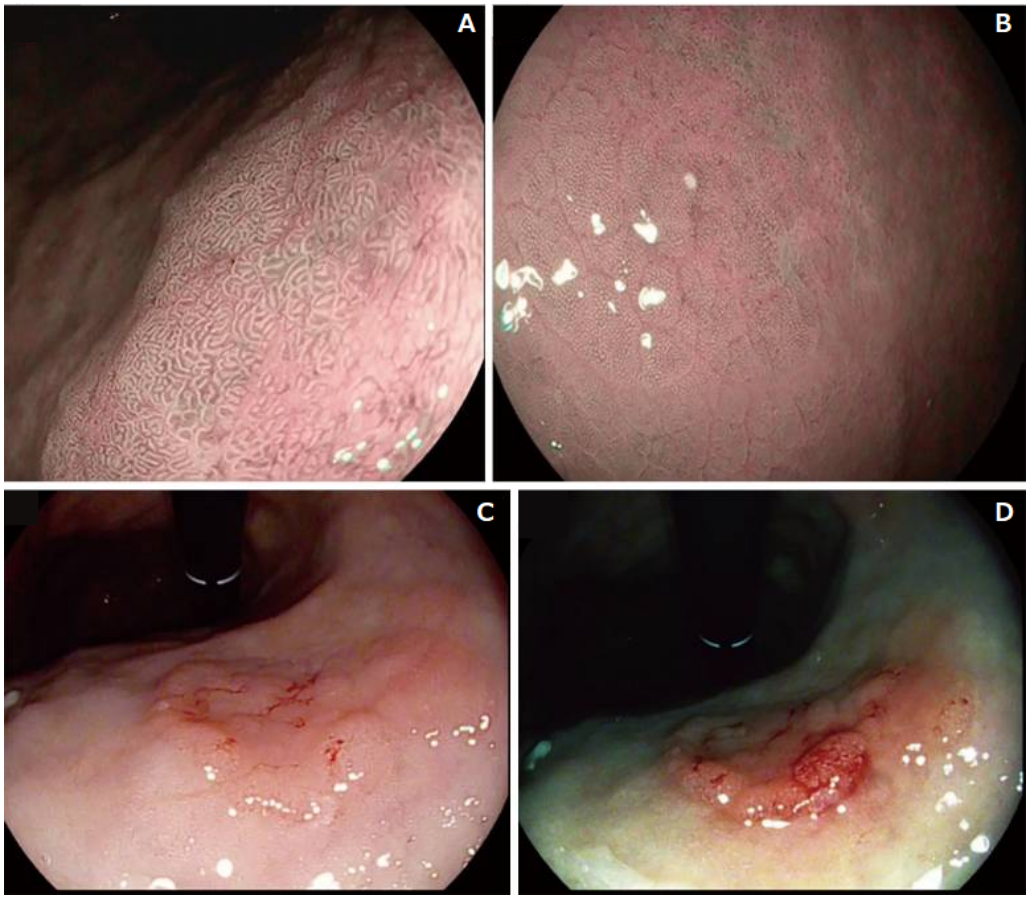



**Figura 2.7: I miglioramenti introdotti dall'NBI.** *Con utilizzo dell'endoscopia con immagini migliorate, in particolare dell'imaging a banda stretta, la sensibilità e la specificità aumentano in maniera significativa; vengono infatti meglio evidenziate le lesioni metaplastiche. (A) Metaplasia intestinale gastrica osservata mediante endoscopia con imaging a banda stretta. (B) Bordo atrofico osservato mediante imaging a banda stretta. (C) Lesioni gastriche precoci osservate con l'endoscopia a luce bianca standard. (D) Lesioni gastriche precoci osservate mediante imaging a banda stretta. Immagini tratte da [21].*

Un altro sistema introdotto in endoscopia è il LASEREO, prodotto della società giapponese Fujifilm Holdings Corporation, è un sitema di miglioramento delle immagini che utilizza una sorgente di luce laser, meglio noto come imaging a laser blu

[33]. Esso è composto da due sorgenti di luce laser, che offrono quattro modalità di osservazione di imaging: *white light imaging* (WLI), *blue laser imaging* (BLI), BLI-*bright* e *linked color imaging* (LCI). Se questo strumento viene associato alle tecnologie di intelligenza artificiale, come la rete neurale convoluzionale profonda, si riescono ad ottenere buoni risultati di classificazione per le patologie gastriche [33]. Le modalità laser di BLI-*bright* e LCI forniscono viste endoscopiche più luminose e consentono di osservare la mucosa gastrointestinale su un'area più ampia, senza utilizzare funzioni di ingrandimento come accade invece con i sistemi di miglioramento delle immagini tradizionale [33].
Nonostante tutti i miglioramenti tecnologici, oltre il 10% dei tumori all'interno del tratto gastro intestinale non viene comunque rilevato dall'endoscopia, per questo motivo una diagnostica assistita da computer potrebbe ridurre al minimo il numero di casi di mancato riconoscimento della patologia.
Esistono algoritmi di rilevamento che utilizzano sia *machine learning* che *deep learning*. Quest'ultimo è stato già usato con successo anche per rilevare e classificare polipi colorettali, predire la neoplasia di Barrett e in generale per migliorare la qualità dell'endoscopia [34]. In generale si tratta di reti neurali convoluzionali complesse, le meglio note CNN (acronimo inglese di *convolutional neural network*). Queste sono state testate di differenti grandezza e profondità, con architetture ResNet e DenseNet, fornendo ottime prospettive future [27].
Sono stati anche creati classificatori che utilizzano gli algoritmi *random forest* e *support-vector machines* (SVM), ottenendo buoni risultati [35]. Nella maggior parte degli studi le sonde endoscopiche utilizzate sono quelle della società giapponese Olympus Corporation, che ha sede a Tokyo [26], [31], [34], [35], [36].
L'uso di queste tecniche per il riconoscimento dell'immagine del cancro gastrico ha prodotto un'accuratezza diagnostica superiore rispetto agli endoscopisti non esperti, paragonabile a quella di endoscopisti esperti [37].
Nonostante i moderni sistemi siano sensibili alle anomalie, essi hanno scarsa specificità nell'individuare cambiamenti rilevanti della mucosa. Utilizzando il *deep learning*, sia la sensibilità che la specificità diagnostica migliorano sensibilmente con l'aumento dei dati [25].
L'intelligenza artificiale garantisce un'elevata riproducibilità, verrebbero ridotti al contempo i costi dovuti alla formazione del personale sanitario e verrebbero eliminati totalmente gli errori dovuti all'affaticamento del medico [36]. Attualmente un computer è in grado di riconoscere immagini contenenti neoplasie con maggiore specificità e sensibilità di un endoscopista [38]. Tramite *deep learning* e CNN è possibile ottenere una sensibilità del 98% su carcinoma a cellule squamose esofagee, anche se con specificità molto bassa [25].
Sono stati presentati inoltre algoritmi che prevedono l'utilizzo di più classificatori di tipo CNN insieme [34]. La prima CNN rilevare le lesioni gastriche, mentre la seconda per diagnosticare le neoplasie separatamente.
Per quanto riguarda invece il rilevamento dell'*Helicobacter pylori*, l'intelligenza artificiale ha ottenuto risultati più sensibili e specifici di un endoscopista umano, con una sensibilità e una specificità del quasi 87%, rispetto al 75% e al 63% ottenuti tramite endoscopia convenzionale. Risultati simili sono stati ottenuti anche nel ca-

so della valutazione della profondità di invasione dei tessuti nel carcinoma gastrico precoce, raggiungendo l'89% di accuratezza e il 76% di specificità [25].
Con il progresso della tecnologia informatica, l'intelligenza artificiale viene applicata sempre più in medicina; in particolare, il *deep learning* [33]. Esso è ampiamente utilizzato nella diagnostica per immagini; imitando il funzionamento della rete neurale umana, può imparare a identificare le caratteristiche specifiche delle immagini, stabilendo automaticamente un protocollo di classificazione [33].
La capacità dell'intelligenza artificiale di analizzare rapidamente le immagini, senza fatica, rende questa tecnologia ideale a tali scopi. Essa potrebbe sensibilmente migliorare la qualità dell'indagine e il rilevamento delle lesioni gastrointestinali poiché ha già dimostrato di superare l'essere umano nel riconoscimento endoscopico delle immagini [25].

# 3. Materiali e metodi

Al giorno d'oggi la procedura più utilizzata per diagnosticare la metaplasia gastro intestinale risulta essere l'endoscopia con immagini ad alta risoluzione combinata con biopsia. Il medico infatti, quando rileva un tessuto che considera anomalo decide di eseguire una biopsia della zona interessata. I tessuti campionati vengono fissati in formalina tamponata, processati per l'inclusione in paraffina, sezionati e colorati con ematossilina ed eosina. Viene inoltre valutata una possibile infezione da *Helicobacter pylori* utilizzando la colorazione di Giemsa; solo a questo punto viene eseguita l'indagine istologica [30]. Questo processo si porta dietro diverse problematiche.
In generale è molto difficile distinguere un tessuto sano da uno patologico, ad un medico occorrono infatti diversi anni di pratica ed esperienza. Nonostante ciò vi è sempre un margine di errore, può infatti capitare che un medico non riconosca subito la patologia, o più sovente, il medico interpreti un tessuto come patologico quando in realtà non lo è. A tutto ciò si somma la componete umana dell'operatore. Infatti dopo diverse ore di lavoro, dopo diverse endoscopie effettuate, il medico potrebbe soffrire di affaticamento, e le sue capacità diagnostiche potrebbero diminuire.
Per questi motivi si preferisce sempre eseguire una biopsia anche per i casi dubbi o di difficile interpretazione. Il problema è che la biopsia, anche se mini invasiva, è sempre un intervento, che quindi si porta dietro lesioni, sanguinamenti, cicatrizzazioni. Di solito si esegue una biopsia multipunto, questa aumenta il trauma gastrico e il rischio di sanguinamento e inoltre non può essere eseguita se il paziente sta assumendo farmaci come l'aspirina [27]. Inoltre per ottenere gli esiti degli esami bioptici occorre spesso molto tempo, mediamente due settimane, spesso anche più di tre. A questi si aggiungono gli esami batteriologici e l'analisi istopatologica. Tutto ciò comporta perdite di tempo che potrebbero risultare fondamentali. L'anatomopatologo analizza ogni giorno decine di vetrini, questo oltre a sovraccaricare di lavoro il sistema sanitario potrebbe portare all'affaticamento del medico, che potrebbe eseguire diagnosi poco accurate.
Nonostante l'endoscopia sia considerata lo strumento migliore per l'individuazione di condizioni premaligne gastriche e tumori precoci, la correlazione tra endoscopia convenzionale e istologia è considerata inadeguata [30]. Infatti solo circa la metà dei vetrini analizzati riconduce veramente ad una metaplasia intestinale, gli altri sono solo falsi positivi. Per questo motivo l'ausilio dell'intelligenza artificiale potrebbe essere una chiave di svolta futura.
In questo lavoro è stata valutata la possibilità di poter assistere il medico nella diagnosi della metaplasia intestinale, con l'ausilio dell'intelligenza artificiale. Agendo solamente sulle immagini endoscopiche standard, evitando biopsia ed esame istologico.

Dalle endoscopie sono state ricavate le caratteristiche delle immagini, basate sulle sole informazioni tissutali di texture e pattern istologico, a livello macroscopico. Le osservazioni ricavate sono state poi utilizzate all'interno di un classificatore binario con lo scopo di discriminare automaticamente le porzioni di tessuto sano da quello patologico.
Tutti gli algoritmi, dal pre-processing all'interfaccia grafica sono stati implementati con la piattaforma di programmazione e calcolo numerico MATLAB.

## 3.1 Il campione e le immagini utilizzate

Il campione di pazienti utilizzato è composto da 25 individui, di sesso maschile e femminile, completamente sani o affetti da metaplasia intestinale gastrica. Il campione è composto da 17 donne e 8 uomini, con età compresa fra 50 e 80 anni, solo una paziente risultava avere un età inferiore a 40 anni (38).
I pazienti sani erano in tutto 6 (4 donne e 2 uomini), mentre i pazienti affetti da metaplasia erano 19 (13 donne e 6 uomini). Dal campione dei pazienti sani sono state catturate 47 immagini, da ogni paziente il medico ha estratto dalle 4 alle 10 immagini, mediamente 7-8. Dal campione dei pazienti affetti da metaplasia intestinale gastrica sono state catturate in tutto 48 immagini endoscopiche, da ogni paziente sono state estratte da 1 a 6 immagini, con una media di circa 2-3 immagini per paziente. Il totale delle immagini, fra individui sani e patologici è quindi pari a 95.
Le immagini provenienti da pazienti sani erano ritenute potenzialmente patologiche da parte del medico, durante l'esame endoscopico. Per questo motivo anche su queste immagini è stata eseguita la biopsia, risultante poi negativa all'esame istologico. Ragion per cui, anche le immagini sane del dataset, risultavano essere dubbie o di difficile interpretazione da parte del gastroenterologo.
Degli individui del campione sono noti solo sesso ed età, mentre non vi sono informazioni su nazionalità, status sociale, stile di vita, fattori di rischio ambientali, storico familiare positivo, presenza di altre patologie.
La distribuzione dei pazienti per sesso e per età è stata rappresentata in **Figura 3.1**.
Le immagini endoscopiche sono state effettuate durante l'arco di un mese circa, da fine aprile 2021 a fine maggio 2021, presso l'azienda Ospedaliera Ordine Mauriziano di Torino, Ospedale Umberto I.

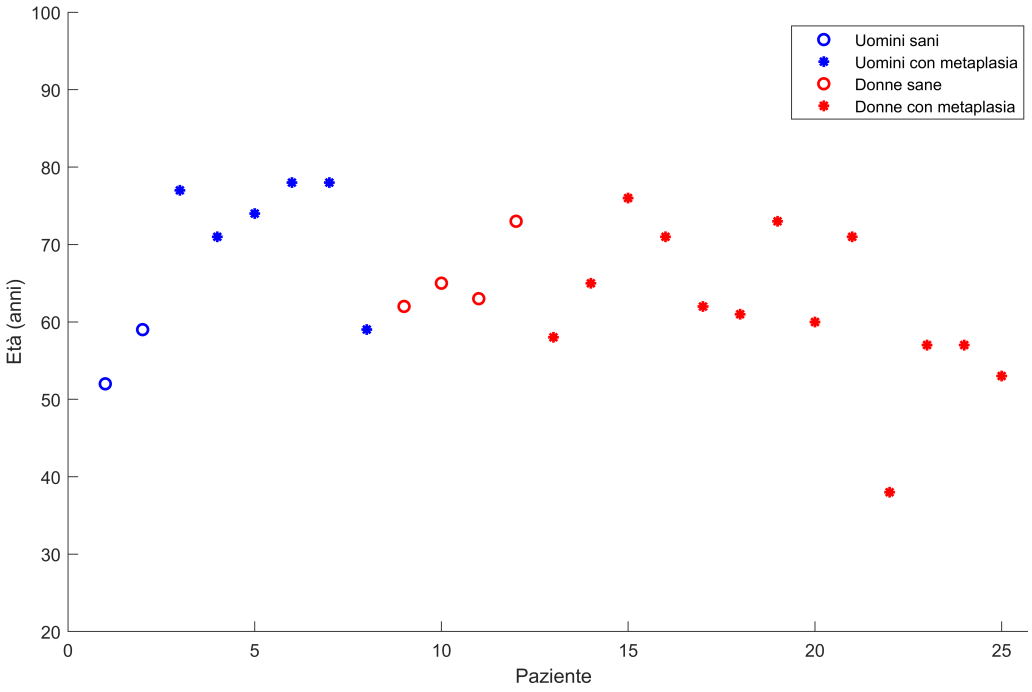



**Figura 3.1: La distribuzione dei pazienti per sesso ed età.** *In figura sono rappresentati i pazienti in base al sesso e all'età. La maggior parte dei pazienti in esame sono donne, ben 17 casi, di cui 4 sono soggetti sani mentre 13 sono soggetti patologici. Gli uomini sono in tutto 8, di cui 2 sono sani mentre 6 sono soggetti patologici. L'età dei pazienti è compresa fra 50 e 80 anni, solo in un caso l'età è inferiore a 40 (una donna di 38 anni affetta da metaplasia intestinale gastrica).*

La immagini sono catture di endoscopie video a colori, modello di colori a 3 layer (RGB), di scala da 0 a 255 e di formato JPG. Sono di dimensioni $1920 \times 1080$ pixel, con risoluzione di 96 dpi e profondità di 24 bit. L'immagine endoscopica principale è nella zona centrale e il resto presenta uno sfondo nero. Esse sono state eseguite tramite sonda endoscopica digitale i-scan della società PENTAX Medical, con sede a Tokyo, in Giappone [39]. Questo strumento offre un'endoscopia digitale a immagini migliorate e, tramite appositi filtri d'immagine, fornisce una cromoendoscopia virtuale dettagliata dei vasi e della mucosa, in tempo reale. I principali filtri utilizzati sono tre:

- Un primo filtro per il miglioramento della superficie che aiuta a visualizzare i bordi delle strutture anatomiche, in grado di delineare meglio le pieghe tissutali e la struttura della mucosa, facendo apparire le strutture elevate e i vasi sanguigni accentuati.

- Un secondo filtro viene utilizzato per il miglioramento del contrasto che aiuta a visualizzare aree depresse mediante colorazione delle aree a bassa densità, focalizza l'aspetto dei vasi superficiali e migliora i dettagli della struttura superficiale della mucosa.

- Un terzo filtro viene utilizzato per il miglioramento del tono che aiuta la diagnosi tramite miglioramento mirato, modificando la colorazione di ciascun pixel,

accentua le strutture vascolari e della mucosa, aiutando a caratterizzare le lesioni [39].

Un volta ricevuto l'esito delle biopsie le immagini sono state classificate come patologiche o come sane. A questo punto il medico ha eseguito una segmentazione manuale delle immagini patologiche, circoscrivendo le aree che risultavano essere visivamente affette da metaplasia intestinale gastrica. Questo tipo di segmentazione ha permesso di distinguere le aree sane da quelle affette da patologia. Tramite l'utilizzo di un puntatore, il medico ha tracciato insiemi chiusi, come isole. Spesso erano presenti più regioni circoscritte per ogni singola cattura, come nell'immagine **Figura 3.2**, che rappresenta un esempio di segmentazione del medico di più regioni patologiche.

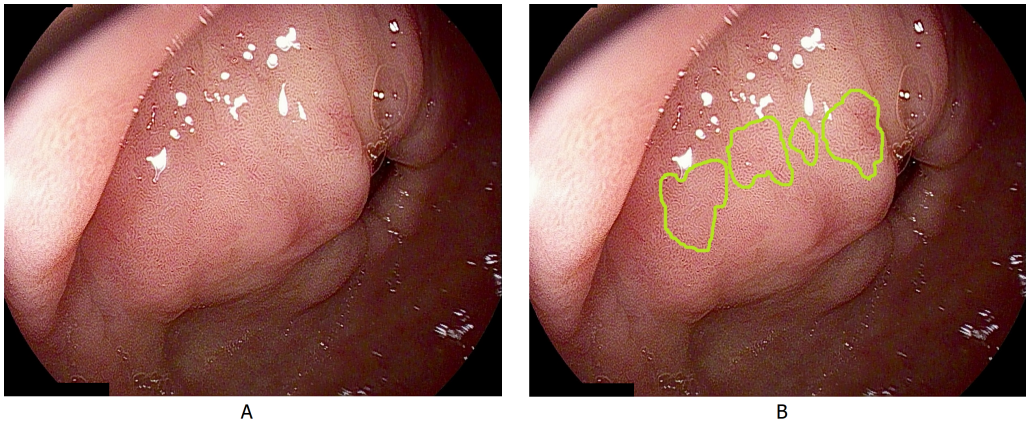



**Figura 3.2: La segmentazione manuale del medico.** *(A) Immagine endoscopica originale. (B) Immagine elaborata al computer manualmente dal medico e rappresenta la segmentazione di 4 regioni patologiche. Per tutte le segmentazioni è stato utilizzato un tratto spesso e di colore verde chiaro. Il colore è stato scelto in vista della fase seguente di pre-processing, risulta essere infatti quello più dissimile dei colori naturali del tessuto.*

Ogni singola cattura elaborata dal medico ha poi subito due processi differenti e separati. Il primo è relativo ad un vero e proprio pre-processing, nella quale l'immagine viene ritagliata e vengono eliminati gli artefatti. Il secondo processo prevede il miglioramento dell'immagine al fine di poter estrarre al meglio le caratteristiche tissutali. Solo nelle immagini provenienti da soggetti patologici, al primo processo se se affianca un secondo, ossia la segmentazione effettuata del medico delle regioni di interesse, le stesse evidenziate manualmente dal medico. Infatti se nelle immagini provenienti da pazienti sani è presente solo tessuto non patologico, nelle immagini provenienti da pazienti affetti da metaplasia sono presenti entrambi i tessuti. Per tal motivo è stato necessario separare le porzioni di tessuto sane da quelle patologiche tramite l'impiego di maschere binarie, seguendo la segmentazione manuale effettuata dal medico.
I due processi di pulizia e segmentazione non sono in cascata, ma in parallelo; le due immagini ottenute vengono poi ricombinate fra di loro. In **Figura 3.3** è rappresentato un diagramma di flusso che evidenzia i due processi paralleli che compongono la fase di preparazione delle immagini patologiche.

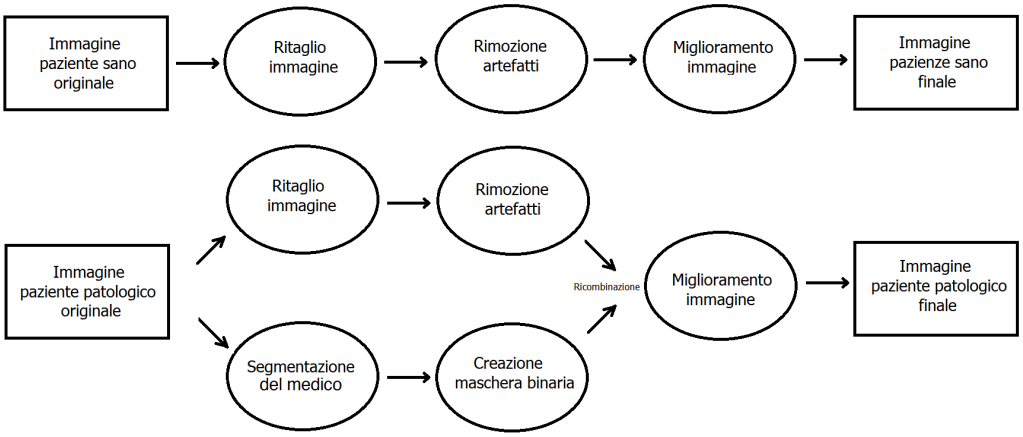



**Figura 3.3: La preparazione delle immagini patologiche.** *Nel primo diagramma in alto è rappresentata la preparazione lineare delle immagine provenienti da pazienti sani. Nel secondo diagramma, in basso, si evidenzia come le immagini patologiche abbiano seguito due processi differenti e paralleli, il primo è mirato alla rimozione degli artefatti, il secondo alla segmentazione delle regioni patologiche. Le due immagini provenienti dai due diversi processi vengono poi ricombinate fra di loro al fine di ottenere un'unica maschera.*

## 3.2 Il pre-processing delle immagini

Una buona parte del lavoro consiste nella preparazione delle immagini prima della loro elaborazione. Come anticipato sopra, la parte di pre-processing è suddivisa in tre processi; la rimozione degli artefatti, il miglioramento dell'immagine e la segmentazione manuale del medico delle zone di interesse. I primi due processi sono comuni ad entrambe le classi, il terzo, invece, è specifico delle immagini provenienti da pazienti patologici.

### 3.2.1 La rimozione degli artefatti

Le immagini catturate tramite sonda endoscopica risultano affette da artefatti e parti non desiderate. Nello specifico la prima fase è stata suddivisa in quattro attività: rimozione di parti di testo non volute, ritaglio dell'immagine di interesse, rimozione delle parti troppo scure, rimozione delle parti troppo chiare.
Nelle immagini originali è presente testo in sovraimpressione, per avere alcune informazioni sul paziente, quali nome, cognome, data di nascita, ID pazienze e il sesso. Risulta presente inoltre una sezione commenti e diversi dettagli su sonda, strumenti e sala medica, sede di esame. Tutte queste informazioni sono spesso sovrapposte all'immagine endoscopica. La loro rimozione è stata implementata tramite tre strategie. In generale è stata presa in considerazione solo la parte sinistra dell'immagine, questo perché solo in quel lato era presente testo in sovraimpressione.

- La prima prevede la rimozione della sezione commenti in basso a sinistra. In questa parte sono presenti informazioni riguardanti la sonda endoscopica, sede e postazione di lavoro dell'operatore. Poiché queste risultavano invariate per tutto

il campione di immagini, è stato scelto di eliminare il testo sovrapponendovi due rettangoli neri, il colore dello sfondo, nello specifico valore 0 su 255.

- La seconda prevede l'utilizzo della funzione *ocr* (acronimo di *optical character recognition*) di MATLAB. Essa è una funzione built-in e utilizza l'intelligenza artificiale basata sull'estrazione di caratteristiche per rilevare porzioni di testo all'interno di un'immagine. Questa è stata applicata nel vertice in alto a sinistra della cattura endoscopica poiché lì risultano presenti anagrafica e ID paziente, valori quindi non prevedibili, e che nella maggior parte dei casi sono in sovraimpressione rispetto all'immagine endoscopica.

  Utilizzando le impostazioni predefinite, questa funzione riconosce automaticamente se nell'immagine è presente del testo e restituisce le posizioni dei vertici dei rettangoli dove sono presenti stringhe di caratteri. Anche in questo caso è stato dato valore 0 a tutti gli elementi presenti all'interno dei rettangoli rilevati. La cattura originale è poi stata ritagliata, estraendo la sola immagine endoscopica, passando quindi da $1920 \times 1080$ pixel a $1226 \times 973$ pixel.

- La terza strategia è di backup, e prevede di porre ad 1 tutti i bit del vertice in alto a sinistra dell'immagine ritagliata. Lo scopo è quello di catturare possibili lettere non riconosciute dall'algoritmo di *ocr*. Per questo è stato creato un triangolo rettangolo, isoscele, di lato obliquo lungo 170 pixel, completamente nero.

Nella **Figura 3.4** è presente un esempio con i principali passi della rimozione del testo non voluto e il ritaglio dell'immagine.

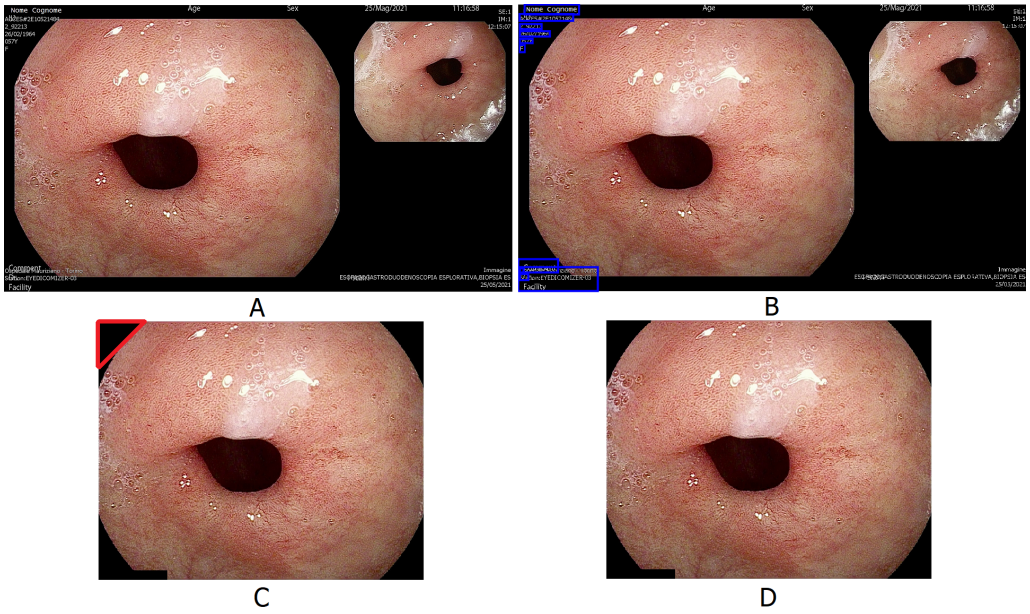



**Figura 3.4: La rimozione del testo e il ritaglio dell'immagine.** *(A) Cattura dell'endoscopia video originale. (B) La rimozione del testo: nei riquadri blu sono presenti le stringhe di testo non volute e che sono state selezionate manualmente o riconosciute dall'algoritmo ocr automatico. Le scritte in alto e a destra sono state deliberatamente ignorate in quanto eliminate tramite ritaglio del fotogramma. (C) La strategia di backup: con il triangolo rosso sono stati eliminati eventuali elementi non riconosciuti dell'algoritmo ocr. (D) Immagine finale ritagliata dopo aver eliminato le zone di testo non desiderate.*

Le immagini cosi ritagliate e ripulite da testo indesiderato risultano ancora inadatte allo studio in quanto sono presenti intere zone troppo scure, dovute all'impossibilità della luce bianca dell'endoscopio di raggiungere tutte le zone riprese dalla telecamera ottica.
Queste sono aree particolarmente cave come l'antro pilorico, oppure più semplicemente aree messe in ombra dalla tuberosità dei tessuti interni dello stomaco, altre ancora sono troppo distanti per essere colpite dalla luce. Alcune aree sono talmente scarsamente illuminate da confondersi con lo sfondo nero della cattura. Poiché risultano essere solo fonte di rumore, tutte le aree che non presentano un'illuminazione sufficiente sono state deliberatamente eliminate. Per eseguire questa operazione è stata utilizzata la funzione built-in di MATLAB *imsegkmeans*.
Essa effettua una segmentazione automatica delle immagini basandosi sulla clusterizzazione di tipo K-means [40]. Lo scopo dell'algoritmo è quello di creare gruppi quanto più omogenei possibile, con una bassa variabilità interna chiamata intracluster e un'alta variabilità fra un cluster e l'altro chiamata variabilità intercluster. In questo algoritmo il prototipo di un cluster viene chiamato centroide, e viene ottenuto come valore medio degli elementi appartenenti al cluster. La variabilità intracluster si ottiene come la somma di tutte le distanze tra tutti gli elementi di uno specifico cluster ed il suo centroide. La distanza intercluster corrisponde invece alla distanza tra i vari centroidi. L'algoritmo è di tipo iterativo e segue il seguente processo:

- Inizia con la scelta di un numero $K$ di cluster, viene definito dunque a priori quanti gruppi si vogliono ottenere.
- Il passo successivo è l'inizializzazione dei centroidi, questo può avvenire in due modalità. Il primo è la scelta casuale di $K$ centroidi, la seconda è l'assegnazione casuale degli elementi ai $K$ cluster per calcolare successivamente i centroidi come media degli elementi del singolo cluster.
- Poi vengono riassegnati gli elementi al cluster avente centroide più vicino ad esso: si calcola quindi la distanza di ogni elemento da ogni centroide e ogni elemento viene assegnato al cluster con il centroide ad esse più vicino.
- Quest'ultimo processo viene ripetuto in maniera iterativa finché ogni elemento raggiunge la stabilità e non cambia più cluster, giungendo così la condizione di stop.

Difficilmente con il K-means si arriva a convergenza con un numero basso di iterazioni, e in genere si imposta un limite massimo di iterazioni che renda accettabile la soluzione trovata in tempi ragionevoli.
Il K-means presenta due svantaggi principali, il primo è che l'inizializzazione dei centroidi avviene sempre in maniera casuale, rendendo la clusterizzazione meno ripetibile e ottenendo sempre risultati differenti, anche con le medesime condizioni iniziali [40]. Il secondo, più importante, è che la scelta del numero dei cluster deve avvenire a priori, facendo variare $K$ ed osservando quale dà migliori risultati in termini di distanza e variabilità.
Ciò è proprio quello che stato fatto per eliminare la aree troppo scure delle immagini endoscopiche, prive di informazioni. Il parametro $K$ è stato fatto variare da 2 a 10, e il valore di cluster prescelto è stato poi $K$ uguale a 4.
Le immagini sono state così suddivise in 4 cluster, in base alle varie intensità di colore e caratteristiche dell'immagine. La funzione *imsegkmeans*, data un'immagine in ingresso RGB, effettua in maniera automatica la segmentazione, accorpando regioni con caratteristiche comuni.
Ai pixel del cluster con intensità di colore minore (calcolata come media dei 3 layer RGB per ogni pixel), è stato assegnato valore 0 su 255, mentre tutti gli altri valore 255 su 255. In questo modo è stata creata una maschera binaria bianco/nero.
Per effettuare la scelta del cluster da eliminare è stata calcolata la media aritmetica dell'intensità di colore di ogni gruppo, ed è stato scelto poi il cluster con valore minore, che presumibilmente sarà quello più scuro. La formula utilizzata è stata la seguente:

$$[\overline{X}_{min}, K_{min}] = min(\frac{1}{N_1}\sum_{i=1}^{N_1} x_i, \frac{1}{N_2}\sum_{i=1}^{N_2} x_i, \frac{1}{N_3}\sum_{i=1}^{N_3} x_i, ..., \frac{1}{N_K}\sum_{i=1}^{N_K} x_i)$$

Dove $N_1, N_2, N_3, ..., N_K$ sono il numero di elementi (pixel), presenti in ogni $K$-esimo cluster, $x_i$ è l'i-esimo elemento presente in ogni cluster $X$, ossia l'intensità del i-esimo pixel, $\overline{X}_{min}$ è la media dei pixel del cluster (che ha ottenuto il valore minimo), $K_{min}$ è il numero del cluster che ha ottenuto il valore medio minimo; in questo caso il numero di cluster $K$ era uguale a 4.

