UNIVERSITÀ
DI PAVIA

FACOLTA' DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

Sviluppo e validazione di un sistema di classificazione di microcalcificazioni mammarie tramite feature radiomiche e descrittori geometrici

Relatore: Prof. Riccardo Bellazzi

Correlatore: Dott. Alberto Luigi Malovini

Candidata: Tetta Federica

Matricola: 544989

A.A. 2024/2025

INDICE

INTRODUZIONE

Il tumore al seno rappresenta una delle neoplasie più diffuse e clinicamente rilevanti nella popolazione femminile, rendendo fondamentale il miglioramento degli strumenti di diagnosi precoce. In questo contesto, la mammografia costituisce tuttora l'esame di riferimento per lo screening, ma la sua interpretazione può risultare complessa, soprattutto in presenza di lesioni di piccole dimensioni e di caratteristiche poco evidenti. Tra queste, le microcalcificazioni mammarie (MC) rivestono un ruolo particolarmente rilevante, poiché possono rappresentare uno dei segni più precoci di neoplasia mammaria, ma allo stesso tempo risultano difficili da analizzare in modo affidabile mediante la sola ispezione visiva.

Negli ultimi anni, l'impiego di tecniche di intelligenza artificiale e machine learning nell'ambito dell'imaging mammografico ha mostrato un potenziale significativo nel supportare il radiologo nelle diverse fasi del processo diagnostico, dalla selezione delle regioni di interesse fino alla classificazione delle lesioni. In particolare, la possibilità di descrivere quantitativamente le MC attraverso feature offre l'opportunità di sviluppare modelli in grado di distinguere pattern benigni e maligni in modo riproducibile e interpretabile.

Pertanto, in questo contesto l'obiettivo del presente lavoro di tesi è lo sviluppo e la validazione di un sistema di classificazione automatica delle MC, basato sull'analisi di feature radiomiche e descrittori geometrici, in grado di distinguere le regioni di interesse (ROI) in benigne e maligne. Il progetto è stato svolto in collaborazione con la Breast Unit di Istituti Clinici Scientifici Maugeri IRCCS di Pavia.

A differenza di molti studi presenti in letteratura, che considerano congiuntamente diverse tipologie di lesioni mammarie, il presente lavoro si concentra esclusivamente sulle MC, che rappresentano un caso particolarmente complesso e clinicamente rilevante. Tale scelta consente di evitare una possibile sovrastima delle prestazioni dei modelli, dovuta alla minore difficoltà intrinseca nella classificazione di altre tipologie di lesioni, come le masse.

Un ulteriore aspetto centrale del lavoro riguarda l'interpretabilità del modello. In ambito clinico, infatti, non è sufficiente ottenere elevate prestazioni predittive, ma è necessario che le decisioni del sistema siano comprensibili e giustificabili dal punto di vista medico. Per questo motivo, è stata dedicata particolare attenzione all'analisi del contributo delle singole features mediante

tecniche di explainability, al fine di verificare la coerenza tra le decisioni del modello e la conoscenza clinica.

Il presente elaborato è organizzato in cinque capitoli, che descrivono in modo progressivo il contesto, la metodologia adottata e i risultati ottenuti.

Il primo capitolo introduce il problema clinico del tumore al seno, il ruolo della mammografia nella diagnosi precoce, le caratteristiche delle microcalcificazioni e i principali concetti delle tecniche di machine learning applicate alle immagini mammografiche.

Il secondo capitolo è dedicato all'analisi della letteratura. In particolare, vengono esaminati i principali approcci utilizzati per l'identificazione, la caratterizzazione e la classificazione delle MC in immagini mammografiche.

Il terzo capitolo descrive i materiali e i metodi impiegati per lo sviluppo della pipeline proposta. Nella prima parte vengono illustrate le tecniche utilizzate per l'identificazione delle MC e l'estrazione di feature radiomiche e geometriche. Nella seconda parte, invece, viene presentata la fase di classificazione con lo sviluppo di modelli per distinguere le ROI tra benigne e maligne.

Il quarto capitolo riporta i risultati sperimentali ottenuti dall'analisi delle immagini mammografiche raccolte presso gli Istituti Clinici Scientifici Maugeri IRCCS di Pavia. In particolare, vengono analizzate le prestazioni dei modelli sia in fase di training, sia sul test set al fine di valutarne la capacità di generalizzazione. Viene inoltre presentata un'analisi di interpretabilità, finalizzata a comprendere il contributo delle diverse features alle decisioni del modello.

Infine, il Capitolo 5 riassume le principali conclusioni del lavoro, discutendo i risultati ottenuti alla luce degli obiettivi iniziali e delineando possibili sviluppi futuri, con particolare attenzione all'integrazione di metodologie più avanzate e alla validazione in ambito clinico.

CAPITOLO 1: CONTESTO E BACKGROUND

1.1 Tumore al seno e importanza della diagnosi precoce

Il cancro al seno è uno dei tumori più comuni nelle donne. Sulla base dei dati GLOBOCAN, nel 2022 è stato il primo o il secondo tumore più comunemente diagnosticato in 183 dei 185 paesi analizzati, con circa 2,3 milioni di nuovi casi e 670.000 decessi. [1]

A livello globale, l'incidenza del cancro al seno ha colpito principalmente donne con età superiore ai 50 anni. La sua distribuzione a livello geografico non è uniforme ma è influenzata dalle condizioni economiche, sociali e sanitarie dei vari territori. In particolare, nei paesi ad alto reddito si osservano generalmente tassi di incidenza più elevati, mentre la mortalità risulta inferiore rispetto ai paesi a basso e medio reddito, dove la diagnosi è spesso tardiva. [1]

Una delle cause principali di queste disparità è legata alla differente offerta sanitaria. Infatti, nei paesi ad alto reddito i sistemi sanitari forniscono l'accesso a diagnosi tempestive e trattamenti di alta qualità. In questo contesto, la diagnosi precoce riveste un ruolo fondamentale, poiché consente di individuare la patologia nelle fasi inziali, migliorando significativamente la sopravvivenza. [1]

Se i tassi attuali rimarranno invariati, si stima che nel 2050 si verificheranno 3,2 milioni di nuovi casi e 1,1 milioni di decessi ad esso correlati, con un aumento rispettivamente del 38% e del 68% rispetto al 2022. [1]

Per garantire la riduzione dei tassi di mortalità sono stati attivati dall' Organizzazione Mondiale della Sanità (OMS) alcuni interventi focalizzati su aspetti come la prevenzione, la diagnosi precoce, la rapidità diagnostica e la qualità del percorso di cura. Il loro obiettivo è di ridurre ogni anno mediamente del 2,5% i tassi di mortalità, evitando un numero di decessi pari a 2,5 milioni tra il 2020 e il 2040. [1]

1.1.1 Mammografia come esame di riferimento nello screening

Nelle linee guida per lo screening del cancro al seno definite dall'Organizzazione Mondiale della Sanità-Agenzia Internazionale per la Ricerca sul Cancro (OMS-IARC), la mammografia è il metodo più comunemente utilizzato. [2] Ciò in parte è legato ai vantaggi che può fornire lo screening mammografico. Infatti, consente di ottenere: una diagnosi precoce di cancro al seno, causa meno effetti collaterali per la popolazione rispetto ad altre tecniche di imaging, fornisce risultati riproducibili e comporta costi sociali accettabili. [3]

Come strumento di screening, la mammografia viene utilizzata in donne sane asintomatiche per individuare i tumori il più precocemente possibile. [4] È consigliata principalmente nelle donne di età superiore ai 40 anni, in particolare in presenza di una storia familiare di cancro al seno, in quanto è in grado di rilevarlo efficacemente in fase iniziale. [5]

La mammografia è una modalità di imaging non invasiva che utilizza raggi X a basso dosaggio per generare immagini ad alta risoluzione del tessuto mammario. Il principio alla base della generazione delle immagini mammografiche è l'attenuazione differenziale dei raggi X. L'immagine generata è una proiezione bidimensionale del tessuto mammario. [5]

L'esame mammografico standard viene eseguito acquisendo immagini radiografiche da 2 angolazioni di ciascun seno, e uno o due radiologi esperti esaminano queste immagini alla ricerca di lesioni maligne mediante un'ispezione visiva. [6]

Numerosi studi hanno analizzato l'efficacia della mammografia come screening del cancro al seno. Tuttavia, la mammografia presenta dei limiti, come la ridotta sensibilità nelle donne con tessuto mammario denso, l'esposizione ripetuta alle radiazioni o la difficoltà nel rilevare tumori di piccole dimensioni. [5]

Per garantire il mantenimento e il miglioramento della qualità delle immagini mammografiche, e delle prestazioni interpretative dei radiologi sono state introdotte linee guida e sistemi di monitoraggio delle performance. La capacità di interpretazione delle mammografie da parte dei radiologi può essere misurata tramite diversi indicatori come: il tasso di rilevamento del cancro, il tasso di richiamo e il tasso di rilevamento del cancro invasivo. Affinché la capacità di interpretazione del radiologo sia considerata ottimale il valore di questi indicatori deve ricadere in un intervallo accettabile prestabilito. [4]

1.1.2 Limiti della lettura radiologica e motivazione dei sistemi CAD

Una volta che le mammografie sono state effettuate, per ricercare eventuali lesioni maligne queste immagini vengono esaminate da uno o più radiologi esperti.[6] Anche se già la lettura delle immagini mammografiche da parte di un unico radiologo è in grado di fornire buone prestazioni, con questo approccio è possibile che una parte dei tumori non venga individuata. Per questo tradizionalmente è buona pratica che un altro radiologo effettui una seconda lettura delle immagini, nota come doppia lettura, che è in grado di ridurre in modo significativo il numero di casi non individuati. [7]

Le prestazioni interpretative dei singoli radiologi, tuttavia, non sono uniformi e possono variare in funzione dell'esperienza, del volume di esami analizzati e per via di fattori fisici come la fatica.[4][7] Inoltre, l'accuratezza dello screening è influenzata anche da altri fattori indipendenti dall'operatore, come la densità mammaria, che può ridurre la sensibilità della mammografia. [4]

In questo contesto, l'intelligenza artificiale rappresenta un valido supporto al processo diagnostico. In particolare, i sistemi di rilevamento assistito da computer, noti come CAD (Computer-Aided Detection/Diagnosis) possono affiancare il radiologo nell'interpretazione delle immagini, fungendo da secondo lettore e contribuendo a ridurre gli errori di rilevamento e aumentare il numero di diagnosi dei tumori effettuate in fase iniziale. [7]

1.2 Microcalcificazioni: definizione e ruolo nella diagnosi precoce

Le microcalcificazioni mammarie sono depositi di calcio con diametro <0,5 mm localizzate all'interno del tessuto mammario. Nelle mammografie le MC sono state descritte per la prima volta negli anni 50 e attualmente rivestono un ruolo cruciale nella diagnosi del tumore al seno, infatti, il 30-50% dei tumori mammari non palpabili può essere individuato esclusivamente grazie alle MC rivelate dalla mammografia. Per questo vengono considerate dei marcatori robusti del cancro al seno (Figura 1). [8]

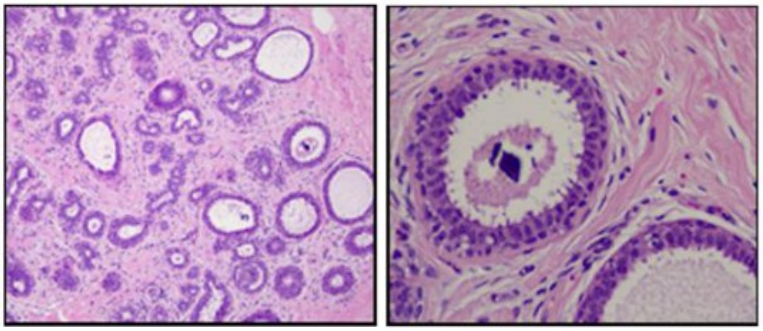

Figura 1: Rappresentazione istopatologica delle MC all'interno di diversi tipi di lesioni mammarie. [8]

A livello biochimico, le MC sono generalmente classificate in due tipi principali: Tipo I, composto da ossalato di calcio (CO) e Tipo II, composto da idrossiapatite (HA). Recentemente, ne è stato identificato un nuovo tipo che presenta una forma diversa di idrossiapatite (HA): idrossiapatite associata al magnesio (Mg-HA). Tra questi, il primo tipo è associato solitamente a condizioni benigne, mentre gli altri rimangono principalmente correlati a lesioni maligne. [8]

Dal punto di vista radiologico nelle mammografie le MC vengono classificate in benigne o sospette. Le prime sono in genere più estese, più grossolane, più rotonde e con margini lisci e più facilmente visibili rispetto a quelle sospette (Figura 2). Le MC associate a malignità, invece, in genere hanno delle dimensioni ridotte tanto da richiedere un ingrandimento per essere ben visualizzate. [8]

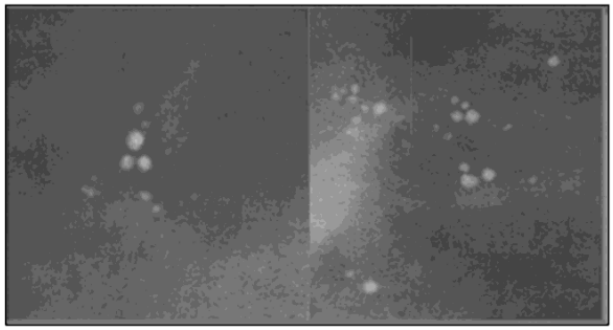

Figura 2: Esempi di calcificazioni tipicamente benigne. [8]

Per classificarle in benigne o sospette, quindi, è necessario descriverne la morfologia e la distribuzione. La morfologia sospetta comprende un aspetto eterogeneo grossolano, natura amorfa, elementi pleomorfi fini (con forma, densità e dimensione variabili) e calcificazioni con forme lineari e ramificate (Figura 3). [8]

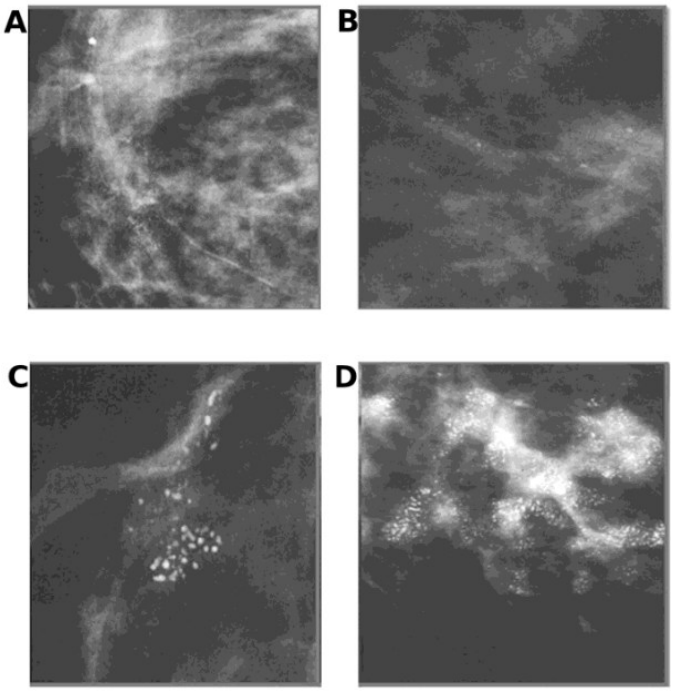

Figura 3: Calcificazioni con morfologia sospetta. (A) eterogeneo grossolano, (B) amorfo, (C) pleomorfo fine e (D) lineare fine o ramificato lineare fine. [8]

Le cinque categorie di distribuzione sono invece diffusa, segmentale, regionale, raggruppata e lineare (Figura 4).