Tutti gli elementi $x$, presenti nel cluster $K_{min}$, sono stati posti a zero. In questo modo le regioni della cattura endoscopica che risultano troppo scure vengono rimosse dall'immagine.
Al fine di rendere i bordi nella maschera più morbidi e regolari e meno dentellati è stato utilizzato l'operatore morfologico di chiusura. Un operatore morfologico è uno strumento matematico che, data un'immagine tenta di modificarla tenendo conto della morfologia degli oggetti presenti in essa. Il fulcro di questa tecnica è l'elemento strutturale, esso infatti definisce come lavorerà l'operatore morfologico quando questo passa sull'immagine. Esso può assumere qualsiasi forma e, una volta creato, scorre sull'immagine come fosse il kernel (risposta all'impulso) di un filtro, generando una convoluzione lineare locale fra elemento strutturale e immagine; in genere l'elemento strutturale è una maschera binaria.
Tramite la funzione *strel* (contrazione di *structuring element*, è stato creato un elemento strutturale non piatto, ma emisferico, di raggio e altezza pari a 7 pixel, che presenta il peso più alto al centro e che si assottiglia verso i bordi (elemento *"ball"* su MATLAB). Questo elemento strutturale è stato preferito rispetto ai più classici elementi 2-D in quanto ha fornito risultati migliori.
Vi sono due tipologie di operatore morfologico, l'erosione e la dilatazione. L'erosione seleziona un certo numero di pixel dell'immagine in ingresso e, mediante l'elemento strutturale, in uscita sceglie il minimo di tutti i pixel selezionati. Per un'immagine in scala di grigi, l'erosione di $f$ da parte di $b$ può essere scritta matematicamente come:

$$(f \ominus b)(x) = \inf_{y \in B}[f(x+y) - b(y)]$$

dove $f(x)$ rappresenta l'immagine e $b(x)$ l'elemento strutturale in scala di grigi, definito nello spazio $B$ [41]. Questo visivamente ha un effetto di erosione degli oggetti nell'immagine.
La dilatazione è il suo duale, esso seleziona un certo numero di pixel dell'immagine in ingresso e, mediante l'elemento strutturale, in uscita sceglie il massimo di tutti i pixel selezionati. Per un'immagine in scala di grigi, la dilatazione di $f$ da parte di $b$ può essere scritta matematicamente come:

$$(f \oplus b)(x) = \sup_{y \in E}[f(y) + b(x-y)]$$

dove $E$ rappresenta lo spazio euclideo definito in $\mathbb{R} \cup \{\infty, -\infty\}$ dove $\mathbb{R}$ è l'insieme dei numeri reali, $f(x)$ rappresenta l'immagine e $b(x)$ l'elemento strutturale in scala di grigi [41]. Questo ha l'effetto visivo di rigonfiamento degli oggetti nell'immagine, ma anche eliminare le discontinuità e chiudere eventuali buchi presenti.
La combinazione di queste due tipologie in cascata può creare due effetti differenti, definendo due nuovi operatori. Nel caso in cui l'erosione sia seguita da dilatazione si parlerà di apertura, nel caso in cui la dilatazione sia seguita da erosione si parlerà di chiusura. In entrambi i casi la prima tipologia di operatore utilizzata nella cascata sarà quella principale. Lo scopo dell'apertura è quello di separare eventuali oggetti sovrapposti o adiacenti, quello della chiusura è di unire oggetti molto vicini e appunto chiudere eventuali buchi.

Per effettuare l'operazione di chiusura è stata utilizzata la funzione built-in di MATLAB *imclose*, con l'elemento strutturale sopra descritto.
La funzione di chiusura è stata applicata singolarmente ad ogni layer RGB, questi sono stati poi ricombinati fra di loro al fine di ottenere un'unica maschera binaria bianco/nero. Moltiplicando punto a punto l'immagine originale ritagliata con la maschera binaria, è stato possibile rimuovere (ponendo a zero) tutti i pixel troppo scuri dell'immagine endoscopica.
In **Figura 3.5** è presente un esempio di rimozione dei pixel troppo scuri, privi di informazione.

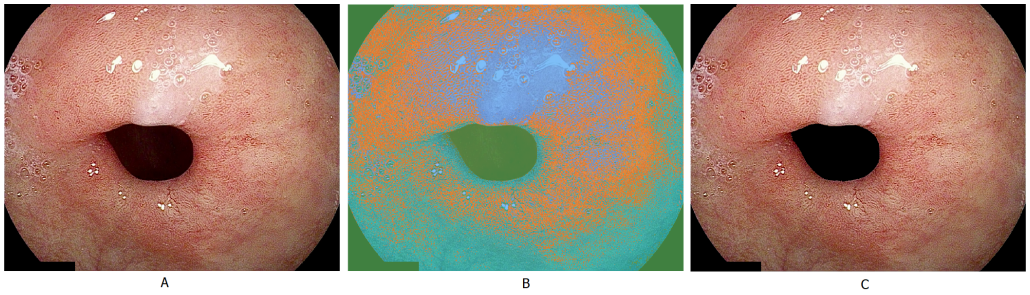



**Figura 3.5: La rimozione delle parti scure dell'immagine.** *(A) L'endoscopia video ritagliata e priva di testo. (B) L'immagine è stata suddivisa in quattro cluster della funzione imsegkmeans. In violetto il cluster dei pixel più chiari, in arancio e celeste i cluster intermedi, in verde il cluster dei pixel più scuri (C) In questa immagine il cluster più scuro è stato rimosso ponendo a 0 tutti i pixel appartenenti a questo cluster. Come si può notare, il colore del piloro è passato da rosso scuro a completamente nero*

L'ultimo elemento di disturbo è rappresentato della aree troppo chiare presenti all'interno delle immagini endoscopiche. Questi sono principalmente riflessi indesiderati dovuti alla presenza della luce bianca della sonda endoscopica. Esse non rappresentano una zona di interesse e pertanto devono essere eliminate dall'immagine.
La parete dello stomaco è ricoperta di secrezioni e questo rende la superficie altamente riflettente alla luce, in particolar modo quella bianca. Per questo motivo è stato necessario rimuovere i pixel troppo chiari, presenti all'interno dell'immagine endoscopica.
Anche in questo caso è stata utilizzata la funzione *imsegkmeans*, ma con numero di cluster pari a 6. Successivamente, di ogni gruppo ottenuto, è stata calcolata la mediana, il cluster scelto è stato poi quello con intensità di colore maggiore, come da formula seguente:

$$[\tilde{X}_{max}, K_{max}] = max(median(X_1), median(X_2), median(X_3), ..., median(X_K))$$

dove $X_1, X_2, X_3, ..., X_K$ sono i $K$-esimi cluster $X$, $\tilde{X}_{max}$ è la mediana dei pixel del cluster (che ha ottenuto il valore massimo), $K_{max}$ è il numero del cluster che ha ottenuto il valore di mediana massimo; in questo caso il numero dei cluster $K$ era pari a 6.
Una volta ottenuto $\tilde{X}_{max}$, ne è stato calcolato il 95%, questo rappresenta il valore di soglia $Th$. Esso è stato ottenuto con la semplice formula:

$$Th = 0.95 \cdot \tilde{X}_{max}$$

Per ogni pixel dell'immagine endoscopica è stata calcolata la media dei tre layer RGB. Tutti i pixel che avevano un valore medio superiore alla soglia $Th$ sono stati posti a zero. A tutti gli altri pixel è stato dato invece valore 255, in modo da creare una maschera binaria bianco/nero.
Al fine di eliminare eventuale rumore, sull'intera maschera è stato applicato un filtro mediana, di finestra $5 \times 5$, esso è un filtro di tipo passa-basso, non lineare [42]. Quando la finestra passa sull'immagine, viene calcolata la mediana di quell'insieme di valori, il corrispondente valore in uscita del filtro sarà appunto la mediana dei pixel selezionati dalla finestra. In questo modo sono stati creati bordi netti nella segmentazione ed è stato rimosso il classico rumore sale e pepe [42]. La maschera binaria così ripulita è stata poi moltiplicata punto a punto con l'immagine RGB ottenuta precedentemente, eliminando così i riflessi di luce indesiderati.
Nella **Figura 3.6** è presente un esempio di rimozione dei pixel troppo chiari, dovuti ai riflessi della luce sulla parete dello stomaco.

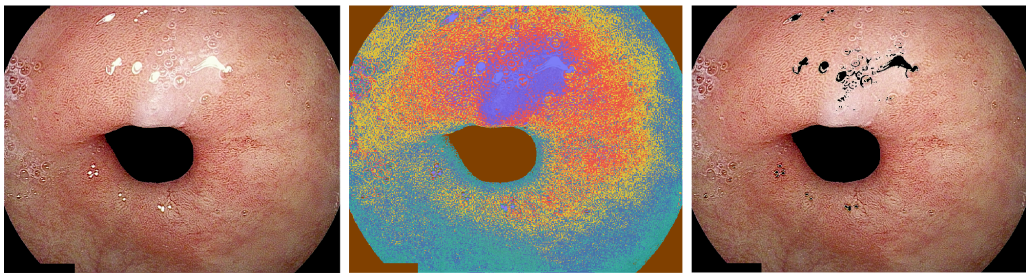



**Figura 3.6: La rimozione delle parti chiare dell'immagine.** *(A) L'endoscopia video ritagliata, priva di testo e di zone scure. (B) L'immagine è stata suddivisa in sei cluster della funzione imsegkmeans, in violetto il cluster dei pixel più chiari. (C) In questa immagine sono stati eliminati tutti i pixel con media di valori RGB superiore alla soglia Th.*

Alla fine di quest'ultimo processo le immagini risultano ritagliate, ripulite da testo non voluto, prive di zone troppo scure e di riflessi di luce. Nella **Figura 3.7** vengono mostrati in successione gli effetti dei quattro principali passi della pulizia e ritaglio della cattura endoscopica.

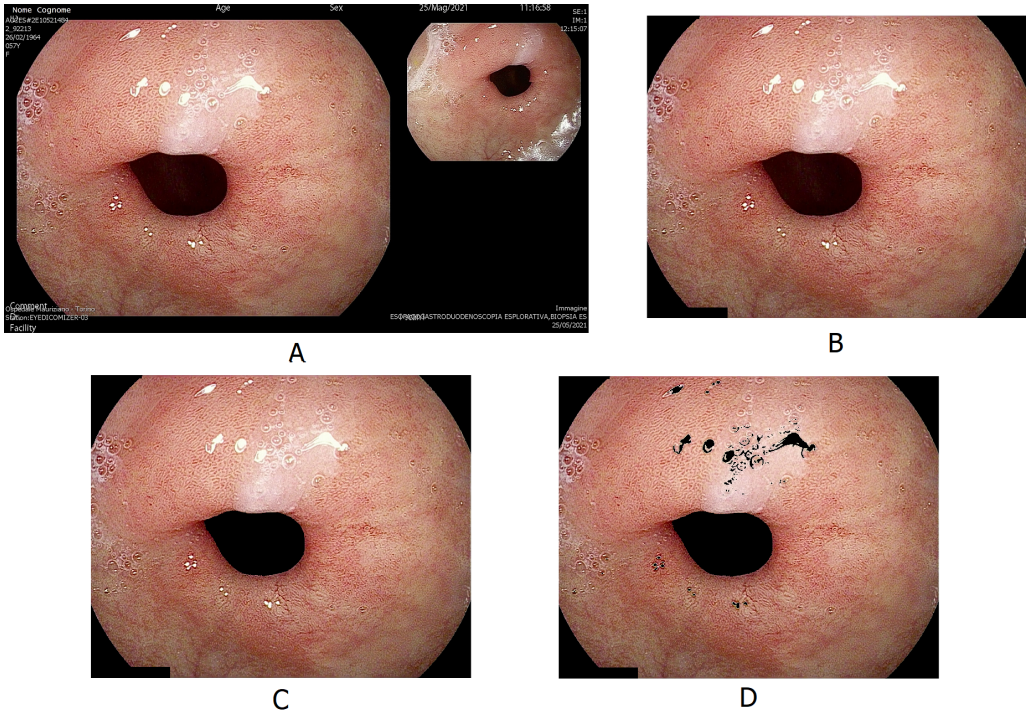



**Figura 3.7: I principali passi della pulizia dell'immagine.** *(A) La cattura endoscopica originale. (B) L'immagine priva di testo non voluto e ritagliata. (C) L'immagine priva di parti troppo scure, poste tutte in nero. (D) L'immagine priva di parti troppo chiare, poste tutte in nero.*

Una volta ritagliata e rimosse tutte le parti indesiderate, l'immagine è pronta per essere processata.

### 3.2.2 Il miglioramento dell'immagine

Le immagini ripulite dagli artefatti e dalle zone non desiderate sono state convertite da RGB a bianco/nero, prima di effettuarne il miglioramento. Quest'ultimo processo, meglio noto in inglese come *image enhancement*, è finalizzato a facilitare l'estrazione delle caratteristiche tissutali.
La classificazione è infatti basata unicamente sulle informazioni estratte dalla texture della parete gastrica. Risulta pertanto fondamentale poter mettere in risalto eventuali rugosità della superficie, o specifici pattern del tessuto o particolari motivi regolari [43].
Data l'eterogeneità del campione di immagini è stata necessaria una sua normalizzazione. Tutte le immagini, sia quelle provenienti da pazienti sani che quelle provenienti da pazienti patologici sono state convertite in scala di grigi. Il passo successivo è stato quello di effettuare una standardizzazione globale del campione, tramite la seguente formula:

$$z = \frac{x - \mu}{\sigma}$$

dove $z$ è l'immagine standardizzata, $x$ è l'immagine originale priva di artefatti, $\mu$ è la media globale del valore di tutti i pixel del campione di immagini e $\sigma$ è la deviazione standard globale del valore di tutti i pixel del campione di immagini.
L'intero campione è stato poi normalizzato fra 0 e 1 dove, in scala di grigi, lo 0 rappresenta il nero e l'1 il totalmente bianco.
Al fine di valorizzare ed esprimere maggiormente la texture del tessuto gastrico è stato applicato un filtro di tipo passa-alto all'immagine. Questo genere di filtri vengono impiegati per riconoscere le discontinuità e matematicamente equivalgono ad effettuare l'operazione di derivata del primo ordine. Per questa ragione i filtri passa-alto vengono anche definiti filtri derivativi.
Essendo l'immagine definita nel discreto, l'operazione matematica effettuata sarà una differenza. Questa viene eseguita tramite una finestra che scorre lungo l'immagine ed effettua una convoluzione bidimensionale. Solitamente la finestra utilizzata ha dimensione dispari in modo da ottenere un punto centrale di simmetria.
In generale l'uscita del filtro derivativo è sempre zero tutte le volte in cui la finestra si trova un una regione omogenea dell'immagine. Nel caso in cui la finestra si trovi a cavallo di una discontinuità, questa darà in uscita un valore diverso da zero.
I filtri derivativi possono essere applicati per qualsiasi direzione dello spazio, in questo lavoro è stato scelto di applicare le sole direzioni $x$ e $y$ ortogonali all'immagine stessa, ottenendo cosi i due gradienti $G_x$ e $G_y$.
I seguenti sono definiti kernel derivativi base, rispettivamente calcolano i gradienti lungo $x$ e lungo $y$:

$$G_x = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$$

$$G_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

Il seguente, più complesso, è l'evoluzione dei kernel precedenti, esso è definito come gradiente di Prewitt, ed è una matrice $3 \times 3$. Questo kernel rispetto al precedente, effettua anche un filtraggio di tipo passa-basso [44]. Esso avviene perché il filtro esegue intrinsecamente anche l'operazione di media, nella direzione ortogonale a quella della derivata. Di seguito i kernel che calcolano i gradienti lungo $x$ e lungo $y$ [44]:

$$G_x = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$G_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Per ovviare al problema della media pesata si può attribuire un peso doppio alla riga o colonna centrale, dipendentemente dalla direzione della derivazione. Questo viene chiamato gradiente di Sobel, esso fornisce prestazioni migliori rispetto ai due

precedenti. Di seguito rispettivamente i kernel che calcolano i gradienti lungo $x$ e lungo $y$ [44]:

$$G_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

$$G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Proprio questo gradiente è stato scelto per evidenziare le discontinuità della superficie dei tessuti gastrici. Per calcolare i gradienti $G_x$ e $G_y$ di Sobel è stata utilizzata la funzione *imfilter* di MATLAB, con il parametro *"circular"* per la gestione dei bordi. In particolare quest'ultimo calcola i limiti estremi dell'immagine assumendo implicitamente che l'input sia periodico.
Una volta ottenuti i due gradienti separatamente, questi sono stati sommati fra di loro, in modo da ottenere un'unica mappa che rappresenti le discontinuità su entrambi gli assi $x$ e $y$.
L'immagine endoscopica in bianco/nero è stana normalizzata fra 0 e 1, passando dal formato *uint8* al formato *double*. A questa è stata sommata la mappa di discontinuità ottenuta precedentemente. La somma delle due immagini è stata poi normalizzata fra 0 e 1, saturando i valori estremi. Nello specifico i valori negativi sono stati portati a 0 e quelli maggiori di 1 al valore 1.
In **Figura 3.8** un diagramma di flusso che rappresenta i principali passaggi dell'estrazione dei gradienti e del miglioramento dell'immagine.
L'immagine cosi ottenuta è stata poi moltiplicata punto a punto per la maschera binaria in bianco/nero, ottenuta nel processo di rimozione degli artefatti, con il fine di eliminare possibili elementi indesiderati.

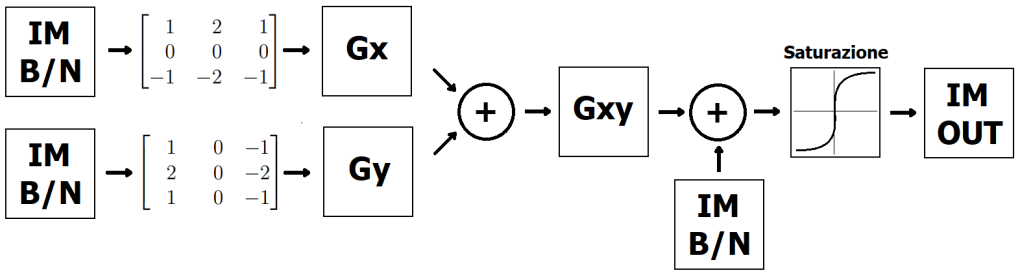



**Figura 3.8: I principali passi del miglioramento dell'immagine.** *L'immagine endoscopica in bianco/nero è stata filtrata con le matrici di Sobel in modo da ottenere i gradienti $G_x$ e $G_y$. Una volta ottenuti separatamente, questi sono stati sommati fra di loro in modo da ottenere un'unica mappa di discontinuità. La mappa così ottenuta è stata poi sommata all'immagine originale in bianco/nero; questo passaggio crea bordi dell'immagine molto accentuati, evidenziando le discontinuità. Una volta sommate le immagini è stato necessario saturare i valori fra 0 e 1, poiché l'immagine sommata presentava sia valori negativi che valori maggiori di 1.*

Tramite questo processo è stato possibile evidenziare le caratteristiche macroscopiche della parete interna dell'organo gastrico. Sono state accentuate le irregolarità della superficie e i motivi caratteristici della texture che altrimenti sarebbero stati piatti e non altrettanto evidenti.
In **Figura 3.9** due esempi di miglioramento dell'immagine tramite la tecnica dell'enfatizzazione delle discontinuità.

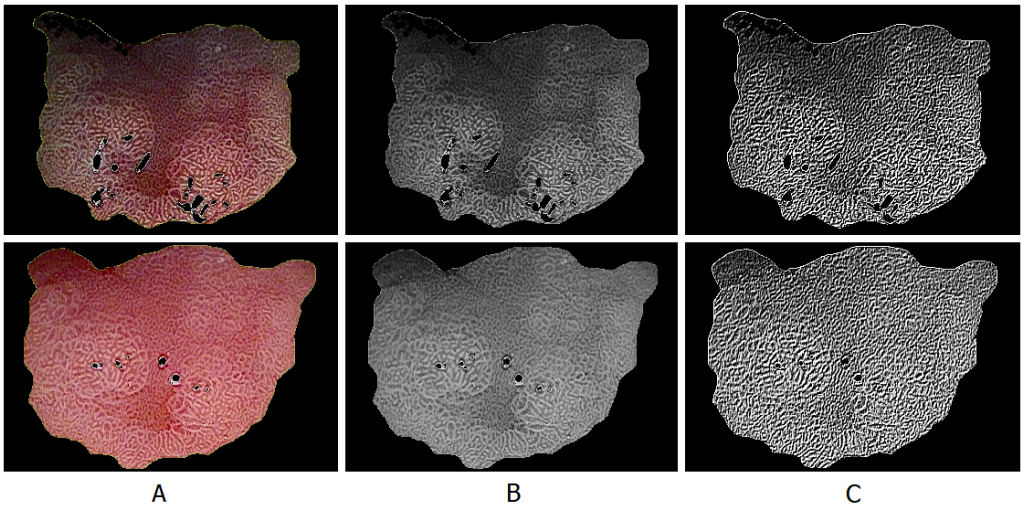



**Figura 3.9: Il miglioramento delle immagini endoscopiche.** *(A) Nelle figure in colonna sono rappresentate due immagini endoscopiche a colori di regioni patologiche dello stesso tessuto, prese da due angolazioni differenti, alla quali sono state rimosse le zone troppo scure e quelle troppo chiare. (B) Lo stesso tessuto ma in scala di colori in bianco/nero, senza il miglioramento dell'immagine. (C) I due tessuti a seguito della normalizzazione della scala di grigi e del miglioramento dell'immagine tramite enfatizzazione dei gradienti.*

La tecnica che è stata utilizzata viene comunemente chiamata *image sharpening*, che in inglese significa affilatura, ma anche nitidezza. Questa tecnica viene appunto utilizzata per rendere i bordi più definiti, quindi appuntiti, accentuati. Esso ha l'effetto visivo opposto dello *smoothing*, ossia la levigatura creata dai filtri passa-basso, per rimuovere il rumore. Questa tecnica riconosce le discontinuità e le enfatizza, contestualmente all'aumento della definizione dei bordi, introduce però un aumento del rumore.
I due processi di rimozione degli artefatti e di miglioramento delle immagini sono comuni sia al dataset di pazienti sani che a quello di pazienti patologici. Il processo seguente, invece, di segmentazione automatica delle regioni patologiche è stato eseguito solo sulle immagini di pazienti affetti da metaplasia intestinale gastriche, appunto le catture endoscopiche sulle quali il medico ha segmentato manualmente le regioni patologhe.

### 3.2.3 La segmentazione delle regioni patologiche

Le immagini provenienti da pazienti patologici sono state segmentate manualmente dal medico. Tramite l'utilizzo di un computer l'operatore ha delimitato manualmen-

te, con un puntatore, le regioni che riteneva essere affette da metaplasia intestinale gastrica.
Il tratto scelto era abbastanza spesso e di colore verde chiaro, questa tonalità è stata scelta poiché risultava la più dissimile rispetto ai tessuti gastrici. Le regioni delimitate hanno un bordo più o meno regolare dipendentemente dalle immagini e un tratto mediamente dentellato.
Le regioni segmentate ricoprono una superficie che va dal 10% dell'immagine endoscopica fino al 90% della stessa. Per ogni immagine vi sono dall'una alle 5 isole patologiche.
Il primo passo è stato ricavare il valore RGB del tratto verde utilizzato dal medico. Questo non aveva un entità unica e precisa ma piuttosto variabile, soprattutto lungo i bordi. Per questa ragione non è stata scelta un'unica tripletta RGB, ma un intervallo di valori.
Per definire l'intervallo è stato preso il valore medio del tratto verde ($R = 175$, $G = 230$, $B = 35$) e per ogni valore, in scala da 0 a 255, è stato presa una fascia $\pm 25$, in particolare:

- **Layer rosso:** $150 \leq R \leq 200$
- **Layer verde:** $205 \leq G \leq 255$
- **Layer blu:** $10 \leq B \leq 60$

Una volta definito l'intervallo è stata creata una prima maschera binaria in cui tutti i pixel dell'immagine erano 0, quindi neri, ad esclusione di quelli appartenenti alla fascia di colori descritta sopra, posti invece a 255, in bianco.
Sopra questa immagine è stata utilizzata la funzione *imfill* di MATLAB. Essa implementa un algoritmo basato sulla ricostruzione morfologica, che, combinato con il paramento *"holes"*, permette di riempire i buchi presenti in un immagine. In questo modo è stato possibile porre in bianco (quindi valore 255) anche i pixel presenti all'interno dell'insieme chiuso ricavato sopra.
In questa fase è stato fondamentale che le segmentazioni effettuate dal medico fossero insiemi perfettamente chiusi, separati e non adiacenti fra di loro.
Tramite questo algoritmo però si è presentato il problema che, all'interno della maschera bianca, erano presenti anche i tratti verdi delineati dal medico. Essendo questi non voluti, è stato applicato nuovamente l'intervallo di valori RGB scelto sopra, ponendo questa volta i bit a 0, quindi rendendoli neri.
Moltiplicando punto a punto le due maschere è stato così rimosso il tratto verde dalle regioni segmentate.
Al fine di rendere i bordi lisci ed omogenei nel passaggio bianco/nero, all'intera maschera è stato poi applicato un filtro mediana $5 \times 5$.
Tramite la moltiplicazione punto a punto dell'immagine endoscopica con la maschera bianco/nero sono state estratte le porzioni di tessuto affette da metaplasia intestinale gastrica.
Dualmente a quanto visto sopra è stato possibile estrarre anche le porzioni di tessuto sano provenienti dalle immagini patologiche. In questo caso in fase di creazione delle maschere è stato invertito il colore bianco con il nero. Non è stato necessario agire sul tratto verde in quanto già rimosso dalla maschera stessa.

In **Figura 3.10** un esempio di immagine endoscopica con due isole patologiche segmentate manualmente dal medico, dalla quale sono state estratte due regioni, una completamente patologica e una regione sana ma proveniente da soggetto patologico.

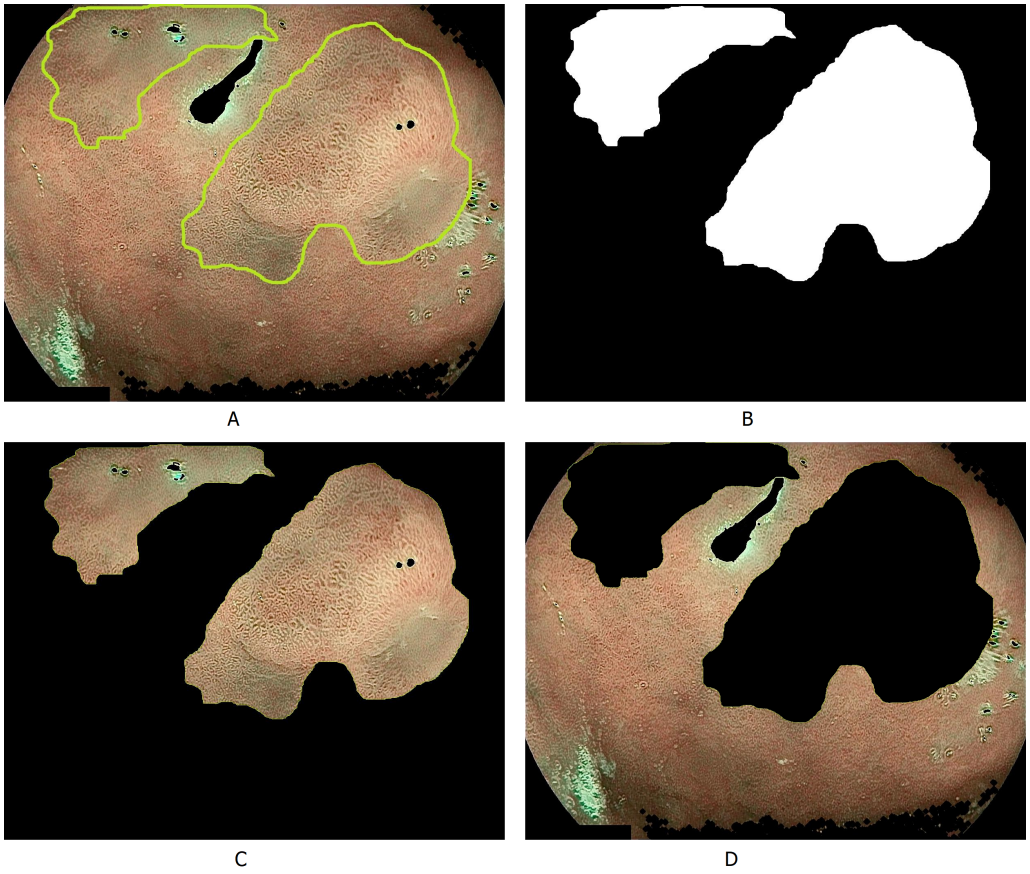



**Figura 3.10: La segmentazione automatica delle regioni patologiche.** *(A) Immagine endoscopica, ripulita da artefatti, zone troppo scure e zone troppo chiare, con due regioni patologiche segmentate manualmente dal medico. (B) La maschera binaria che rappresenta in bianco le regioni affette da metaplasia intestinale gastrica. (C) Le regioni patologiche separate tramite l'algoritmo di segmentazione automatica. (D) Porzione di tessuto sano separato dalle isole patologiche, queste regioni sono state categorizzate come tessuto sano ma proveniente da un soggetto patologico.*

## 3.3 Le regioni di interesse

Le immagini endoscopiche, per essere elaborate, sono state suddivise in piccole finestre, campioni dell'immagine stessa note come regioni di interesse (meglio conosciute come ROI, dall'inglese *region of interest*). Ogni immagine è stata suddivisa in più campioni, circa 600-700 regioni di interesse per ogni cattura endoscopica. Le ROI scelte sono di forma quadrata e di ampiezza $50 \times 50$ pixel. Queste non risultano contigue fra di loro ma sono sovrapposte sia orizzontalmente che verticalmente del 40%, quindi 20 pixel, la sovrapposizione delle finestre viene comunemente chiamato *overlap*.

La tecnica dell'*overlap* è stata utilizzata per aumentare il numero dei campioni per ogni immagine, essendo il numero delle immagini endoscopiche limitato. In questo modo, una porzione dell'immagine può essere analizzata fino a quattro volte, come mostrato in **Figura 3.11**

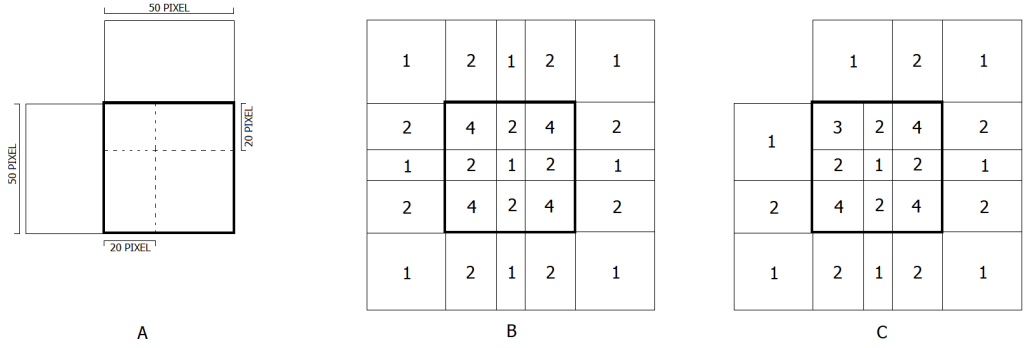



**Figura 3.11: L'overlap delle finestre.** *(A) Le finestre che campionano le immagini endoscopiche hanno ampiezza* $50 \times 50$ *pixel, con un overlap di 20 pixel sia verticalmente che orizzontalmente. (B) Una ROI nel caso in cui sia presente overlap su tutti e 4 lati e tutti e 4 gli spigoli. I numeri da 1 a 4 rappresentano il numero di ROI che sono sovrapposte contemporaneamente in quella porzione (C) Una ROI nel caso in cui lo spigolo in alto a sinistra non sia sovrapposto da un'altra finestra, in quel caso nello spigolo vi sono 3 finestre sovrapposte contemporaneamente.*

L'ampiezza delle finestre è stata scelta tale, non troppo grande, in modo da ottenere una buona quantità di dati ma al contempo abbastanza estesa per poter osservare correttamente le strutture tipiche del tessuto in esame. Un finestra più piccola infatti avrebbe sicuramente generato un maggior numero di campioni, ma non avrebbe permesso di apprezzare ottimamente la texture del tessuto gastrico.
Durante la fase di pre-processing delle catture endoscopiche, le parte delle immagini non desiderate sono state poste volutamente in nero, con valore di intensità di colore a 0. Queste parti sono state perciò assimilate alle parti nere nelle maschere binarie, appunto porzioni di immagine non volute.
La scelta di porre queste parti a valore 0 è stata fondamentale per la successiva fase di estrazione delle caratteristiche. Infatti tutte le ROI che avevano una quantità di pixel neri superiore all'1% non sono state prese in considerazione dall'algoritmo. Ciò significa che sono state scartate tutte le finestre che possedevano più di 25 pixel neri. Questo perché era importante calcolare le caratteristiche solo su finestre che fossero interamente popolate da tessuto gastrico, e non da parti indesiderate.
L'estrazione delle caratteristiche è stata effettuata quindi solo sulle ROI che possedevano una percentuale di pixel di tessuto maggiore o uguale al 99% della totale (2500 pixel).
La scelta di porre un margine dell'1% è stata effettuata perché l'intera cattura endoscopica risultava puntinata da pixel neri, dovuti a rumore sale e pepe e dall'eliminazione dei pixel troppi chiari in fase di pre-processing. Ponendo un valore del 100% sarebbero state eliminate troppe ROI, riducendo al minimo l'ampiezza del dataset.

### 3.3.1 L'incremento delle ROI

Con l'algoritmo di selezione delle finestre descritto sopra, sono state ricavate circa 12.200 ROI utili di tessuto patologico e 33.200 ROI utili di tessuto sano. Questo è dovuto al fatto che le porzioni di tessuto patologico risultavano essere nettamente più piccole rispetto all'intera immagine endoscopica, a volte anche solo il 10% del totale.
Per ovviare a questo squilibrio delle due classi sono state utilizzate due strategie. La prima è stata quella di ruotare le immagini endoscopiche di soggetti patologici di 90°, in senso antiorario. Grazie a questo stratagemma le ROI di tessuto sano sono più che raddoppiate. Questo è stato possibile in quanto la texture e il pattern del tessuto interno dello stomaco non ha un asse di simmetria o una specifica orientazione, ma risulta piuttosto caotico e casuale. Le nuove finestre ottenute sono state circa 12.500, portando il totale delle ROI patologiche a circa 24.700 unità.
La seconda strategia ha invece una doppia valenza. Si è assunto che le segmentazioni effettuate dal medico non fossero perfette ma piuttosto affette da errori soggettivi di valutazione. Esse sono state segmentante da un unico operatore e le diagnosi non sono state confrontate con altri medici, per un secondo parere. Quindi in realtà le zone che il medico ha segnalato come sane, all'interno di catture endoscopiche patologiche, possono in realtà contenere piccole porzioni di tessuto patologico. L'algoritmo utilizzato per separare queste due classi verrà descritto in maniera più approfondita nella sottosezione 3.5.2.
Questo ha permesso in primis di ampliare il numero di ROI in generale, ma anche di aumentare significativamente il numero di ROI di tessuto affetto da metaplasia intestinale gastrica.
Grazie a questa seconda strategia sono state estratte ulteriori 27.800 osservazioni, che sono state categorizzate come ROI sane ma provenienti da immagini patologiche. Con le due strategie il dataset è stato ampliato di 40.300 osservazioni, portando il totale delle ROI da 45.500 elementi a 85.800.

## 3.4 L'estrazione delle caratteristiche

L'obiettivo di un classificatore è trovare un modo sistematico di prevedere un fenomeno, dato un insieme di misure. Nell'ambito del *machine learning* con apprendimento supervisionato, questo obiettivo è formulato come il compito di dedurre dai dati raccolti un modello che preveda il valore di una variabile di output, in base ai valori osservati delle variabili di input. In quanto tale, trovare un modello appropriato si basa sul presupposto che la variabile di output non assuma il suo valore a caso e che esista una relazione tra gli input e l'output.
Per dare una formulazione generica, si assume che i valori di input siano $x_1, x_2, ..., x_n$, dove $xj \in X_j$, (per $j = 1, 2, ..., n$) corrisponde al valore della variabile di ingresso $X_j$.
Insieme, i valori di input $(x_1, x_2, ..., x_n)$ formano un vettore di input n-dimensionale $x$, che assume i suoi valori in $X_1 \times ... \times X_n = X$ dove $X$ è definito come lo spazio di input. Allo stesso modo, si definiscono $y \in Y$ il valore della variabile di output $Y$, dove $Y$ è definito come lo spazio di output.

Per definizione, si presume che sia lo spazio di input che quello di output contengano rispettivamente tutti i possibili vettori di input e tutti i possibili valori di output. Si noti che le variabili di input sono talvolta note come caratteristiche, i vettori di input come campioni e la variabile di output come target.
Tutte le variabili si distinguono in due tipologie, variabili qualitative i cui valori sono simbolici, come il genere o la condizione, e variabili quantitative i cui valori sono numeri reali. In questo lavoro di tesi, per le variabili di input, sono state utilizzate solo quelle di tipo quantitativo e da qui in poi verranno chiamate variabili o caratteristiche.
Per ogni ROI $50 \times 50$ sono state estratte 35 variabili, dove per ogni caratteristica è stata creata una funzione di estrazione, utilizzando l'algoritmo descritto nella sezione 3.3. Per ogni ROI è stato estratto un solo valore numerico, e l'output della funzione di estrazione è una matrice composta da due vettori colonna, il primo contenente le coordinate del pixel posto al centro della ROI, espresso in indici lineari, il secondo contenente il valore numerico relativo a quella caratteristica, in quella specifica ROI. Sulle righe, sono quindi presenti le osservazioni, il numero delle righe è quindi variabile per ogni immagine, dipendentemente dal numero di caratteristiche estratte.

### 3.4.1 Le caratteristiche del primo ordine

La texture di un'immagine è determinata dal modo in cui i livelli di grigio dei pixel sono distribuiti in un'immagine [45]. Non esiste una definizione univoca di texture, ma ci si limita a descrivere se l'immagine ha un aspetto fine o grossolano, liscio o irregolare, omogeneo o disomogeneo [45]. Quindi le sue caratteristiche vengono espresse quantificando le proprietà dei livelli di grigio dei pixel di una data immagine.
Le caratteristiche del primo ordine prendono in considerazione le intensità di grigio dei pixel, senza entrare nel merito della loro distribuzione spaziale, ma solo in maniera assoluta.
Data una variabile casuale $i$ che rappresenta i livelli di grigio della regione dell'immagine, si definisce $P(i)$ come l'istogramma di primo ordine. Esso rappresenta il numero dei pixel per ogni livello di grigio $i$, e descrive la distribuzione di questi a seconda della loro intensità luminosa.
Si definiscono inoltre $N_p$ come il numero di elementi presenti in una ROI, $N_g$ come il numero di livelli discreti di grigio (in questo caso 256 livelli di grigio), $X$ come la totalità degli elementi all'interno della ROI, $X(i)$ come l'i-esimo elemento presente all'interno della regione di interesse.
Si definisce $p(i)$ come l'istogramma di primo ordine normalizzato, ed è descritto come:

$$p(i) = \frac{P(i)}{N_p}$$

In **Figura 3.12** un esempio di istogramma delle luminosità di una finestra $50 \times 50$ con intensità di livelli di grigio dei pixel avente distribuzione gaussiana.

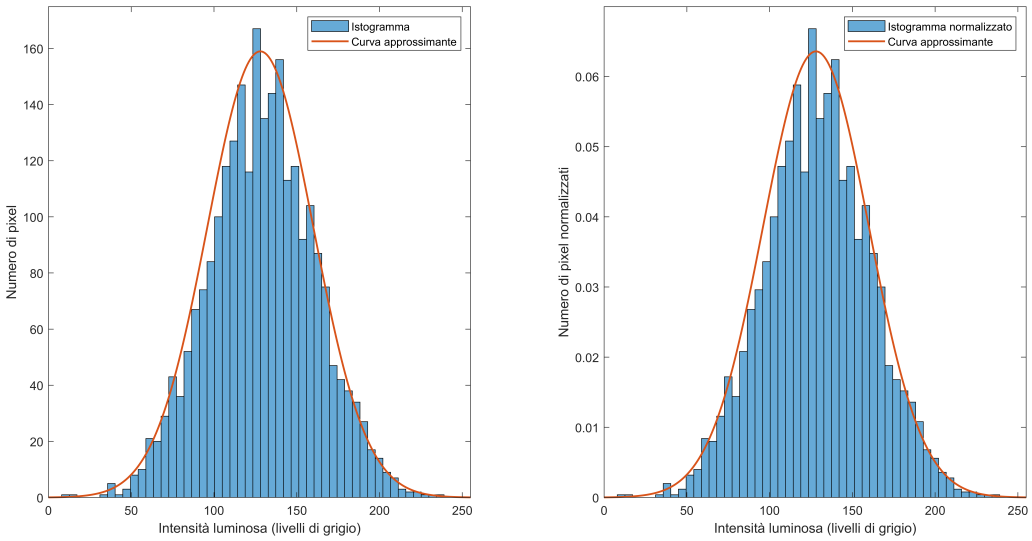



**Figura 3.12: L'istogramma delle luminosità.** *Nella figura di sinistra è rappresentato l'istogramma dell'immagine del primo ordine, sulle ascisse il valore di intensità in scala di grigio, sulle ordinate il numero di pixel presenti nella ROI, per ogni intensità di colore specifica. In rosso la curva approssimante dell'istogramma, in questo caso di andamento gaussiano. Tramite l'istogramma e la sua curva approssimante è possibile calcolare tutte le caratteristiche del primo ordine. Nella figura di destra è rappresentato il medesimo istogramma dell'immagine ma normalizzato rispetto al numero totale di pixel presenti nella ROI, in questo caso 2500 (essendo la finestra di dimensioni* $50 \times 50$*). Quest'ultimo grafico viene chiamato istogramma dell'immagine del primo ordine normalizzato.*

Grazie a questi strumenti è stato possibile estrarre le seguenti 15 caratteristiche del primo ordine.
Come prime caratteristiche sono stati calcolati il minimo e il massimo valore di intensità di grigio presente all'interno della ROI, a seguire la differenza fra le due, chiamata *range*:

$$minimo = min(X)$$

$$massimo = max(X)$$

$$range = max(X) - min(X)$$

A seguire sono stati calcolati la media, definito come il primo momento statistico, e la mediana:

$$media = \frac{1}{N_p} \sum_{i=1}^{N_p} X(i)$$

$$mediana = median(X)$$

Successivamente sono stati calcolati, il decimo percentile, il novantesimo percentile e la differenza fra il venticinquesimo e il settantacinquesimo percentile, nota come differenza interquartile. Il percentile rappresenta una misura statistica per indicare il minimo valore sotto al quale ricade una precisa percentuale di elementi sotto osservazione.

$$P_{10} = prctile(X, 10)$$

$$P_{90} = prctile(X, 90)$$

$$interquartile = P_{75} - P_{25}$$

Sono state inoltre calcolate le caratteristiche più legate alla distribuzione delle intensità di luminosità in toni di grigio. In particolare la deviazione assoluta media o MAD (dall'inglese *mean absolute deviation*) e la media quadratica o RMS (dall'inglese *root mean square*).
Il primo è calcolato come la distanza media di tutti i valori di intensità di grigio, rispetto al valore medio dell'immagine [46] e rappresenta una misura di accuratezza della previsione calcolando la media dell'errore presunto, ossia il valore assoluto di ciascun errore [47].
Il secondo è calcolato come la radice quadrata della media di tutti i valori di intensità al quadrato e rappresenta una misura di grandezza dei valori dell'immagine. Inoltre è stato calcolato lo scarto quadratico medio, noto anche come deviazione standard o STD (dall'inglese *standard deviation*), esso è un indice statistico di dispersione che stima la variabilità di una popolazione o di una variabile casuale, esprimendo la dispersione dei dati intorno ad un indice. Per ultimo è stata calcolata la varianza anche chiamato momento del secondo ordine. Essa fornisce una misura della variabilità dei valori assunti dalla variabile stessa e misura quanto essi si discostano quadraticamente dalla media aritmetica o dal valore atteso. Si noti come la STD sia esattamente la radice quadrata della varianza. Nelle formule seguenti $\overline{X}$ rappresenta il valore medio di tutte le osservazioni, in questo caso la media di intensità tutti i pixel della ROI.

$$MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} |X(i) - \overline{X}|$$

$$RMS = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} [X(i)]^2}$$

$$STD = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} [X(i) - \overline{X}]^2}$$

$$varianza = \frac{1}{N_p} \sum_{i=1}^{N_p} [X(i) - \overline{X}]^2$$

Come caratteristica strettamente collegata all'istogramma dell'immagine, è stata calcolata l'uniformità. Essa è una misura della somma dei quadrati di ciascun valore di intensità [48]. Calcola l'omogeneità dei pixel dell'immagine, dove ad una grande omogeneità corrisponde una grande uniformità e vice versa. Un alto indice di uniformità può essere ottenuto anche se l'intervallo di intensità dei pixel ha un'ampiezza di valori ristretta. Di seguito la sua formula calcolata a partire dell'istogramma dell'immagine.

$$uniformità = \sum_{i=1}^{N_g} p(i)^2$$

Come ultime caratteristiche del primo ordine sono state estratte la *skewness* e la *kurtosis*, rispettivamente momento del terzo e del quarto ordine. Il primo è un indice di asimmetria di una distribuzione e fornisce una misura della sua mancanza di simmetria. Il secondo è un indice relativo alla forma della distribuzione [49]. In una funzione di densità restituisce una misura dello spessore delle code e del grado di appiattimento della distribuzione stessa.

$$skewness = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}[X(i) - \overline{X}]^3}{(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}[X(i) - \overline{X}]^2})^3}$$

$$kurtosis = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}[X(i) - \overline{X}]^4}{(\frac{1}{N_p}\sum_{i=1}^{N_p}[X(i) - \overline{X}]^2)^2}$$

A differenza delle caratteristiche del primo ordine, quelle di second'ordine non prendono in considerazione solo i valori delle intensità dei livelli di grigio dei pixel, ma anche la loro distribuzione spaziale.

### 3.4.2 Le GLCM

Il termine GLCM deriva dall'inglese ed è l'acronimo di *gray level co-occurrence matrix* e grazie a questo strumento è possibile derivare diverse caratteristiche relative alla texture di un'immagine. Queste caratteristiche strutturali di co-occorrenza sono facilmente calcolabili e descrittive, basate su dipendenze spaziali dei livelli di grigio [50].
Possono essere calcolati diversi parametri statistici, come entropia, omogeneità, energia, secondo momento angolare, contrasto, correlazione, varianza e prominenza. Supponendo di analizzare un'immagine quadrangolare di dimensioni $x$ e $y$, si definisce $N_x$ il numero di pixel di risoluzione in direzione orizzontale e $N_y$ il numero di pixel di risoluzione in direzione verticale. Si definisce inoltre $N_g$ il numero di livelli di grigio quantizzati; in questo caso specifico $N_x = N_y = 50$ e $N_g = 256$.
La GLCM è una tecnica statistica del dominio dello spazio che calcola le statistiche di secondo ordine e superiori per le occorrenze di coppie di pixel $(i, j)$, per le quali un livello di grigio $i$ è distanziato da un livello di grigio $j$, di una distanza $\delta$ e lungo una direzione $\theta$ [51]. Si definisce quindi come $P(i, j|\delta, \theta)$ la funzione di probabilità congiunta di secondo ordine di una regione dell'immagine.

Le GLCM non sono altro che matrici di dimensioni $N_g \times N_g$, i cui elementi dipendono dalla relazione angolare tra pixel vicini di un'immagine e dalla distanza tra di loro. Utilizzando una distanza in pixel e angoli quantizzati a intervalli di 45° si possono ricavare quattro matrici: orizzontale, quindi 0°, prima diagonale, a 45°, verticale, a 90°, e seconda diagonale, a 135° [50]. Fissate le variabili di distanza ed angolo, la matrice delle frequenze relative diventa semplicemente $P(i, j)$, che rappresenta la matrice di co-occorrenza per un valore arbitrario di distanza $\delta$ e angolo $\theta$ [52]. In questo caso è stata scelta una distanza in pixel $\delta = 1$ e un angolo $\theta = 0°$.
Dividendo la matrice delle frequenze relative per la somma dei suoi elementi si ricava la matrice normalizzata delle co-occorrenze [52] ed è definita come:

$$p(i, j) = \frac{P(i, j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j)}$$

Si definiscono le proprietà marginali di riga $p_x(i)$ e le proprietà marginali di colonna $p_y(j)$ come segue [52].

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

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

Si definiscono inoltre $\mu_x$ e $\mu_y$, rispettivamente la media di intensità dei livelli di grigio di $p(x)$ e di $p(y)$ e sono descritti dalle formule seguenti [52].

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

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

Infine di definiscono $\sigma_x$ e $\sigma_y$. Questi ultimi rappresentano la deviazione standard di $p(x)$ e di $p(y)$ [52].
La matrici GLCM sono state generate tramite la funzione built-in di MATLAB *graycomatrix*. Essa calcola la GLCM da una versione in scala dell'immagine. Se in ingresso riceve un'immagine binaria, questa la ridimensiona a due livelli di grigio [53]. Se, come in questo caso, l'immagine presenta più livelli di intensità di grigio, la funzione *graycomatrix* la ridimensiona a otto livelli ($N_g = 8$).
Tramite la funzione *graycoprops* sono state calcolate le seguenti quattro caratteristiche statistiche che forniscono informazioni sulla texture dell'immagine [54]. Essa normalizza la matrice GLCM in modo che la somma dei suoi elementi sia uguale a 1. In questo modo ogni elemento $(r, c)$ normalizzato è la probabilità congiunta di occorrenza di coppie di pixel, aventi una relazione spaziale definita con valori di livello di grigio $r$ e $c$ [54].

La prima caratteristica misura le variazioni locali nella matrice GLCM e restituisce una misura dell'intensità del contrasto tra un pixel e il suo vicino. L'output ha un intervallo di valori che va da 0, per un'immagine costante, a $size(GLCM, 1) - 1)^2$, cioè il numero delle righe della GLCM meno 1, al quadrato; in questo caso quindi 49. La proprietà del contrasto è anche nota come varianza o inerzia ed è espresso dalla seguente formula [54].

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

La seconda misura la probabilità congiunta di occorrenza delle coppie di pixel specifiche. Restituisce una misura di quanto sia correlato un pixel rispetto al suo vicino e ha un intervallo di valori che va da $-1$ a $+1$. Quando la correlazione assume valore $-1$ si dice che l'immagine è correlata negativamente, quando invece assume $+1$ si dice che è correlata positivamente. Nel caso di immagine costante la funzione restituisce *NaN* (acronimo inglese di *not a number*), è rappresentato dalla seguente espressione [54].

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

La terza fornisce la somma degli elementi quadrati nella GLCM ed è anche conosciuta come uniformità, uniformità di energia o secondo momento angolare. Restituisce un intervallo di valori che va da 0 a 1, dove vale 1 per un'immagine costante, è descritta come segue [54].

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

La quarta e ultima caratteristica misura la vicinanza della distribuzione degli elementi nella GLCM alla diagonale della GLCM. Assume un intervallo da 0 ad 1, dove 1 rappresenta una matrice GLCM diagonale, ed è calcolato come segue [54].

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

### 3.4.3 Le GLRLM

Un'altra importante classe di caratteristiche di texture dell'immagine è rappresentata dalle GLRLM, acronimo inglese di *gray level run length matrix*. Tramite questa matrice è possibile calcolare numerose caratteristiche di texture, anch'esse basate sull'intensità dei livelli di grigio e sulla loro distribuzione spaziale.
Per *gray level run* si intende un insieme di punti dell'immagine collineari, consecutivi, e aventi lo stesso valore del livello di grigio [55]. Per *length*, invece, la lunghezza di questo insieme, espresso in numero di punti, quindi di pixel.

Si può calcolare una GLRLM di una data immagine, o finestra di osservazione, per tutte le sequenze aventi una specifica direzione $\theta$ [55]. Generalmente si scelgono angoli quantizzati a intervalli di 45°, ricavando quindi quattro matrici: orizzontale, a 0°, prima diagonale, a 45°, verticale, a 90°, e seconda diagonale, a 135°. La GLRLM è definita quindi come $P(i, j|\theta)$ dove l'elemento di matrice $(i, j)$ specifica il numero di volte in cui l'immagine contiene una sequenza di lunghezza $j$, nella arbitraria direzione $\theta$, costituita da punti aventi livello di grigio $i$, o giacenti nell'intervallo di livelli di grigio $i$ [55].
Definita una specifica direzione $\theta$ si ottiene una matrice $P(i, j)$ di dimensioni $N_g \times N_r$ dove $N_g$ è il numero discreto di livelli di grigio e $N_r$ è il numero massimo di *run length* (differenti) che occorrono nell'immagine o finestra di osservazione [55], [56].
Si definiscono inoltre $N_p$ come il numero di pixel presenti nell'immagine, $N_r(\theta)$ come il numero di *run length* presenti nell'immagine, lungo la direzione $\theta$. Questo assume un valore compreso nel seguente intervallo $1 \leq N_r(\theta) \leq N_p$, ed è definito come segue [57].

$$N_r(\theta) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i, j|\theta)$$

Si definisce inoltre $p(i, j|\theta)$ come la GLRLM normalizzata, ossia la $P(i, j|\theta)$ normalizzata rispetto al numero di pixel presenti nell'immagine $N_r(\theta)$, ed è descritta quindi come:

$$p(i, j|\theta) = \frac{P(i, j|\theta)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i, j|\theta)}$$

Tramite gli strumenti sopra descritti, è stato possibile ricavare 13 caratteristiche relative alla texture del tessuto, definite appunto caratteristiche GLRLM, definite come segue.
La prima caratteristica calcolata è la SRE (acronimo inglese di *short run emphasis*). Essa è una misura della distribuzione delle *run length* corte [55].
Questa funzione divide ogni valore di *run length* per il quadrato della sua lunghezza, al denominatore invece il numero totale di *run length* nell'immagine e funge da fattore di normalizzazione [55].
Un valore alto di SRE indica la presenza di trame a tessitura più fine, ed è definita come:

$$SRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta)}{j^2}}{N_r(\theta)}$$

Le seconda caratteristica è la LRE (acronimo inglese di *long run emphasis*) ed è una misura della distribuzione delle *run length* lunghe [55].
Questa funzione moltiplica ogni valore di *run length* per la lunghezza della stessa al quadrato. Il denominatore è un fattore di normalizzazione, come sopra [55].
Un valore alto di LRE indica la presenza di trame più lunghe e forme strutturali più grossolane, ed è definita come:

$$LRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta) j^2}{N_r(\theta)}$$

La GLN (acronimo inglese di *gray level nonuniformity*) misura la somiglianza dei valori di intensità del livello di grigio nell'immagine.
Questa funzione eleva al quadrato il numero di *run length* per ciascun livello di grigio. La somma dei quadrati viene quindi divisa per il solito fattore di normalizzazione. Ciò dovrebbe misurare la non uniformità dei livelli di grigio dell'immagine [55].
Un basso valore di GLN è correlato a una maggiore somiglianza nei valori di intensità, cioè quando le sequenze di pixel sono equamente distribuite nei livelli di grigio [55].

$$GLN = \frac{\sum_{i=1}^{N_g} (\sum_{j=1}^{N_r} P(i,j|\theta))^2}{N_r(\theta)}$$

La GLNN (acronimo inglese di *gray level nonuniformity normalized*) misura la somiglianza dei valori di intensità dei livelli di grigio nell'immagine. Un valore basso di GLNN è associato a una maggiore somiglianza nei valori di intensità. Non è altro che la versione normalizzata della formula GLN.

$$GLNN = \frac{\sum_{i=1}^{N_g} (\sum_{j=1}^{N_r} P(i,j|\theta))^2}{N_r(\theta)^2}$$

La RLN (acronimo inglese di *run length nonuniformity*) è una misura di non uniformità, valuta la somiglianza delle *run length* in tutta l'immagine. Questa funzione eleva al quadrato il numero delle sequenze per ogni lunghezza, la somma dei quadrati viene quindi divisa per il fattore di normalizzazione. Quando assume un valore basso indica una maggiore omogeneità [55].

$$RLN = \frac{\sum_{i=1}^{N_r} (\sum_{j=1}^{N_g} P(i,j|\theta))^2}{N_r(\theta)}$$

La RLNN (acronimo inglese di *run length nonuniformity normalized*) misura la somiglianza delle *run length* in tutta l'immagine, con un valore basso indica una maggiore omogeneità. Non è altro che è la versione normalizzata della formula RLN.

$$RLNN = \frac{\sum_{i=1}^{N_r} (\sum_{j=1}^{N_g} P(i,j|\theta))^2}{N_r(\theta)^2}$$

La RP (acronimo inglese di *run percentage*) misura la grossolanità della texture valutando il rapporto tra il numero di *run length* e il numero di pixel nella ROI.
Assume valori compresi nell'intervallo $\frac{1}{N_p} \leq RP \leq 1$, dove valori più alti indicano che una porzione maggiore del ROI è costituita da *run length* piccole, individua perciò una trama più fine [55].

$$RP = \frac{N_r(\theta)}{N_p}$$

La LGLRE (acronimo inglese di *low gray level run emphasis*) misura la distribuzione dei valori di livello di grigio più bassi, quando assume un valore più alto indica una maggiore concentrazione di bassi livelli di grigio nell'immagine [58].

$$LGLRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta)}{i^2}}{N_r(\theta)}$$

La HGLRE (acronimo inglese di *high gray level run emphasis*) misura la distribuzione dei valori di livello di grigio più alti, quando assume un valore più alto indica una maggiore concentrazione di alti livelli di grigio nell'immagine [58].

$$HGLRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta) i^2}{N_r(\theta)}$$

La SRLGLE (acronimo inglese di *short run low gray level emphasis*) misura la distribuzione congiunta di *run length* più brevi, con valori di livello di grigio più bassi [59].

$$SRLGLE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta)}{i^2 j^2}}{N_r(\theta)}$$

La SRHGLE (acronimo inglese di *short run high gray level emphasis*) misura la distribuzione congiunta di *run length* più brevi, con valori di livello di grigio più elevati [59].

$$SRHGLE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta) i^2}{j^2}}{N_r(\theta)}$$

La LRLGLRE (acronimo inglese di *long run low gray level emphasis*) misura la distribuzione congiunta di *run length* più lunghe, con valori di livello di grigio più bassi [59].

$$LRLGLRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta) j^2}{i^2}}{N_r(\theta)}$$

La LRHGLRE (acronimo inglese di *long run high gray level emphasis*) misura la distribuzione congiunta di *run length* più lunghe, con valori di livello di grigio più elevati [59].

$$LRHGLRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta) i^2 j^2}{N_r(\theta)}$$

### 3.4.4 Le caratteristiche nel dominio della frequenza

Le ultime 3 caratteristiche utilizzate come input del classificatore sono state ricavate analizzando le ROI nel dominio delle frequenze, un'analisi basata sulla scomposizione dei segnali immagine in funzioni seno e coseno, utilizzando la trasformata di Fourier [60].
Sebbene l'analisi spettrale di Fourier sia nata per valutare segnali monodimensionali, tipicamente nel dominio del tempo, sia continui che discreti, è possibile applicare i medesimi algoritmi anche a variabili bidimensionali, come per l'elaborazione di immagini digitali.
Per i segnali nel dominio del tempo la frequenza è rappresentata dall'inverso del periodo ($f = 1/\tau$). Nei segnali nel dominio dello spazio, supponendo di avere un segnale 1D di lunghezza in campioni $M$, con una risoluzione spaziale espressa in punti $p$ per centimetro, gli corrisponde un intervallo di campionamento $\tau = 1/p$. Si ottiene così una frequenza fondamentale $f = p/M$ dove $p$ rappresenta la risoluzione spaziale espressa in punti per centimetro e $M$ il numero di campioni totali del segnale monodimensionale [60].
Poiché le immagini sono segnali discreti non viene applicata la trasformata di Fourier semplice ma la trasformata di Fourier discreta, meglio nota come DFT (acronimo inglese di *discrete Fourier transform*). Per un segnale discreto $g(u)$ di lunghezza $M(u = 0...M-1)$ è definita come segue, dove $i$ è l'unità immaginaria [60].

$$G(m) = \frac{1}{\sqrt{M}} \sum_{u=0}^{M-1} g(u) \cdot [cos(2\pi \frac{mu}{M}) - i \cdot (sin(2\pi \frac{mu}{M})]$$

ossia

$$G(m) = \frac{1}{\sqrt{M}} \sum_{u=0}^{M-1} g(u) \cdot e^{-i2\pi \frac{mu}{M}}$$

Applicando quanto visto sopra ad un segnale 2D è possibile analizzare immagini digitali nel dominio delle frequenze. La trasformata di Fourier infatti non è definita solo per segnali monodimensionali ma per segnali di dimensioni arbitrarie [60]. Per un segnale 2D discreto $g(u, v)$ di lunghezza $M \times N$, la DFT è definita come segue [60].

$$G(m, n) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} g(u, v) \cdot e^{-i2\pi \frac{mu}{M}} \cdot e^{-i2\pi \frac{nv}{N}}$$

ossia

$$G(m, n) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} g(u, v) \cdot e^{-i2\pi (\frac{mu}{M} + \frac{nv}{N})}$$

con le coordinate spettrali $m = 0, ..., M-1$ e $n = 0, ..., N-1$.
In via generale l'implementazione diretta della DFT risulta computazionalmente molto intensiva e non è possibile calcolare questa forma di trasformata di Fourier in

tempi sufficientemente brevi su computer standard. Fortunatamente esistono algoritmi veloci nei quali i risultati intermedi vengono calcolati una sola volta e riutilizzati in modo ottimale più volte. Questo è il caso della trasformata di Fourier veloce, meglio nota come FFT (dall'inglese *fast Fourier transform*).
Essa riduce la complessità temporale del calcolo da $O(M^2)$ a $O(M \log_2 M)$, nonostante questa differenza sostanziale, la DFT e la FFT forniscono esattamente lo stesso risultato [60].
Tramite la funzione built-in di MATLAB chiamata *fft2* è stata calcolata la FFT bidimensionale delle ROI di ampiezza $50 \times 50$. La trasformata $Y$ di una matrice $X$ di dimensioni $m \times n$ viene implementata come segue [61].

$$Y_{p+1,q+1} = \sum_{j=0}^{m-1} \sum_{k=0}^{n-1} \omega_m^{jp} \omega_n^{kq} X_{j+1,k+1}$$

Dove $\omega_m$ e $\omega_n$ sono radici complesse di unità

$$\omega_m = e^{-2\pi i/m}$$

$$\omega_n = e^{-2\pi i/n}$$

e $i$ è l'unità immaginaria, $p$ e $j$ sono indici che vanno da 0 a $m-1$, e $q$ e $k$ sono indici che vanno da 0 a $n-1$.
Il teorema di Parseval afferma che la somma, o meglio l'integrale, del quadrato di una funzione è uguale alla somma, o integrale, del quadrato della sua trasformata [62].
Ciò significa che la distribuzione dell'energia di un segnale alle diverse frequenze è uguale alla densità spettrale di energia. Essa è matematicamente uguale al modulo quadro della sua trasformata di Fourier. Ciò significa che l'energia $E_x$ di un segnale è uguale a

$$E_x = \int_{-\infty}^{+\infty} \|X(f)\|^2 \, df = \int_{-\infty}^{+\infty} \|x(t)\|^2 \, dt$$

Per questo motivo la FFT è stata prima messa in valore assoluto e poi elevata al quadrato, e la densità spettrale di energia $E(f)$ è stata calcolata come:

$$E(f) = |X(f)|^2$$

dove $X(f)$ è la FTT di un generico segnale $x(t)$, in questo caso però nel dominio dello spazio.
Non esiste un metodo semplice per visualizzare funzioni a valori complessi bidimensionali. Un'alternativa consiste nel visualizzare singolarmente le parti reali e immaginarie come superfici 2D. Per questo motivo, la densità spettrale di energia è stata preliminarmente ridimensionata e successivamente centrata in 0.
Nella maggior parte delle immagini naturali, la densità spettrale di energia si concentra alle frequenze più basse con un picco massimo, netto, al centro delle coordinate,

tipicamente in (0,0). I valori dello spettro di energia di solito coprono un ampio intervallo e la loro visualizzazione lineare spesso rende non visibili i valori più piccoli. Per mostrare l'intera gamma di valori spettrali, in particolare i valori più piccoli per le alte frequenze, è comune visualizzare la radice quadrata o il logaritmo dello spettro di energia [60].
In questo caso è stato scelto di utilizzare il logaritmo naturale, avendo l'accortezza di aggiungere un offset minimo pari ad 1 per evitare la presenza di frequenze uguali a 0, che avrebbero portato i valori del logaritmo verso $-\infty$. Una somma di un valore pari ad 1 non va ad inficiare i calcoli successivi poiché l'ordine di grandezza del picco è mediamente di $10^6$-$10^7$, commettendo quindi un errore inferiore alla parte per milione.
Per eseguire invece la centratura in 0 è stata utilizzata la funzione built-in di MATLAB chiamata *fftshift*. Essa prende in ingresso una variabile e la restituisce centrata in 0, spostando la componente a frequenza zero al centro del vettore, ma poiché in ingresso ho una variabile bidimensionale la funzione scambia il primo quadrante di della matrice con il terzo e il secondo quadrante con il quarto [63].
L'ultima operazione è stata trasformare la variabile bidimensionale così ottenuta in un'altra monodimensionale, più facilmente interpretabile. Per eseguire ciò, banalmente, sono state sommate per colonna, tutte le righe della matrice, appiattendo così la superficie e creando una curva 1D. La curva così ottenuta ha un andamento simil gaussiano, e di questa sono stati calcolati,rispettivamente, il valore massimo, la *kurtosis* e la banda, quest'ultimo proprio come fosse uno spettro di un segnale monodimensionale.
In figura **Figura 3.13** i principali passi del calcolo della curva monodimensionale dalla quale sono state estratte le 3 caratteristiche nel dominio della frequenza.