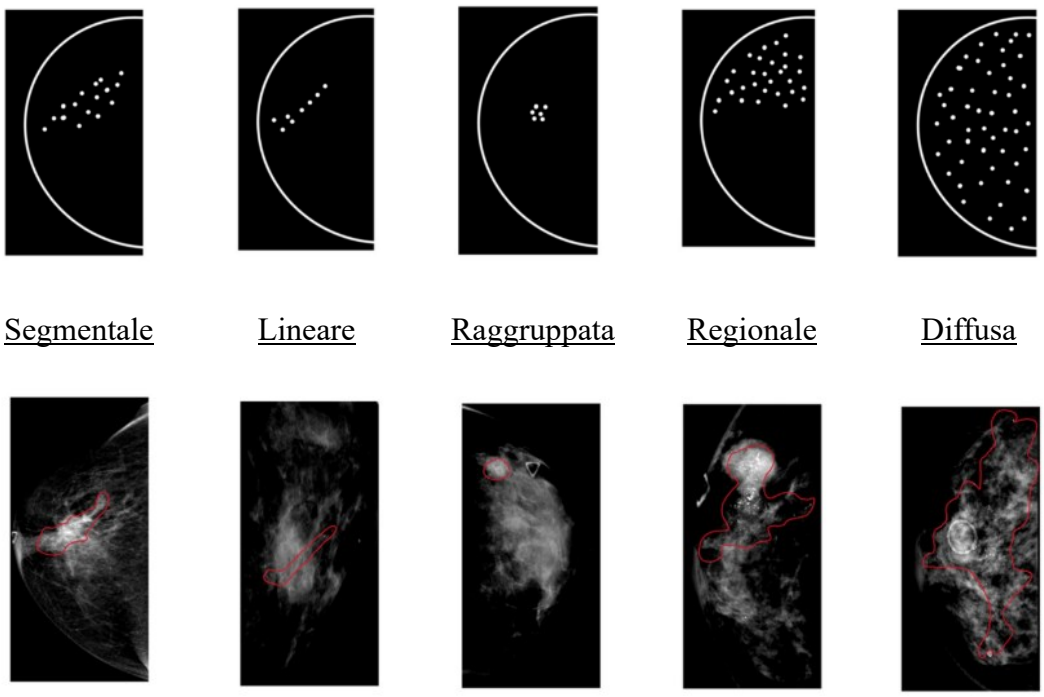

Figura 4: Esempi dei descrittori di distribuzione delle calcificazioni. Per ciascuna categoria, la parte superiore mostra un'illustrazione schematica, mentre la parte inferiore riporta un caso reale. Le calcificazioni sono evidenziate con contorni rossi. [9]

In generale, le calcificazioni sono alterazioni estremamente frequenti, osservate nell'80% delle mammografie, che riflettono per lo più un processo benigno. Tuttavia, quando si presentano come piccole particelle, raggruppate e polimorfiche, possono essere associate a malignità. [8]

1.2.1 Le MC come caso di studio per la diagnosi assistita

Nel 50% dei tumori al seno non palpabili la diagnosi di neoplasia viene effettuata esclusivamente attraverso l'individuazione di pattern di MC. Le MC mammarie, infatti, possono presentare sulle mammografie dimensioni, forma, estensione, densità e distribuzione diversa. Attraverso lo studio quantitativo delle immagini, è possibile ricavare queste caratteristiche e sfruttarle per assistere il medico nella fase di diagnosi. Dal punto di vista clinico, solitamente viene richiesta l'esecuzione di una biopsia su una regione sospetta esclusivamente in seguito a una valutazione effettuata da parte dei radiologi. I radiologi, quindi, osservano la morfologia e la forma delle MC individuate sulle mammografie e le classificano

secondo il sistema BI-RADS. Spesso, però nonostante l'efficacia della mammografia, vengono commessi ancora numerosi errori con tassi di biopsie false positive per le calcificazioni che variano dal 37% al 80%. Dei fattori che possono influenzarne le interpretazioni sono la presenza di un basso contrasto e di un tessuto mammario denso perché rendono più complicata la localizzazione delle MC nelle immagini mammografiche. Infatti, la sensibilità dello screening per rilevare calcificazioni maligne rimane relativamente bassa. Molte calcificazioni rilevabili non vengono immediatamente segnalate per effettuare ulteriori indagini, ma vengono invece identificate durante cicli di screening successivi, quando la malattia è già progredita fino a uno stadio invasivo. Per migliorare il processo diagnostico del medico è possibile introdurre un approccio quantitativo per descrivere le caratteristiche morfologiche delle MC. I metodi di apprendimento automatico si sono rivelati preziosi in questo contesto. [10]

1.3 Caratterizzazione delle MC

I sistemi automatici per il riconoscimento delle lesioni mammarie costituiscono strumenti utili che possono essere sfruttati per assistere il medico nell'interpretazione delle mammografie. Lo sviluppo di tali sistemi prevede l'esecuzione di due task principali. Il primo riguarda l'individuazione delle ROI nelle mammografie che includono lesioni sospette, in questo caso le MC. Il secondo riguarda la caratterizzazione delle MC tramite features che consentano di distinguerle tra maligne e benigne. Queste caratteristiche possono poi essere fornite in input a un sistema in grado di effettuare classificazione. Esistono diverse tipologie di features che possono essere estratte come feature di texture, intensità e forma, ciascuna con specifici vantaggi e limitazioni. Tra queste quelle morfologiche sono quelle più comunemente utilizzate per via della loro somiglianza con le caratteristiche considerate dai radiologi stessi. [11] [12]

1.3.1 Le feature radiomiche

Tra i diversi approcci per la caratterizzazione quantitativa delle MC, un ruolo rilevante è svolto dalla radiomica. La radiomica è un approccio innovativo che permette di convertire le immagini in dati che possono essere utilizzati per condurre indagini cliniche. Consente, infatti, di convertire le regioni di interesse (ROI) in caratteristiche quantitative. Le feature vengono calcolate attraverso formule matematiche applicate all'istogramma dei livelli di grigio, alle matrici che definiscono la texture o alla forma delle ROI. Le caratteristiche estratte hanno un significato noto (caratteristiche intelligibili), per questo è possibile utilizzarle per addestrare

modelli di apprendimento automatico con lo scopo di trarne conclusioni rilevanti dal punto di vista clinico. [10]

Le feature radiomiche sono genericamente raggruppabili in tre macro-categorie che descrivono proprietà diverse dell'immagine: feature morfologiche (shape), statistiche di primo ordine e di texture. [13] Sommando le 29 feature morfologiche, le 50 feature di primo ordine e le 95 feature di texture, si ottiene un totale complessivo di 174 caratteristiche radiomiche. Questo conteggio fa riferimento a un protocollo di standardizzazione specifico: [14]

1) **Feature morfologiche (shape)- 29 feature:** Descrivono le caratteristiche geometriche e spaziali della ROI, come ad esempio volume, area, sfericità. Sono fondamentali per quantificare la forma dell'oggetto analizzato, fornendo informazioni indipendenti dall'intensità dei pixel. [13][14]
2) **Feature di Primo ordine – 50 feature:** Analizzano la distribuzione dei livelli di grigio all'interno della ROI, senza considerare le relazioni spaziali tra i pixel. Alcuni esempi solo la media, la mediana, la deviazione standard ed entropia, che forniscono un'indicazione globale dell'intensità e dell'eterogeneità dell'area. [15][16][14]
3) **Feature di texture – 95 feature:** Rappresentano statistiche di ordine superiore e descrivono la relazione spaziale tra i pixel (o voxel), quantificando l'eterogeneità intra-lesione. [16] Queste feature vengono suddivise in diverse matrici di texture, nel complesso forniscono 95 feature: [14]
    - **GLCM (Gray Level Co-occurrence Matrix) – 25 feature:** Descrive la relazione tra coppie di pixel, quantificando la frequenza con cui coppie di livelli di grigio compaiono a una certa distanza e in una certa direzione. Utile per caratterizzare il contrasto, l'omogeneità e la regolarità della texture. [13][15] [14]
    - **GLRLM (Gray Level Run-length Matrix) – 16 feature:** Analizza la lunghezza di sequenze consecutive di pixel con lo stesso livello di grigio lungo una direzione specifica, fornendo informazioni sulla granulosità della texture e sulla presenza di strutture lineari. [13][15][14]
    - **GLSZM (Gray Level Size Zone Matrix) – 16 feature:** Valuta le dimensioni delle aree contigue di pixel con lo stesso livello di intensità, indipendentemente

dalla direzione. Utile per caratterizzare se la texture è composta da zone piccole e frammentate o da regioni più estese e omogenee. [13][14]

- **GLDZM (Gray Level Dependence Zone Matrix) – 16 feature:** Quantifica il numero di pixel connessi che dipendono da un pixel centrale, basandosi su una soglia definita dalla differenza assoluta dei livelli di grigio. Questa matrice descrive il grado di dipendenza locale e la complessità strutturale dell'immagine. [13][14]
- **NGTDM (Neighborhood Gray Tone Difference Matrix) – 5 feature:** Descrive la differenza tra il valore di un pixel e la media dei suoi vicini in un intorno definito, risultando utile per identificare la presenza di dettagli fini o di irregolarità spaziali nell'immagine. [13][14]
- **NGLDM (Neighboring gray level dependence matrix) – 17 feature**: Quantifica la dipendenza di un pixel (o voxel) rispetto ai suoi vicini, entro una certa distanza, catturando la complessità e coesione della texture. [14]

In particolare, il numero e il tipo di variabili estratte dipendono dal software utilizzato: ad esempio, la libreria Pyradiomics, standard di riferimento in ambiente Python, consente l'estrazione di un set di feature che può differire, per numero e tipologia, dallo schema teorico sopra citato. Tali valori, quindi, possono variare in base alla versione del software, ai parametri di configurazione impostati o all'utilizzo di altri pacchetti analitici. [13]

1.3.2 Analisi della distribuzione delle MC

Le MC sono piccoli depositi di calcio che, all'interno delle mammografie, appaiono come punti con livelli di grigio più elevati per via del maggiore coefficiente di attenuazione del calcio ai raggi x rispetto al tessuto mammario normale. Nel processo di diagnosi, però è importante valutare non tanto le singole MC in sé perché singolarmente non sono altamente indicative della presenza di cancro al seno, ma piuttosto verificare se si aggregano a formare dei cluster. La presenza di cluster costituiti da 3 o più MC, infatti, costituisce un importante segno precoce di cancro al seno. Per questo la loro individuazione supporta il processo di diagnosi contribuendo ad individuare il tumore durante le fasi iniziali. [8] [12]

In particolare, diversi studi riportano che le lesioni maligne tendono a presentare MC più piccole, numerose e densamente concentrate, mentre quelle MC benigne sono in genere più grandi, meno numerose e più sparse. [17]

1.3.3 Descrittori Geometrici

Dato che la formazione di cluster di MC è considerata un segnale di maggiore sospetto oncologico rispetto alle singole calcificazioni isolate, ne viene studiata la morfologia complessiva e la distribuzione spaziale. Dopo aver raggruppato le MC in cluster è possibile caratterizzare gli stessi estraendo descrittori geometrici che ne quantificano la forma e la dispersione. [11]

Esempi di descrittori utilizzati in letteratura includono:

- **Numero di MC per cluster e densità** (MC per cm^2), utili per distinguere raggruppamenti piccoli e numerosi (tipici dei quadri maligni) da cluster con poche calcificazioni più grandi e disperse (più frequentemente benigni). [18]
- **Dimensione e forma del cluster**: come area, rapporto tra asse maggiore e minore, circolarità, che caratterizzano se il cluster è più rotondeggiante o allungato. [18]
- **Distribuzione interna delle MC** rispetto al baricentro del cluster (come, ad esempio, la distanza media dal centro), che permette di descrivere quanto le MC siano concentrate o distribuite in modo regolare. [18]

Le caratteristiche geometriche estratte dai cluster, spesso combinate con altre feature morfologiche, di intensità o texture, vengono utilizzate come input di sistemi di classificazione, per discriminare tra loro casi benigni e maligni, supportando il radiologo nella formulazione della diagnosi finale. [18]

Tuttavia, la robustezza di tali descrittori è strettamente legata alla qualità dell'identificazione a monte delle MC e dei relativi cluster: una delimitazione imprecisa può alterare drasticamente il valore dei parametri, compromettendo l'intera fase di classificazione. [13]

1.4 CAD in mammografia: Pipeline

La discriminazione delle lesioni identificate nelle mammografie di solito viene modellizzata come un problema di classificazione a due classi (benigno/maligno). In generale gli approcci seguono una pipeline standard che consta di una serie di step. Il punto di partenza è una ROI,

ossia una regione di interesse che contiene le lesioni da classificare, in questo caso le MC. La ROI può essere definita manualmente da un radiologo o identificata in maniera automatica da un sistema CAD. In genere, si tratta di una immagine rettangolare ritagliata da una mammografia. Una volta definita la regione di interesse gli step previsti sono 4: [19]

1) Segmentazione delle lesioni
2) Estrazione delle caratteristiche
3) Selezione delle caratteristiche
4) Classificazione

Nella figura 5 viene mostrato un diagramma di flusso di un tipico sistema CAD. Spesso uno o più stadi del processo illustrato vengono omessi. Inoltre, alcuni approcci utilizzano più ROI contenenti la lesione ritagliata da diverse proiezioni mammografiche o da modalità aggiuntive (come l'ecografia) o da esami precedenti. In questi casi, la segmentazione della lesione e l'estrazione delle caratteristiche relative viene eseguita in modo indipendente in tutte le ROI che la contengono. Infine, alcuni approcci CAD utilizzano delle caratteristiche aggiuntive, che possono essere ad esempio dati clinici come l'età del paziente, i quali non vengono estratti dalle ROI ma a partire da annotazioni. [19]

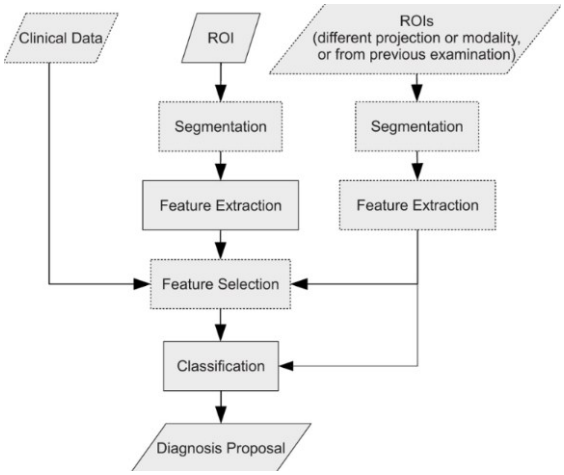

Figura 5: diagramma di flusso di un sistema CAD. [19]

1.5 Machine learning applicato all'analisi di dati di mammografie

La mammografia rappresenta il metodo standard per la diagnosi del tumore al seno, tuttavia, presenta alcuni limiti come una bassa sensibilità in presenza di tessuto mammario denso. Questi limiti vengono superati mediante l'utilizzo di sistemi CAD basati su tecniche di machine

learning (ML), che supportano il radiologo nelle diverse fasi di analisi delle immagini, tra cui il rilevamento delle lesioni, la segmentazione, l'estrazione delle caratteristiche e la classificazione. [12]

Il Ml è un sottocampo dell'intelligenza artificiale, che consente di sviluppare modelli in grado di apprendere dai dati e di generalizzare su nuovi esempi. I principali approcci includono l'apprendimento supervisionato, basato su dati etichettati, e quello non supervisionato che individua automaticamente le strutture e le regolarità presenti nei dati a partire da esempi privi di etichette. [12]

Fanno parte dell'apprendimento supervisionato anche gli algoritmi di apprendimento ensemble. Questi sono metodi in grado di combinare due o più algoritmi di ML per ridurre la varianza e gli errori di bias. I metodi di apprendimento ensemble possono essere classificati in metodi di boosting e bagging. I metodi ensemble di bagging addestrano diversi classificatori in modo indipendente e poi combinano le loro previsioni. Tra i metodi di bagging rientra il Random Forest (RF). I metodi ensemble di boosting, invece, addestrano i modelli in maniera iterativa, in modo che ciascun modello successivo impari a correggere gli errori commessi da quello precedente. Ne è un esempio il Gradient Boosting. [20]

Nonostante i notevoli progressi ottenuti, i sistemi di ML presentano ancora alcune limitazioni. Per garantire risultati affidabili, è fondamentale eseguire una fase di pre-elaborazione delle immagini mammografiche, finalizzata a migliorarne il contrasto, alla riduzione del rumore e al mantenimento delle informazioni rilevanti. [12]

1.5.1 Metodi di clustering

I metodi di clustering sono algoritmi di analisi dei dati che si inseriscono prevalentemente nell'ambito dell'apprendimento non supervisionato. Lo scopo principale di tali tecniche è quello di raggruppare i dati in cluster sulla base della somiglianza delle caratteristiche e delle proprietà dei punti considerati, senza l'ausilio di etichette di classe fornite esternamente (Figura 6). [21]

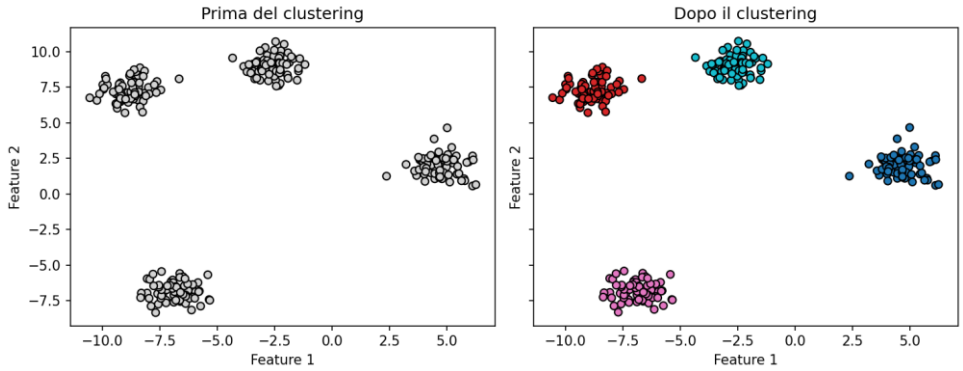

Figura 6: Visualizzazione del processo di clustering. A sinistra: configurazione iniziale dei dati; A destra: raggruppamento dei dati in cluster distinti mediante apprendimento non supervisionato.

Esistono diverse famiglie di algoritmi di clustering, tra cui metodi basati su centroidi, metodi gerarchi e metodi basati sulla densità. Questi ultimi risultano particolarmente adatti per l'imaging medicale in quanto consentono di individuare cluster di forma arbitraria senza richiedere la definizione a priori del numero di gruppi. Tra gli algoritmi più utilizzati in questo ambito vi sono DBSCAN (Density-Based Spatial Clustering of Applications with Noise) DBSCAN e OPTICS (Ordering Points To Identify the Clustering Structure), impiegati per identificare gruppi di punti spazialmente vicini e separare eventuali osservazioni isolate. [21]

1.5.2 Metodi di classificazione

Una volta estratte le feature che descrivono le MC è possibile sfruttarle per addestrare un algoritmo di classificazione. Infatti, la discriminazione delle lesioni individuate nelle mammografie viene modellizzata come un problema di classificazione a due classi che mira a distinguere le lesioni in benigne e maligne. [19] Per svolgere questo task possono essere utilizzati diversi algoritmi di classificazione. In particolare, la scelta dell'algoritmo supervisionato per svolgere il compito di classificazione dipende da una serie di fattori. Tra questi, ad esempio, ricordiamo il numero di esempi nel dataset di addestramento, le dimensioni dello spazio delle caratteristiche, la presenza di caratteristiche correlate o l'overfitting. [24]

1.5.2.1 Support Vector Machines

Le Support Vector Machines (SVM) rappresentano una famiglia di algoritmi di apprendimento supervisionato, noti per le loro elevate prestazioni nei compiti di classificazione e regressione. Il principio fondamentale alla base delle SVM è la costruzione di un iperpiano in uno spazio

delle caratteristiche che massimizzi la separazione tra i punti appartenenti alle diverse classi all'interno di un dataset. I punti più vicini al confine decisionale, detti Support Vector, definiscono tale iperpiano. L'obiettivo è massimizzare il margine ossia la distanza tra i Support Vector e il confine decisionale. Dal punto di vista matematico, quindi, viene risolto un problema di ottimizzazione che mira a determinare i coefficienti ottimali dell'iperpiano, garantendo così una corretta classificazione e la massimizzazione del margine (Figura 7). [25]

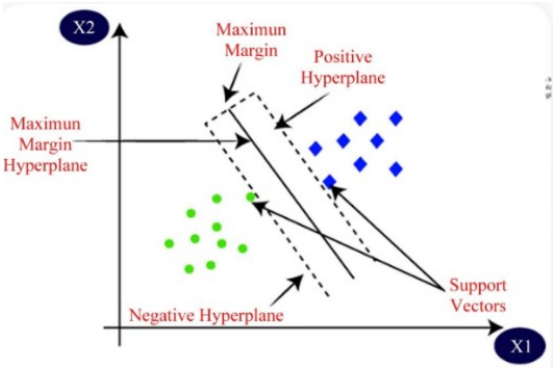

Figura 7: Rappresentazione funzionamento SVM in 2D. [24]

Le SVM presentano diversi vantaggi: sono efficienti nella gestione di dati ad alta dimensionalità, sono ideali per applicazioni come la classificazione di immagini perché in grado di modellare relazioni complesse nello spazio delle caratteristiche e gestiscono bene relazioni non lineari tra le feature. Grazie all'uso delle funzioni kernel, inoltre possono mappare i dati in spazi di dimensione superiore, consentendo la cattura di confini decisionali complessi. [25]

Quando i dati non sono linearmente separabili nello spazio originale, si introduce quindi, una funzione Kernel che realizza il prodotto scalare in uno spazio di caratteristiche (anche di dimensione infinita nel caso di kernel RBF), consentendo comunque di trovare un iperpiano ottimale. [26] Tipicamente si impiegano kernel lineari, polinomiali e radiali di base (RBF), ciascuno caratterizzato da specifici iperparametri: il parametro di regolarizzazione C, condiviso da tutti i kernel, controlla il compromesso tra massimizzazione del margine ed errori di classificazione, mente parametri come il grado d (del polinomiale) e gamma (per il kernel RBF) regolano la forma e sinuosità del confine decisionale, consentendo al classificatore di adattarsi con maggiore o minore precisione alla struttura locale dei dati. [27]

1.5.2.2 Random Forest

Il Random Forest è un modello di apprendimento supervisionato progettato per superare le limitazioni strutturali dei singoli alberi decisionali. A differenza di questi ultimi, spesso soggetti a overfitting, il RF riduce tale rischio combinando numerosi alberi decisionali: si ottiene così un modello robusto e adattabile. [25]

Proprio perché si basa sulla generazione di un insieme di alberi decisionali, il RF fa parte dei metodi di apprendimento ensemble. In particolare, ciascun albero viene addestrato su una porzione casuale del dataset, garantendo che ognuno si concentri su aspetti differenti dei dati, diminuendo così la probabilità di overfitting. In questo modo il modello diventa più robusto e con una maggiore capacità di generalizzazione. La previsione finale nei problemi di classificazione viene poi ottenuta combinando le uscite di tutti gli alberi mediante voto di maggioranza (Figura 8). [25]

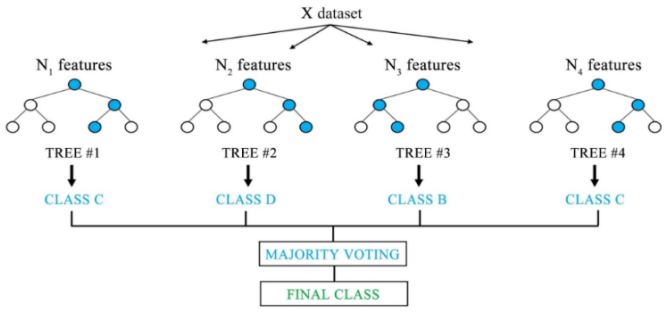

Figura 8: Funzionamento dell'algoritmo RF [24]

Le sue prestazioni e la sua capacità di generalizzazione dipendono in modo cruciale dalla corretta calibrazione di specifici iperparametri. La configurazione ottimale di tali iperparametri permette di bilanciare il bias e la varianza del modello, garantendo una maggiore robustezza. In particolare, i principali iperparametri che regolano il comportamento del modello sono: [20] [24] il numero degli alberi, la profondità massima degli alberi decisionali, il numero minimo di osservazioni richieste dei nodi terminali e il numero di variabili considerate casualmente a ogni suddivisione. [28]

Il RF presenta diversi vantaggi: è robusto al rumore, infatti, grazie all'approccio ensemble l'influenza dai dati rumorosi o dagli outlier diminuisce; è adatto alla gestione di grandi dataset grazie alla costruzione di numerosi alberi decisionali; fornisce una misura dell'importanza delle

feature che è utile per identificare le variabili più rilevanti. Per le sue caratteristiche è un algoritmo che viene spesso utilizzato nell'ambito della ricerca biomedica per l'analisi di dati complessi. Tuttavia, la sua efficacia non è da considerarsi universale: le prestazioni del modello dipendono strettamente dalla qualità del dataset di partenza e dalla pertinenza delle feature selezionate, rendendo essenziale una valutazione critica dei risultati ottenuti in ogni specifico contesto diagnostico. [24][25]

1.5.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) è un algoritmo di apprendimento supervisionato appartenente alla famiglia dei metodi ensemble con approccio sequenziale. Esso si basa sul paradigma del Gradient Boosting, in cui un modello predittivo viene costruito combinando più modelli di base, tipicamente alberi decisionali, addestrati in maniera iterativa affinché ciascun modello successivo corregga gli errori dei precedenti. [20]

Il processo di apprendimento si fonda sulla minimizzazione di una funzione di perdita: a ogni iterazione viene addestrato un nuovo modello sui residui del modello corrente, definiti come il gradiente negativo della funzione di perdita. XGBoost introduce una funzione obiettivo regolarizzata, che consente di controllare la complessità del modello e ridurre il rischio di overfitting. Inoltre, utilizza un'approssimazione al secondo ordine basata sullo sviluppo di Taylor, che sfrutta sia il gradiente sia l'hessiana per valutare in modo più accurato la qualità delle suddivisioni durante la costruzione degli alberi, mediante una metrica nota come Structure Score (Figura 9). [20][29]

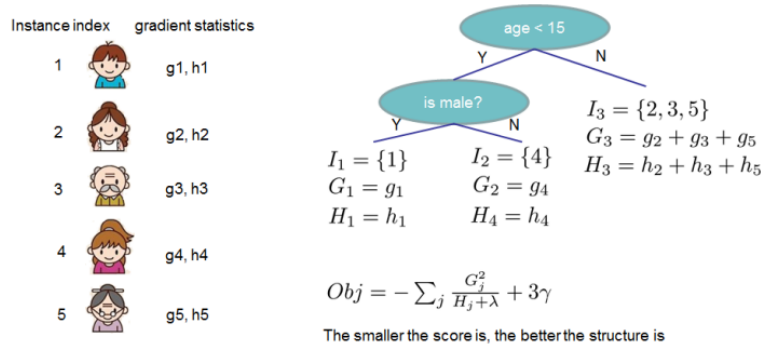

Figura 9: Schema del calcolo dello Strcuture Score in XGBoost. L'algoritmo aggrega i gradienti g_i e le hessiane h_i di ciascuna istanza per definire la qualità dello split, permettendo di minimizzare la funzione obbiettivo. [29]

Dal punto di vista operativo, XGBoost offre diversi vantaggi, tra cui la gestione autonoma dei dati mancanti, la ridotta necessità di pre-elaborazione e la possibilità di stimare l'importanza delle feature. Tuttavia, il modello presenta numerosi iperparametri, come il learning rate, max_depth e subsampling che richiedono un'attenta ottimizzazione per ottenere prestazioni ottimali. [20]