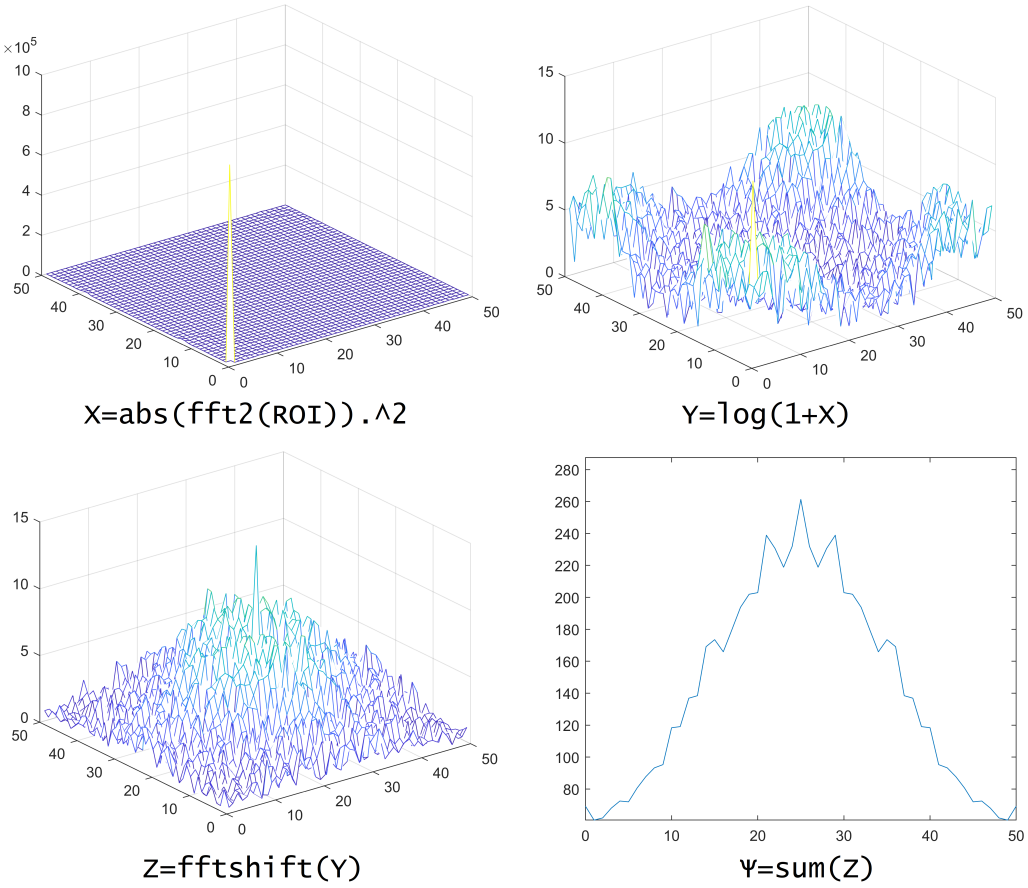



**Figura 3.13: Il calcolo della densità spettrale di energia monodimensionale.** *In figura in alto a sinistra, è rappresentata la densità spettrale di energia di una ROI di ampiezza* $50 \times 50$*, calcolata come il modulo quadro della trasformata di Fourier, la variabile ricavata è stata chiamata* $X$*, sulla base la formula utilizzata su MATLAB per calcolarla. Nella figura in alto a destra il logaritmo della densità spettrale di energia, le basse frequenze vengono così enfatizzate, la variabile ricavata è stata chiamata* $Y$*. In basso a sinistra la traslazione in zero di* $Y$*, la nuova variabile è stata chiamata* $Z$*. In figura in basso a destra la densità spettrale di energia monodimensionale, tramite la funzione somma si è passati da una funzione 2D ad una 1D, la nuova variabile così ottenuta è stata chiamata* $\Psi$*.*

Si definisce $\Psi(f)$ la densità spettrale di energia di un i-esima ROI, dopo averne ricavato il logaritmo, averla traslata in zero e dopo averne ridotto la dimensionalità ad 1D.
La prima caratteristica ricavata è stato il massimo valore, tramite la funzione built-in di MATLAB.

$$fmax = max(\Psi(f))$$

La seconda caratteristica ricavata è stata la *kurtosis* della densità spettrale di energia, con l'apposita funzione di MATLAB.

$$fkurtosis = kurtosis(\Psi(f))$$

La terza e ultima caratteristica calcola la lunghezza di banda del segnale, quest'ultima è stata calcolata tramite la funzione built-in di MATLAB chiamata *powerbw*.

$$bandwidth = powerbw(\Psi(f))$$

Per determinare la larghezza di banda la funzione calcola la differenza di frequenza tra i punti in cui lo spettro scende di almeno 3 dB, al di sotto del livello di riferimento [64]. Nel caso in cui il segnale raggiunge uno dei suoi estremi prima di scendere di 3 dB, *powerbw* utilizza gli stessi estremi per calcolare la differenza [64].
In figura **Figura 3.14** la rappresentazione grafica delle tre caratteristiche in frequenza.

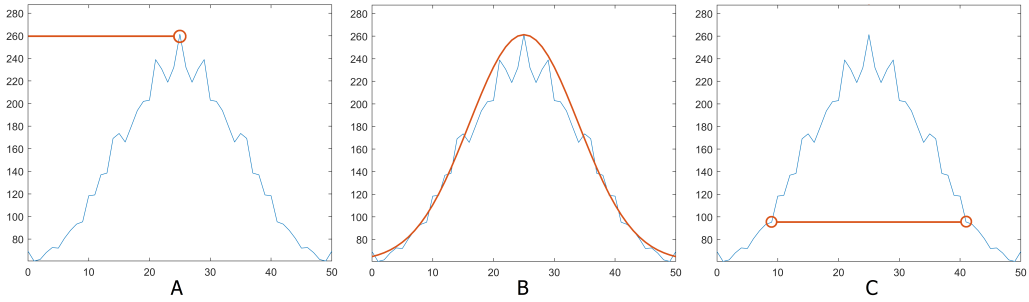



**Figura 3.14: Le tre caratteristiche in frequenza.** *(A) La prima caratteristica calcolata è il massimo valore che assume lo spettro. (B) La seconda caratteristica è la kurtosis dello spettro, in rosso una curva gaussiana approssimante. (C) La terza e ultima caratteristica è la banda passante del segnale* $\Psi$.

## 3.5 La composizione del dataset

Nel complesso sono state estratte 15 caratteristiche del primo ordine, 4 di GLCM, 13 di GLRLM e 3 caratteristiche nel dominio della frequenza. Tutte le variabili estratte sono state riunite in una matrice $n \times N$, dove $n$ rappresenta il numero delle osservazioni e $N$ il numero delle caratteristiche, spesso questa matrice, o dateset, utilizzato per l'addestramento di un classificatore, viene anche definito *learning set*. Prima di essere processate, le variabili sono state normalizzate fra 0 e 1, utilizzando il metodo noto come *min-max scaling*, che segue la seguente formula:

$$x' = \frac{x - min(X)}{max(X) - min(X)}$$

dove $x'$ rappresenta la caratteristica normalizzata, $x$ rappresenta la caratteristica estratta dall'i-esima ROI e $X$ è il vettore contenente tutte le osservazioni della caratteristica. Vista la natura eterogenea del campione di immagini, è stata necessaria un'elaborazione delle osservazioni e un'analisi delle caratteristiche. Queste *feature* infatti, poiché molto simili fra di loro, potevano risultare ridondanti o addirittura non vantaggiose per la classificazione.

### 3.5.1 La selezione delle caratteristiche

Durante la fase di composizione del dataset è stato scelto di non utilizzare algoritmi di riduzione della dimensionalità o di selezione delle caratteristiche.
L'unico intervento che è stato svolto sulle caratteristiche estratte è stato quello di una loro valutazione in termini di distribuzione e variabilità. Risulta essere fondamentale la ricerca delle migliori caratteristiche in modo che definiscano in modo più efficiente l'insieme di dati, al fine di ridurre l'impatto della dimensionalità [65]. Occorre infatti selezionare solo le caratteristiche più importanti e rilevanti, rimuovendo quindi quelle ridondanti e irrilevanti [65].
Lo scopo principale della selezione delle caratteristiche è quello di costruire un sottoinsieme di caratteristiche il più piccolo possibile ma che rappresenti l'intero dataset di input delle caratteristiche [65]. Questa selezione offre numerosi vantaggi, oltre la semplice riduzione delle dimensioni dei dati e dello spazio di archiviazione. Infatti migliora l'accuratezza delle previsioni, evita l'overfitting e riduce i tempi di esecuzione e addestramento [65].
Lo strumento utilizzato per valutare le caratteristiche è stato il diagramma a scatola e baffi, meglio noto come *boxplot*. Esso è uno strumento semplice ma potente per la visualizzazione di un singolo insieme di dati [66].
Viene utilizzato per visualizzare i dati, studiarne la simmetria, la forma e la lunghezza delle code della distribuzione e confrontare in parallelo dataset differenti [66].
Il *boxplot* si presenta come un rettangolo orientato con gli assi di un sistema di coordinate in cui l'asse verticale ha la scala del set di dati. La sua parte superiore e quella inferiore sono disegnate nei quartili superiore e inferiore dell'insieme di dati [66]. Questa casella è tagliata da una linea orizzontale in corrispondenza della mediana, i segmenti estremi superiori e inferiori, definiti come baffi, delimitano invece il valore minimo e il valore massimo dell'insieme [66].
Questo diagramma rende rapidamente disponibili informazioni come la posizione della mediana, la diffusione, l'asimmetria della mediana rispetto al centro, la lunghezza del baffo superiore rispetto alla lunghezza di quello inferiore, la lunghezza delle code del dataset, la distanza tra le estremità dei baffi rispetto alla lunghezza della scatola [66]. Generalmente le informazioni più importanti e interessanti si trovano proprio analizzando gli estremi delle code e grazie a questo strumento è possibile analizzare e confrontare le distribuzioni di molti set di dati contemporaneamente, visualizzando i vari *boxplot* fianco a fianco [66].
Questo è proprio quello che è stato fatto per valutare le caratteristiche estratte. Sono stati cioè affiancati i diagrammi dei dati estratti da immagini patologiche con quello dei dati estratti da immagini sane.
Così facendo è stato valutato che le caratteristiche estratte di *minimo*, *massimo* e *range* erano non efficaci e anzi svantaggiose. Questo è dovuto al fatto che in tutte le finestre il valore minimo era quasi sempre rappresentato dallo 0 mentre il valore massimo era quasi sempre 1. Cosi facendo, anche la caratteristica *range* risulta non utile in quanto differenza delle prime due e quindi assumeva quasi sempre il valore 1.
A seguito di questa valutazione è stato scelto di eliminare le variabili di *minimo*, *massimo* e *range* dall'insieme delle caratteristiche, portando il totale delle variabili

da 35 a 32. I diagrammi a scatola e baffi di tutte le caratteristiche estratte sono consultabili nella sezione Appendice.

### 3.5.2 La struttura del dataset

Le immagini dalle quali sono state estratte le caratteristiche sono molto differenti fra di loro, sia perché provengono da endoscopie diverse, sia perché rappresentano porzioni di stomaco differenti.
Questo comporta un'alta variabilità all'interno del dataset, anche a livello di numerosità dei campioni. Infatti come già accennato, le osservazioni estratte da immagini patologiche sono molto inferiori rispetto alle quelle provenienti da immagini sane.
A causa di ciò, per la classificazione binaria, non è stato possibile semplicemente dividere in due il dataset delle caratteristiche ma è stata necessaria una sua elaborazione preliminare.
La rielaborazione delle immagini ha seguito tre fasi principali. La prima è relativa alla creazione di due dataset, uno di caratteristiche sane e uno di caratteristiche patologiche. La seconda è relativa alla gestione delle caratteristiche estratte da porzioni di tessuto sane ma provenienti da immagini patologiche. La terza e ultima fase è relativa alla ricombinazione delle caratteristiche provenienti dalle due fasi precedenti, al fine di ottenere un unico dataset.
Nella prima fase, il dataset di caratteristiche estratte da immagini patologiche è stato considerato interamente e non elaborato, data la sua ampiezza ridotta. Al contrario è stato rielaborato quello proveniente da immagini sane. Lo scopo era quello di mantenere le dimensioni dei due dataset ragionevolmente equilibrate, cercando quindi di ottenere un rapporto 1:1 fra osservazioni sane e osservazioni patologiche. La scelta di quali osservazioni mantenere e quali invece eliminare non è stata totalmente casuale, ma ponderata.
Il dataset delle ROI sane è stato infatti suddiviso in 3 cluster, tramite la funzione *kmeans* di MATLAB, che utilizza il già citato algoritmo di clusterizzazione K-means per suddividere un generico insieme in più cluster. La scelta di suddividerli in 3 gruppi è stata fatta per mantenere le dimensioni dei cluster sufficientemente equilibrate fra di loro, con un'alta distanza intercluster e una bassa distanza intracluster. Dei tre gruppi così suddivisi ne è stato preso solo il 45% dei campioni. Questo valore è stato scelto al fine di ottenere un equilibrio fra il numero di osservazioni di ROI sane e di ROI patologiche, creando quindi due dataset di dimensioni confrontabili. Le tre frazioni selezionate sono state poi riunite al fine di creare un unico dataset, successivamente le osservazioni sono state rimescolate fra di loro in maniera casuale. Lo scopo di ciò è di creare un dataset più piccolo ma che sia il più possibile rappresentativo del dataset originale. In questo modo è stata ricercata la conservazione delle proporzioni fra le osservazioni, senza quindi effettuare una pura scelta casuale dei campioni. In **Figura 3.15** un'immagine che riassume schematicamente i passi della creazione del dataset delle ROI sane.

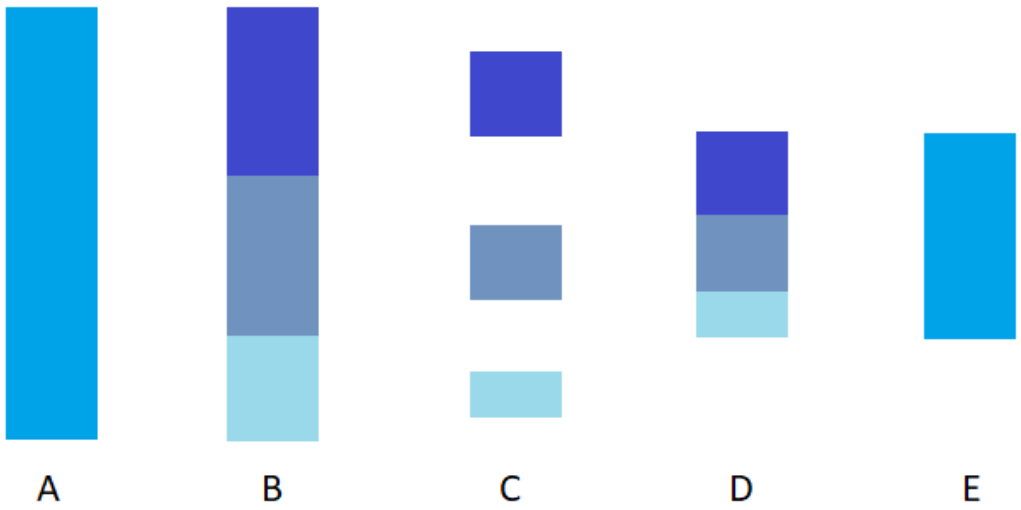



**Figura 3.15: La creazione del dataset di ROI sane.** *(A) Il dataset di ROI originale, completo. (B) Il dataset è stato diviso in 3 cluster tramite l'algoritmo K-means. (C) Di ogni cluster ne è stato selezionato solo il 45%, in maniera casuale. (D) I 3 cluster selezionati sono stati poi uniti fra di loro in un unico dataset ridotto. (E) I dataset ridotto è stato rimescolato in modo da ottenere un campione significativo del dataset originale.*

La seconda fase prevede la gestione delle caratteristiche estratte da ROI sane ma provenienti da pazienti patologici. Queste non sono state classificate né come ROI completamente sane, né come ROI patologiche. A questo gruppo di osservazioni non è stata neanche dedicata una classe a se stante. Sono state invece gestite come un mescolanza di ROI sane e patologiche. Questa supposizione è stata fatta anche sulla base di possibili errori di segmentazione da parte del medico, poiché dettate in buona parte dell'esperienza dell'operatore. L'onere della scelta di come classificare una ROI, se sana o patologica, è stata assegnata all'algoritmo di apprendimento automatico a vettori di supporto, meglio noto come SVM, acronimo inglese di *support-vector machines*. Per per un approfondimento sulle SVM, si veda la sezione Appendice.
Come input alla funzione sono stati inseriti l'intero dataset di ROI sane e patologiche, con le corrispettive etichette di categoria, e come algoritmo di ottimizzazione è stato utilizzato quello di default, l'IDSA. La probabilità a priori della classe non è stata alterata, utilizzando l'impostazione predefinita *"empirical"*. Il classificatore binario così creato è stato utilizzato per classificare le ROI sane ma provenienti da pazienti patologici. Tramite la funzione *predict* è stato discriminato se un ROI era più simile ad una struttura sana, oppure più simile ad una affetta da metaplasia intestinale gastrica.
L'algoritmo di previsione ha classificato il 30-35% circa delle ROI come patologiche, mentre le restanti come sane. Le nuove osservazioni così ottenute sono andate poi ad aggiungersi ai dataset originali, sani o patologici, rispettivamente nel gruppo di appartenenza. Questo processo ha avuto una doppia valenza, sia aumentare il numero delle osservazioni patologiche, visto che avevano una bassa numerosità, e sia aiutare nella classificazione finale dell'endoscopia gastrica.
Riassumendo, il dataset finale presenta quindi la seguente composizione: il 45% di osservazioni provenienti delle ROI sane, selezionate tramite K-means (circa 14.900),

tutte le osservazioni patologiche (circa 12.200), tutte le osservazioni patologiche provenienti da immagini ruotate (circa 12.500), tutte le osservazioni di regioni sane ma provenienti da pazienti patologici e selezionate tramite SVM (circa 27.900 delle quali 18.200 sane e 9.700 patologiche).
Questo ha portato il totale delle osservazioni del dataset a circa 67.500, delle quali 33.200 sane e 34.300 patologiche. Quindi il dataset finale risulta leggermente sbilanciato, come numerosità, verso le osservazioni patologiche, ma in misura trascurabile (49,2%-50,8%).

## 3.6 La random forest

L'intelligenza artificiale è guidata dall'ambizione di comprendere e scoprire relazioni complesse nei dati, ricercando modelli che possano produrre previsioni accurate. La ricerca sull'intelligenza artificiale ha dato origine a numerosi algoritmi e metodi di apprendimento, più o meno complessi. Tuttavia, i metodi basati sugli alberi rappresentano uno dei metodi più efficaci e utili, in grado di produrre risultati affidabili e comprensibili, praticamente su qualsiasi tipo di dati [67]. In particolare, la tecnica è utile per problemi di classificazione in cui si ha un insieme di variabili $X$ e una variabile a risposta singola $Y$ [68]. Per queste ragioni l'algoritmo scelto per classificare le immagini endoscopiche è stato proprio uno basato sugli alberi decisionali: la foresta casuale, meglio conosciuta in inglese come *random forest*.
L'elemento base di questo metodo di classificazione è appunto l'albero, esso consiste in una raccolta di molte regole, determinate da una procedura nota come partizionamento ricorsivo [68].
Questa forma di classificazione, o regola di previsione, è molto diversa da quella fornita da modelli più classici, dove la modalità principale per esprimere le relazioni tra variabili sono le combinazioni lineari. Questa differenza rappresenta sia il punto di forza che quelli di debolezza di questo metodo [68].
Nelle applicazioni in cui l'insieme di predittori contiene un insieme di variabili e fattori numerici, i modelli basati su alberi sono spesso più facili da interpretare e discutere rispetto ai modelli lineari. Questo perché sono invarianti rispetto alle espressioni monotone delle variabili predittrici, in questo modo la forma precisa in cui queste appaiono in una formula del modello è irrilevante [68]. Inoltre il trattamento dei valori mancanti risulta più soddisfacente, cosi come sono più abili nel catturare il comportamento non additivo. il modello lineare standard infatti non consente interazioni tra variabili a meno che non siano prespecificate e di una particolare forma moltiplicativa [68]. Questi modelli sono così chiamati perché il metodo principale per visualizzare l'adattamento si trova sotto forma di un albero binario.
Un albero decisionale consiste in un grafico costituito da un insieme di nodi, che possono essere di tre tipi differenti. Esistono nodi decisionali, solitamente rappresentati come quadrati, nodi casuali, tipicamente rappresentati da cerchi e nodi terminali, solitamente rappresentati da triangoli [69]. Per ogni spigolo si definiscono nodi genitori e nodi figli.
I tre tipi di nodi rappresentano diverse fasi di un problema decisionale sequenziale. In un nodo di decisione, il decisore seleziona un'azione, cioè uno degli spigoli che derivano da questo nodo, ossia uno degli spigoli che ha come genitore il nodo in

questione. In un nodo casuale viene selezionato casualmente uno degli archi che ne derivano, ossia si ottiene una reazione. Infine i nodi terminali rappresentano la fine di una sequenza di azioni e reazioni nel problema decisionale [69].
Un albero decisionale è una struttura simile ad un diagramma di flusso, dove ogni ramo rappresenta il risultato di un test e ogni nodo foglia rappresenta un etichetta di una classe. I percorsi dalla radice alla foglia rappresentano le regole di classificazione [70].
Viene costruito suddividendo il dataset di origine, che costituisce il nodo radice dell'albero, in sottoinsiemi più piccoli, chiamati invece figli successori. La suddivisione si basa sull'insieme delle regole di suddivisione, fondate sulle caratteristiche di classificazione. Tramite il partizionamento ricorsivo questo processo viene ripetuto su ogni sottoinsieme derivato [71]. La ricorsione si dice completata quando il sottoinsieme in un nodo ha tutti gli stessi valori della variabile target o quando la divisione non aggiunge più valore alle previsioni [71].
Gli algoritmi per la costruzione di alberi decisionali funzionano dall'alto verso il basso, scegliendo ad ogni passaggio una variabile che suddivide al meglio l'insieme di elementi [72].
Vi sono diversi algoritmi che stimano quale sia la struttura migliore. Essi utilizzano diverse metriche che generalmente misurano l'omogeneità della variabile target all'interno dei sottoinsiemi. Queste metriche vengono applicate a ciascun sottoinsieme candidato e i valori risultanti vengono combinati per fornire una misura della qualità della suddivisione. A seconda della metrica utilizzate le prestazioni di vari algoritmi euristici per l'apprendimento dell'albero decisionale possono variare in modo significativo [72].
Tra le varie metriche più utilizzate vi è ad esempio la stima dei corretti positivi. Una metrica semplice ed efficace utilizzata per identificare il grado in cui i veri positivi superano i veri negativi, sfruttando la matrice di confusione, meglio nota in inglese come *confusion matrix*. Un'altra metrica è la cosiddetta impurità di Gini, noto anche come indice di diversità di Gini. Esso è una misura di quanto frequente un elemento scelto a caso dall'insieme verrebbe etichettato in modo errato se fosse etichettato in modo casuale, in base alla distribuzione delle etichette nel sottoinsieme [73].
Altre metriche sono basate sul concetto di entropia, come nel caso della metodo del guadagno di informazione che si basa sia sul concetto di entropia che sul contenuto di informazione della teoria dell'informazione [74].
La metrica della riduzione della varianza viene spesso impiegata nei casi in cui la variabile target risulta continua. Essa è definita come la riduzione totale della varianza della variabile target dovuta alla divisione di uno specifico nodo [75].
Un ultimo esempio di metrica è la cosiddetta misura della bontà. Essa è una funzione che cerca di ottimizzare l'equilibrio tra la capacità di una scissione candidata di creare figli puri con la sua capacità di creare figli della stessa taglia. Viene ripetuto questo processo per ogni nodo impuro finché l'albero non è completo [76].
Un aspetto importante dei modelli basati su alberi decisionali è che questi possono essere semplificati senza sacrificare la bontà dell'addestramento. Poiché la dimensione dell'albero non è intenzionalmente limitata nel processo di crescita, si può verificare un certo grado di overfitting [68]. Ci sono due modi per affrontare questo problema e per scegliere quale utilizzare occorre decidere se si vuole ottenere una

descrizione parsimoniosa oppure una previsione accurata [68]. I due modi prendono il nome di *pruning* e *shrinking*.
Nel primo caso si dice che l'albero viene potato, cioè vengono tagliate ricorsivamente le divisioni meno importanti. Nel secondo caso invece l'albero viene rimpicciolito, cioè non modifica la struttura ad albero, ma la regolarizza restringendo la previsione su ogni nodo verso le medie campionarie dei suoi antenati [77]. Non esiste un metodo migliore di un altro, dipende infatti dal caso in esame, la preferenza dipende da quale priorità si è scelta, semplicità contro accuratezza [68].
Gli alberi decisionali hanno diversi vantaggi, fra i quali la semplicità di comprensione e interpretazione, la possibilità di gestire dati sia numerici che categorici, richiedono poca preparazione dei dati [67]. Ma anche possibilità di convalidare un modello utilizzando test statistici, non ha difficoltà nel gestire dataset di grandi dimensioni, in generale è un algoritmo che rispecchia il processo decisionale umano [67].
Di contro a volte può risultare poco robusto, un piccolo cambiamento nei dati di addestramento può comportare un grande cambiamento nell'albero e di conseguenza nelle previsioni finali [67]. Gli alberi decisionali soffrono anche di overfitting, in quanto tendono ad adattare eccessivamente i dati dell'addestramento, quindi ha un'elevata varianza che generalizza male su nuovi dati di test [67].
Questo problema viene mitigato utilizzando gli alberi decisionali all'interno di un insieme. Esistono due modi per creare insieme di alberi, il metodo di potenziamento o *boosting*, e il metodo di *bagging*, questi vengono appunto chiamati metodi d'insieme [78].
Il metodo di *boosting* sfrutta il processo di costruzione incrementale di un insieme, addestra ogni nuova istanza per enfatizzare le istanze di addestramento precedentemente mal modellate [78]. Questo metodo migliora l'accuratezza del modello, basato sull'idea che è più facile trovare e calcolare la media di molte regole empiriche approssimative, piuttosto che trovare una singola regola di previsione altamente accurata [79].
Gli algoritmi di *boosting* creano ed uniscono, in maniera sequenziale ed iterativa, i risultati di più modelli in modo graduale. Inoltre variano nel modo in cui quantificano la mancanza di adattamento e selezionano le impostazioni per l'iterazione successiva [80]. Gli algoritmi come AdaBoost, ad esempio, applicano pesi alle osservazioni, enfatizzando quelle scarsamente modellate, quindi si tende a discutere il potenziamento in termini di variazione dei pesi [80].
I campioni utilizzati in ogni passaggio non sono tutti estratti allo stesso modo dalla stessa popolazione, ma piuttosto ai casi previsti in modo errato da un dato passaggio viene dato un peso maggiore durante il passaggio successivo [81].
Il secondo metodo d'insieme è quello del *bagging*, il termine viene dalla contrazione dei termini *bootstrap aggregating*. Esso crea più alberi decisionali ricampionando ripetutamente i dati di addestramento con sostituzione e votando gli alberi per una previsione di consenso [82].
Con il termine *bootstrap* si intende una metrica che utilizza il campionamento casuale con sostituzione e rientra nella classe più ampia dei metodi di ricampionamento. Esso assegna misure di accuratezza come bias, varianza, intervalli di confidenza , errore di previsione, alle stime campionarie [83].

La tecnica di *bagging* riduce la varianza associata alla previsione e quindi migliora il processo di previsione. È un'idea relativamente semplice nella quale vengono estratti molti campioni dai dati disponibili, viene applicato un metodo di previsione a ciascun campione, quindi i risultati vengono combinati calcolando la media per la regressione e votando semplicemente per la classificazione [81]. In questo modo, con la tecnica dell'*averaging*, si ottiene la previsione complessiva, con la varianza ridotta [81].
A differenza del *boosting* che incorpora pesi in maniera iterativa, la tecnica del *bagging* è basata su una semplice media delle previsioni [81].
Un classificatore a foresta casuale è un caso specifico di *bagging* [82]. Infatti nella *random forest*, oltre a ciascun albero che esamina solo un set di campioni di *bootstrap*, viene considerato solo un numero ridotto, ma consistente, di caratteristiche uniche. Ciò significa che ogni albero conosce solo i dati relativi a un piccolo numero costante di caratteristiche e un numero variabile di campioni inferiore o uguale a quello del set di dati originale. Di conseguenza, è più probabile che gli alberi restituiscano una gamma più ampia di risposte, derivate da conoscenze più diverse [84].
Essa possiede numerosi vantaggi rispetto ad un singolo albero decisionale generato senza casualità. In una foresta casuale infatti ogni albero sceglie se classificare o meno un campione come positivo in base alle sue caratteristiche, in base al voto di maggioranza [84].
La *random forest* combina diversi alberi decisionali in maniera casuale e aggrega le loro previsioni tramite la media [85]. Ha ottime prestazioni quando è presente un elevato numero di variabili e risulta molto versatile. Può essere applicato a problemi su larga scala e si adatta facilmente a vari compiti di apprendimento [85].
Le foreste casuali sono state ideate da Leo Breiman nei primi anni 2000 [86] e ad oggi rappresenta uno dei metodi di maggior successo. Si adatta al volume delle informazioni e al contempo mantiene un'efficienza statistica elevata [85].
Questa procedura, opera secondo il semplice ma efficace principio del *divide et impera*: ossia campiona le frazioni dei dati, fa crescere un predittore ad albero in maniera casuale su ogni piccola porzione, e infine aggrega insieme questi predittori [85]. Il suo successo è dovuto anche al fatto che, nonostante la sua versatilità, ha pochi parametri da dover regolare. Inoltre può essere applicato ad un'ampia gamma di problemi di previsione, è semplice da usare, è generalmente apprezzato per la sua accuratezza e la sua capacità di gestire campioni di piccole dimensioni con un elevato numero di variabili [85].
Oltre che per compiti di classificazione l'algoritmo *random forest* può essere utilizzato anche per scopi predittivi, e nel tempo è stato utilizzato in abiti economici, scientifici, sociologici [85], [87].
Una foresta casuale è un predittore costituito da una raccolta di $M$ alberi dove per il j-esimo albero della famiglia, il valore previsto nel punto di interrogazione $x$ è definito da $m_n(x; \Theta_j, D_n)$, dove $D_n$ è un campione di addestramento definito come $Dn = ((X_1, Y_1), ..., (X_n, Y_n))$, composto da variabili casuali indipendenti distribuite come coppia di prototipi indipendenti $(X, Y)$. I valori $\Theta_1, ..., \Theta_M$ sono invece variabili casuali indipendenti, distribuite come una generica variabile casuale $\Theta$ e indipendente da $D_n$. In pratica, la variabile viene utilizzata per ricampionare il set di addestramento prima della crescita dei singoli alberi e per selezionare le direzioni successive per la divisione [85].

La *random forest* viene ottenuta tramite un voto di maggioranza tra gli alberi di classificazione, ovvero utilizzando la tecnica del *majority voting* [85], in termini matematici può essere scritta come segue:

$$m_{M,n}(x;\Theta_1,...,\Theta_M,Dn) = \begin{cases} 1 & se \ \ \frac{1}{M}\sum_{j=1}^{M} m_n(x;\Theta j, D_n) > 1/2 \\ 0 & altrimenti \end{cases}$$

Se una foglia rappresenta la regione $A$, allora un classificatore ad alberi casuali assume la seguente forma [85]:

$$m_n(x;\Theta_j,Dn) = \begin{cases} 1 & se \sum\limits_{i\in D_n^\star(\Theta_j)} 1_{X_i\in A, Y_i=1} > \sum\limits_{i\in D_n^\star(\Theta_j)} 1_{X_i\in A, Y_i=0}, x\in A \\ 0 & altrimenti \end{cases}$$

Dove $D_n^\star(\Theta_j)$ contiene i punti selezionati nella fase di ricampionamento. Ovvero, in ogni foglia, viene preso un voto di maggioranza su tutti i valori $(X_i, Y_i)$, per cui $X_i$ appartiene alla stessa regione [85], per convenzione i pareggi vengono assegnati a favore della classe 0 [85].
L'algoritmo può essere facilmente adattato per eseguire la classificazione in due classi senza modificare il criterio di CART-*split*.
Esso viene utilizzato nella costruzione dei singoli alberi per scegliere i migliori tagli perpendicolari agli assi. Infatti ad ogni nodo di ciascun albero, viene selezionato il taglio migliore ottimizzando così il criterio CART-*split*. Per i compiti di classificazione questo è basato sulla già citata di impurità di Gini, mentre per i compiti di regressione è basato sull'errore di previsione quadratico [88].
Il criterio prende origine dal più famoso algoritmo CART (acronimo inglese di *Classification and Regression Trees*) di Leo Breiman [89], pietra miliare nell'evoluzione dell'intelligenza artificiale, dell'apprendimento automatico, delle statistiche non parametriche e dell'estrazione dei dati.
Per osservare ciò, si prende $Y \in \{0,1\}$ e si considera un singolo albero senza il passaggio di sottocampionamento. Per qualsiasi cella generica $A$ si definisce $p_{0,n}(A)$ come la probabilità empirica [85], dato un punto in una cella $A$, che abbia classe 0. Facendo notare che $\overline{Y}_A = p_{1,n}(A) = 1 - p_{0,n}(A)$, il criterio di classificazione CART-*split* recita che, per ogni $(j,z) \in C_A$:

$$\begin{aligned} L_{class,n}(j,z) = p_{0,n}(A)p_{1,n}(A) &- \frac{N_n(A_L)}{N_n(A)} \times p_{0,n}(A_L)p_{1,n}(A_L) \\ &- \frac{N_n(A_R)}{N_n(A)} \times p_{0,n}(A_R)p_{1,n}(A_R) \end{aligned}$$

Questo criterio è basato sulla già citata misura dell'impurità di Gini [85], che ha la seguente interpretazione. Per classificare un elemento che ricade nella cella $A$, si usa la regola di assegnare l'elemento all'etichetta $l$ di probabilità $p_{l,n}(A)$, per $j \in \{0,1\}$, uniformemente selezionato da $\{X_i \in A : (X_i, Y_i) \in D_n\}$.
La probabilità stimata che l'elemento abbia effettivamente un'etichetta $l$ è chiamata $p_{l,n}(A)$. Pertanto l'errore stimato secondo questa regola è $2p_{0,n}(A)p_{1,n}(A)$, che viene definito appunto impurità o l'indice di Gini [85].

La strategia di previsione è diversa nella classificazione e nella regressione. Nel regime di classificazione infatti, ogni albero utilizza un voto di maggioranza locale, mentre nella regressione invece la previsione è ottenuta mediante una media locale [85].
Il valore del numero di alberi nella foresta $M$ è spesso scelto arbitrariamente grande, dipendentemente dalle risorse di calcolo disponibili. Ha senso, da un punto di vista della modellazione, lasciare che $M$ tenda all'infinito [85].
Nelle foreste originali di Leo Breiman ogni nodo di un singolo albero è associato ad una cella iperrettangolare [86]. L'algoritmo funziona facendo crescere $M$ diversi alberi casuali e, prima della costruzione di ogni albero, $a_n$ osservazioni vengono estratte a caso, con o senza sostituzione dal dataset originale. Queste, e solo queste, osservazioni vengono prese in considerazione nella costruzione dell'albero [85]. A seguito di ciò in ogni cella di ciascun albero, viene eseguita una divisione massimizzando il criterio sopra descritto di CART. Infine, la costruzione di singoli alberi viene interrotta quando ogni cella contiene un numero di punti inferiore al valore desiderato [85].
I metodi d'insieme sono una tecnica relativamente semplice ed efficace per migliorare il metodo degli alberi decisionali. Tutto ciò va discapito della perdita della struttura semplice e interpretabile degli alberi. I guadagni però sono notevoli, soprattutto per quanto riguarda la precisione [82].

### 3.6.1 La cross-validation

Al fine di addestrare il classificatore, l'insieme delle caratteristiche estratte, descritto nella sezione 3.5, è stato suddiviso in due dataset. Il primo è quello dell'addestramento vero e proprio del classificatore, il cosiddetto *training set*, il secondo è dedicato alla valutazione dell'addestramento, chiamato appunto *test set*.
Infatti, quando non si dispone di un set di dati particolarmente ampio, risulta essere quasi proibitivo avere sia set di validazione che set di test; questo ridurrebbe notevolmente la grandezza del set di addestramento. Di solito si preferisce disporre di più dati possibile per addestrare il modello, dividendo quindi il dataset nei soli set di addestramento e set di test.
La divisione è stata eseguita scegliendo le osservazioni in maniera casuale, con le seguenti proporzioni: 75% di *training set* e 25% di *test set*. Di conseguenza il dataset di addestramento risulta composto da circa 50.600 osservazioni, mentre quello di test da circa 16.900 osservazioni.
In questi casi si utilizza la convalida incrociata sul set di addestramento per simulare un set di validazione, questa tecnica viene comunemente chiamata *cross-validation* [90].
La tecnica di convalida incrociata più utilizzata è la *K-fold cross-validation*. Essa fornisce una stima accurata dell'errore reale senza sprecare troppi dati. Nella *K-fold cross-validation* il set di addestramento originale viene suddiviso in $K$ sottoinsiemi, definiti appunto *fold* di dimensioni uguali $m/K$, dove $m$ è il numero totale delle osservazioni [71].
Viene scelto preliminarmente il numero $K$ di sottoinsiemi, nella maggior parte dei casi viene scelto $K = 5$ (ma a volte anche $K = 10$, dipendentemente dal caso in

esame). Nel caso di $K = 5$, il dataset di addestramento viene diviso cinque volte, in maniera casuale $\{F_1, F_2, ..., F_5\}$.
Ogni $F_k$, con $(k = 1, ..., 5)$, contiene quindi il 20% dei dati di addestramento. Verranno quindi addestrati cinque modelli come segue, per addestrare il primo modello, $f_1$, si utilizzano tutti i *fold* $F_2$, $F_3$, $F_4$, $F_5$ come set di addestramento e l'insieme $F_1$ come set di validazione.
Per addestrare il secondo modello, $f_2$, si utilizzano gli elementi dei *fold* $F_1, F_3, F_4, F_5$ come set di addestramento e gli elementi di $F_2$ come set di validazione. Si continua a creare modelli in modo iterativo in questo modo e si calcola il valore della metrica di interesse su ogni set di validazione, da $F_1$ a $F_5$. Si fa quindi la media dei cinque valori ottenuti per calcolare il valore finale della metrica di interesse.
Il caso particolare di $K = m$ è un metodo alternativo di validazione e viene chiamato *leave-one-out* (LOO). Brevemente, si lascia fuori dal set di addestramento un solo elemento e su quest'ultimo si fa poi la validazione [90].
Per dare una definizione matematica, chiamiamo $R$ il set di addestramento e $V$ il set di validazione, chiamiamo inoltre $D$ il set generato dalla loro unione $D = R \cup V$; questi devono essere uniti senza sovrapposizione (quindi $R \cap V = \emptyset$).
Dopo l'addestramento, si valutano le prestazioni del predittore $f$ sul set di validazione $V$, in questo caso calcolando l'errore quadratico medio. Più precisamente, per ogni partizione $k$ i dati di addestramento $R_k$ producono un predittore $f_k$, che viene poi applicato all'insieme di validazione $V_k$ per calcolare cosiddetto il rischio empirico $R(f_k, V_k)$ [91]. Poiché non è possibile sapere a priori come si comporterà l'algoritmo di classificazione su un dataset ignoto, si assume che le sue prestazioni si possano misurare su un dataset di addestramento noto, questo viene appunto chiamato rischio empirico.
Quindi la *cross-validation* approssima l'errore di generalizzazione previsto, come segue:

$$E_v[R(f, V)] \approx \frac{1}{K} \sum_{k=1}^{K} R(f_k, V_k)$$

dove $R(f_k, V_k)$ è il rischio calcolato sul set di validazione $V_k$, per il predittore $f_k$ [91]. L'approssimazione è originata da due fonti: la prima è a causa del set di addestramento finito, la seconda a causa dell'insieme di validazione finito, questo si traduce in una stima imprecisa del rischio $R(f_k, V_k)$.
Grazie alla convalida incrociata si evitano fenomeni dannosi di *overfitting*, a discapito però di un più elevato costo computazionale [91].

### 3.6.2 L'addestramento del classificatore

La foresta casuale, e il suo addestramento, sono stati implementati grazie al *Classification Learner app* di MATLAB. Essa è un'applicazione built-in che permette di addestrare modelli per classificare dati [92]. Fornendo un set noto di dati di input e di riposte, quindi le osservazioni e le etichette, è possibile eseguire l'apprendimento automatico supervisionato con differenti tipi di modelli. I dati vengono utilizzati per

eseguire il training di un modello che genera previsioni per la risposta a nuovi dati [92].
L'applicazione prende in ingresso il dataset di variabili, appunto il dataset di addestramento, e le corrispettive etichette della classe di appartenenza. Una volta selezionato il tipo di classificatore desiderato, singolo o d'insieme è possibile definirne le caratteristiche e i parametri. Questi possono essere scelti attraverso due strade. La prima è in maniera arbitraria, ossia scelti dall'utente con valori ragionevoli per poi magari essere successivamente adattati al sistema in esame, con cosiddetto il metodo del *tuning*. La seconda è quella di lasciare che queste caratteristiche e parametri, definiti iperparametri, vengano ottimizzati dall'applicazione di MATLAB. In questo lavoro è stato scelto di lasciare il compito di assegnare il valore degli iperparametri al *Classification Learner app*. Le uniche scelte effettuate arbitrariamente sono state il metodo d'insieme, appunto il metodo di *bagging*, e il numero di predittori da campionare, dove in questo caso è stato scelto di utilizzare tutto il dataset di addestramento.
Sono stati invece lasciati ottimizzare il numero massimo di divisioni e il numero dei classificatori. Il primo è stato fatto variare da 1 a 50.600 (ossia il numero totale delle osservazioni) e ha ottenuto un valore di 3.739 come numero massimo di divisioni, mentre il numero totale di classificatori è risultato essere 489, facendo variare il numero da 10 a 500. Il numero totale di iterazioni è stato lasciato quello predefinito, ossia 30.
L'algoritmo di ottimizzazione scelto è stato quello basato sulla probabilità bayesiana. Esso tenta di minimizzare una funzione obiettivo scalare $f(x)$ per $x$ in un dominio limitato [93]. In quest'algoritmo la funzione può essere deterministica o stocastica e i componenti di $x$ possono essere reali continui, interi o categorici. L'algoritmo valuta la funzione $y_i = f(x_i)$ per un numero definito di punti $x_i$ presi a caso all'interno dei limiti della variabile [93]. Consiste in un algoritmo iterativo che processa il modello gaussiano di $f(x)$ per ottenere una distribuzione a posteriori sulle funzioni definita come $Q(f|x_i, y_i \; per \; i = 1, ..., t)$. Successivamente trova il nuovo punto $x$ che massimizza la funzione di acquisizione $a(x)$ [93], ossia una funzione che valuta la bontà di un punto $x$ in base alla funzione di distribuzione a posteriori $Q$, basata sul modello gaussiano di $f$ [93].
L'algoritmo si arresta dopo aver raggiunto uno dei seguenti criteri: un numero massimo di iterazioni (in questo caso 30), oppure un tempo massimo di calcolo (in questo caso non è stato fissato nessun limite di tempo, l'ottimizzazione ha richiesto circa 48 ore di calcolo).

## 3.7 Creazione e affinamento delle maschere

Dopo aver addestrato il classificatore, questo è stato validato e successivamente testato sul dataset di test. Una volta ottenuto un buon livello di classificazione, la *random forest* è stata testata su immagini endoscopiche intere, prese singolarmente. In particolare, il classificatore è stato testato sulle immagini endoscopiche originali, ripulite da artefatti e ritagliate. Esso ha valutato ogni singola ROI e gli ha attribuito valore 0 se considerata sana e valore 1 se considerata patologica.

Come descritto nella sezione 3.3, le ROI $50 \times 50$ risultano sovrapposte sia orizzontalmente che verticalmente del 40% (ossia 20 pixel). Questo comporta che, per la stessa porzione dell'immagine, possono esservi sovrapposte fino a quattro ROI contemporaneamente. Nel caso di classificazione discordante delle ROI, nella regione sovrapposta, è stata scelta la via più conservativa e ridondante: se almeno una delle quattro ROI adiacenti risulta patologica, allora tutta la regione sovrapposta assume valore 1, cioè si assume che la porzione con *overlap* sia affetta da metaplasia intestinale gastrica. In **Figura 3.16**, una rappresentazione grafica di come viene gestita la sovrapposizione delle ROI con classificazione discordante; per semplicità di disegno sono sovrapposte al massimo due ROI contemporaneamente nella stessa porzione (1 sana e 1 patologica), ma il risultato sarebbe identico anche con tre e quattro ROI.

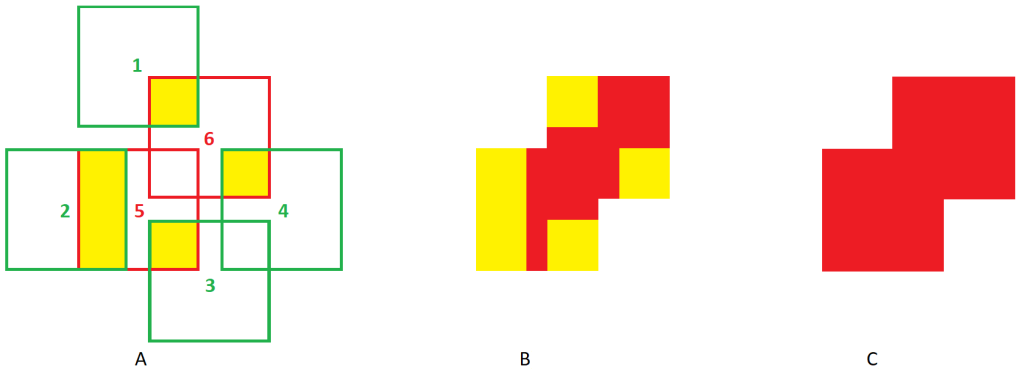



**Figura 3.16: La gestione delle ROI con classificazione discordante.** *(A) Rappresentazione di 6 ROI con più sovrapposizioni: quelle in verde, numerate dalla 1 alla 4 rappresentano ROI sane, quelle in rosso, la 5 e la 6, rappresentano due ROI patologiche. In giallo sono evidenziate le porzioni sovrapposte con classificazione discordante (massimo 2 ROI per porzione sovrapposta). (B) Le due ROI patologiche: gli overlap con classificazione discordante sono stati evidenziati in giallo. (C) La maschera binaria risultante: vengono mantenute solo le ROI patologiche; le zone sovrapposte, con classificazione discordante, vengono contrassegnate come patologiche.*

Sono state così create delle vere e proprie maschere binarie, dove sono state segnate in bianco tutte le regioni classificate come affette da metaplasia intestinale gastrica, in nero tutte le altre.
Le maschere in uscita dal classificatore presentano però dei valori anomali, in particolare alcune aree isolate e circoscritte risultano classificate come patologiche, nonostante siano in realtà sane. Graficamente questo errore ha l'aspetto di puntinatura sparsa, di colore bianco, su un sfondo nero; questa non è sempre presente e risulta più meno marcata dipendentemente dai casi in esame.
Per risolvere questo problema è stata realizzata una strategia di rimozione dei valori anomali (*outlier*) che è stata implementata con l'utilizzo del metodo di clustering noto come DBSCAN. Il termine deriva dall'acronimo inglese *density-based spatial clustering of applications with noise* e come suggerisce il nome è un metodo di *clustering* basato sulla densità spaziale dei punti. Esso è stato proprosto la prima volta nel 1996 da Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu [94].

Si tratta di un algoritmo non parametrico di clustering che, dato un insieme di punti in uno spazio, raggruppa punti che sono molto vicini tra loro, contrassegnando come valori anomali i punti che si trovano isolati, nelle regioni a bassa densità.
Esso è un algoritmo di clustering come il già citato K-means, ma a differenza di quest'ultimo, non è basato sui centroidi ma sul concetto di densità. Quindi non è necessario conoscere a priori il numero $K$ di cluster ma in questo caso occorre scegliere due differenti valori, la distanza e la numerosità.
Comunemente, per immagini digitali, la distanza viene espressa in numero di punti (pixel) e viene definita $\epsilon$, essa è un parametro che specifica il raggio di un intorno rispetto ad un punto. La numerosità $n$, invece, è il numero minimo di punti vicini, ed è anch'esso un valore arbitrario.
Esistono tre tipi di punti, punti centrali, punti direttamente raggiungibili e valori anomali (*outlier*).
Un punto $p$ è un punto centrale se almeno $n$ punti si trovano entro la distanza $\epsilon$ da esso, incluso $p$. Un punto $q$ è direttamente raggiungibile da $p$ se il punto $q$ si trova entro la distanza $\epsilon$ dal punto centrale $p$, per definizione i punti sono direttamente raggiungibili solo dai punti centrali [94].
Un punto $q$ è raggiungibile da $p$ se esiste un percorso $p_1, ..., p_n$ con $p1 = p$ e $p_n = q$, dove ogni $p_i + 1$ è direttamente raggiungibile da $p_i$. Questo implica che il punto iniziale e tutti i punti sul percorso devono essere punti centrali, con la possibile eccezione di $q$. Tutti i punti non raggiungibili da nessun altro punto sono valori anomali o punti di disturbo [95].
Se $p$ è un punto centrale, forma un cluster insieme a tutti i punti che sono raggiungibili da esso. Ogni cluster contiene almeno un punto centrale, mentre i punti non centrali possono far parte di un cluster, ma ne costituiscono il bordo poiché non possono essere utilizzati per raggiungere più punti [96].
Un cluster soddisfa quindi due proprietà, tutti i punti all'interno del cluster sono reciprocamente connessi in base alla densità. Se un punto è raggiungibile in densità da qualche punto del cluster, fa parte anch'esso del cluster [97].
L'algoritmo parte selezionando un campione casuale $x$ dal set di dati e lo assegna al *clusterA*. Quindi conta quanti campioni hanno distanza da $x$ minore o uguale a $\epsilon$. Se questa quantità è maggiore o uguale a $n$, allora mette tutte queste osservazioni $\epsilon$-vicine nello stesso *clusterA*. Successivamente esamina ogni membro del *clusterA* e trova i rispettivi $\epsilon$-vicini [90].
Se un membro del *clusterA* ha $n$ o più $\epsilon$-vicini, espande il *clusterA* inserendo quei punti $\epsilon$-vicini nel *cluster* e continua ad espandere il *clusterA* fino a quando non ci sono più campioni da inserire [90].
In quest'ultimo caso, sceglie dal set di dati un'altra osservazione che non appartiene a nessun cluster e la mette nel *clusterB*, poi continua così finché tutte le osservazioni non appartengono a qualche cluster, oppure sono contrassegnate come valori anomali (*outlier*) ossia osservazioni il cui $\epsilon$-vicinato contiene meno di $n$ campioni [90].
Il fatto di poter scegliere questi due iperparametri è il punto di forza di questo algoritmo, ma anche la sua debolezza. Infatti se personalizzare l'algoritmo può essere utile ai fini della clusterizzazione, trovarne i valori corretti può essere spesso difficoltoso [90].

La funzione MATLAB che esegue l'implementazione dell'algoritmo di cluster si chiama *dbscan*. Essa prende in ingresso una matrice $x$ di ampiezza $r \times c$, la distanza $\epsilon$ e il numero minimo di campioni $n$. La funzione restituisce un vettore $r \times 1$ contenente gli indici dei cluster di ciascuna osservazione [98].
Come tipo di distanza è stata scelta quella predefinita, ossia quella euclidea, e come valore $\epsilon$ è stato scelto quello della diagonale della finestra di osservazione, la diagonale della ROI, ossia: $\epsilon = \sqrt{50^2 + 50^2} \simeq 70,71$. In questo caso il valore 50 rappresenta la lunghezza, in pixel, di un lato della ROI.
Come numero minimo di punti $n$ è stato scelto il valore 9, esso rappresenta infatti un gruppo di ROI quadrato di dimensioni $3 \times 3$ (1 centrale e 8 sovrapposti fra spigoli e lati), il cluster di dimensioni minime accettato.
L'algoritmo DBSCAN è stato eseguito su tutte le maschere binarie, sia su immagini sane che immagini provenienti da pazienti patologici. Sulle prime ha avuto un effetto assimilabile ad un filtro passa-basso; su queste maschere binarie infatti erano presenti alcune ROI classificate erroneamente come patologiche, creando sulla maschera un aspetto grafico simile ad un rumore (puntinatura bianca, sparsa, su sfondo nero). L'algoritmo DBSCAN ha completamente ripulito l'immagine binaria da queste ROI patologiche inattese, essendo queste sporadiche ed isolate.
Sulle seconde ha avuto un effetto assimilabile ad affinamento delle maschere, quest'ultime risultano infatti quasi totalmente ripulite da *outlier*, lasciando spazio alle sole segmentazioni più estese e con tratti ben delineati; è possibile vedere un esempio di maschera binaria ripulita in **Figura 3.17**, dove è rappresentato un esempio di metaplasia costituita de quattro isole metaplastiche distinte.


**Figura 3.17: La rimozione degli *outlier* dalle maschere binarie.** *(A) La maschera binaria segmentata manualmente dal medico. (B) La maschera binaria ottenuta in uscita dal classificatore Random Forest, affetta da errori di classificazione. (C) La maschera binaria ripulita da outlier a valle dell'algoritmo DBSCAN: i margini sono nettamente più nitidi e la forma della maschera risulta meglio delineata.*

Le maschere binarie così costituite, risultano un ottimo strumento per valutare, successivamente, la bontà del classificatore. Sono inoltre una rappresentazione grafica di rapida e semplice intuizione, per compiti di categorizzazione.

# 4. L'interfaccia grafica

Tramite l'applicazione built-in di MATLAB *App Designer* è stata implementata una semplice interfaccia grafica (anche nota come GUI, dall'inglese *graphical user interface*), che permette ad un operatore l'esecuzione degli algoritmi di classificazione e visualizzarne i risultati. L'idea nasce dal concetto che un medico possa caricare un'endoscopia gastrica sul PC per poi ottenere una valutazione oggettiva dell'immagine in pochi secondi. Questo potrebbe aiutare l'operatore nella diagnosi, riducendo al contempo il numero delle biopsie.

## 4.1 La mappa di calore

Per la visualizzazione delle aree potenzialmente metaplastiche è stata generata una mappa di calore (meglio conosciuta in inglese come *heatmap*). Essa è una tecnica di visualizzazione dei dati che mostra l'entità di un fenomeno sotto forma di colore, in due dimensioni. La variazione di colore può essere di tonalità o intensità, fornendo evidenti segnali visivi al lettore su come il fenomeno è raggruppato oppure come varia nello spazio [99]. In questo caso è stata scelta una scala che va dal giallo al rosso, al crescere della intensità nella predizione della metaplasia.
La risoluzione della mappa è di $50 \times 50$ pixel (la dimensione della ROI), con sovrapposizione delle celle, verticale e orizzontale di 20 pixel.
Proprio la sovrapposizione delle celle risulta essere la chiave di questa mappa di calore. Infatti ogni ROI presenta, al più, 8 ROI sovrapposte ad essa. Quindi l'elemento base della mappa di calore risulta composto da 9 ROI (1 ROI principale al centro e 8 ROI secondarie sovrapposte).
Con questa struttura la ROI principale sarà composta da tre tipi di aree diverse: un'area centrale, quattro aree spigolo e quattro aree lato. L'area centrale della ROI principale non presenta alcuna ROI sovrapposta, ed è di dimensioni $10 \times 10$ pixel. Le aree spigolo avranno, al più, quattro ROI sovrapposte (una centrale, una spigolo e una lato), e saranno di dimensioni $20 \times 20$ pixel. Le aree lato invece avranno, al più, due ROI sovrapposte (una centrale e una lato), e saranno di dimensioni $10 \times 20$ pixel.
Ad una ROI considerata patologica viene assegnato valore 1, ad una ROI sana viene assegnato valore 0. Il valore risultante delle sovrapposizioni è ottenuto della media delle classificazioni di ogni singola ROI. Quindi, nel caso in cui risultino tutte patologiche, alla sovrapposizione viene assegnato valore 1 (rosso), nel caso risultino tutte sane viene assegnato valore 0 (nessun colore). Per tutti i valori compresi fra 0 e 1 è presente la scala di colore giallo-rosso.

Nel caso in cui la ROI adiacente sovrapposta risulti vuota (per esempio se rimossa dall'algoritmo di pulizia delle parti troppo scure descritto nella sezione 3.2), a questa non viene assegnato alcun valore, per cui non prenderà parte al calcolo della media. Una combinazione possibile potrebbe essere quindi una ROI sana e tre patologiche, con risultato finale 0,75. Oppure un'altra combinazione potrebbe essere la sovrapposizione di due ROI sane e due ROI patologiche, con risultato finale 0,50 e cosi via. Di seguito la scala di colori in base al risultato ottenuto:

- **Colore rosso:** 1
- **Colore arancione:** 0,75
- **Colore ocra:** 0,50
- **Colore giallo:** 0,25

Naturalmente l'effetto grafico desiderato è che, più una determinata area presenta regioni di colore rosso-arancione, più risulta interessata da metaplasia intestinale gastrica, le aree di colore ocra-giallo risultano essere invece regioni di transizione fra i tessuti sani e quelli patologici.
Nella **Figura 4.1** una rappresentazione grafica delle tre aree delle ROI e della definizione di aree spigolo e aree lato.

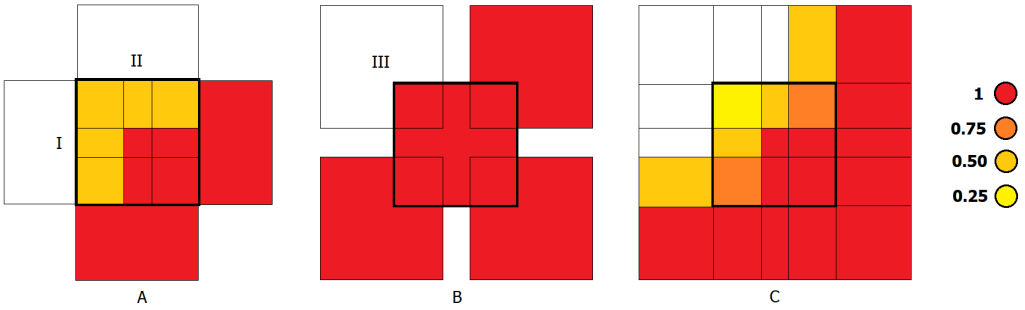



**Figura 4.1: La sovrapposizione delle ROI e l'heatmap.** *(A) La gestione dei lati. La ROI principale con quattro sovrapposizioni ai lati, in questo caso la I e la II rappresentano due ROI classificate come sane, le altre invece tutte patologiche; in questa configurazione le aree sovrapposte ottengono un valore medio di 0,5. (B) La gestione degli spigoli. La ROI principale con quattro sovrapposizioni agli spigoli, in questo caso la III rappresenta una ROI assente, le altre invece tutte patologiche; in questa configurazione le aree sovrapposte ottengono un valore medio di 1. (C) La creazione della heatmap. La mappa di calore rappresentata viene dalla sovrapposizione delle ROI descritte in A e B, in questa configurazione le aree sovrapposte ottengono valori medi di 0,25-0,50-0,75-1.*

La mappa di calore è una tecnica molto semplice ed intuitiva che fornisce al medico un'indicazione sulle aree maggiormente considerate anomale dall'algoritmo di classificazione.

## 4.2 L'indicazione semaforica

All'interno dell'interfaccia grafica è stata inserita anche un'indicazione semaforica. Essa ha lo scopo di fornire un'indicazione al medico sull'entità della lesione metaplastica. L'indicazione dell'entità della lesione è basata sull'estensione della metaplasia, rilevata dall'algoritmo di classificazione.
In particolare viene calcolato il rapporto fra le ROI classificate come patologiche ($ROI_d$), sul totale delle ROI valutate dell'immagine ($ROI_t$), in percentuale.
Per totale delle ROI si intende tutte quelle che rappresentano solo il tessuto gastrico e che sono non nulle, mentre per ROI patologiche sono state prese in considerazione solo quelle in uscita dall'algoritmo di affinamento delle maschere.
La percentuale delle ROI patologiche ($ROI_\%$) è stata calcolata con la seguente, semplice, formula.

$$ROI_\% = \frac{ROI_d}{ROI_t} \cdot 100$$

In base al risultato ottenuto corrisponde una determinata classe di appartenenza, esplicitata con un colore caratteristico. Di seguito la legenda dell'indicazione semaforica:

- **Colore verde:** $ROI_\% = 0\%$
- **Colore giallo:** $0\% < ROI_\% < 5\%$
- **Colore arancione:** $5\% \leq ROI_\% < 10\%$
- **Colore rosso:** $ROI_\% \geq 10\%$

Il colore verde indica che non è stata rilevata nessuna ROI patologica, il colore rosso indica che è stato rilevato un numero di ROI uguale o superiore al 10% del numero totale di ROI dell'immagine. Questo valore è stato scelto tale poiché la più piccola maschera segmentata dal medico aveva proprio estensione circa 10% rispetto al totale. I colori giallo e arancione non sono altro che dei punti intermedi fra verde e rosso, posti simmetricamente.
Per indicare in quale livello di interessamento ricade l'immagine, all'interno dell'endoscopia (esattamente in alto a destra), è stato posto un riquadro colorato, appunto del colore di appartenenza, con all'interno la percentuale di estensione dell'anomalia.
Come già detto, l'indicazione semaforica ha lo scopo di dare un'informazione oggettiva, al medico, sull'estensione della lesione metaplastica, rilevata dal classificatore. Essa riporta il grado di interessamento della patologia, che potrebbe essere riassunto in:

- **Colore verde:** nessuna anomalia rilavata
- **Colore giallo:** basso riscontro di anomalie
- **Colore arancione:** medio riscontro di anomalie
- **Colore rosso:** significativo riscontro di anomalie

In base a quanto visto sopra, con una semplice informazione basata sui colori, il medico riceve un'indicazione oggettiva sul grado di estensione delle anomalie rilevate.

## 4.3 L'interazione dell'operatore

L'interfaccia grafica si compone di un'unica finestra nella quale è possibile eseguire tre semplici macro attività. L'operatore infatti, può caricare un'immagine endoscopica, analizzarla ed infine esportare i risultati.
L'applicazione si avvia con una pagina principale nella quale è presente un bottone con il quale poter caricare un'immagine, catturata dalla sonda endoscopica digitale i-scan della società PENTAX Medical. Una volta caricata l'immagine questa compare a schermo intero nella finestra principale. Sull'immagine originale è possibile zoomare e navigare, così l'operatore può già eseguire una sua prima analisi.
Una volta terminata l'analisi preliminare l'operatore può scegliere di farla analizzare dall'algoritmo di classificazione. L'algoritmo la ritaglia, la ripulisce da artefatti, ne estrae le caratteristiche, classifica le ROI e restituisce un'immagine con indicazione semaforica e un'*heatmap* delle zone ritenute anomale, se presenti; l'intero processo richiede circa un minuto di elaborazione per ogni singola immagine.
A questo punto l'operatore può zoomare l'immagine ottenuta e navigare su di essa, al fine di valutare l'analisi effettuata dall'algoritmo.
Sull'interfaccia è presente una semplice barra di strumenti da disegno così che il medico possa eventualmente appuntare qualche nota o effettuare qualche semplice disegno grafico.
A seguito della sua ulteriore analisi può decidere se esportare l'immagine analizzata dall'algoritmo (corredata di *heatmap* e indicazione semaforica), oppure caricare una nuova immagine da analizzare.
In **Figura 4.2** sono presenti due immagini relative all'interfaccia grafica dove è possibile osservare la mappa di calore e l'indicazione semaforica.