1.5.2.4 Metriche di valutazione

La fase di valutazione delle prestazioni di un classificatore è di grande importanza per giudicarne la sua utilità. Nella classificazione binaria, i dati vengono suddivisi in due classi diverse: positivi (P) e negativi (N). Questa classificazione produce quattro tipi di risultati: due tipi di classificazione corretta ossia: veri positivi (TP) e veri negativi (TN), e due tipi di classificazione errata: falsi positivi (FP) e falsi negativi (FN). [30]

Per valutare le prestazioni degli algoritmi di AI per lo screening mammografico vengono utilizzate diverse metriche di valutazione. La scelta della metrica di valutazione dipende dal compito specifico e dagli obiettivi del modello. Alcune delle metriche più comunemente utilizzate sono: [5]

- **Accuratezza:** misura la percentuale di previsioni corrette fatte dal modello. [5]
- **Sensibilità:** misura la percentuale di predizioni vere e positive su tutti i casi effettivamente positivi del dataset. [5]
- **Precisione (Valore predittivo positivo, PPV):** misura la percentuale di previsioni positive vere su tutte le previsioni positive fatte dal modello. [5]
- **Specificità:** misura la capacità del modello di identificare correttamente i casi negativi. [31]
- **Valore predittivo negativo (NPV):** indica la probabilità che un soggetto risulti veramente negativo dato un esito negativo del test. [31]
- **Punteggio $\mathbf{F_1}$**: una metrica che bilancia precisione e sensibilità. [5]
- **Punteggio $\mathbf{F_\beta}$**: una metrica che generalizza il punteggio F_1, permettendo di modulare l'importanza relativa tra precisione e sensibilità tramite il parametro β. [32]
- **Area sotto la curva ROC (AUCROC):** distingue tra punti positivi e negativi in un intervallo di valori soglia. La curva ROC viene ricavata tracciando il tasso di veri

positivi rispetto al tasso di falsi positivi per diversi valori di soglia. L'AUCROC è definita come l'area sotto questa curva. [5]

L'equazione comunemente utilizzata per il calcolo dell'accuratezza è:

$$Accuratezza = \frac{TP + TN}{TP + TN + FP + FN} \quad [5]$$

La precisione, la sensibilità, la specificità, il valore predittivo negativo e il punteggio F1 e F_β, sono invece definite con le formule seguenti:

$$Precisione = \frac{TP}{TP + FP} \quad [5]$$

$$Sensibilità = \frac{TP}{TP + FN} \quad [5]$$

$$Specificità = \frac{TN}{TN + FP} \quad [31]$$

$$\text{Valore predittivo negativo} = \frac{TN}{TN + FN} \quad [31]$$

$$F_1 = \frac{2(Precisione + Sensibilità)}{Precisione + Sensibilità} \quad [5]$$

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \quad [32]$$

È stato osservato, però, che metriche come l'accuratezza o l'AUCROC possono fornire un'informazione fuorviante in presenza di dataset fortemente sbilanciati. In questi casi la classe positiva costituisce in genere una piccola frazione dei dati. In particolare, l'AUCROC in presenza di dataset fortemente sbilanciati tende a rimane elevata anche quando un classificatore commette molti errori nella classe di minoranza. In queste situazioni metriche robuste che possono essere utilizzate al loro posto sono il punteggio F1 e il Coefficiente di Correlazione di Matthews (MCC) il cui valore è compreso tra [-1,1]. È definito dalla formula seguente: [33]

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

MCC è in grado di trattare in modo simmetrico entrambe le classi, permettendo di valutare l'accuratezza anche quando il dataset risulta fortemente sbilanciato. [33]

Tutte queste misure ad eccezione dell'AUCROC, però, sono misure a soglia singola quindi definite per singole soglie di punteggio di un classificatore. Di conseguenza non possono fornire

una panoramica dell'intervallo di prestazioni con soglie variabili. Poiché la scelta del valore di soglia non è ovvia, una soluzione efficace è quella di utilizzare misure senza soglia come i grafici ROC (Receiver Operating Characteristic) e PRC (Precision-Recall Curve). In particolare, nel caso di dataset sbilanciati la metrica più adatta è la PRC che mostra i valori di precisione per i corrispondenti valori di sensibilità. In maniera analoga al grafico ROC, la PRC fornisce una valutazione dell'intero modello, definendo una metrica chiamata AUC della PRC (PR-AUC). La PR-AUC si ottiene calcolando l'area sottesa a tale curva e sintetizza in un unico valore la capacità del modello di mantenere un'elevata precisione al crescere della sensibilità, garantendo allo stesso tempo una rilevazione esaustiva dei casi positivi. A differenza dell'AUCROC, l'PR-AUC risulta essere la metrica più adatta nel caso di dataset fortemente sbilanciati, frequenti in ambito clinico in cui la classe "benigna" è preponderante. [30]

1.5.2.5 Spiegabilità del modello

Nel contesto dei sistemi di supporto alle decisioni cliniche basati sul ML, un aspetto cruciale è la loro spiegabilità (explainability), intesa come la capacità di comprendere in che modo le variabili contribuiscano alle decisioni del modello. I modello complessi, vengono spesso definiti "black-box", poiché le loro rappresentazioni interne non hanno un immediato significato clinico e loro le predizioni non sono facilmente riconducibili a pattern intuitivi per il medico. In ambito di imaging mammografico, diversi lavori sottolineano come l'adozione di tecniche di Explainable AI (XAI) sia essenziale per promuovere l'adozione clinica dei sistemi automatici, consentendo di monitorare il focus decisionale del modello e di rilevare eventuali errori sistematici nell'apprendimento. [34]

Tra i metodi di XAI, un ruolo di rilievo è ricoperto da SHAP (Shapley Additive exPlanations): questo approccio associa a ciascuna feature un contributo additivo alla predizione, permettendo sia di ordinare le variabili per importanza globale, sia di valutarne la direzione dell'effetto (ovvero se valori superiori o inferiori aumentino la probabilità di appartenenza alla classe positiva). SHAP quindi consente di analizzare il contributo delle singole features sulle decisioni del modello, verificando se i pattern appresi dal modello siano in linea con le aspettative cliniche e le evidenze scientifiche sui criteri di malignità. [34]

CAPITOLO 2 – ANALISI DELLA LETTERATURA

Il presente capitolo analizza lo stato dell'arte per l'analisi automatica delle MC nelle mammografie digitali. L'analisi viene articolata esaminando in ordine: la selezione delle ROI, il rilevamento delle MC, la loro caratterizzazione e la relativa classificazione.

2.1 Selezione delle ROI

La mammografia digitale produce immagini di grandi dimensioni che includono oltre al tessuto mammario, anche ampie porzioni di sfondo che non sono rilevanti dal punto di vista clinico. Di conseguenza, l'elaborazione dell'intera immagine non risulta necessaria ai fini diagnostici e comporta inoltre criticità in termini di risorse computazionali e di memoria. Per questo, un passaggio preliminare che viene effettuato consiste nella definizione delle cosiddette regioni di interesse (ROI). Operativamente, una ROI è una sotto-immagine (o patch) che contiene una o più strutture di interesse (in questo caso le MC) e che si ottiene a partire dall'immagine originale. L'utilizzo delle ROI, per questo, consente di concentrare l'analisi sulle aree in cui le lesioni sono presenti o sospette ottimizzando le fasi successive di estrazione delle caratteristiche e di classificazione. [34]

Dal punto di vista metodologico, si distinguono due approcci principali per la generazione delle ROI:

- **Approcci manuali o semi-automatici:** basati sull'intervento diretto del radiologo per isolare le aree sospette. Questi metodi sono accurati ma risultano poco scalabili e introducono una componente di soggettività inter-operatore che può influenzare le performance del sistema.
- **Approcci automatici:** ampiamente utilizzati nei modelli di Deep Learning più recenti, in cui le ROI vengono generate automaticamente tramite suddivisione dell'immagine in patch o mediante l'utilizzo di modelli di segmentazione che individuano direttamente le regioni candidate.

Concentrando il processo di apprendimento esclusivamente sulle regioni contenenti le lesioni è possibile migliorare significativamente l'efficacia dell'estrazione delle caratteristiche e la robustezza del classificatore finale. [34]

Tra gli approcci automatici più recenti, la pipeline DeepMica, rappresenta una soluzione efficace per l'analisi delle MC. In questo contesto, la selezione delle ROI avviene attraverso una fase di pre-elaborazione delle immagini seguita da una segmentazione automatica basata su architettura di tipo UNet. In particolare, dopo una fase iniziale di normalizzazione, l'immagine viene suddivisa in patch di dimensione 256 x 256 pixel. Tali patch costituiscono l'input per la rete UNet. [34] La rete UNet è un'architettura caratterizzata da una struttura simmetrica dotata di un encoder (che cattura il contesto globale dell'immagine) e di un decoder (che consente di ricostruire i dettagli). Il suo utilizzo permette di generare una maschera di segmentazione che evidenzia le aree contenenti MC. Di tutte le aree segmentate dall'algoritmo DeepMica si considereranno nelle analisi successive solamente quelle localizzate entro il perimetro delle ROI evidenziate dai radiologi. Tali regioni segmentate costituiscono le ROI effettive su cui verranno applicate le successive tecniche di identificazione e analisi delle MC (Figura 10). [34]

Il metodo DeepMica dimostra prestazioni robuste nel delineare le MC dalle scansioni mammografiche. In particolare, sul test set sono state ricavate l'Intersection over Union (IOU) e l'Accuracy (ACC) calcolando la media delle metriche su tutte le immagini presenti al suo interno. Si osserva, quindi, un valore di IoU pari al 74% che indica una buona sovrapposizione tra le maschere predette e la ground truth, ovvero le annotazioni di riferimento realizzate da esperti e utilizzate come standard per la valutazione; e un'accuratezza dell'83% che a sua volta conferma l'affidabilità del modello. L'accuratezza a livello di pixel invece è stata espressa in termini di AUCROC e PR-AUC rispettivamente pari al 95% e al 78%. Questi valori indicano che il modello è in grado di distinguere efficacemente, a livello di singolo pixel, tra le regioni di interesse e lo sfondo, classificando correttamente i pixel appartenenti alle diverse classi. [34]

Dal punto di vista computazionale, i tempi di inferenza misurati sull'hardware NVIDIA 3070Ti utilizzato nello studio per la segmentazione risultano contenuti: si registrano circa 8.8 ms per immagine su GPU, mentre su CPU i tempi di elaborazione sono significativamente più elevati attestandosi a circa 218.6 ms. [34]

Nel complesso, questi risultati dimostrano che il modello raggiunge un buon equilibrio tra accuratezza della segmentazione ed efficienza computazionale, risultando efficace per il task considerato. L'automazione della selezione delle ROI consente quindi di ottenere regioni

candidate in modo coerente e riproducibile, riducendo la dipendenza dall'intervento umano e fornendo un input standardizzato per le fasi successive della pipeline. [34]

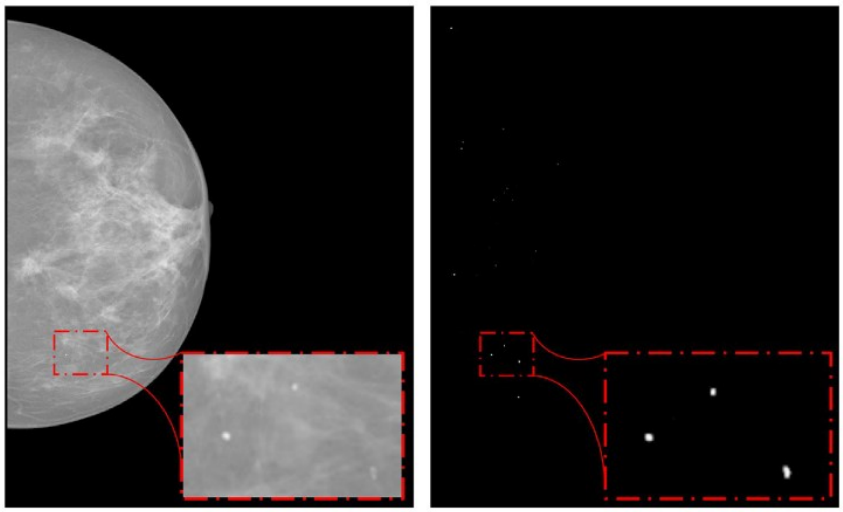

Figura 10: Esempio di scansione e relativa maschera di riferimento con annotazioni delle MC a livello di pixel. Data la dimensione estremamente ridotta delle lesioni, per una migliore visualizzazione è riportato uno zoom di un'area limitata dell'immagine originale e della maschera corrispondente. [34]

2.2 Identificazione delle MC

Una volta individuate le regioni dell'immagine che potenzialmente contengono MC è necessario identificare le singole MC come oggetti distinti. Il punto di partenza è rappresentato dai risultati forniti dai modelli di segmentazione che spesso producono una maschera in scala di grigio o binaria che evidenzia le zone in cui sono presenti le MC, senza però distinguerle tra loro.

In letteratura esistono diverse strategie per passare da una maschera che evidenzia genericamente la presenza di MC alla loro identificazione come oggetti distinti. Di seguito vengono analizzati i principali approcci proposti.

Un approccio classico per ottenere oggetti distinti a partire da una maschera binaria consiste nell'analisi delle componenti connesse. [35] [36] In questo contesto, il lavoro di Yapa e Harada [37] introduce un algoritmo di etichettatura che opera direttamente su immagini in scala di grigi, evitando la necessità di una binarizzazione preliminare. L'algoritmo esegue una scansione

dell'immagine: quando incontra un pixel appartenente a una nuova componente, ne traccia completamente il contorno e successivamente propaga l'etichetta ai pixel interni. In questo modo ogni oggetto viene identificato completamente nel momento in cui viene analizzato (Figura 11). Dal punto di vista applicativo, questo approccio rappresenta lo schema più semplice e diffuso: infatti a partire da una maschera applicando l'analisi delle componenti connesse è possibile ottenere una lista di oggetti distinti. Tuttavia, il suo limite principale è che non è in grado di separare oggetti molto vicini o fusi**,** poiché ogni regione connessa viene trattata come un unico elemento.

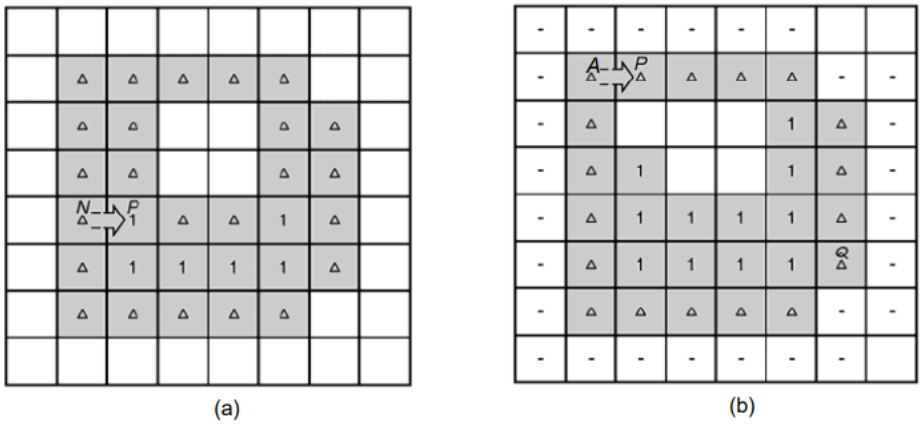

Figura 11: Meccanismo di scansione e tracciamento dell'algoritmo di Yapa e Harada [37]. l pannello (a) illustra la fase di scansione raster e l'individuazione di un contorno, mentre il pannello (b) evidenzia il processo di tracciamento del bordo e l'assegnazione dell'etichetta ai pixel interni.

Per superare questo limite, in particolare nei casi in cui le MC siano ravvicinate o parzialmente sovrapposte, diversi lavori introducono tecniche basate su morfologia matematica e sulla segmentazione Watershed, una strategia ampiamente validata in diverse applicazioni di imaging biomedico per la separazione di oggetti adiacenti o sovrapposti. [38] [39]

Ciecholewski [17] propone un approccio in cui, a partire da una mappa delle MC viene applicato un algoritmo di segmentazione Watershed controllato da marker per identificare separatamente le singole MC. La mappa iniziale consente di individuare le aree contenenti MC, ma non separa ancora le singole strutture. In questo studio per ottenere una segmentazione più precisa, vengono definiti dei marker interni, derivati dalle regioni candidate che cadono

all’interno delle MC (ad esempio minimi regionali filtrati), e dei marker esterni, che rappresentano lo sfondo. Questi marker fungono da semi da cui l’algoritmo Watershed inizia il processo di allungamento, facendo crescere le regioni fino a incontrarsi e generare linee di separazione tra strutture adiacenti. In questo modo, non viene segmentata indiscriminatamente l’intera immagine, ma il processo è vincolato dai marker a seguire solo le frontiere rilevanti tra MC e sfondo. Il risultato è un’identificazione più accurata delle MC, in grado di separare strutture adiacenti, superando così i limiti della sola analisi delle componenti connesse (Figure 12-13).

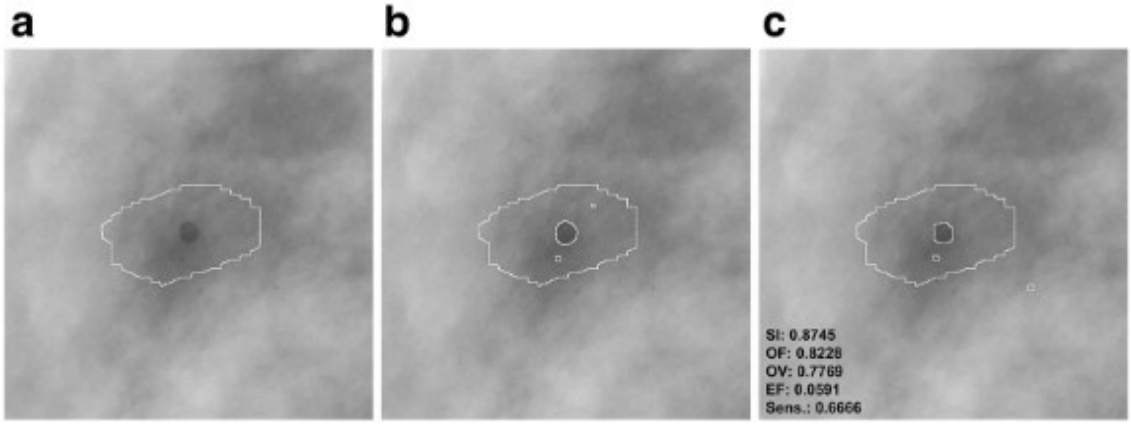

Figura 12: Esempio su caso benigno. Confronto tra: (a) contorno di riferimento (GTA), (b) riferimento del radiologo e (c) output dell'algoritmo. Performance: SI (Similarity Index) = 0.8745, Sensibilità = 0.6666. [17]

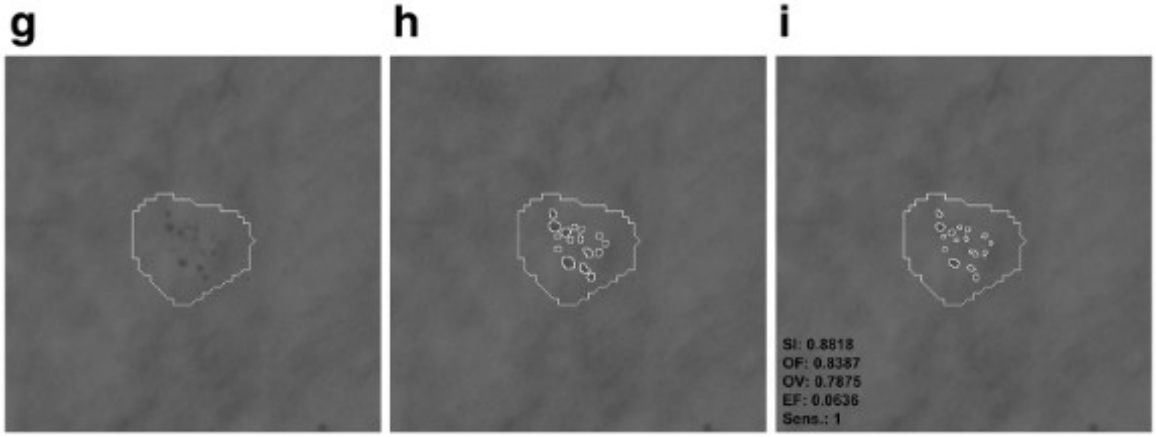

Figura 13: Esempio su caso maligno. Confronto tra: (g) contorno di riferimento (GTA), (h) riferimento del radiologo e (i) output dell'algoritmo. Performance: SI (Similarity Index) = 0.8818, Sensibilità = 1. [17]

2.3 Caratterizzazione delle MC

In letteratura, la caratterizzazione delle MC viene generalmente effettuata mediante l'estrazione di feature quantitative, che consentono di descrivere in modo oggettivo le proprietà delle lesioni .[10] [11] [40] [41] [42] Diversi studi hanno affrontato il problema della classificazione delle MC utilizzando differenti tipologie di caratteristiche, tra cui intensità, texture e forma, ciascuna con specifici vantaggi e limitazioni. In particolare, le feature di intensità descrivono i valori dei livelli di grigio dei pixel, fornendo informazioni sulla luminosità e sul contrasto della lesione rispetto allo sfondo; le feature di texture descrivono la distribuzione spaziale dei livelli di intensità nell'immagine, quindi, sono relative a come i pixel sono organizzati tra loro fornendo informazioni sulla struttura interna della lesione. [10] Mentre le feature di forma descrivono la geometria della lesione, inclusi parametri come area, perimetro e circolarità al fine di discriminare la natura della MC. [11]

Nel lavoro di Prinzi et. Al [10] sono state estratte 93 caratteristiche radiomiche suddivise in caratteristiche di primo ordine (intensità) e texture. Le prime descrivono la distribuzione dei livelli di grigio all'interno di una ROI, mentre le seconde vengono calcolate a partire da diverse matrici statistiche, tra cui GLCM, GLRLM, NGTDM, GLSZM, e GLDM. In questo studio, le caratteristiche di forma 2D, invece, non sono state utilizzate, in quanto considerate fortemente dipendenti dalla qualità della segmentazione. Infatti, gli autori sottolineano che la possibile differenza dimensionale tra lesioni maligne e benigne avrebbe potuto introdurre un bias, portando il modello a basarsi su informazioni geometriche piuttosto che su reali proprietà del tessuto. [10]

Analogamente, anche nello studio di Hu et al. [40] gli autori privilegiano l'uso di caratteristiche di texture, ritenute più robuste rispetto a quelle morfologiche, soprattutto a causa delle piccole dimensioni delle MC e del basso contrasto nelle immagini mammografiche.

Diversamente, nello studio di Papadopoulos et al. [11] viene adottato un approccio più completo, che combina caratteristiche morfologiche e di texture. In questo caso, sono state estratte 54 feature che descrivono sia le singole MC sia il cluster nel suo insieme. Le caratteristiche includono proprietà geometriche (area, eccentricità, orientamento, solidità), misure di intensità e contrasto, e informazioni sulla distribuzione spaziale delle MC all'interno del cluster, come la distanza dal centroide e la dispersione. [11]

Un approccio intermedio è proposto da Marasinou et al. [41] in cui sono state estratte 31 caratteristiche divise in due gruppi principali: caratteristiche globali del cluster, come area dell'involucro convesso, eccentricità e orientamento; e caratteristiche delle singole MC, tra cui area, intensità massima, minima e media al loro interno o lunghezza degli assi maggiore e minore. Le caratteristiche individuali vengono poi integrate statisticamente utilizzando la media e la deviazione standard per rappresentare ciascuna ROI. [41]

Infine, il sistema proposto dallo studio di Singh et al. [42] integra sia caratteristiche di forma che di texture, al fine di sfruttare in modo complementare le informazioni morfologiche e strutturali. Le caratteristiche di forma descrivono la morfologia del cluster, trattato come un unico oggetto, e includono inizialmente 14 feature geometriche (come area, perimetro, orientamento, compattezza). A queste si aggiungono 28 caratteristiche ottenute mediante il metodo del centroide gerarchico, basato sul sottocampionamento ricorsivo, che consente una rappresentazione più dettagliatata della struttura del cluster. Inoltre, vengono estratte diverse caratteristiche di texture che descrivono la distribuzione dei livelli di intensità nella regione del cluster. [42]

Nel complesso, questi studi evidenziano come diverse tipologie di feature di primo ordine, di texture e morfologiche forniscano informazioni complementari per la caratterizzazione delle MC. In particolare, mentre le caratteristiche di intensità e texture descrivono le proprietà dei livelli di grigio e la distribuzione dei pixel all'interno della regione, quelle di forma permettono di catturare aspetti geometrici e strutturali sia delle singole MC sia dei cluster. Ciò giustifica l'adozione di approcci integrati, in grado di combinare queste informazioni per ottenere una rappresentazione più completa e migliorare le prestazioni di classificazione.

2.4 Algoritmi di clustering applicabili alle MC

In letteratura sono stati proposti diversi approcci per l'identificazione di cluster di MC, che si differenziano sia per il tipo di informazioni utilizzate nel processo di raggruppamento, sia per il principio di funzionamento.

Un approccio avanzato per lo studio della formazione dei cluster di MC è rappresentato dall'algoritmo Fuzzy-C-Means (FCM). Il FCM standard è un'estensione del K-means e si basa sul principio che ogni elemento, e quindi in questo caso ogni MC può appartenere a più cluster contemporaneamente, con un grado di appartenenza variabile in modo continuo tra 0 e 1. Maggiore è il grado di appartenenza al cluster, maggiore è il livello di confidenza con cui si

ritiene che l'oggetto appartenga a quel cluster. Questo approccio risulta particolarmente utile nel caso delle MC, dove i cluster potrebbero essere contigui o sovrapposti, rendendo difficile l'assegnazione deterministica delle MC ai gruppi. [43]

Una sua variante è rappresentata nello studio di Vivona et al. [43], che utilizza l'algoritmo denominato Fuzzy C-Means implemented with Features (FCM-WF). In questo caso, il clustering non si basa esclusivamente sulla posizione spaziale delle MC, ma su un insieme più ampio di caratteristiche descrittive. In particolare, vengono considerate sette caratteristiche che aggiungono informazioni sulla struttura di ogni MC: tre relative alla forma geometrica (area, perimetro, ed eccentricità, che misura la deviazione di una regione dalla forma circolare perfetta, un valore pari a 0 indica una geometria circolare, mentre valori crescenti indicano una progressiva alterazione della circolarità), e quattro legate alla distribuzione dell'intensità dei pixel (media, deviazione standard, curtosi ed eccentricità dell'intensità, che descrive l'asimmetria della distribuzione dei livelli di grigio). In questo modo la clusterizzazione viene eseguita in uno spazio multidimensionale delle feature, anziché nel solo spazio euclideo delle coordinate spaziali. Le caratteristiche vengono inoltre normalizzate nell'intervallo [0,1], così da garantire un contributo equilibrato di ciascuna variabile nel processo di clustering. [43]

Un limite noto dell'algoritmo FCM è la necessità di specificare a priori il numero di cluster. Per affrontare questo problema, nello studio seguente viene adottato un approccio iterativo basato sull'ipotesi che ogni cluster debba contenere almeno tre MC, consentendo così una stima automatica del numero di cluster senza fissarlo esplicitamente a priori. [43]

Nonostante questi miglioramenti, la necessità di definire (anche indirettamente) il numero di cluster rappresenta ancora una criticità. Per questo motivo, in letteratura sono stati proposti metodi alternativi basati sulla distanza e sulla densità, che non richiedono questa informazione a priori.

Tra questi, uno degli algoritmi più utilizzati è DBSCAN, adottato nello studio Maya et al. [44] Questo metodo raggruppa le MC in base alla loro densità spaziale, in particolare due MC vengono considerate appartenenti allo stesso cluster se la loro distanza è inferiore a una soglia prefissata (nel caso specifico posta a 4.1 mm), e se il numero minimo di elementi in una regione densa supera un valore stabilito (impostato nello studio a 4). Le MC isolate vengono invece classificate come rumore ed escluse dal clustering. Questo approccio presenta diversi vantaggi tra cui: non richiede la definizione a priori del numero di cluster, è robusto alla presenza di

rumore e consente di individuare cluster di forma arbitraria, caratteristiche particolarmente rilevanti nel caso delle MC, che possono presentare distribuzioni spaziali irregolari. [44]

2.5 Modelli di classificazione

In letteratura, diversi studi si sono occupati della classificazione delle MC a partire da feature descrittive, utilizzando principalmente approcci di machine learning supervisionato. I modelli più frequentemente impiegati includono SVM, RF, metodi di boosting e reti neurali artificiali, con prestazioni generalmente promettenti.