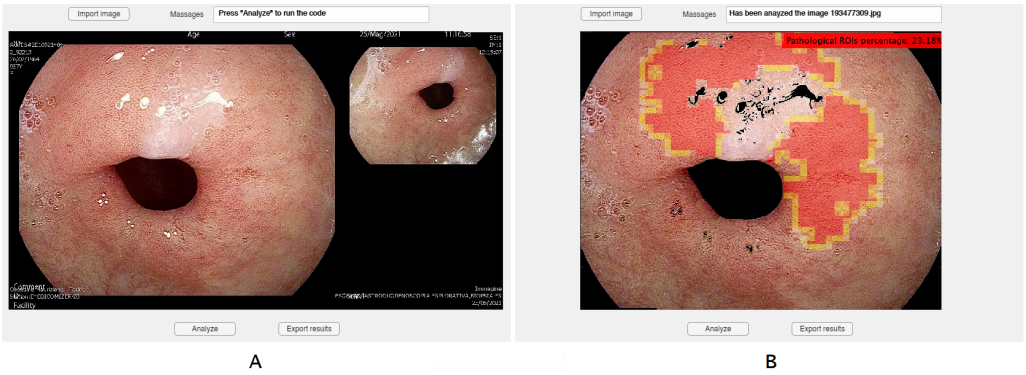



**Figura 4.2: L'utilizzo dell'interfaccia grafica.** *(A) Esempio di immagine endoscopica originale una volta che è stata importata dall'operatore nell'interfaccia grafica. In alto a sinistra della GUI è presente un bottone che permette il caricamento di una qualsiasi immagine tramite ricerca di file. In alto al centro è presente una barra di testo al cui interno compare un messaggio sullo stato d'avanzamento del lavoro: caricamento immagine, analisi, esportazione risultati. In basso a sinistra è presente un bottone che permette di analizzare l'immagine endoscopica una volta caricata. In basso a destra è presente un bottone che permette di esportare i risultati ottenuti dal classificatore, sotto forma d'immagine. (B) Immagine endoscopica analizzata dall'algoritmo di classificazione. Essendo un'immagine patologica è dotata di heatmap che rappresenta le regioni classificate come metaplastiche. In alto a destra dell'immagine è presente l'indicazione semaforica di colore rosso, che indica un riscontro di anomalie significativo; questo poiché la regione patologica ha un'ampiezza maggiore del 10% del totale, come mostrato all'interno dello stesso riquadro (23,18%) .*

# 5. Risultati

Vista la complessità dell'algoritmo di classificazione, e data la natura eterogenea del dataset, la sezione relativa ai risultati è stata suddivisa in tre parti differenti.
La prima è relativa alla classificazione puntuale delle ROI, dove quindi verranno analizzati i risultati su scala microscopica.
La seconda è relativa all'efficacia delle maschere e sul grado di affidabilità di queste di eseguire una classificazione, i risultati verranno quindi analizzati su scala intermedia.
La terza è relativa alla categorizzazione dell'intera immagine endoscopica, verrà analizzata cioè la capacità dell'algoritmo di classificare se un'immagine proviene da un paziente sano o ad uno patologico; i risultati verranno quindi analizzati su scala macroscopica.

## 5.1 La classificazione delle ROI

La classificazione binaria è uno degli studi più frequenti nei problemi di apprendimento automatico, per questo motivo sono state implementate centinaia di metriche per valutare le prestazioni di classificazione [100].
Spesso i termini misure, metriche e indicatori vengono utilizzati in modo intercambiabile fra di loro, in realtà fra questi vi sono nette differenze. Le misure costituiscono la base, sono valori numerici con poco o nessun contesto, le metriche sono al di sopra delle misure e possiedono una raccolta di misure all'interno del proprio contesto, gli indicatori si trovano in cima e sono un confronto fra le varie metriche [100].
L'elemento chiave delle misure sono le matrici di confusione (meglio conosciute in inglese come *confusion matrix*). Essa è una vera e propria matrice dove le colonne rappresentano i valori predetti, mentre le righe rappresentano i valori reali.
I risultati con esito "vero" della classificazione o le corrispondenze fra previsione e realtà si trovano sulla diagonale principale della matrice di confusione, mentre i risultati con esito "falso", le mancate corrispondenze o gli errori si trovano fuori dalla diagonale.
Nel caso specifico di classificazione binaria le righe e le colonne diventano due, quindi la matrice avrà dimensioni $2 \times 2$. Da qui si ottengono le prime quattro misure di base di in una classificazione binaria, con approccio di apprendimento supervisionato [100].
In alto a sinistra della matrice si trovano i veri positivi (TP), in alto a destra i falsi positivi (FP), in basso a sinistra i falsi negativi (FN), in basso a destra i veri negativi (TN). In via generale, nei temi di ingegneria e medicina, gli errori di tipo

falsi negativi, sono di solito più gravi o peggiori degli errori di tipo falsi positivi [100].
Al di sopra delle misure di base vi sono le misure di performance di primo livello, esse sono la condizione positiva (P) e la condizione negativa (N). Rappresentano rispettivamente il numero totale di casi realmente positivi e il numero totale di casi realmente negativi.
A queste si aggiungono il numero totale di casi predetti positivi (OP) e il numero totale di casi predetti negativi (ON). Queste quattro misure non sono altre che le marginalità delle matrici della teoria della probabilità [100].
Grazie alle misure di base e a quelle di primo livello si possono ricavare le metriche di base [100]. Di seguito verranno presentate quelle maggiormente impiegate e conosciute [101], utilizzate anche in questo lavoro per valutare le prestazioni del classificatore.
L'accuratezza fa parte delle metriche base ed è una delle più popolari nella classificazione, viene calcolata direttamente dalla matrice di confusione [102]. La formula dell'accuratezza presenta al numeratore la somma degli elementi veri positivi e veri negativi, mentre al denominatore presenta la somma di tutti gli elementi della matrice di confusione [101]. In altre parole, considerando di scegliere un'unità casuale e di prevederne la classe, l'accuratezza è la probabilità che la previsione del modello sia corretta [102], ed è definita come segue:

$$Accuratezza = p(\widehat{C} = C) \cong \frac{TP + TN}{TP + TN + FP + FN}$$

dove $\widehat{C}$ è la classe predetta e $C$ la classe vera. L'accuratezza restituisce una misura complessiva di quanto il modello prevede correttamente sull'intero set di dati [102]. L'elemento base della metrica sono i singoli individui nel dataset dove ogni unità ha lo stesso peso e contribuisce in egual misura al valore di accuratezza [102]. La metrica è molto intuitiva e facile da capire, essa assume valori compresi tra 0 e 1, dove 1 indica 100% di previsioni corrette; la quantità restante per raggiungere il valore 1 è chiamata tasso di errata classificazione (meglio conosciuta in inglese come *misclassification rate*): $p(\widehat{C} \neq C) = 1 - p(\widehat{C} = C)$ [102].
Un'altra metrica molto utilizzata è la sensibilità, essa misura l'accuratezza predittiva del modello per la classe positiva, ossia misura la capacità del modello di trovare tutte le unità positive nel set di dati. Essa rappresenta la frazione di elementi TP, divisa per il numero totale di unità classificate positivamente, misura la capacità del classificatore di trovare tutte le unità positive nel dataset. Viene anche comunemente chiamata *Recall* o *true positive rate* (TPR) [101].

$$Sensibilità = p(\widehat{C} = P \mid C = P) \cong \frac{TP}{TP + FN}$$

Anche questa metrica varia fra 0 e 1 e spesso si esprime in percentuali. In questo caso, un test che ottiene il 100% di sensibilità significa che il classificatore riconosce tutti i pazienti con la malattia. Da notare però che la sensibilità non tiene conto dei falsi positivi. La corrispettiva negativa viene chiamata FNR (acronimo inglese di *false negative rate*), ed è definita come $1 - TPR$.

In modo duale alla precedente, viene calcolata anche la metrica della specificità. Essa è la probabilità di un risultato negativo del test, condizionata al fatto che l'elemento sia veramente negativo.

$$Specificità = p(\widehat{C} = N \mid C = N) \cong \frac{TN}{TN + FP}$$

Come le precedenti, anche questa metrica varia fra 0 e 1 e si esprime in percentuali. In questo caso, un test che ottiene il 100% di specificità significa che il classificatore riconosce tutti i pazienti senza la malattia. Da notare però che la specificità non tiene conto dei falsi negativi.
Fra le metriche di base più importanti vi sono poi il PPV e il NPV.
Il PPV, acronimo inglese di *positive predictive value*, rappresenta la frazione di elementi TP divisa per il numero totale di unità previste positivamente. Questa metrica esprime la proporzione di unità che il classificatore afferma essere positive e che in realtà sono positive [102].

$$PPV = p(C = P \mid \widehat{C} = P) \cong \frac{TP}{TP + FP}$$

Il valore ideale del PPV, con un test perfetto, vale 1 (o 100%), mentre il peggior valore possibile è invece zero. Esprime quanto sia possibile fidarsi del classificatore quando prevede un elemento come positivo. Viene comunemente chiamata *Precision.*[101].
Il NPV, acronimo inglese di *negative predictive value*, rappresenta la frazione di elementi TN divisa per il numero totale di unità previste negativamente. Questa metrica esprime la proporzione di unità che il classificatore afferma essere negative e che in realtà sono negative [102].

$$NPV = p(C = N \mid \widehat{C} = N) \cong \frac{TN}{TN + FN}$$

Il valore ideale del NPV, con un test perfetto che non restituisce falsi negativi, risulta uguale a 1 (o 100%). Mentre il peggior valore possibile risulta essere invece uguale a zero, con un test che non restituisce veri negativi. Questa metrica esprime quanto sia possibile fidarsi del classificatore quando prevede un elemento come negativo [101].
Fra le metriche di base, la curva ROC è considerata la migliore per stabilire le prestazioni, essa è una tecnica grafica utilizzata per valutare la capacità diagnostica di un classificatore binario al variare della sua soglia di discriminazione. Essa viene creata tracciando il tasso di veri positivi (TPR) rispetto al tasso di falsi positivi (FPR) a varie impostazioni di soglia. La curva è ottenuta quindi calcolando la sensibilità e la specificità del test in ogni possibile punto di *cut-off*, e tracciando la sensibilità in confronto alla 1−specificità. Le curve ROC consentono analisi visive dei compromessi tra la sensibilità e la specificità di un test rispetto ai vari *cut-off* che possono essere utilizzati. [103].
Per convenzione, la sensibilità, ossia la proporzione di risultati veri positivi, è mostrata sull'asse delle ordinate, e va da 0 a 1 (0–100%). Mentre la 1−specificità, ossia la proporzione di risultati falsi positivi, è mostrata sull'asse delle ascisse, e va anch'essa da 0 a 1 (0–100%) [103].

La curva ROC si ottiene facendo variare la soglia rispetto alla quale si confronta la metrica scelta con il test. Se la metrica è maggiore della soglia, allora si dichiara il campione positivo. I valori ottimali di soglia ottenuti nel test hanno coordinate (0,11; 0,89).
La forma di una curva ROC e l'area sotto la curva (meglio nota come AUC dell'inglese *area under the curve*) ci aiutano a stimare quanto sia alto il potere discriminante di un test [104]. Più la curva è vicina all'angolo in alto a sinistra e più grande è l'area sotto la curva, migliore è il test nel discriminare tra malato e non malato. L'area sotto la curva può assumere qualsiasi valore compreso tra 0 e 1 ed è un buon indicatore della bontà del test. Un test diagnostico perfetto ha un valore di AUC uguale ad 1, mentre un test inutile ha un valore di 0,5 [104]. Un modo per interpretare l'area sotto la curva ROC è che un test con un'area maggiore di 0,9 ha un'accuratezza elevata, mentre 0,7-0,9 indica un'accuratezza moderata, 0,5-0,7 una precisione bassa e 0,5 un risultato casuale [103]. L'AUC funge da singola misura, indipendente dalla prevalenza, che riassume la capacità discriminativa di un test attraverso l'intera gamma di *cut-off*. Maggiore è l'area sotto la curva ROC, migliore sarà il risultato del test [103]. L'AUC è una misura globale dell'accuratezza diagnostica e non ci dice nulla sui singoli parametri di sensibilità e di specificità [104].
In **Figura 5.1** è rappresentata la curva ROC relativa al *test set*, ottenuta tramite il *Classification Learner app* di MATLAB.

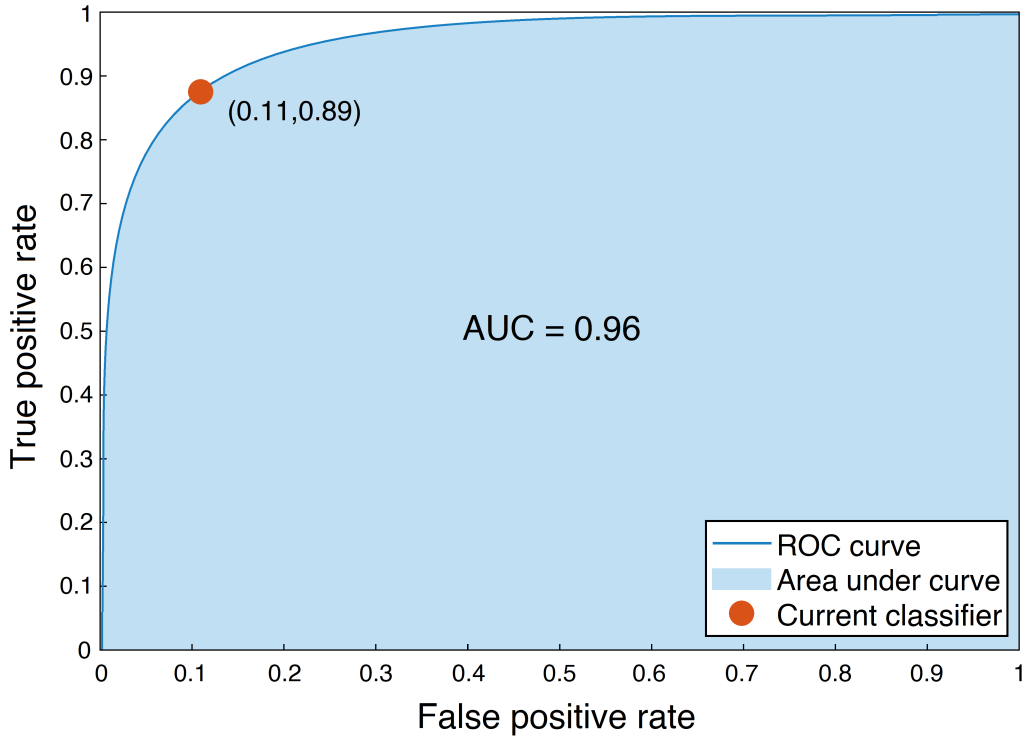



**Figura 5.1: La curva ROC del set di test.** *La curva ROC ottenuta dal set di test, risulta abbastanza aderente all'angolo in alto a sinistra, indice che la classificazione ha ottenuto buoni risultati. Sull'asse delle ordinate si trova il TPR (true positive rate) o sensibilità, mentre sull'asse delle ascisse si trova il FPR (false positive rate) o 1−specificità. Anche l'area sotto la curva mostra prestazioni di classificazione ottime, il valore raggiunto è infatti di 0,96. Esso corrisponde all'integrale dei valori TPR della curva ROC, rispetto a FPR, da FPR=0 a FPR=1 [105].*

Dopo le metriche di base ci sono quelle di primo livello, fra queste la metrica $F_1$-*score* è sicuramente una delle più importanti [102].
Essa valuta le prestazioni del modello di classificazione partendo dalla matrice di confusione, aggrega le metriche sopra descritte di *Precision* e *Recall*, sotto il concetto di media armonica. La metrica $F_1$-*score* può essere interpretata come una media ponderata tra queste due metriche, e raggiunge il suo valore migliore a 1 e il punteggio peggiore a 0 [102]. Esso è uguale al contributo relativo di *Precision* e *Recall* e la media armonica è utile per trovare il miglior compromesso tra le due grandezze.

$$F_1\text{-}score = \left(\frac{2}{Precision^{-1} + Recall^{-1}}\right) = 2 \cdot \left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$$

Da notare che sia *Precision* che *Recall* assumono valori nell'intervallo 0-1. Quando uno di essi assume valori prossimi allo 0, la metrica $F_1$-*score* subisce un enorme calo, infatti la media armonica tende a dare maggior peso ai valori più bassi [102].
Una delle metriche di primo livello più utilizzata è il cosiddetto Kappa di Cohen, esso è un coefficiente statistico che mostra il grado di affidabilità e accuratezza in una classificazione [106]. Il coefficiente Kappa di Cohen è un indice che tiene conto della probabilità di concordanza casuale, calcolato in base al rapporto tra l'accordo in

eccesso rispetto alla probabilità di concordanza casuale e l'eccesso massimo ottenibile [106].
Il Kappa di Cohen si basa sull'idea di misurare la concordanza tra le etichette predette e le etichette vere, entrambe considerate variabili categoriche casuali. Tramite la matrice di confusione è possibile confrontare le due variabili categoriali, calcolando le distribuzioni delle righe marginali e delle colonne marginali [102].
Gli indicatori Kappa di Cohen possono essere visti come i valori di valutazione della dipendenza, o indipendenza, tra la predizione del modello e la classificazione effettiva. La distribuzione delle colonne marginali può essere considerata come la distribuzione dei valori predetti, mentre le righe marginali rappresentano la distribuzione delle classi reali [102].
Cohen nel 1960 ha definito il coefficiente Kappa come segue:

$$K = \frac{P_o - P_e}{1 - P_e}$$

dove $P_o = p(\widehat{C} = C)$ è la proporzione dell'accordanza osservata, in altre parole è l'accuratezza raggiunta dal modello [107]. In un caso di classificazione binaria, con matrice di confusione $2 \times 2$, si ha:

$$P_o = \frac{TP + TN}{TP + FP + FN + TN} = p(\widehat{C} = C)$$

$P_e$ è invece l'accuratezza nel caso in cui $\widehat{C}$ sia statisticamente indipendente da $C$. Si ha:

$$\begin{aligned} P_e &= p(C = P, \widehat{C} = P) + p(C = N, \widehat{C} = N) \\ &= p(C = P)p(\widehat{C} = P) + p(C = N)p(\widehat{C} = N) \end{aligned}$$

che utilizzando gli elementi della matrice di confusione vale:

$$P_e \cong \left(\frac{TP + FP}{N} \cdot \frac{TP + FN}{N}\right) + \left(\frac{TN + FP}{N} \cdot \frac{TN + FN}{N}\right)$$

Il massimo valore di $P_o$ è 1 (classificatore perfetto) e $1 - P_e$ è dunque il massimo valore del numeratore di $K$, e $K$ ha come valore massimo 1. In particolare, $K = -1$ se $P_o = 0$ (il classificatore sbaglia sempre) e $P_e = 1/2$. Di conseguenza:

$$K = \frac{2 \cdot (TP \cdot TN - FP \cdot FN)}{(TP + FP) \cdot (FP + TN) + (TP + FN) \cdot (FN + TN)}$$

Il risultato Kappa può assumere valori da $-1$ a $+1$ e viene interpretato come segue [102]. I valori $\leq 0$ indicano nessuna accordanza, 0,01–0,20 da nulla a lieve, 0,21–0,40 accordanza discreta, 0,41–0,60 moderata, 0,61–0,80 sostanziale e 0,81–1,00 accordanza quasi perfetta [106]. I valori negativi indicano che l'accordanza osservata è peggiore di quanto ci si aspetterebbe per un evento casuale. Un'interpretazione alternativa afferma che valori di Kappa inferiori a 0,60 indicano un significativo livello di disaccordo [102].

La metrica di secondo livello più ampiamente utilizzata è il coefficiente quadratico medio di contingenza, meglio conosciuto come MCC (acronimo inglese di *Mattheus correlation coefficient*) [102] che negli anni 2000 è diventato una metrica ampiamente utilizzata per testare le prestazioni delle tecniche di *machine learning* [102].
Esso è stato sviluppato da Brian W. Matthews nel 1975, ed è uguale al coefficiente di correlazione $\rho$ di Karl Pearson per la coppia di variabili aleatorie $\alpha$ e $\widehat{\alpha}$ definite come segue:

$$\alpha = 1 \text{ se la classe vera è } P \text{ e } \alpha = 0 \text{ se la classe vera è } N$$
$$\widehat{\alpha} = 1 \text{ se la classe stimate è } P \text{ e } \widehat{\alpha} = 0 \text{ se la classe stimata è } N$$

con

$$\rho_{\alpha,\widehat{\alpha}} = \frac{\mathbb{E}\left\{(\alpha - \mu_\alpha)(\widehat{\alpha} - \mu_\alpha)\right\}}{\sigma_\alpha \sigma_{\widehat{\alpha}}}$$

allora

$$\mu_\alpha = 1 \cdot p(C = P) + 0 \cdot p(C = N) = P(C = P)$$
$$\mu_{\widehat{\alpha}} = p(\widehat{C} = P)$$

$$\mathbb{E}\left\{\alpha^2\right\} = 1^2 \cdot (C = P) + 0^2 \cdot (C = N) = p(C = P)$$
$$\mathbb{E}\left\{\widehat{\alpha}^2\right\} = p(\widehat{C} = P)$$

$$\sigma_\alpha^2 = \mathbb{E}\left\{\alpha^2\right\} - \mu_\alpha^2 = p(C = P) - (p(C = P))^2 = p(C = P)(1 - P(C = P))$$
$$= p(C = P)p(C = N)$$

$$\sigma_{\widehat{\alpha}}^2 = p(\widehat{C} = P)[1 - p(\widehat{C} = P)] = p(\widehat{C} = P)p(\widehat{C} = N)$$

con

$$\rho = \frac{\mathbb{E}\left\{\alpha\widehat{\alpha}\right\} - \mu_\alpha \mu_{\widehat{\alpha}}}{\sqrt{\sigma_\alpha^2 \sigma_{\widehat{\alpha}}^2}}$$

e

$$\mathbb{E}\left\{\alpha\widehat{\alpha}\right\} = 1 \cdot p(\alpha = 1, \widehat{\alpha} = 1) =$$
$$= p(C = P, \widehat{C} = P)$$

allora

$$\rho = \frac{p(C = P, \widehat{C} = P) - p(C = P)p(\widehat{C} = P)}{\sqrt{p(C = P)p(C = N)p(\widehat{C} = P)p(\widehat{C} = N)}}$$

con

$$p(C = P, \widehat{C} = P) \cong \frac{TP}{N}$$
$$p(C = P) \cong \frac{TP + FN}{N}$$
$$p(\widehat{C} = P) \cong \frac{TP + FP}{N}$$

Il coefficiente di correlazione di Matthews assume valori in un intervallo da $-1$ a $+1$. Valori prossimi ad 1 indicano una previsione molto buona, ciò significa che esiste una forte correlazione positiva tra la previsione e le etichette reali. Questo implica che le due variabili concordano fortemente, quindi i valori previsti saranno molto simili alla classificazione effettiva. Al contrario, quando il coefficiente di correlazione è prossimo allo 0, significa che non vi è correlazione tra le variabili. In questo caso indica che il classificatore assegna casualmente le unità alle classi, senza alcun legame con il loro vero valore di classe [102]. Il coefficiente di correlazione può anche essere negativo, in questo caso la relazione tra classi vere e predette è di tipo inverso. Una forte correlazione inversa significa che il modello ha imparato a classificare i dati, ma cambia sistematicamente tutte le etichette [102].
Il coefficiente di correlazione di Mattheus potrebbe essere visto come il coefficiente Phi applicato a problemi di classificazione binaria, ed ha la seguente formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

Come di può notare dalla formula, il coefficiente di correlazione di Mattheus tiene conto di tutte le celle della matrice di confusione [102]. Ciò significa che esso è una misura bilanciata che può essere utilizzata nella classificazione binaria anche se le classi hanno dimensioni molto diverse.
Tuttavia il coefficiente ha anche dei punti deboli. Infatti il valore finale mostra fluttuazioni molto ampie, nel caso in cui vi siano risultati sbilanciati nella previsione del modello [102].
Tutte le metriche sopra descritte sono state calcolate sul *test set* per valutare le prestazioni del classificatore. Una volta registrati i valori ottenuti, sulle maschere binarie è stato applicato l'algoritmo di pulizia delle immagini descritto nella sezione 3.7. Una volta ripulite le maschere, e quindi un dataset di test, sono state ricalcolate le medesime metriche.
In **Tabella 5.1** sono riassunti tutti i risultati delle metriche ottenute, sia in uscita dal classificatore che a valle dell'algoritmo di pulizia delle immagini. L'ultima colonna a destra rappresenta la differenza fra i valori delle due metriche, prima e dopo l'algoritmo.
Dalla **Tabella 5.1** si osservano buoni risultati di classificazione, in particolare per accuratezza, sensibilità e PPV che sfiorano il 90%.
L'algoritmo di pulizia delle maschere ha rimosso 470 ROI che erano state classificate come patologiche; tra queste, 129 risultavano essere realmente patologiche, mentre 341 risultavano essere sane. È da ricordare che, per come è stato implementato l'algoritmo, la pulizia agisce solo sulle ROI classificate come positive, quindi patologiche.

| Metrica | Pre algoritmo [%] | Post algoritmo [%] | Δ [%] |
|---|---|---|---|
| Accuratezza | 89,3 | 90,5 | +1,2 |
| Sensibilità | 89,7 | 88,8 | −0,9 |
| Specificità | 88,8 | 92,3 | +3,5 |
| PPV | 89,9 | 88,3 | −1,6 |
| NPV | 88,6 | 92,7 | +4,1 |
| $F_1$-score | 89,3 | 90,3 | +1,0 |
| K | 78,5 | 81,0 | +2,5 |
| MCC | 78,5 | 81,1 | +2,6 |

**Tabella 5.1: Le prestazioni del classificatore.** *La tabella riassume i risultati ottenuti in base alla metrica di riferimento. Nella seconda colonna sono presenti i risultati effettivi ottenuti dal classificatore. Nella terza colonna sono presenti i risultati ottenuti a valle dell'algoritmo di pulizia delle maschere. La quarta e ultima colonna rappresenta la differenza fra i valori della seconda e della terza colonna. In sei metriche su otto utilizzate, l'algoritmo di pulizia delle maschere ha migliorato le prestazioni di classificazione, in due casi invece (sensibilità e PPV) l'algoritmo di pulizia ha peggiorato le prestazioni.*

In questo modo i veri positivi (TP) sono diminuiti leggermente, passando da 7572 ROI a 7443 (−1,7%), mentre i falsi positivi (FP) hanno subito una diminuzione sostanziale, passando da 958 ROI a 617 (−35,6%).
Nella Appendice sono consultabili le matrici di confusione ottenute dalla classificazione del *test set*, calcolate sia prima che dopo l'algoritmo di pulizia delle maschere. Sono presenti due tipologie di matrici, quelle con valore numerico e quelle con valore percentuale, normalizzato per riga.
Grazie all'algoritmo di pulizia delle maschere sono aumentate le prestazioni di quasi tutte le metriche. In particolare è aumentata l'accuratezza generale (+1,2%), ma soprattutto sono migliorate le prestazioni di specificità (+3,5%) e NPV (+4,1%).
Inoltre gli indici $K$ e $MCC$ sono saliti entrambi a oltre l'80%, mostrando ottime prestazioni di classificazione.
Di contro, come da attese, sono diminuiti la sensibilità (−0,9%) e la PPV (−1,6%), ma in misura nettamente inferiore rispetto ai miglioramenti ottenuti.

## 5.2 L'efficacia delle maschere

Le prestazioni dell'algoritmo di classificazione sono state valutate anche in base alla capacità della foresta casuale di creare maschere binarie che circoscrivano le aree potenzialmente patologiche.
Nella sezione 3.7, è stato descritto in che modo le ROI predette, in uscita dal classificatore, sono state convertite in maschere binarie. Tramite la valutazione di quest'ultime è possibile calcolare la capacità del classificatore di identificare intere aree realmente patologiche.
In particolare, le maschere binarie di cui sopra sono state confrontate con quelle segmentate manualmente dal medico, descritte nella sezione 3.1.

Il metodo utilizzato per valutare la sovrapposizione delle due maschere binarie è il coefficiente di similarità di Dice. Viene spesso chiamato anche coefficiente o indice Sørensen–Dice, o più semplicemente coefficiente di Dice.
Venne ideato nel 1945 dal genetista Lee Raymond Dice per studi di ecologia [108], successivamente nel 1948 il professore Thorvald Sørensen ne creò uno simile, in maniera indipendente, per studi di botanica [109], a seguito di ciò la misura venne chiamata coefficiente di similarità Sørensen–Dice [109].
Il DSC (acronimo inglese di *Dice similarity coefficient*) è un coefficiente statistico utilizzato per valutare la somiglianza di due campioni [110]. Presenta un intervallo di valori che va da 0 ad 1, dove 0 significa assenza di somiglianza ed 1 significa perfetta somiglianza.
Il coefficiente di similarità di Dice per due campioni $A$ e $B$ è definito come segue:

$$DSC = \frac{2\,|A \cap B|}{|A| + |B|}$$

Nel caso particolare di classificazione binaria, il coefficiente può essere descritto anche tramite gli elementi della matrice di confusione [111]:

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

In **Figura 5.2** il significato grafico dell'indice Sørensen–Dice.

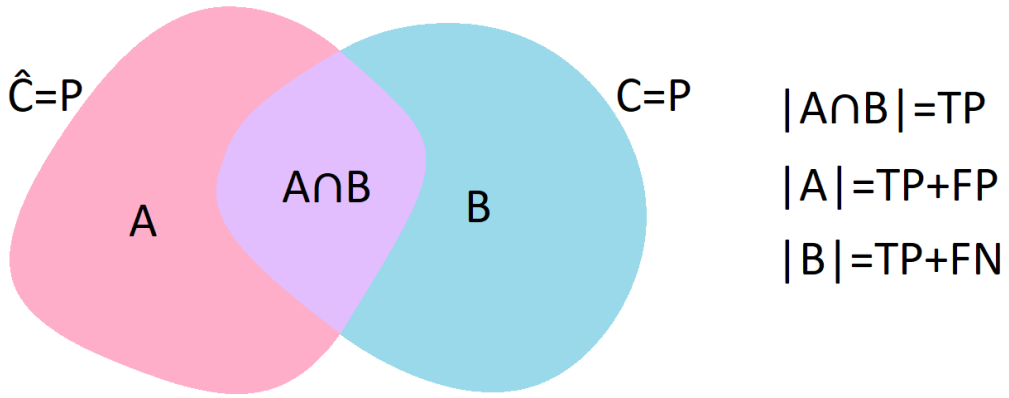



**Figura 5.2: Il significato grafico dell'indice Sørensen–Dice.** *Una rappresentazione grafica dell'indice Sørensen–Dice. L'insieme $A$, di colore rosa, è distinto dall'insieme $B$, di colore celeste, la loro intersezione $A \cap B$ è rappresentata di colore lilla.*

Le maschere binarie in uscita dal classificatore sono state quindi confrontate, tramite l'indice Sørensen–Dice, con le maschere segmentate manualmente dal medico. In **Figura 5.3** è presente un grafico a dispersione che mostra i risultati ottenuti per ogni singola immagine patologica.

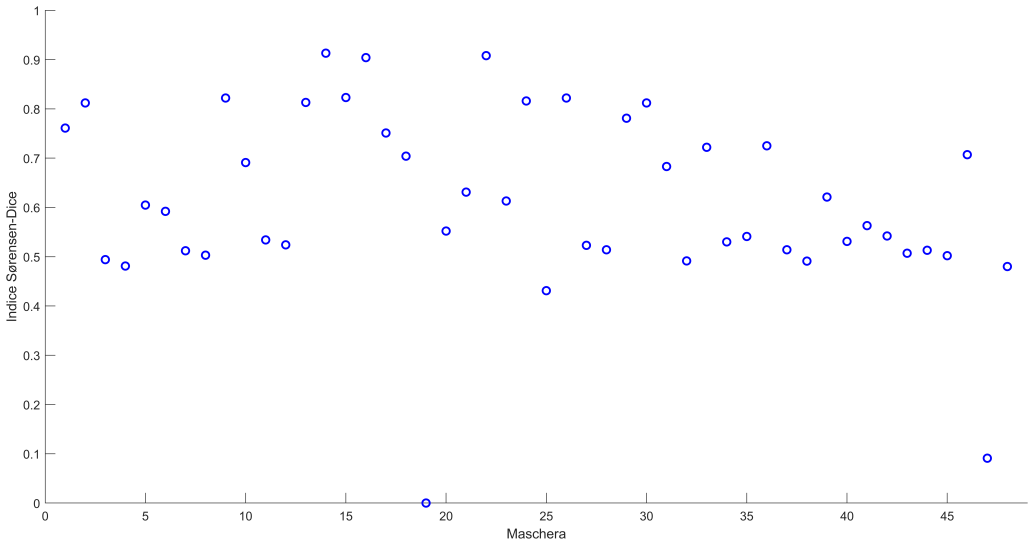



**Figura 5.3: Grafico a dispersione dell'indice Sørensen-Dice.** *Nel grafico a dispersione sono presenti i risultati dell'indice Sørensen-Dice, per tutte le 48 immagini patologiche. La maggior parte delle maschere ha ottenuto un valore superiore allo 0.5, solo in due casi (immagini 19 e 47), è stato ottenuto un valore molto basso, inferiore allo 0.10.*

In oltre l'85% dei casi (41 immagini su 48) l'indice Sørensen-Dice ha ottenuto un valore superiore allo 0.5. La quasi totalità delle maschere (46 immagini su 48) ricade nella fascia $0,431 \leq DSC \leq 0,913$. Solo due immagini risultano avere valori anomali, la numero 19 e la numero 47, che hanno ottenuto rispettivamente valori di $DSC = 0$ e $DSC = 0,091$.
Queste ultime due maschere possono essere considerate dei veri e propri *outlier*, in **Tabella 5.2** sono riportati il valore dell'indice di Sørensen-Dice minimo, massimo, medio e deviazione standard, sia considerando l'intero dataset che rimuovendo i due *outlier*. I valori sono stati calcolati dopo la fase di *post-processing* delle maschere, ossia dopo la loro pulizia con DBSCAN.

| Indice Sørensen-Dice | Intero dataset | Senza outlier |
|---|---|---|
| Minimo | 0 | 0,431 |
| Massimo | 0,913 | 0,913 |
| Medio | 0,612 | 0,637 |
| Deviazione standard | 0,182 | 0,141 |

**Tabella 5.2: I risultati dell'indice Sørensen-Dice.** *In tabella sono rappresentati i valori minimo, massimo, medio e deviazione standard dell'indice Sørensen-Dice. Nella colonna centrale sono riportati i valori ottenuti dall'intero dataset, nella colonna di destra sono riportati i valori ottenuti a seguito dell'eliminazione delle immagini 19 e 47. Come si può notare, rimuovendo queste due immagini, sale sia il valore minimo che la media, si riduce invece il valore della deviazione standard.*

Facendo riferimento alla **Tabella 5.2**, l'indice Sørensen-Dice medio dell'intero dataset è uguale a 0,612. Considerando invece le immagini 19 e 47 come valori anomali,

la media sale a 0,637. In **Figura 5.4** sono rappresentate le maschere vere e predette delle immagini 19 e 47.