Nel lavoro di Prinzi et. Al [10] sono stati confrontati diversi classificatori, tra cui SVM (con kernel radiale), RF e XGBoost. In questo studio non viene affrontato un problema di classificazione binario ma multi-classe a 3 livelli: sano, benigno e maligno. Nello specifico, la classe sano comprende le ROI che, a seguito dell'analisi delle caratteristiche radiomiche vengono riconosciute come tessuto normale (assenza di MC). Con benigno si intendono le ROI contenenti MC e di tipo benigno. Con maligno si intendono le ROI contenenti MC e di tipo maligno. Tra i modelli testati, il modello XGBoost ha mostrato le prestazioni migliori. In particolare, utilizzando una strategia di valutazione One-vs-Rest, il modello ha raggiunto valori di AUCROC pari a 0.830 per la classe sano, 0.856 per la classe benigno e 0.876 per la classe maligno. Ciascuno di questi valori esprime la capacità del classificatore di discriminare la classe di riferimento rispetto all'insieme delle restanti. Poiché il modello si basa su caratteristiche radiomiche, questi risultati evidenziano come i metodi di boosting basati su alberi decisionali siano particolarmente efficaci quando applicati a feature radiomiche estratte e organizzate in formato tabellare. [10]

Marasinou et al. [41] invece propone un approccio basato su gradient boosting, in cui le caratteristiche estratte vengono utilizzate per risolvere un problema di classificazione binaria e di conseguenza distinguere i casi tra benigni e maligni. Il metodo ha ottenuto prestazioni in termini di AUCROC pari a 0.763 mostrando un miglioramento rispetto allo stato dell'arte pari a 0.710. Tuttavia, questo studio mostra delle limitazioni, come un numero ancora elevato di falsi positivi e difficoltà legate alla generalizzazione su dataset diversi. [41]

Nel lavoro di Fanizzi et al. [45], la classificazione delle ROI contenenti MC è stata effettuata utilizzando un modello RF, addestrato su feature selezionate. Il modello ha garantito prestazioni molto elevate nella distinzione tra lesioni benigne e maligne, raggiungendo un valore di AUCROC del 92.08% con un'accuratezza del 88.46%. [45]

Uno studio più articolato, invece, è quello di Papadopoulos et al. [11], in cui sono stati confrontati diversi approcci di classificazione, tra cui sistemi basati su regole, reti neurali artificiali (ANN) e SVM. I risultati mostrano che le SVM e le ANN offrono prestazioni migliori rispetto ai sistemi basati su regole. In particolare, le SVM hanno raggiunto valori di AUCROC pari a 0.81, mentre le ANN hanno ottenuto valori leggermente inferiori, arrivando a 0.78. [11]

Infine, nello studio di Singh et al. [25], viene adottato un approccio basato sull'utilizzo del SVM, scelto per la sua capacità di generalizzazione e per le buone prestazioni su dati non visti. In particolare, sono stati testati diversi kernel: lineare, gaussiano e polinomiale ma è stato il kernel lineare quello che ha mostrato i risultati migliori, confermando l'efficacia delle SVM nella classificazione binaria tra cluster benigni e maligni. Inoltre, la combinazione di caratteristiche di forma e texture consente di ottenere prestazioni migliori rispetto all'utilizzo di singole tipologie di feature. Complessivamente, il sistema raggiunge un'accuratezza del 94.25% e un'AUCROC pari a 0.9307**,** dimostrando l'efficacia dell'approccio proposto nella distinzione tra cluster benigni e maligni. [25]

Nel complesso, gli studi analizzati mostrano come diversi modelli di classificazione possano raggiungere prestazioni elevate nella distinzione tra MC benigne e maligne. In particolare, approcci basati su SVM, RF e metodi di boosting si dimostrano efficaci, pur presentando alcune limitazioni, come la presenza di falsi positivi e difficoltà di generalizzazione su dataset eterogenei, evidenziando la necessità di soluzioni sempre più robuste e generalizzabili.

CAPITOLO 3 – MATERIALI E METODI

3.1 Dataset e definizione del problema

Il dataset utilizzato nel presente studio è costituito da immagini mammografiche raccolte presso gli Istituti Clinici Scientifici Maugeri IRCCS di Pavia. Le ROI analizzate sono state ottenute mediante l'approccio DeepMica, già descritto nel Capitolo 2. [34] Nella Tabella 1 è riportata la distribuzione dei dati iniziali e delle ROI estratte.

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
202	398	1251	1090 (87.1%)	161 (12.9%)

Tabella 1: Numero di dati disponibili inizialmente, in termini di pazienti, immagini e numero di ROI ottenute mediante la pipeline DeepMica, con distinzione tra casi benigni e maligni.

In particolare, per ogni paziente possono essere disponibili diverse proiezioni mammografiche: obliqua (OBL), medio-laterale (ML) o cranio-caudale (CC). A sua volta ogni mammografia può essere caratterizzata da più di una ROI, ciascuna classificabile come benigna o maligna, indipendentemente dall'appartenenza alla stessa proiezione o allo stesso soggetto. A ciascuna immagine mammografica del dataset di partenza, quindi, sono state associate una o più maschere, con lo scopo di identificare le ROI contenenti MC da analizzare. Ciascuna mammografia è identificata mediante una codifica composta da un codice numerico associato al paziente e da un suffisso alfabetico che indica la proiezione. Le maschere associate alle immagini sono identificate mediante una codifica coerente con quella delle mammografie di origine, generalmente ottenuta aggiungendo un identificativo numerico al nome del file. Questo consente di stabilire una corrispondenza univoca tra ciascuna immagine e le relative ROI.

In questo contesto, l'obiettivo del presente lavoro consiste nello sviluppo di un modello in grado di classificare automaticamente le ROI sulla base della natura delle MC, distinguendo i casi tra benigni e maligni.

3.2 Definizione della pipeline

Lo studio è stato condotto attraverso una pipeline composta da una serie di step sequenziali, schematizzata nella Figura 14. Il punto di partenza è costituito da immagini mammografiche, ciascuna associata a una maschera. Le maschere indentificano le ROI contenenti le MC e una stessa immagine può contenere più ROI, anche appartenenti a classi diverse (benigna o maligna).

A partire dal dataset a disposizione, sono stati eseguiti i seguenti step principali:

1) Identificazione delle MC
2) Estrazione delle feature
3) Classificazione

Nella prima fase, a partire dalle immagini e dalle relative maschere, sono state individuate le singole MC presenti all'interno delle ROI. Successivamente, ciascuna MC è stata caratterizzata mediante l'estrazione di feature descrittive di varia natura, utilizzate come input per la fase di classificazione. Infatti, attraverso l'addestramento di algoritmi di ML sulla base delle feature estratte, è stato sviluppato un modello in grado di classificare automaticamente le ROI, supportando il processo diagnostico.

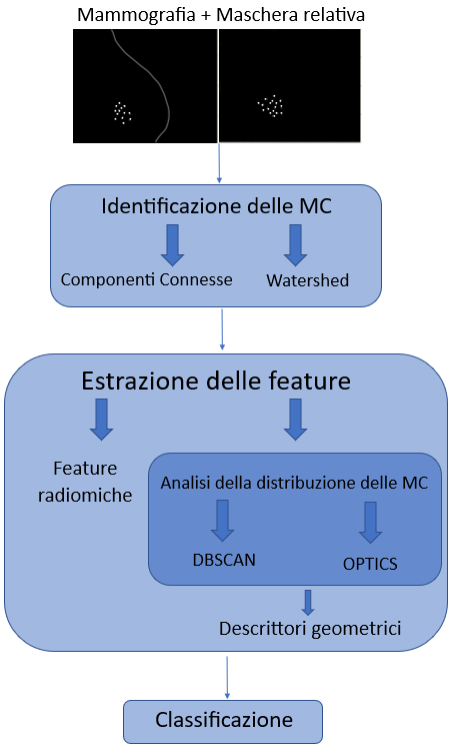

Figura 14: Diagramma di flusso della pipeline di analisi e classificazione delle MC.

3.3 Identificazione delle MC

Poiché le maschere non distinguono tra loro i pixel relativi a ciascun MC, il primo step della pipeline consiste nel separare, al loro interno, i pixel relativi a MC diverse. A tal fine sono stati adottati due metodi:

1) Componenti Connesse
2) Watershed

Il primo metodo, denominato **Componenti Connesse**, si basa sul principio secondo cui pixel adiacenti appartengono allo stesso oggetto. Tuttavia, questo approccio tende a identificare come un'unica componente anche oggetti tra loro adiacenti o parzialmente sovrapposti.

Per superare questa limitazione, è stato utilizzato il metodo **Watershed**, che consente di separare anche regioni connesse, permettendo di distinguere MC vicine che altrimenti verrebbero considerate come un unico oggetto. Tale problematica risulta particolarmente rilevante nel caso di immagini bidimensionali, come le mammografie, in cui la proiezione su un piano può portare a sovrapposizioni apparenti tra MC.

Entrambi i metodi sono stati implementati in ambiente Python 3.7 utilizzando principalmente le librerie SimpleITK, scikit-image, SciPy, NumPy e OpenCV. Per ciascuna MC identificata è stato inoltre calcolato il centroide, espresso in coordinate pixel, successivamente utilizzato per l'analisi della distribuzione spaziale mediante algoritmi di clustering.
A scopo illustrativo, la Figura 15 riporta un esempio di maschera relativa a un caso di studio.

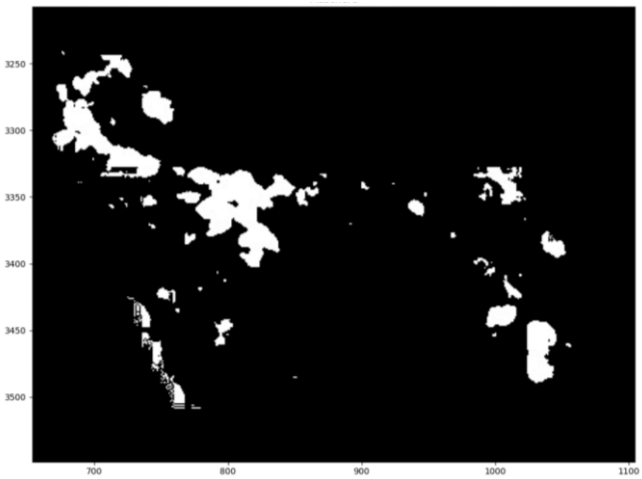

Figura 15: Maschera di segmentazione delle microcalcificazioni, con ingrandimento della regione di interesse.

3.3.1 Metodo delle componenti connesse

Il metodo delle Componenti Connesse consente di identificare i pixel appartenenti alle singole MC sulla base del concetto di vicinanza. Nel presente studio è stata adottata una connettività a 8 pixel, che considera adiacenti anche i pixel disposti lungo le direzioni diagonali, oltre a quelli nelle direzioni orizzontale e verticale (Figura 16).

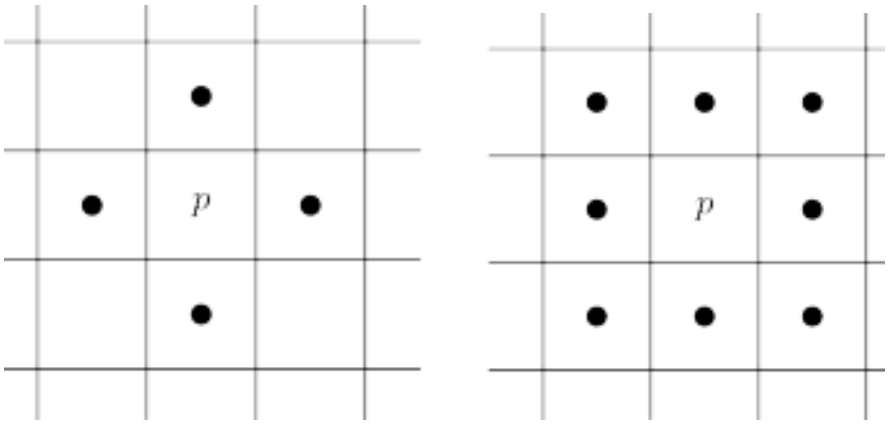

Figura 16: Rappresentazione della connettività a 4 e a 8 pixel. [46]

Nel dataset a disposizione, le maschere associate alle mammografie sono fornite in scala di grigio: i pixel neri rappresentano le regioni non appartenenti alle MC (background della maschera), mentre i pixel bianchi identificano le regioni di interesse contenenti le MC. In una fase preliminare, le maschere sono state binarizzate al fine di separare i pixel appartenenti alle MC dalle regioni restanti.

Successivamente, attraverso una funzione della libreria scikit-image (measure.label), è stato applicato il metodo delle componenti connesse utilizzando una connettività a 8 pixel, con l'obiettivo di identificare ed assegnare un'etichetta univoca a ciascun gruppo di pixel connessi. Come mostrato nella Figura 17, questa operazione produce una mappa di etichette (label map), in cui ogni MC è rappresentata da un identificatore numerico distinto, utilizzato nelle fasi successive per l'estrazione delle feature.

Per migliorare la qualità dell'identificazione, sono state rimosse le componenti con estensione spaziale molto limitata, sulla base del numero di pixel occupati nelle direzioni x e y. In particolare, sono state eliminate le componenti con dimensioni inferiori a 2 pixel lungo entrambe le direzioni, in quanto tali regioni risultano generalmente associate a rumore piuttosto che a vere MC. Regioni di dimensioni così ridotte, infatti, non consentono un'estrazione

affidabile delle feature, in quanto il numero limitato di pixel rende instabili o non significative le caratteristiche estratte, introducendo quindi criticità dal punto di vista computazionale.

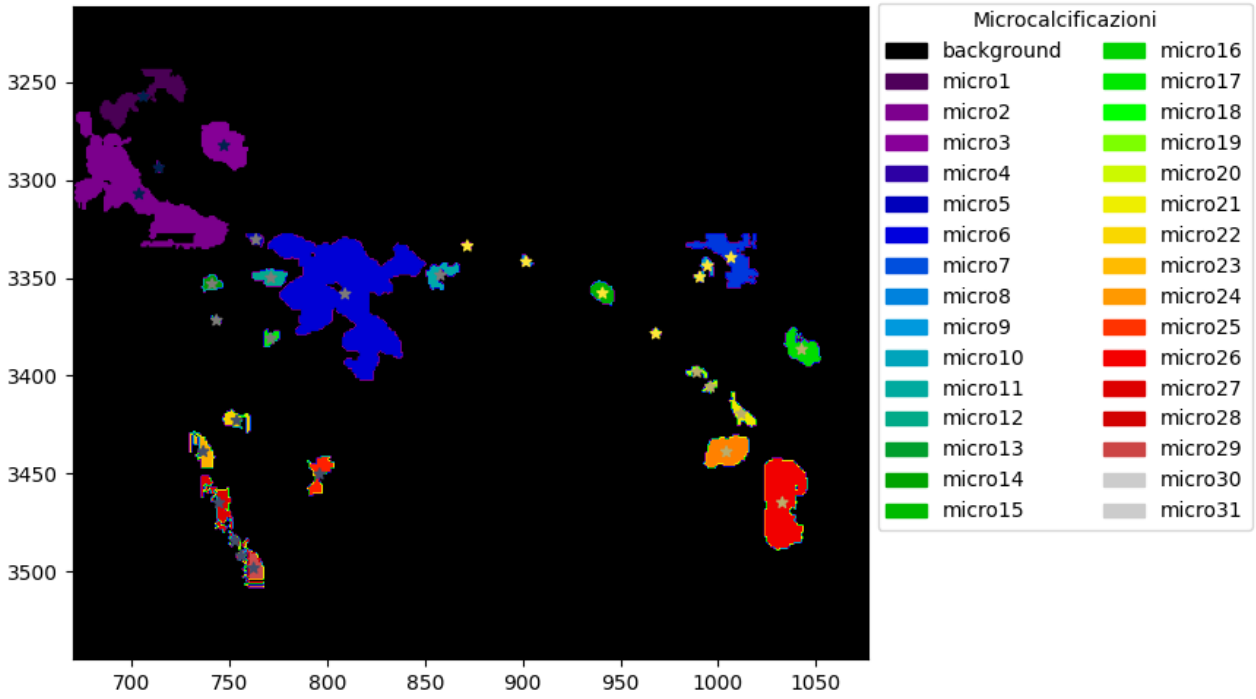

Figura 17: Identificazione delle microcalcificazioni mediante il metodo delle Componenti Connesse.

3.3.2 Metodo Watershed

L'implementazione del metodo Watershed, analogamente a quanto descritto per le Componenti Connesse, parte dalle maschere che, inizialmente in scala di grigio, sono state binarizzate per separare i pixel appartenenti alle MC dalle restanti regioni dell'immagine.

Successivamente, sulla maschera binaria è stata calcolata la trasformata di distanza euclidea, che associa a ciascun pixel appartenente alle regioni connesse la distanza dal bordo più vicino della componente, definito come il confine tra i pixel appartenenti alle MC e il background della maschera. In questo modo, i pixel più interni alle regioni assumono valori di distanza più elevati, mentre quelli prossimi al bordo presentano valori più bassi. Di conseguenza, i centri delle MC corrispondono a massimi locali della mappa delle distanze.

A partire da questa rappresentazione, sono stati individuati i massimi locali mediante l'operatore **h-maxima**, che consente di selezionare solo i picchi sufficientemente distinti dal contesto locale. In questo contesto, è stato introdotto il parametro: h_prominence che definisce

la differenza minima tra il valore di un massimo nella mappa delle distanze e il valore dei pixel circostanti, permettendo di mantenere solo i picchi che emergono in modo significativo rispetto alle regioni vicine.

Valori elevati di questo parametro tendono a ridurre il numero di marcatori, favorendo fenomeni di sotto-segmentazione e l'unione tra regioni adiacenti. Al contrario, valori più bassi aumentano il numero di marcatori, con il rischio di sovra-segmentazione e una maggiore separazione tra strutture contigue.

Poiché il valore ottimale di questo parametro non è noto a priori, esso è stato determinato empiricamente attraverso una serie di prove preliminari, valutando visivamente i risultati dell'identificazione ottenuta per diversi valori su un'immagine di riferimento. Il valore finale è stato selezionato in modo da ottenere un compromesso ottimale tra l'individuazione delle MC e la riduzione dei fenomeni di sovra-segmentazione. In seguito a tale processo, il parametro h_prominence è stato impostato a 2.

I massimi così ottenuti, che approssimano i centri delle MC, sono stati utilizzati come marcatori iniziali per guidare il processo di separazione. I marcatori sono stati etichettati tramite analisi delle componenti connesse con connettività a 8 pixel e forniti in ingresso all'algoritmo Watershed, applicato alla mappa delle distanze con segno invertito.

L'inversione della mappa delle distanze è necessaria poiché l'algoritmo Watershed individua naturalmente i minimi locali: invertendo i valori, i massimi della distanza vengono trasformati in minimi permettendo l'espansione delle regioni a partire dai marcatori fino all'individuazione dei confini tra aree adiacenti. Come illustrato nella Figura 18, questo consente di separare efficacemente strutture contigue appartenenti alla stessa componente connessa.

In assenza di marcatori significativi, la separazione è stata effettuata direttamente mediante analisi delle componenti connesse sulla maschera binaria. Successivamente, come nel caso precedente per migliorare la qualità dell'identificazione, sono state eliminate le componenti con estensione spaziale molto limitata.

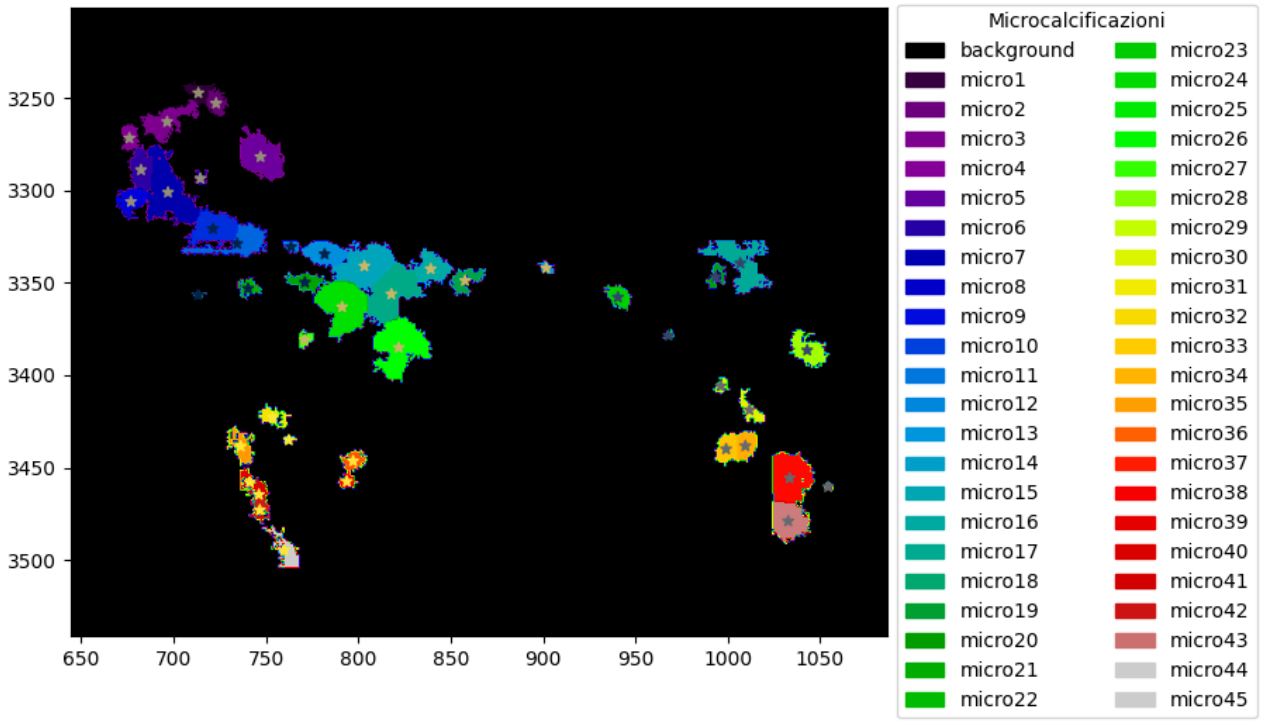

Figura 18: Identificazione delle microcalcificazioni mediante il metodo Watershed.

3.4 Estrazione delle feature

Una volta identificate le singole MC mediante i metodi descritti in precedenza, il secondo step della pipeline consiste nell'estrazione delle feature utilizzate per la successiva classificazione.

In particolare, sono state considerate due tipologie di feature:

1) **feature radiomiche**: volte a descrivere le caratteristiche delle singole MC;
2) **descrittori geometrici**: utilizzati per caratterizzare la distribuzione spaziale delle MC.

È stata quindi effettuata un'analisi di clustering con l'obiettivo di individuare eventuali gruppi di MC, sui quali sono stati successivamente calcolati i descrittori geometrici.

Le feature ricavate a livello di singola MC e di singolo cluster poi sono state aggregate a livello di ROI. In particolare, per ciascuna feature sono stati calcolati, all'interno della ROI, i principali indicatori statistici quali: media, deviazione standard, valore massimo e valore minimo.

Ad esempio, considerando una ROI contenente tre MC, supponendo di estrarre per ciascuna di esse l'area, si ottiene un insieme di tre valori. Tali valori vengono quindi sintetizzati calcolandone media, deviazione standard, massimo e minimo, ottenendo quattro feature

rappresentative della ROI. Lo stesso approccio è stato adottato anche per le feature calcolate a livello di cluster.

Si osserva infine che alcune ROI possono contenere una sola MC o un unico cluster. In tali casi, essendo disponibile un solo valore per ciascuna feature, la variabilità non può essere stimata in modo informativo. Di conseguenza la deviazione standard risulta non definita ed è rappresentata nel dataset come valore nullo (Null).

3.4.1 Feature radiomiche

Per l'estrazione delle feature radiomiche è stata utilizzata la libreria di Python PyRadiomics. Tale libreria consente di calcolare, a partire da un'immagine di riferimento e da una maschera associata, un insieme di feature descrittive relative alla regione selezionata.

Nel presente studio l'estrazione delle feature radiomiche è stata effettuata per ciascuna ROI a livello di singola MC, utilizzando le label map ottenute nella fase di identificazione. L'estrazione è stata eseguita considerando ciascuna etichetta presente nella label map, ottenendo per ogni MC un vettore di feature dedicato. Tali informazioni sono state successivamente aggregate a livello di ROI secondo i criteri descritti nel paragrafo precedente.

Mediante l'utilizzo del pacchetto Python PyRadiomics è possibile estrarre diverse categorie di feature radiomiche. Nel presente lavoro sono state considerate le seguenti famiglie di feature radiomiche:

1) Shape
2) First Order
3) GLCM (Gray Level Co-occurrence Matrix)
4) GLRLM (Gray Level Run Length Matrix)
5) GLSZM (Gray Level Size Zone Matrix)
6) GLDM (Gray Level Dependence Matrix)
7) NGTDM (Neighboring Gray Tone Difference Matrix)

Tra le categorie considerate, le feature di shape e di first order sono state definite nello studio come feature radiomiche di base, in quanto l'informazione fornita da queste è assimilabile a quella interpretabile durante la valutazione radiologica tradizionale.

3.4.2 Analisi della distribuzione delle MC

Oltre allo studio delle caratteristiche delle singole MC è stata inoltre analizzata la distribuzione spaziale delle MC tramite clustering.

L'analisi è stata effettuata utilizzando le coordinate spaziali dei centroidi delle MC, precedentemente calcolati durante la fase di identificazione. Tali coordinate sono state utilizzate come input per gli algoritmi di clustering.

In particolare, per l'identificazione dei gruppi sono stati adottati due metodi di clustering basati sulla densità: DBSCAN e OPTICS, che consentono di individuare insiemi di punti spazialmente vicini e distinguere eventuali MC isolate.

L'algoritmo DBSCAN definisce i cluster come regioni ad alta densità di punti separate da aree a bassa densità. Si basa su due parametri principali: il raggio di vicinato ε (epsilon), e N ovvero il numero minimo di punti richiesto per definire una regione densa. [21] In base a tali criteri, i punti vengono classificati come core points (punto centrale), punti di bordo o rumore, e i cluster vengono costruiti mediante un processo iterativo di espansione a partire dai punti centrali. [22] [23]

Tuttavia, la scelta di un unico valore di ε può risultare limitante in presenza di distribuzioni con densità variabili. Per questo motivo è stato utilizzato anche l'algoritmo OPTICS, che rappresenta un'estensione di DBSCAN. A differenza di quest'ultimo OPTICS, non restituisce direttamente una partizione dei dati, ma produce un ordinamento dei punti basato sulla densità locale, attraverso il calcolo della core-distance (che rappresenta una misura della densità locale attorno a un punto) e della reachability-distance (che quantifica quanto un punto sia accessibile rispetto ai suoi vicini). Tale approccio consente di individuare cluster a diversi livelli di densità, risultando più flessibile nell'analisi di strutture complesse. [23]

Gli algoritmi di clustering sono stati implementati mediante la libreria scikit-learn.

Lo schema della pipeline, che integra i risultati dell'identificazione delle MC con le tecniche adottate per l'analisi della distribuzione delle MC (clustering) è riportato in Figura 19.

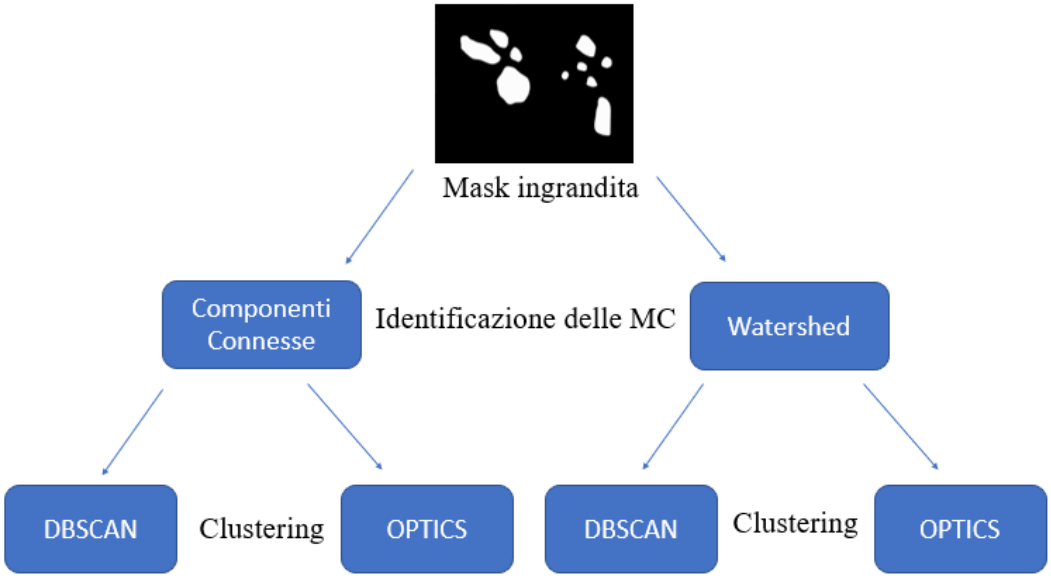

Figura 19: Schema della pipeline: identificazione delle microcalcificazioni (Componenti Connesse, Watershed) e successiva analisi di clustering (DBSCAN, OPTICS).

I metodi di clustering sono stati applicati ai risultati ottenuti da entrambi gli approcci di identificazione delle MC. Di conseguenza, sono state considerate quattro diverse combinazioni per la definizione delle feature, come riportato nella Tabella 2.