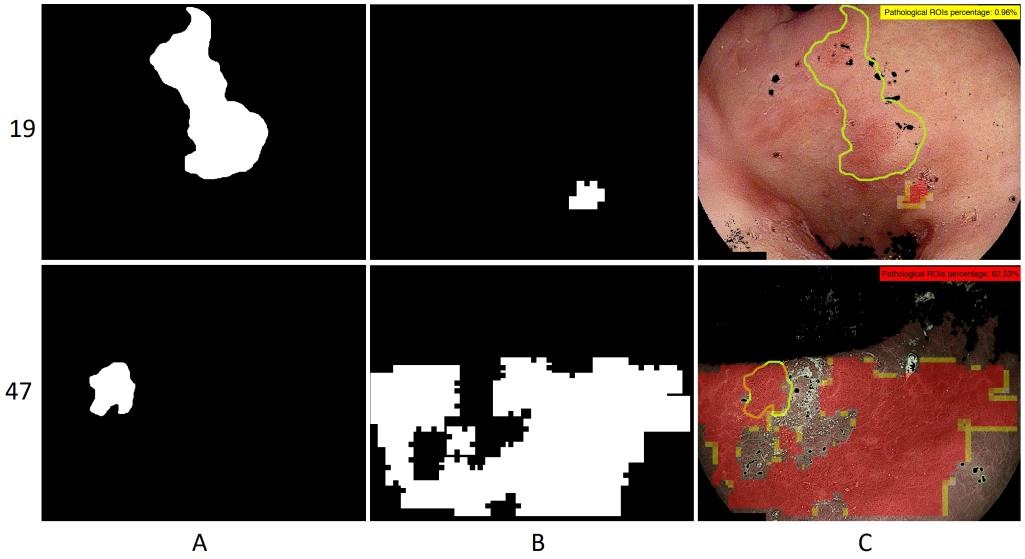



**Figura 5.4: Le maschere vere e predette delle immagini 19 e 47.** *(A) Le maschere segmentate dal medico delle immagini 19, prima riga, e 47, seconda riga. (B) Le maschere ottenute dall'algoritmo di classificazione. (C) Le immagini in uscita dal sistema di diagnosi automatizzata; risulta evidente la sovrapposizione non efficace delle maschere vere e predette.*

La differenza di risultati ottenuti fra le metriche della scala microscopica (e.g. sensibilità $\cong 89\%$) e quelle della scala intermedia ($DSC \cong 64\%$) sono riconducibili al differente numero di ROI impiegate per il calcolo. Infatti le metriche per la scala microscopica sono calcolate solo sulle ROI effettivamente analizzate, ossia con immagini ripulite da parti indesiderate e artefatti. L'efficacia delle maschere è stata invece calcolata sulle immagini endoscopiche intere, ossia comprendenti anche le aree non di interesse. Per rendere meglio l'idea, si pensi che su un'immagine di dimensioni $1226 \times 973$ pixel, come quelle utilizzate in questo lavoro, è possibile estrarre fino a 1302 ROI di dimensioni $50 \times 50$ con *overlap* bidirezionale del 40%, mentre per ogni immagine sono state estratte, e quindi analizzate, mediamente 700 ROI.
Oltre la sovrapposizione delle maschere è stata valutata anche la loro ampiezza; questa è stata calcolata in base al numero di pixel posseduti da ciascuna. Il valore di ampiezza ottenuto con l'algoritmo di classificazione è stato poi confrontato con quello delle maschere segmentate manualmente dal medico, secondo la seguente formula:

$$\Delta_{\%} = \frac{N_{RF} - N_M}{N_M} \cdot 100$$

Dove $\Delta_{\%}$ è la differenza di pixel percentuale, $N_{RF}$ è il numero di pixel della regione patologica della maschera automatica e $N_M$ è il numero di pixel della regione patologica della maschera segmentata manualmente dal medico.
Sono state considerate trascurabili le differenze di ampiezza inferiori o uguali al 20%, in base al valore $\Delta_{\%}$ ottenuto sono state determinate le seguenti tre classi:

- **Ampiezza sovrastimata:** $\Delta_{\%} > +20\%$
- **Ampiezza confrontabile:** $-20\% \leq \Delta_{\%} \leq +20\%$
- **Ampiezza sottostimata:** $\Delta_{\%} < -20\%$

Con questa configurazione, in 18 casi l'algoritmo di classificazione ha sovrastimato l'ampiezza della regione metaplastica, in 22 casi la maschera binaria presenta una superficie confrontabile con quella segmentata dal medico, in 8 casi l'algoritmo di classificazione ha sottostimato la regione metaplastica.
Si può quindi affermare che, nella maggior parte dei casi, l'algoritmo di classificazione, ha stimato correttamente l'ampiezza della regione metaplastica. L'algoritmo al più ha sovrastimato l'area potenzialmente anomala; solo in pochi casi sottostima l'ampiezza della lesione.

## 5.3 La classificazione delle immagini

L'algoritmo di classificazione implementato, basato sull'intelligenza artificiale, si pone l'intento di rilevare la presenza di metaplasia intestinale gastrica, durante un esame endoscopico.
Si vuole, data un'immagine gastrica, poter supportare il medico nella diagnosi della patologia, fornendo una valutazione oggettiva. Questo sistema, oltre ad assistere il medico nell'analisi, potrebbe evitare inutili biopsie.
Il classificatore deve stimare se un'immagine endoscopica in ingresso proviene da un paziente sano o da uno patologico.
La classificazione delle immagini è stata implementata come segue. Sono state classificate come sane solo le immagini nelle quali non è stata riscontrata alcuna anomalia. Sono state classificate come patologiche tutte le altre immagini, divise su tre livelli di gravità basati sull'estensione dell'anomalia.
Valendosi dell'indicazione semaforica descritta nella sezione 4.2 è stato valutato l'intero dataset di immagini (sia immagini sane che patologiche) ottenendo i seguenti risultati.
Le immagini endoscopiche che erano provenienti da pazienti affetti da metaplasia intestinale gastrica (patologia accertata tramite biopsia) sono state classificate come segue:

- **Rosso:** 43 immagini su 48
- **Arancione:** 4 immagini su 48
- **Giallo:** 1 immagine su 48
- **Verde:** 0 immagini su 48

Secondo quanto descritto sopra, 48 immagini su 48 sono state classificate come provenienti da pazienti patologici. Quindi il 100% delle immagini è stato classificato nella giusta categoria.

Di queste, il 90% è stato classificato come livello rosso, ossia aventi un significativo riscontro di anomalie. Solo 1 immagine su 48 è stata posta nel livello giallo, cioè il più basso livello di interessamento della lezione metaplastica.
Le immagini endoscopiche provenienti da pazienti sani sono state invece classificate come segue:

- **Rosso:** 0 immagini su 47
- **Arancione:** 0 immagini su 47
- **Giallo:** 1 immagine su 47
- **Verde:** 46 immagini su 47

Un totale di 46 immagini su 47 è stato classificato come proveniente da pazienti sani. Questo significa che circa il 98% delle immagini è stato classificato nella giusta categoria.
Nel dataset delle immagini sane è stato commesso solo il 2% d'errore, ossia 1 immagine, classificata come livello giallo, ovvero con basso riscontro di anomalie. Nessuna immagine è stata classificata con livello arancione o rosso.
Prendendo in considerazione l'intero dataset di immagini sane e patologiche, si possono riassumere i seguenti risultati.
Sono state classificate correttamente 94 immagini su 95 totali. Questo significa che è stata ottenuta un'accuratezza di circa il 99%. Solo un'immagine è stata classificata nella categoria sbagliata. Nella Appendice sono consultabili quattro esempi di output del sistema di diagnosi automatizzata: due immagini sane, una classificata correttamente e una sbagliata (l'unica del dataset) e due immagini patologiche.

# 6. Conclusioni

L'utilizzo dell'intelligenza artificiale nella diagnostica per immagini è un'area di ricerca in rapida crescita, dato il suo enorme potenziale e la sua versatilità. Viene applicata nella diagnosi di numerose patologie, e in letteratura si trovano diversi lavori relativi all'analisi di immagini endoscopiche con l'utilizzo di vari modelli computazionali.
In questa tesi è stata analizzata la possibilità di utilizzare l'intelligenza artificiale come supporto al medico nella diagnosi della metaplasia intestinale gastrica. Essa può rappresentare il passaggio istologico precedente allo sviluppo di una displasia ed è direttamente associata alla comparsa del carcinoma gastrico.
Oggigiorno il mezzo più efficace per contrastare lo sviluppo del carcinoma gastrico è una diagnosi precoce della patologia. L'intelligenza artificiale potrebbe ridurre notevolmente il tempo necessario alla diagnosi, assistendo il medico nell'identificare la metaplasia intestinale gastrica. In questo, l'assistenza di un computer, potrebbe apportare diversi benefici significativi.
Attualmente il processo di diagnosi prevede un esame endoscopico gastrico al quale segue, nel caso si riscontri possibile presenza di metaplasia, una biopsia locale. Questo processo si porta dietro diverse criticità. Tra le più importanti sono sicuramente i tempi di attesa; fra esame endoscopico e diagnosi finale può intercorrere diverso tempo, di solito nell'ordine dei giorni o settimane. Alla biopsia segue infatti un esame istologico e questo, oltre a rallentare notevolmente il processo di diagnosi, può essere anch'esso affetto da errore di valutazione da parte dell'operatore.
Un computer che si avvale dell'intelligenza artificiale, in questo, potrebbe fornire diversi vantaggi. Un singolo computer potrebbe analizzare migliaia di immagini endoscopiche al giorno riducendo notevolmente i tempi di attesa. Ma soprattutto senza andare incontro ai fenomeni di variabilità intra- e inter- operatore. Infatti, oltre alla soggettività dettata dall'esperienza, la diagnosi è condizionata anche dall'affaticamento del medico. Tutte problematiche risolte tramite l'utilizzo di un computer.
La metaplasia intestinale gastrica è inoltre una patologia molto difficile da diagnosticare. Occorrono molti anni di esperienza e tanta pratica da parte di un medico prima di raggiungere un alto livello di accuratezza nella diagnosi. Nonostante ciò, in tanti casi, il risultato istologico della biopsia gastrica confuta la valutazione preliminare effettuata dall'operatore sanitario. Ciò significa che la diagnosi della metaplasia, allo stato attuale, è fortemente dipendente dall'operatore, per cui medici diversi potrebbero fornire valutazioni preliminari differenti. Una diagnosi oggettiva dell'immagine, effettuata da un computer, potrebbe fornire un'indicazione non dipendente dalla soggettività del medico. L'intelligenza artificiale potrebbe per esem-

pio richiamare l'attenzione dell'operatore su porzioni di tessuto gastrico che erano state precedentemente ignorate o valutate come sane. Oppure, una volta ricevuto un esito discordante, il medico potrebbe effettuare una seconda analisi più mirata e approfondita.
Ci si chiede dunque se l'intelligenza artificiale possa avere prestazioni di classificazione pari o superiori ad un medico esperto.
La fase di sviluppo del progetto è stata preceduta da diversi incontri con medici specializzati dell'azienda Ospedaliera Ordine Mauriziano di Torino. Con essi ci si è confrontati sulle strategie di diagnosi, utilizzate in endoscopia, per valutare la metaplasia gastrica.
Sulla base di questi confronti, è stato scelto un approccio di classificazione incentrato sull'estrazione di caratteristiche in bianco-nero. Si tratta quindi di intelligenza artificiale basata su *machine learning* ad apprendimento supervisionato.
Una volta ripulite le immagini dagli artefatti ed estratte le caratteristiche, è iniziata la parte più impegnativa del lavoro, ossia la creazione del modello.
Molta attenzione è stata dedicata alla creazione del dataset di caratteristiche, utilizzato per l'addestramento del classificatore. Questo risultava fortemente sbilanciato, così è stato necessario creare un set di dati che fosse significativo per l'addestramento ma al contempo equilibrato. La creazione di un dataset *ad hoc* si è dimostrata molto vantaggiosa, influenzando notevolmente i risultati ottenuti.
Il core del progetto è stata la scelta, l'adattamento e l'ottimizzazione del classificatore. Data la complessità del compito richiesto e data l'eterogeneità del dataset d'immagini, è stato necessario effettuare numerose prove e diversi tentativi.
Sono stati vagliati diversi modelli, come reti neurali, alberi decisionali e differenti metodi d'apprendimento d'insieme. La scelta è ricaduta poi sul classificatore d'insieme delle foreste casuali, vantaggioso per robustezza e resistenza all'*overfitting*.
Inoltre è stata valutata la possibilità di ridurre la dimensionalità del dataset con il metodo della analisi delle componenti principali (meglio nota come PCA dall'inglese *principal component analysis*). Quest'ipotesi è stata poi declinata data la limitatezza dei risultati ottenuti. La scelta di non alterare il dataset ha fornito la soluzione più complessa dal punto di vista computazionale, ma al contempo quella più efficace.
A causa dell'incertezza nella classificazione si è resa necessaria l'implementazione di una breve parte di *post-processing* delle maschere, questa ha completato l'algoritmo di classificazione e contribuito significativamente a migliorarne le *performance*.

## 6.1 Discussione

Nel complesso l'algoritmo di classificazione ha ottenuto buone prestazioni di categorizzazione. Su scala microscopica, a livello quindi delle ROI, il modello ha raggiunto un'accuratezza generale superiore al 90%. L'algoritmo pecca un po' di precisione ma ha attenuto un buon livello di specificità, oltre il 92%. Questo si traduce in un'elevata capacità di rilevare i veri negativi, ossia le regioni di tessuto sane.
Su scala intermedia, a livello quindi delle maschere di segmentazione, il modello ha attenuto risultati incerti ma apprezzabili. In particolare, la maschera creata dall'algoritmo di segmentazione automatica non è spesso aderente a quella segmentata dal medico. Nonostante ciò circoscrive sempre aree ben delimitate e definite,

spesso comparabili per dimensione e posizione, con quelle tracciate manualmente dall'operatore.
Su scala macroscopica, a livello quindi di intero dataset di immagini, il modello ha ottenuto risultati soddisfacenti. I risultati sono avvalorati dal fatto che le immagini endoscopiche erano di per sé dubbie per il medico, poiché in una diagnosi preliminare erano state valutate come potenzialmente patologiche ma con l'esame istologico sono state poi classificate come sane. Su questa scala il modello ha infatti ottenuto un'accuratezza del 99%.
Si può dunque affermare che, in via generale, il modello ha ottenuto buone capacità di categorizzazione, con prospettive promettenti. L'algoritmo analizza e valuta una singola immagine endoscopica in circa un minuto, dunque i tempi di calcolo sono brevi ma comunque non trascurabili.
Le funzioni e gli algoritmi implementati potrebbero risultare inaccessibili per un utente non esperto in programmazione e calcolo numerico. A tale proposito è stata creata un'interfaccia grafica, semplice ed intuitiva.
Con la creazione dell'interfaccia grafica è stato implementato un vero e proprio prototipo di sistema per diagnosi assistita da computer. Questi strumenti informatici, noti anche come CAD (dall'inglese *computer aided diagnosis*), vengono al giorno d'oggi sempre più spesso impiegati dai medici e sembrerebbero, tutt'ora, rappresentare il futuro della medicina.
Nonostante i buoni risultati ottenuti, emergono alcuni limiti dell'algoritmo di classificazione.
I limiti principali sono legati al dataset di immagini. Sin dalla fase preliminare del progetto è apparso subito chiaro che il numero esiguo di immagini a disposizione fosse il principale collo di bottiglia. In altri lavori simili, il dataset di pazienti e immagini era molto più ampio e vario. Per citarne un paio, Yan et al. hanno lavorato su 1880 immagini endoscopiche provenienti da 80 pazienti [35], Lin et al. su addirittura 7037 immagini provenienti da 2741 pazienti [112].
Per estrarre correttamente le caratteristiche di texture e pattern è stato necessario utilizzare finestre di acquisizione piuttosto grandi, questo ha ridotto notevolmente il numero delle ROI. Anche utilizzando altre tecniche di aumento del dataset, come ruotare le immagini con un angolo casuale, specchiarle verticalmente e orizzontalmente, traslazione laterale, ingrandimento, il dataset sarebbe comunque limitato.
Ciò ha reso l'algoritmo di classificazione fortemente dipendente dalle immagini di addestramento, riducendo così le capacità di generalizzazione del classificatore.
Inoltre, allo stato attuale, l'algoritmo riesce ad analizzare solo catture video della sonda endoscopica digitale i-scan della società Pentax. Qualsiasi altra immagine endoscopica gastrica non verrebbe analizzata.

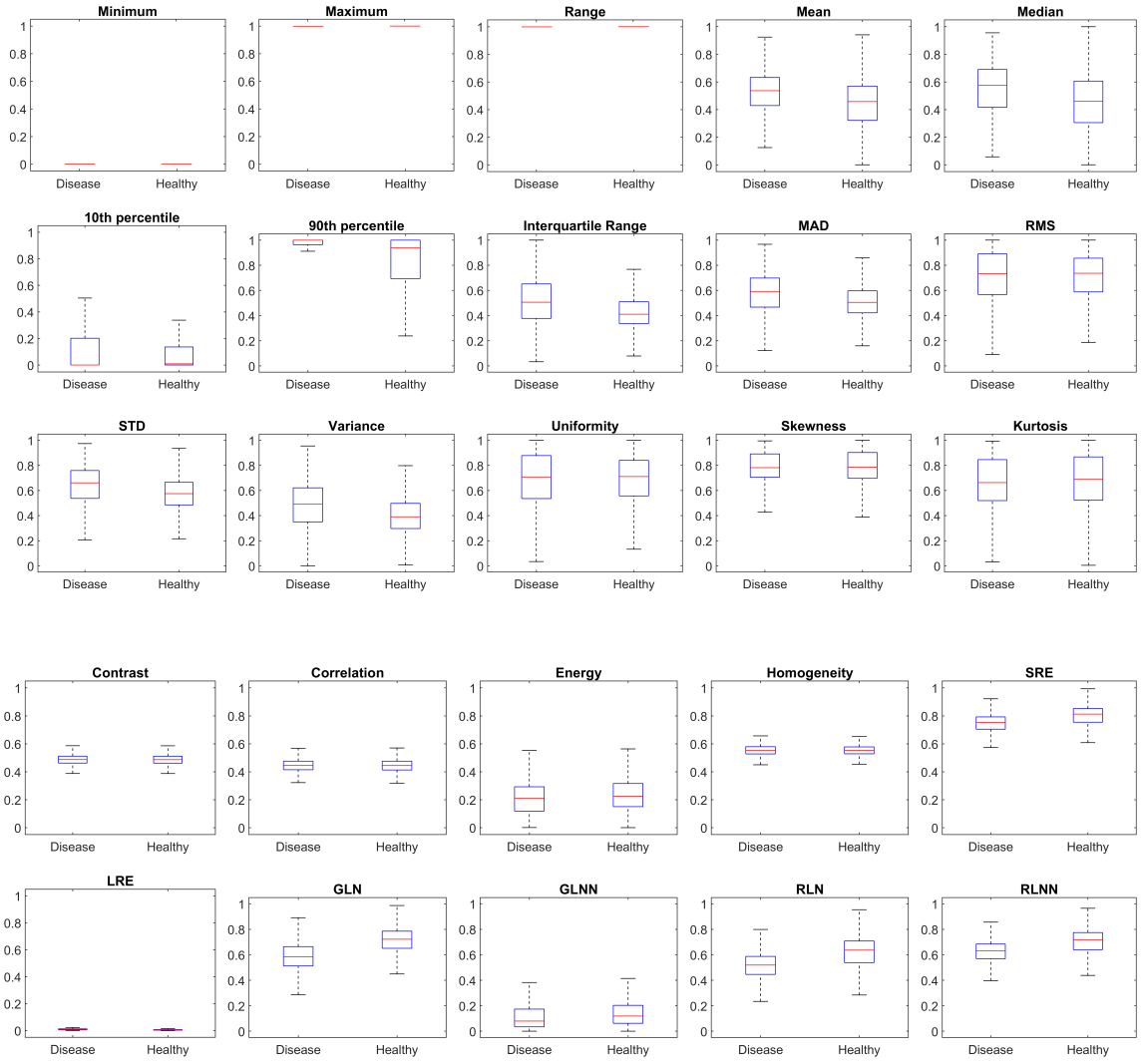
## 6.2 Possibili sviluppi futuri

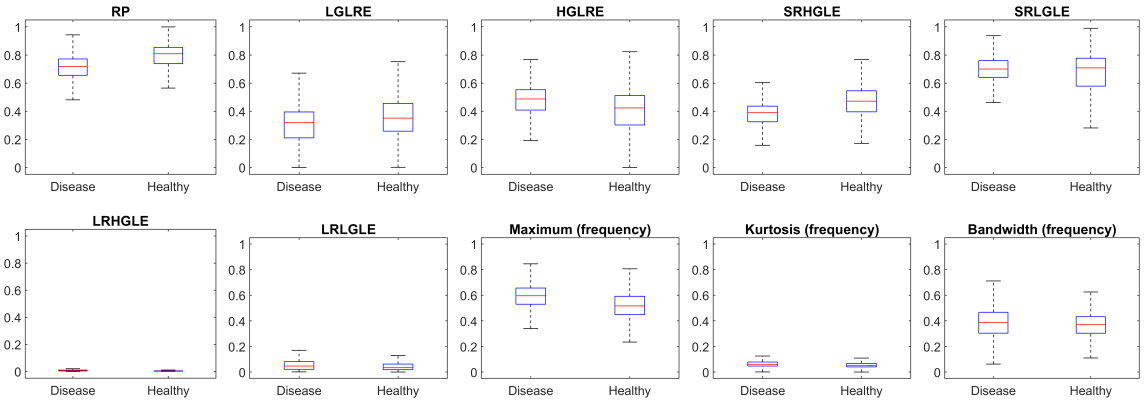
Sono stati identificati alcuni possibili sviluppi del lavoro, in ottica di miglioramento del classificatore e per fornire nuovi spunti per progetti futuri.
L'algoritmo di classificazione ha ampi margini di miglioramento. Si potrebbe aumentare il numero delle caratteristiche estratte, come ad esempio aggiungendo ca-

ratteristiche basate sul colore. Un numero maggiore di *feature*, potrebbe essere poi gestito con una adeguata selezione delle caratteristiche, con approcci lineari e non.
Si potrebbero inoltre esplorare altri tipi di classificatori, diversi dalle foreste casuali, come l'apprendimento profondo o *deep learning*. Uno strumento molto potente e versatile sono sicuramente le reti neurali convoluzionali, le cosiddette CNN. Sempre più spesso vengono impiegate nella diagnostica per immagini, ottenendo buoni risultati. Il sistema per diagnosi assistita implementato potrebbe essere la base di partenza per applicazioni ben più complesse.
Occorrerebbe migliorare la realizzazione delle maschere di segmentazione, che risultano troppo legate alla forma quadrangolare delle ROI. Sarebbe quindi conveniente implementare una segmentazione basata su molteplici fattori.
Sarebbe opportuno sviluppare un metodo più intelligente per sviluppare l'indicazione semaforica, non fondato sulla sola estensione dell'anomalia ma che prenda in considerazione più aspetti della lezione, come forma, intensità, posizione anatomica. Inoltre una scala basata su 4 colori è sicuramente intuitiva ma non troppo adeguata all'ambito sanitario. Sarebbe più idoneo assegnare un punteggio complessivo all'intera immagine endoscopica, fornendo una indicazione della probabilità di presenza di lesioni metaplastiche.
Si potrebbe infine sviluppare un sistema di analisi di endoscopia video in *real time*, che fornisca al medico un'analisi immediata del tessuto gastrico in esame. Questo ridurrebbe fortemente i tempi di diagnosi della patologia, e offrirebbe al contempo un supporto diagnostico più efficace agli operatori sanitari.

# 7. Appendice

Diagrammi a scatola e baffi delle caratteristiche estratte






L'apprendimento automatico a vettori di supporto (SVM)

La definizione classica di SVM è quella di un classificatore a margine massimo, cioè un classificatore la cui funzione decisionale è un iperpiano che separa al massimo i campioni di classi diverse [113]. Un SVM esegue la classificazione costruendo un iperpiano in dimensioni superiori. Un piano decisionale è quello che distingue e separa un gruppo di dati di un tipo, da un altro [114].
La SVM ricerca quei punti vettoriali, indicati come vettore di supporto, che definiscono il confine decisionale e danno la grande separazione marginale tra le classi, e che quindi separa le classi con distanza marginale massima nel piano decisionale [114].
La linea centrale rappresenta l'iperpiano a margine massimo. In altre parole, si selezionerebbe la linea di confine che separa le due classi a una distanza massima dal punto dati più vicino [114]. Le macchine a vettori di supporto sono dei modelli di apprendimento supervisionato, utilizzati spesso per compiti di classificazione binaria, ossia dove vi sono solo due possibili esiti [91].
In questo caso il margine è la più piccola distanza perpendicolare al punto dati dall'iperpiano e, quello marginale massimo, è il piano di decisione in cui il margine è maggiore. La SVM seleziona proprio il margine massimo che separa l'iperpiano [114].
Durante la classificazione può capitare che alcuni dati si trovino nella zona sbagliata della separazione, per risolvere questo tipo di casi, l'algoritmo SVM ha introdotto il nuovo termine noto come *soft margin*, appunto margine leggero. Mentre nel caso in cui i dati presentino rumore e valori anomali questo problema viene risolto utilizzando il concetto di variabili di tipo *slack* [114].
Un potente strumento utilizzato dalle SVM sono le funzione di kernel. Esse sono un metodo matematico utilizzato per la mappatura non lineare per dati di dimensioni superiori [114].
Attraverso questo strumento è possibile risolvere la classificazioni dimensionali superiori, calcolando il valore del prodotto scalare mappato dei dati, nello spazio delle caratteristiche. Il vantaggio della funzione del kernel è che la complessità del problema dipende solo dalla dimensionalità dello spazio di input piuttosto che dallo spazio delle caratteristiche [114].
La funzione built-in di MATLAB che addestra un classificatore binario SVM è chiamata *fitcsvm*. Essa supporta la mappatura dei dati del predittore utilizzando le fun-

zioni del kernel e supporta l'ottimizzazione minima sequenziale, meglio conisciuta come SMO (acronimo inglese di *sequential minimal optimization*), l'algoritmo iterativo ISDA (acronimo inglese di *iterative single data algorithm*) o la minimizzazione del margine leggero $L1$, tramite la programmazione quadratica per la minimizzazione della funzione obiettivo [115].
Questa funzione di MATLAB ricerca un iperpiano ottimale che separi i dati in due classi. Per fare ciò massimizza un margine, ossia uno spazio che non contiene alcuna osservazione, creando confini per le classi positive e negative [115]. Per le classi inseparabili, l'obiettivo è lo stesso, ma l'algoritmo impone una penalità sulla lunghezza del margine per ogni osservazione che si trova dalla parte sbagliata del suo confine di classe [115]. Essa utilizza la seguente funzione di punteggio:

$$f(x) = x'\beta + b$$

dove $x$ è l'osservazione, Il vettore $\beta$ contiene i coefficienti che definiscono un vettore ortogonale all'iperpiano dove, per dati linearmente separabili, la lunghezza ottimale del margine è $2/\left\|\beta\right\|$. La variabile $b$ rappresenta invece il termine di polarizzazione [115].
La radice di $f(x)$ definisce un iperpiano per specifici coefficienti, uno specifico iperpiano $f(z)$ è la distanza dal punto $z$ all'iperpiano. L'algoritmo cerca la lunghezza massima del margine, mantenendo separate le osservazioni nelle classi positive $(y = 1)$ e negative $(y = -1)$ [115].
Per le classi separabili, l'obiettivo è quello di minimizzare $\left\|\beta\right\|$ rispetto a $\beta$ e $b$, soggetti a $y_j f(x_j) \geq 1$, per ogni $j = 1, .., n$ [116].
Per le classi inseparabili, l'algoritmo utilizza le variabili *slack* note come $\xi_j$ che consentono la classificazione errata di qualche punto. Esse penalizzano la funzione obiettivo per le osservazioni che attraversano il limite del margine per la loro classe [116]. Nel dettaglio assumono valore $\xi_j = 0$ per le osservazioni che non attraversano il limite del margine per la loro classe, altrimenti $\xi_j \geq 0$ [115], [117].
In questo caso l'obiettivo è minimizzare $0.5\left\|\beta\right\|^2 + C\sum \xi_j$ rispetto a $\beta$, $b$ e $\xi_j$, soggetti a $y_j f(x_j) \geq 1 - \xi_j$ e $\xi_j \geq 0$ per ogni $j = 1, .., n$, e per un vincolo scalare positivo C, quest'ultimo è un parametro che controlla la penalità massima imposta alle osservazioni che violano i margini, il che aiuta a prevenire l'overfitting [115].
L'algoritmo di MATLAB utilizza il metodo dei moltiplicatori di Lagrange per ottimizzare l'obiettivo, che introduce $n$ coefficienti $\alpha_1, ..., \alpha_n$ [115], e che per classi separabili minimizza la seguente funzione:

$$0.5\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\alpha_k y_j y_k x_j' x_k - \sum_{j=1}^{n}\alpha_j$$

rispetto a $\alpha_1, ..., \alpha_n$ soggetto a $\sum \alpha j y j = 0$ con $\alpha \geq 0$ per ogni $y = 1, ..., n$ e sotto le condizioni di complementarietà di Karush-Kuhn-Tucker, meglio note come condizioni KKT [115]. Esse sono vincoli di ottimizzazione richiesti per soluzioni ottimali di programmazione non lineare, e nel caso dell'SVM queste sono:

$$\begin{cases} \alpha_j[y_j f(x_j) - 1 + \xi_j] = 0 \\ \xi_j(C - \alpha_j) = 0 \end{cases}$$

Per ogni $j = 1, ..., n$, dove $f(x_j) = \phi(x_j)'\beta + b$. Dove $\phi$ è la funzione di kernel e $\xi_j$ è la variabile *slack* [117]. Se le classi sono perfettamente separabili, allora $\xi_j = 0$ per ogni $j = 1, ..., n$ [115].
Per le classi inseparabili, l'obiettivo è lo stesso delle classi separabili, ad eccezione della condizione aggiuntiva $0 \leq \alpha_j \leq C$ per ogni $j = 1, .., n$. La funzione obiettivo risultante è la seguente:

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\alpha}_j y_j x' x_j + \widehat{b}$$

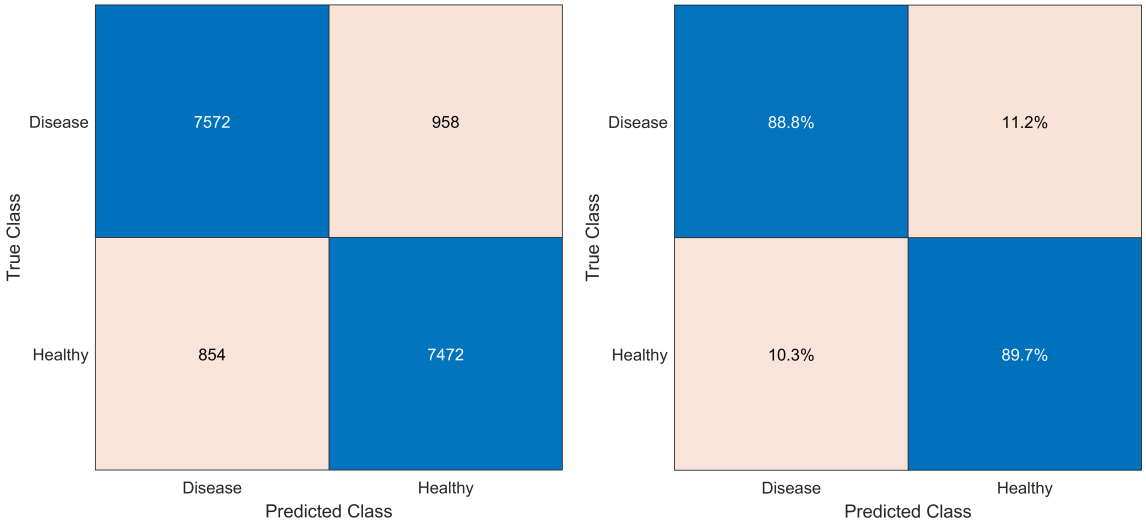
dove $\widehat{b}$ è la stima del bias e $\widehat{\alpha}_j$ è la j-esima stima del vettore $\widehat{\alpha}_j = 1, ..., n$ [118]. Scritta in questo modo, la funzione obiettivo è libera dalla stima di $\beta$ come risultato della formalizzazione primaria [115].
L'algoritmo SVM classifica una nuova osservazione $z$ usando $sign(\widehat{f}_z)$. In alcuni casi le classi sono separate da un limite non lineare, in questo caso la SVM funziona in uno spazio predittivo trasformato per trovare un iperpiano di separazione ottimale [115].
Nel caso in cui le classi siano separate da un limite non lineare, la SVM funziona in uno spazio predittivo trasformato per trovare un iperpiano di separazione ottimale. In quel caso la funzione obiettivo diventa:

$$0.5 \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k G(x_j, x_k) - \sum_{j=1}^{n} \alpha_j$$

rispetto ad $\alpha_1, ...\alpha_n$ soggetto a $\sum \alpha_j y_j = 0$ e $0 \leq \alpha_j \leq C$ per ogni $j = 1, ..., n$ e sotto le condizioni di complementarietà di KKT [115]. $G(x_j, x_k)$ sono invece gli elementi della matrice di Gram, essa è un insieme di $n$ vettori $\{x_1, ..., x_n; x_j \in R^p\}$ [118]. Consiste in una matrice $n \times n$ di elementi $(j, k)$ e definita come $G(x_j, x_k) = \langle \phi x_j, \phi x_k \rangle$, un prodotto scalare dei predittori, trasformati utilizzando la funzione kernel $\phi$ [115].
Nel caso di SVM non lineare, l'algoritmo forma una matrice di Gram utilizzando le righe dei predittori di $X$. La doppia formalizzazione sostituisce il prodotto interno delle osservazioni in $X$ con gli elementi corrispondenti della matrice di Gram risultante [117]. Di conseguenza, la SVM non lineare, per trovare un iperpiano di separazione, opera nello spazio predittore trasformato [115].
Quindi nel caso di limite non lineare la funzione obiettivo diventa la seguente:

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\alpha}_j y_j G(x, x_j) + \widehat{b}$$

Matrici di confusione ottenute dalla classificazione del test set




Matrici di confusione ottenute dopo l'algoritmo di pulizia delle maschere

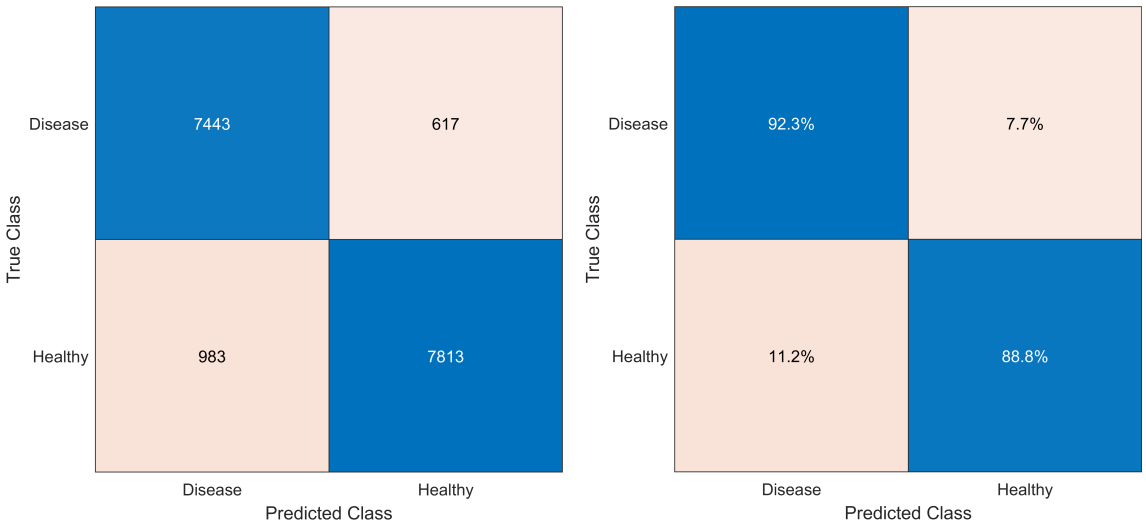


Immagine sana classificata correttamente (verde)

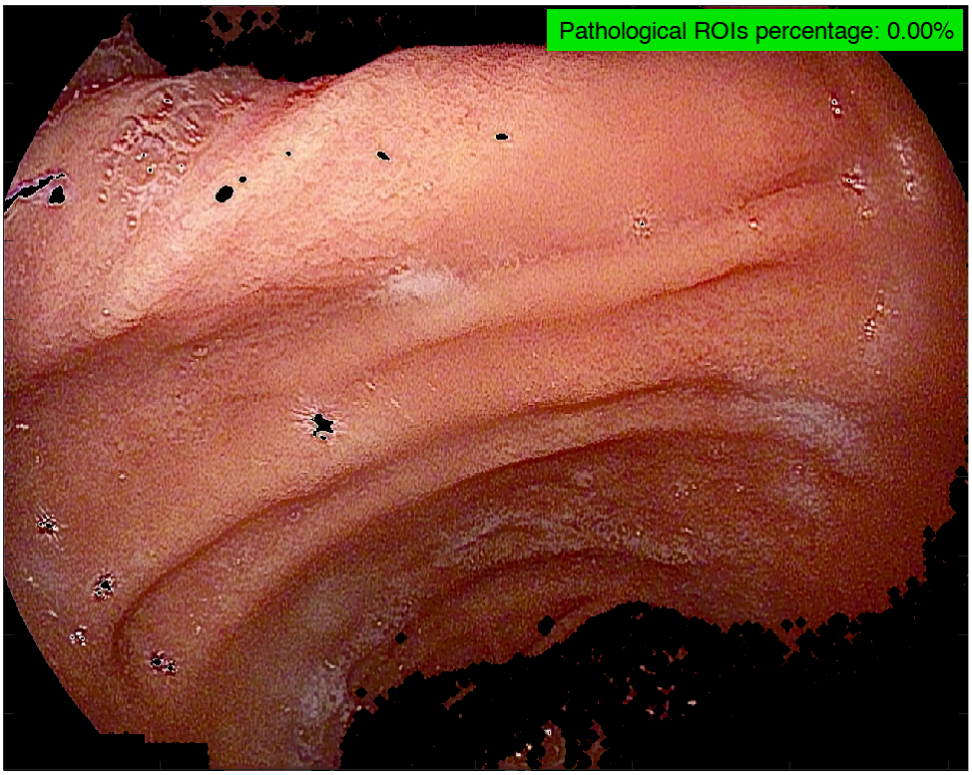



Immagine sana classificata non correttamente (giallo)

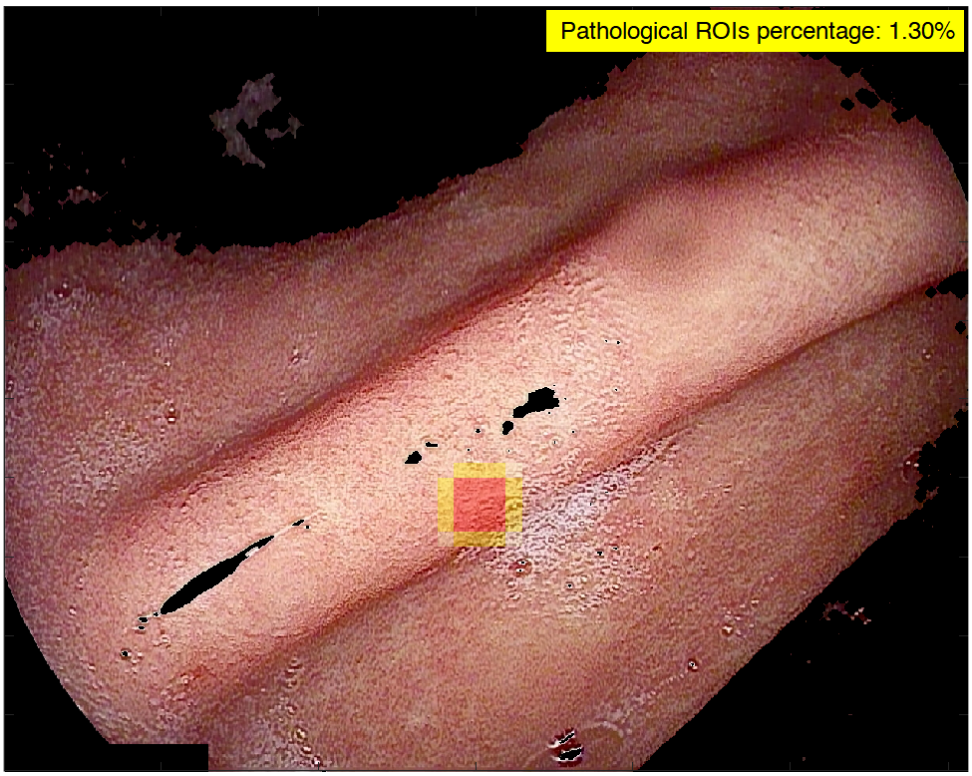


Immagine patologica classificata correttamente (arancione)

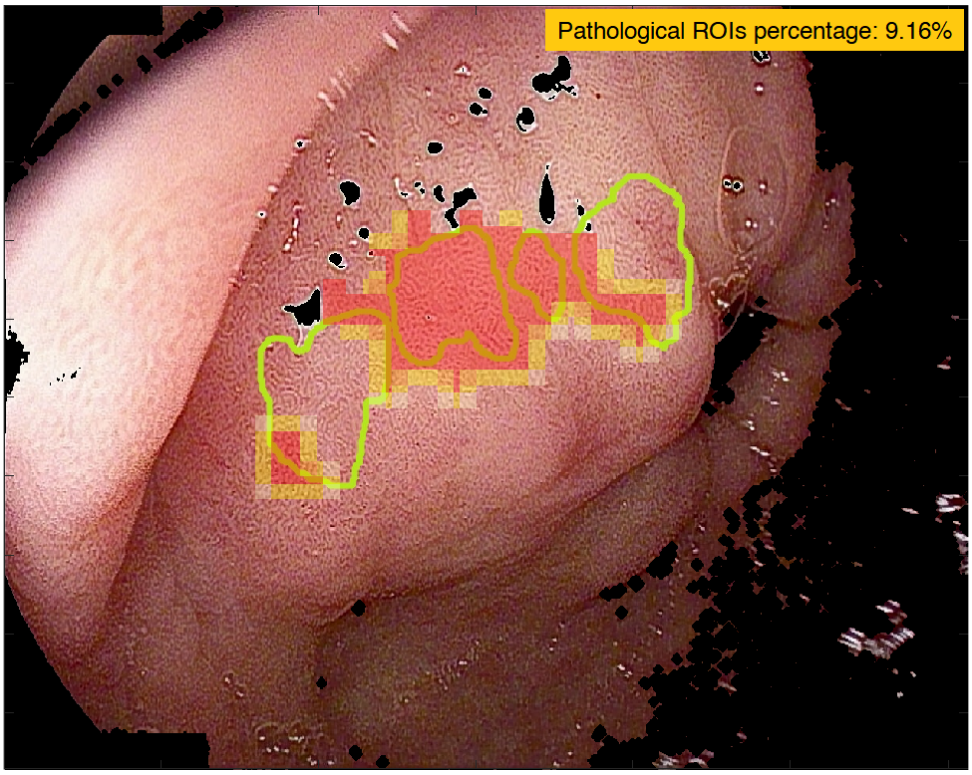
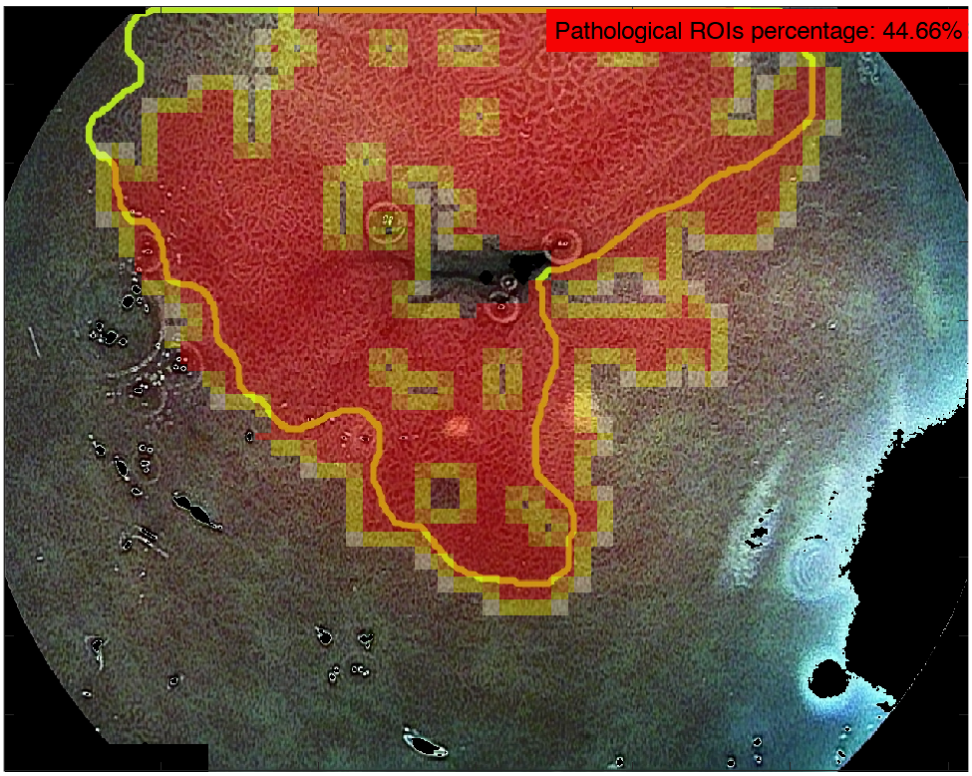



Immagine patologica classificata correttamente (rosso)



# 8. Bibliografia


[1] Claudius Galen. *On the natural faculties*. Dalcassian Publishing Company, 2019.

[2] Anton Sebastian. *A dictionary of the history of medicine*. CRC Press, 2018.

[3] Giuseppe Anastasi et al. *Trattato di anatomia umana sistematica e funzionale - Volume secondo*. Edi.Ermes s.r.l, Milano, 2020.

[4] Vincenzo Mezzogiorno e Antonio Mezzogiorno. *Compendio di anatomia umana*. Piccin Nuova Libraria S.p.A., Padova, 1994.

[5] Giulia d'Amati e Carlo Della Rocca. «Anatomia Patologica. La sistematica - II ed.» In: Edra. 2018.

[6] Jacques Ferlay et al. «Cancer statistics for the year 2020: An overview». In: *International journal of cancer* 149.4 (2021), pp. 778–789.

[7] J Ferlay et al. «Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018». In: *European journal of cancer* 103 (2018), pp. 356–387.

[8] Milena Ilic e Irena Ilic. «Epidemiology of stomach cancer». In: *World Journal of Gastroenterology* 28.12 (2022), p. 1187.

[9] The International Agency for Research on Cancer (IARC). *Global Cancer Observatory — gco.iarc.fr*. `https://gco.iarc.fr/`. [Accessed 21-Aug-2022].

[10] Jalal Poorolajal et al. «Risk factors for stomach cancer: a systematic review and meta-analysis». In: *Epidemiology and health* 42 (2020).

[11] Shailja C Shah et al. «Histologic subtyping of gastric intestinal metaplasia: overview and considerations for clinical practice». In: *Gastroenterology* 158.3 (2020), pp. 745–750.

[12] Kouji Banno et al. «Carcinogenic mechanisms of endometrial cancer: involvement of genetics and epigenetics». In: *Journal of Obstetrics and Gynaecology Research* 40.8 (2014), pp. 1957–1967.

[13] Eric Van Cutsem et al. «Gastric cancer». In: *The Lancet* 388.10060 (nov. 2016), pp. 2654–2664. DOI: `10.1016/s0140-6736(16)30354-3`. URL: `https://doi.org/10.1016/s0140-6736(16)30354-3`.

[14] Giuseppe Mario Pontieri. *Elementi di patologia generale*. Piccin Nuova Libraria S.p.A., Padova, 2018.

[15] Xinqi He et al. «Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter diagnostic study (with videos)». In: *Gastrointestinal Endoscopy* 95.4 (2022), pp. 671–678.

[16] JMW Slack. «Epithelial metaplasia and the second anatomy». In: *The Lancet* 328.8501 (1986), pp. 268–271.

[17] Veronique Giroux e Anil K Rustgi. «Metaplasia: tissue injury adaptation and a precursor to the dysplasia–cancer sequence». In: *Nature Reviews Cancer* 17.10 (2017), pp. 594–604.

[18] Stuart Jon Spechler e Raj K Goyal. «Barrett's esophagus». In: *New England Journal of Medicine* 315.6 (1986), pp. 362–371.

[19] Stuart Jon Spechler. «Barrett's esophagus». In: *Principles of Deglutition* (2013), pp. 723–738.

[20] Stuart Jon Spechler. «Barrett's esophagus». In: *New England Journal of Medicine* 346.11 (2002), pp. 836–842.

[21] Jonathan R White e Matthew Banks. «Identifying the pre-malignant stomach: from guidelines to practice». In: *Translational Gastroenterology and Hepatology* 7 (2022).

[22] William K Hirota et al. «Specialized intestinal metaplasia, dysplasia, and cancer of the esophagus and esophagogastric junction: prevalence and clinical data». In: *Gastroenterology* 116.2 (1999), pp. 277–285.

[23] Pelayo Correa, M Blanca Piazuelo e Keith T Wilson. «Pathology of gastric intestinal metaplasia: clinical implications». In: *The American journal of gastroenterology* 105.3 (2010), p. 493.

[24] WK Leung e JJY Sung. «Intestinal metaplasia and gastric carcinogenesis». In: *Alimentary pharmacology & therapeutics* 16.7 (2002), pp. 1209–1216.

[25] Wojciech Marlicz, Xuyang Ren e Alexander et al. Robertson. «Frontiers of robotic gastroscopy: a comprehensive review of robotic gastroscopes and technologies». In: *Cancers* 12.10 (2020), p. 2775.

[26] Nobuhiro Fukuta et al. «Endoscopic diagnosis of gastric intestinal metaplasia: a prospective multicenter study». In: *Digestive Endoscopy* 25.5 (2013), pp. 526–534.

[27] Yaqiong Zhang et al. «Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence». In: *Digestive and Liver Disease* 52.5 (2020), pp. 566–572.

[28] Shyam Menon e Nigel Trudgill. «How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis». In: *Endoscopy international open* 2.02 (2014), E46–E50.

[29] Jorge Lage et al. «Light-NBI to identify high-risk phenotypes for gastric adenocarcinoma: do we still need biopsies?» In: *Scandinavian journal of gastroenterology* 51.4 (2016), pp. 501–506.

[30] Pedro Pimentel-Nunes et al. «A multicenter prospective study of the real-time use of narrow-band imaging in the diagnosis of premalignant gastric conditions and lesions». In: *Endoscopy* 48.08 (2016), pp. 723–730.

[31] Abd El-Rahiem Ahmed et al. «Enhanced endoscopy for the diagnosis of gastricantralintestinal metaplasia». In: *Al-Azhar Medical Journal* 50.1 (2021), pp. 643–654.

[32] Gianluca Esposito et al. «Endoscopic grading of gastric intestinal metaplasia (EGGIM): a multicenter validation study». In: *Endoscopy* 51.06 (2019), pp. 515–521.

[33] Hirotaka Nakashima et al. «Artificial intelligence diagnosis of Helicobacter pylori infection using blue laser imaging-bright and linked color imaging: a single-center prospective study». In: *Annals of gastroenterology* 31.4 (2018), p. 462.

[34] Lianlian Wu et al. «Real-time artificial intelligence for detecting focal lesions and diagnosing neoplasms of the stomach by white-light endoscopy (with videos)». In: *Gastrointestinal Endoscopy* 95.2 (2022), pp. 269–280.

[35] Tao Yan et al. «Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images». In: *Computers in Biology and Medicine* 126 (2020), p. 104026.

[36] Ming Xu et al. «Artificial intelligence in the diagnosis of gastric precancerous conditions by image-enhanced endoscopy: a multicenter, diagnostic study (with video)». In: *Gastrointestinal Endoscopy* 94.3 (2021), pp. 540–548.

[37] Huiyan Luo et al. «Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study». In: *The Lancet Oncology* 20.12 (2019), pp. 1645–1654.

[38] Albert J de Groof et al. «Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking». In: *Gastroenterology* 158.4 (2020), pp. 915–929.

[39] Helmut Neumann et al. «Learning curve of virtual chromoendoscopy for the prediction of hyperplastic and adenomatous colorectal lesions: a prospective 2-center study». In: *Gastrointestinal Endoscopy* 78.1 (2013), pp. 115–120.

[40] Song Yuheng e Yan Hao. «Image segmentation algorithms overview». In: *arXiv preprint arXiv:1707.02051* (2017).

[41] Henk JAM Heijmans e Christian Ronse. «The algebraic basis of mathematical morphology I. Dilations and erosions». In: *Computer Vision, Graphics, and Image Processing* 50.3 (1990), pp. 245–295.

[42] Uğur Erkan, Levent Gökrem e Serdar Enginoğlu. «Different applied median filter in salt and pepper noise». In: *Computers & Electrical Engineering* 70 (2018), pp. 789–798.

[43] Gaurav Kumar e Pradeep Kumar Bhatia. «A detailed review of feature extraction in image processing systems». In: *2014 Fourth international conference on advanced computing & communication technologies*. IEEE. 2014, pp. 5–12.

[44] Ahmed Shihab Ahmed. «Comparative study among Sobel, Prewitt and Canny edge detection operators used in image processing». In: *J. Theor. Appl. Inf. Technol* 96.19 (2018), pp. 6517–6525.

[45] Namita Aggarwal e RK Agrawal. «First and second order statistics features for classification of magnetic resonance brain images». In: (2012).

[46] Hiroshi Konno e Tomoyuki Koshizuka. «Mean-absolute deviation model». In: *Iie Transactions* 37.10 (2005), pp. 893–900.

[47] Ummul Khair et al. «Forecasting error calculation with mean absolute deviation and mean absolute percentage error». In: *Journal of Physics: Conference Series*. Vol. 930. 1. IOP Publishing. 2017, p. 012002.

[48] Marco Ravanelli et al. «Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: prediction of the response to the first-line chemotherapy». In: *European radiology* 23 (2013), pp. 3450–3455.

[49] Kevin P Balanda e HL MacGillivray. «Kurtosis: a critical review». In: *The American Statistician* 42.2 (1988), pp. 111–119.

[50] Abdulrahman Al-Janobi. «Performance evaluation of cross-diagonal texture matrix method of texture analysis». In: *Pattern Recognition* 34.1 (2001), pp. 171–180.

[51] Igor Pantic, Senka Pantic e Gordana Basta-Jovanovic. «Gray level co-occurrence matrix texture analysis of germinal center light zone lymphocyte nuclei: physiology viewpoint with focus on apoptosis». In: *Microscopy and Microanalysis* 18.3 (2012), pp. 470–475.

[52] Robert M Haralick, Karthikeyan Shanmugam e Its' Hak Dinstein. «Textural features for image classification». In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

[53] *Create gray-level co-occurrence matrix from image - MATLAB graycomatrix — mathworks.com*. `https://www.mathworks.com/help/images/ref/graycomatrix.html`. [Accessed 27-Nov-2022].

[54] *Properties of gray-level co-occurrence matrix - MATLAB graycoprops — mathworks.com*. `https://www.mathworks.com/help/images/ref/graycoprops.html`. [Accessed 20-Nov-2022].

[55] Mary M Galloway. «Texture analysis using gray level run lengths». In: *Computer graphics and image processing* 4.2 (1975), pp. 172–179.

[56] Xiaoou Tang. «Texture information in run-length matrices». In: *IEEE transactions on image processing* 7.11 (1998), pp. 1602–1609.

[57] Deepalakshmi Balasubramanian, Poonguzhali Srinivasan e Ravindran Gurupatham. «Automatic classification of focal lesions in ultrasound liver images using principal component analysis and neural networks». In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2007, pp. 2134–2137.

[58] A Chu, Chandra M Sehgal e James F Greenleaf. «Use of gray value distribution of run lengths for texture analysis». In: *Pattern recognition letters* 11.6 (1990), pp. 415–419.

[59] Belur V Dasarathy e Edwin B Holder. «Image characterizations based on joint gray level—run length distributions». In: *Pattern Recognition Letters* 12.8 (1991), pp. 497–502.

[60] Wilhelm Burger e Mark J Burge. *Digital Image Processing*. 3ª ed. Texts in computer science. Cham, Switzerland: Springer International Publishing, lug. 2022.

[61] *2-D fast Fourier transform - MATLAB fft2 — mathworks.com*. `https://www.mathworks.com/help/matlab/ref/fft2.html`. [Accessed 18-Dec-2022].

[62] Hussein Ahmad et al. «Harmonic components of leakage current as a diagnostic tool to study the aging of insulators». In: *Journal of Electrostatics* 66.3-4 (2008), pp. 156–164.

[63] *Shift zero-frequency component to center of spectrum - MATLAB fftshift — mathworks.com*. `https://www.mathworks.com/help/matlab/ref/fftshift.html`. [Accessed 18-Dec-2022].

[64] *Power bandwidth - MATLAB powerbw — mathworks.com*. `https://www.mathworks.com/help/signal/ref/powerbw.html`. [Accessed 22-Dec-2022].

[65] Rizgar Zebari et al. «A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction». In: *Journal of Applied Science and Technology Trends* 1.2 (2020), pp. 56–70.

[66] Yoav Benjamini. «Opening the box of a boxplot». In: *The American Statistician* 42.4 (1988), pp. 257–262.

[67] Gareth James et al. *An introduction to statistical learning*. en. 1ª ed. Springer texts in statistics. New York, NY: Springer, giu. 2013.

[68] Linda A Clark e Daryl Pregibon. «Tree-based models». In: *Statistical models in S*. Routledge, 2017, pp. 377–419.

[69] Bogumił Kamiński, Michał Jakubczyk e Przemysław Szufel. «A framework for sensitivity analysis of decision trees». In: *Central European journal of operations research* 26 (2018), pp. 135–159.

[70] J. Ross Quinlan. «Simplifying decision trees». In: *International journal of man-machine studies* 27.3 (1987), pp. 221–234.

[71] Shai Shalev-Shwartz e Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014. ISBN: 978-1-107-05713-5.

[72] Lior Rokach e Oded Maimon. «Top-down induction of decision trees classifiers-a survey». In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35.4 (2005), pp. 476–487.

[73] B Chandra e P Paul Varghese. «Fuzzifying Gini Index based decision trees». In: *Expert Systems with Applications* 36.4 (2009), pp. 8549–8559.

[74] Sebastian Nowozin. «Improved information gain estimates for decision tree induction». In: *arXiv preprint arXiv:1206.4620* (2012).

[75] Thomas G Dietterich e Eun Bae Kong. *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Rapp. tecn. Citeseer, 1995.

[76] John Mingers. «An empirical comparison of selection measures for decision-tree induction». In: *Machine learning* 3 (1989), pp. 319–342.

[77] Abhineet Agarwal et al. «Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods». In: *arXiv preprint arXiv:2202.00858* (2022).

[78] Jerome H Friedman. «Stochastic gradient boosting». In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[79] Robert E Schapire. «The boosting approach to machine learning: An overview». In: *Nonlinear estimation and classification* (2003), pp. 149–171.

[80] Jane Elith, John R Leathwick e Trevor Hastie. «A working guide to boosted regression trees». In: *Journal of animal ecology* 77.4 (2008), pp. 802–813.

[81] Clifton D Sutton. «Classification and regression trees, bagging, and boosting». In: *Handbook of statistics* 24 (2005), pp. 303–329.

[82] Leo Breiman. «Bagging predictors». In: *Machine learning* 24 (1996), pp. 123–140.

[83] Robert J Tibshirani e Bradley Efron. «An introduction to the bootstrap». In: *Monographs on statistics and applied probability* 57.1 (1993).

[84] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[85] Gérard Biau e Erwan Scornet. «A random forest guided tour». In: *Test* 25 (2016), pp. 197–227.

[86] Leo Breiman. «Random forests». In: *Machine learning* 45 (2001), pp. 5–32.

[87] Mario Molina e Filiz Garip. «Machine learning for sociology». In: *Annual Review of Sociology* 45 (2019), pp. 27–45.

[88] Erwan Scornet, Gérard Biau e Jean-Philippe Vert. «Consistency of random forests». In: (2015).

[89] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[90] Andriy Burkov. *The hundred-page machine learning book*. Vol. 1. Andriy Burkov Quebec City, QC, Canada, 2019.

[91] Marc Peter Deisenroth, A. Aldo Faisal e Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge: Cambridge University Press, 2020. ISBN: 978-1-108-47004-9.

[92] *Train models to classify data using supervised machine learning - MATLAB*. https://www.mathworks.com/help/stats/classificationlearner-app.html. [Accessed 11-Feb-2023].

[93] *Bayesian Optimization Algorithm - MATLAB; Simulink — mathworks.com*. https://www.mathworks.com/help/stats/bayesian-optimization-algorithm.html. [Accessed 10-Mar-2023].

[94] Martin Ester et al. «A density-based algorithm for discovering clusters in large spatial databases with noise.» In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[95] Li Meng'Ao et al. «Research and improvement of DBSCAN cluster algorithm». In: *2015 7th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE. 2015, pp. 537–540.

[96] Slava Kisilevich, Florian Mansmann e Daniel Keim. «P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos». In: *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*. 2010, pp. 1–4.

[97] Marzena Kryszkiewicz e Piotr Lasek. «TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality». In: *Rough Sets and Current Trends in Computing: 7th International Conference, RSCTC 2010, Warsaw, Poland, June 28-30, 2010. Proceedings 7*. Springer. 2010, pp. 60–69.

[98] *Density-based spatial clustering of applications with noise (DBSCAN) - MATLAB dbscan — mathworks.com*. https://www.mathworks.com/help/stats/dbscan.html. [Accessed 15-Mar-2023].

[99] Leland Wilkinson e Michael Friendly. «The history of the cluster heat map». In: *The American Statistician* 63.2 (2009), pp. 179–184.

[100] Gürol Canbek et al. «Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights». In: *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE. 2017, pp. 821–826.

[101] Miho Ohsaki et al. «Confusion-matrix-based kernel logistic regression for imbalanced data classification». In: *IEEE Transactions on Knowledge and Data Engineering* 29.9 (2017), pp. 1806–1819.

[102] Margherita Grandini, Enrico Bagli e Giorgio Visani. «Metrics for multi-class classification: an overview». In: *arXiv preprint arXiv:2008.05756* (2020).

[103] Anthony K Akobeng. «Understanding diagnostic tests 3: receiver operating characteristic curves». In: *Acta paediatrica* 96.5 (2007), pp. 644–647.

[104] Ana-Maria Šimundić. «Measures of diagnostic accuracy: basic definitions». In: *ejifcc* 19.4 (2009), p. 203.

[105] *Visualize and Assess Classifier Performance in Classification Learner - MATLAB amp; Simulink — mathworks.com*. https://www.mathworks.com/help/stats/assess-classifier-performance.html. [Accessed 17-Apr-2023].

[106] Mary L McHugh. «Interrater reliability: the kappa statistic». In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[107] Davide Chicco, Matthijs J Warrens e Giuseppe Jurman. «The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment». In: *IEEE Access* 9 (2021), pp. 78368–78381.

[108] Lee R Dice. «Measures of the amount of ecologic association between species». In: *Ecology* 26.3 (1945), pp. 297–302.

[109] Laurie Butgereit e Reinhardt A Botha. «A comparison of different calculations for N-gram similarities in a spelling corrector for mobile instant messaging language». In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. 2013, pp. 1–7.

[110] Aaron Carass et al. «Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis». In: *Scientific reports* 10.1 (2020), p. 8242.

[111] Xiaoya Li et al. «Dice loss for data-imbalanced NLP tasks». In: *arXiv preprint arXiv:1911.02855* (2019).

[112] Ne Lin et al. «Simultaneous recognition of atrophic gastritis and intestinal metaplasia on white light endoscopic images based on convolutional neural networks: a multicenter study». In: *Clinical and translational gastroenterology* 12.8 (2021).

[113] Sancho Salcedo-Sanz et al. «Support vector machines in engineering: an overview». In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.3 (2014), pp. 234–267.

[114] Mayank Arya Chandra e SS Bedi. «Survey on SVM and their application in image classification». In: *International Journal of Information Technology* 13 (2021), pp. 1–11.

[115] *Train support vector machine (SVM) classifier for one-class and binary classification - MATLAB fitcsvm — mathworks.com*. `https://www.mathworks.com/help/stats/fitcsvm.html`. [Accessed 29-Jan-2023].

[116] John Paul Mueller e Luca Massaron. *Machine Learning For Dummies*. New York: John Wiley  Sons, 2016. ISBN: 978-1-119-24577-3.

[117] Mehryar Mohri, Afshin Rostamizadeh e Ameet Talwalkar. *Foundations of Machine Learning*. Cambridge: MIT Press, 2012. ISBN: 978-0-262-30473-3.

[118] Richard E. Neapolitan e Xia Jiang. *Artificial Intelligence - With an Introduction to Machine Learning, Second Edition*. Boca Raton, Fla: CRC Press, 2018. ISBN: 978-1-351-38438-4.