Identificazione delle MC	Clustering
Componenti Connesse	DBSCAN
Componenti Connesse	OPTICS
Watershed	DBSCAN
Watershed	OPTICS

Tabella 2: Combinazioni tra metodi di identificazione delle MC e algoritmi di clustering.

3.4.2.1 DBSCAN

Il primo metodo di clustering adottato è DBSCAN, impiegato per individuare gruppi di MC caratterizzati da elevata densità all'interno delle ROI.

In particolare, DBSCAN è stato applicato utilizzando i parametri epsilon e min_samples, quest'ultimo impostato pari a 4. Per quanto riguarda il parametro epsilon, esso è stato definito in pixel. In assenza di metadati DICOM relativi allo spacing delle immagini, non disponibili in

quanto le immagini utilizzate sono in formato JPEG, è stato considerato un valore medio pari a circa 0.085 mm/pixel, utilizzato per convertire la distanza in unità di pixel.

I punti non assegnati ad alcun cluster sono stati automaticamente identificati dall'algoritmo come rumore. Tuttavia, ai fini dell'analisi successiva, tali punti sono stati considerati come cluster singoli, in modo da mantenere informazioni su tutte le MC presenti nella ROI.

Nelle due figure sono riportati i risultati del clustering ottenuti applicando DBSCAN alle MC identificate mediante i metodi di Componenti Connesse (Figura 20) e Watershed (Figura 21). In queste rappresentazioni, le etichette numeriche in rosso indicano l'identificativo univoco assegnato a ciascuna MC identificata.

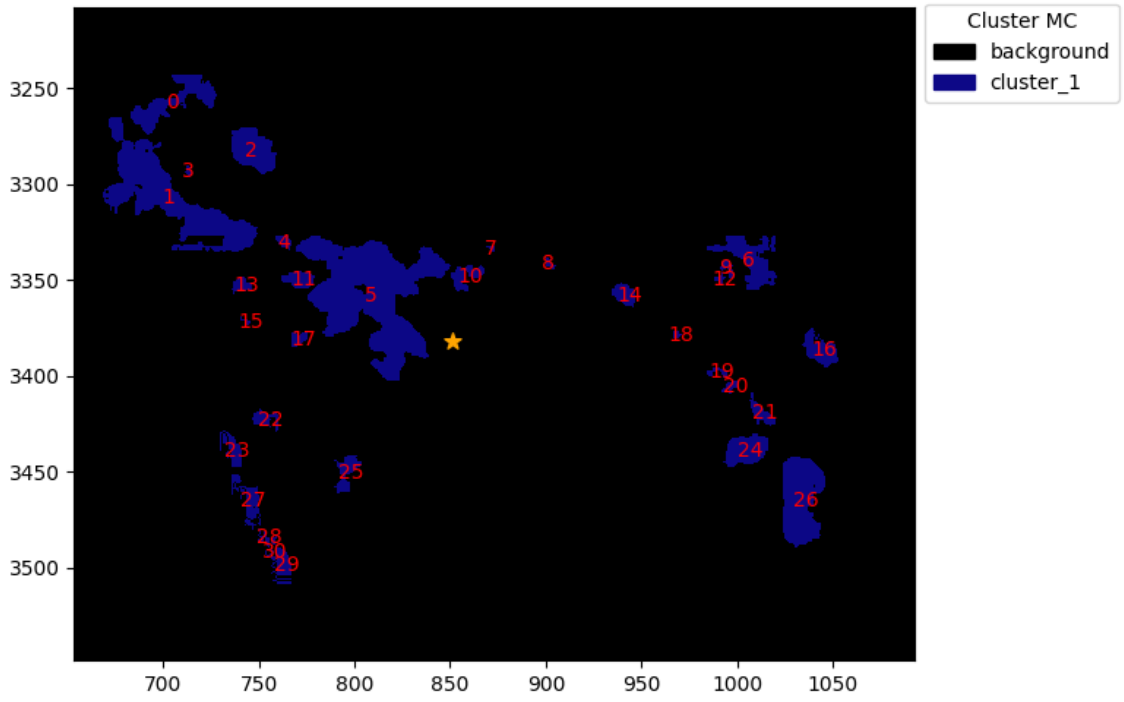

Figura 20: Cluster ottenuti con DBSCAN sulle MC identificate mediante Componenti Connesse.

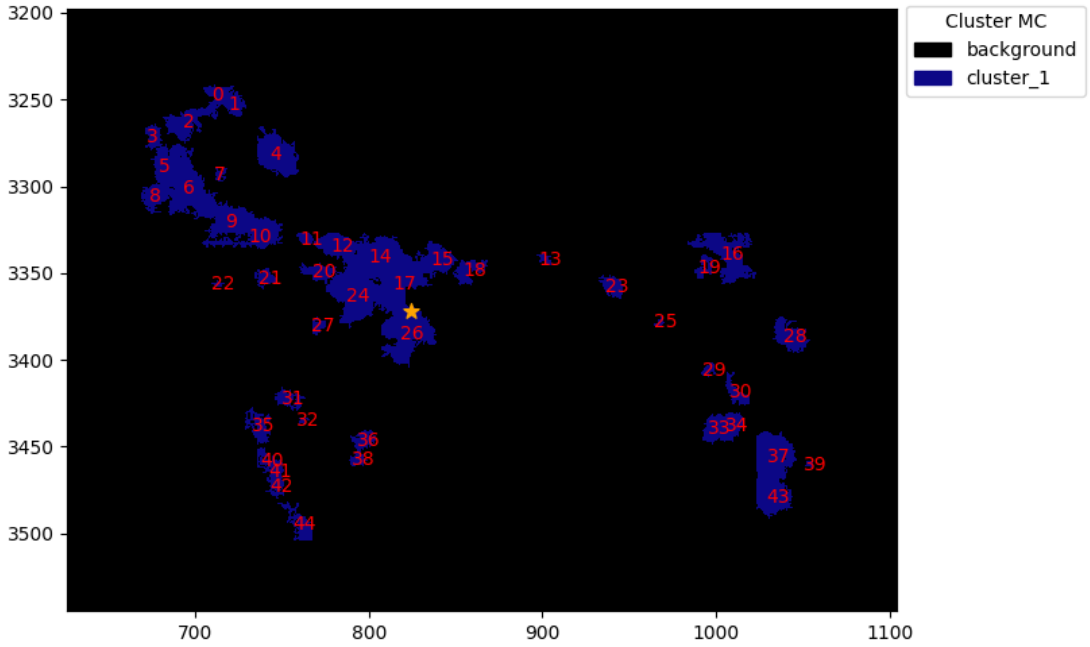

Figura 21: Cluster ottenuti con DBSCAN sulle MC identificate mediante Watershed.

Si osserva che, pur identificando in entrambi i casi un unico cluster, la posizione del suo centroide può variare leggermente tra i risultati ottenuti con Componenti Connesse e Watershed. Tale differenza è dovuta alle diverse modalità di identificazione delle MC adottate dai due metodi, che possono portare a variazioni nella posizione e nel numero delle MC individuate.

3.4.2.2 OPTICS

Il metodo di clustering OPTICS è anch'esso un algoritmo basato sulla densità, adottato in quanto, rispetto a DBSCAN, non richiede la definizione di una soglia di distanza fissa per l'identificazione dei cluster. Questa caratteristica risulta particolarmente vantaggiosa nel presente studio, in cui non sono disponibili informazioni precise sullo spacing delle immagini.

Anche in questo caso, l'algoritmo è stato parametrizzato imponendo come dimensione minima del cluster 4 MC. Nel caso in cui il numero di MC presenti nella ROI fosse inferiore a 4, non essendo soddisfatto il numero minimo di elementi richiesti per la formazione di un cluster, tutte le MC sono state assegnate ad un unico gruppo. I punti non assegnati ad alcun cluster, invece, sono stati identificati automaticamente dall'algoritmo come rumore, tuttavia, come nel caso precedente, tali MC sono state considerate come singoli cluster.

In maniera analoga a DBSCAN nelle figure 22-23 sono riportati i risultati del clustering ottenuti applicando OPTICS alle MC identificate mediante i metodi di Componenti Connesse (Figura 22) e Watershed (Figura 23) con le etichette numeriche in rosso che indicano l'identificativo univoco assegnato a ciascuna MC identificata.

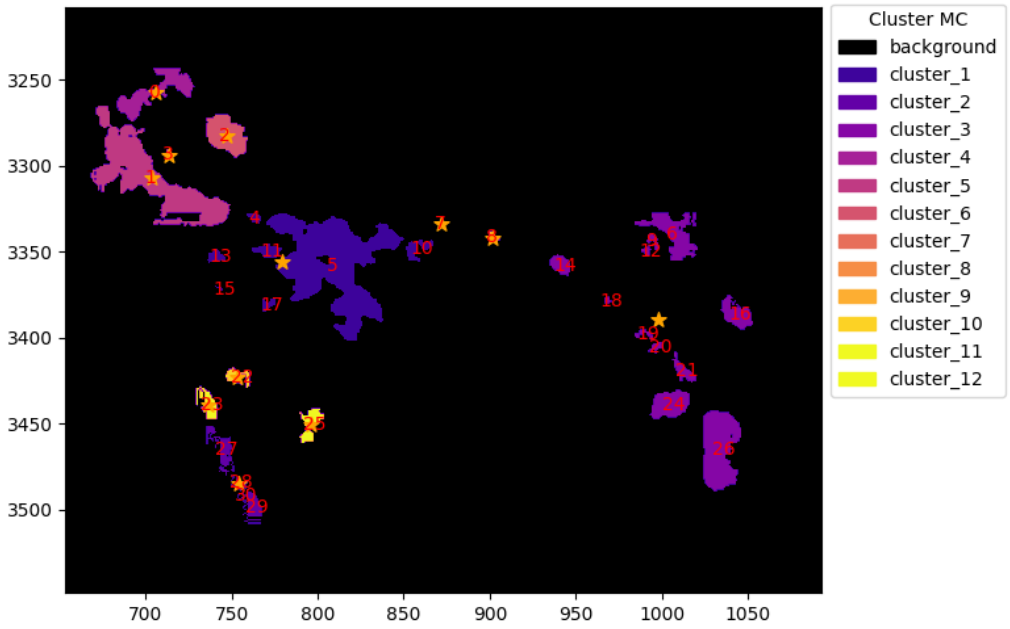

Figura 22: Cluster ottenuti con OPTICS sulle microcalcificazioni identificate mediante Componenti Connesse.

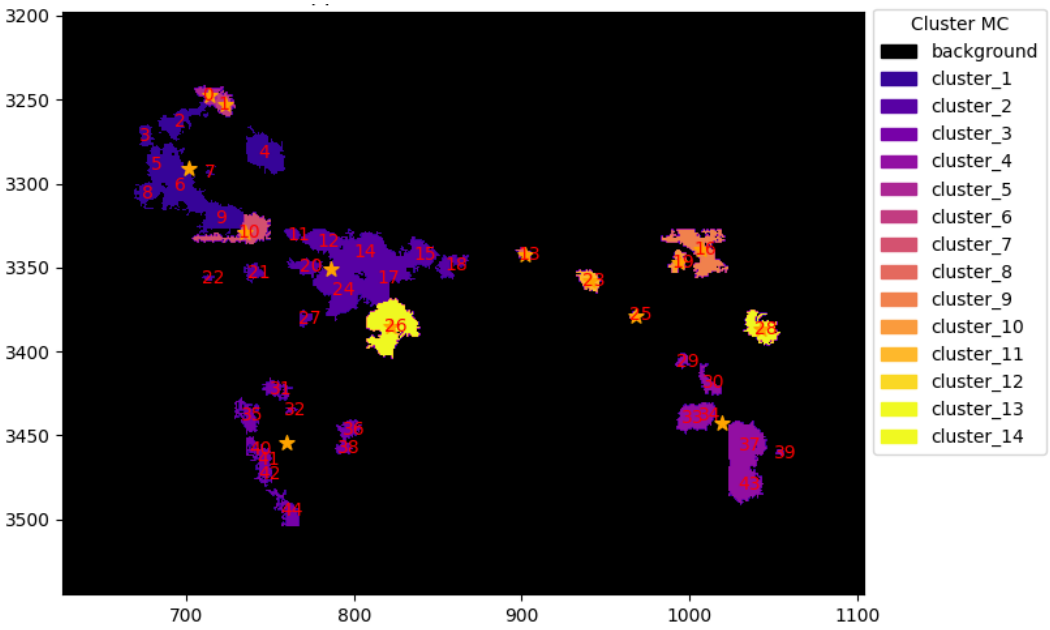

Figura 23: Cluster ottenuti con OPTICS sulle microcalcificazioni identificate mediante Watershed.

3.4.2.3 Descrittori Geometrici

A partire dai cluster individuati mediante i metodi DBSCAN e OPTICS, sono stati estratti descrittori geometrici con l'obiettivo di caratterizzare la distribuzione spaziale delle MC all'interno di ciascun gruppo.

Per ciascun cluster sono stati analizzati due diversi tipi di contorno: un contorno convesso e un contorno concavo. Il contorno convesso è stato ricavato mediante il metodo ConvexHull, che restituisce il più piccolo inviluppo convesso contenente tutti i punti del cluster. Il contorno concavo è stato invece ottenuto mediante il metodo AlphaShape, che consente una rappresentazione più aderente alla reale distribuzione delle MC, preservando eventuali irregolarità del profilo. In tale procedura, viene utilizzato un parametro detto α (alpha) per controllare il grado di concavità del contorno: valori più piccoli producono contorni maggiormente aderenti ai punti, mentre valori più elevati tendono a restituire forme più regolari e prossime all'inviluppo convesso.

Poiché il valore ottimale non è noto a priori, esso è stato determinato empiricamente attraverso prove preliminari condotte su un'immagine di riferimento, valutando visivamente i contorni ottenuti per i diversi valori del parametro. Il valore finale è stato scelto in modo da ottenere un compromesso tra aderenza alla distribuzione spaziale delle MC e stabilità geometrica del contorno. Sulla base di tale procedura, il parametro è stato fissato a $\alpha = 0.01$.

A partire da tali contorni sono stati calcolati due principali descrittori geometrici:

1) Il primo è l'indice di circolarità, definito a partire da area e perimetro del contorno, che rappresenta una misura della compattezza e regolarità della forma del cluster. È definito dalla seguente formula:

$$indice\_circolarità = \frac{4\pi \cdot Area}{perimetro^2}$$

   Il suo valore è compreso pertanto tra 0 e 1: un valore pari a 1 indica che il cluster presenta una forma perfettamente circolare, mentre valori inferiori, prossimi a 0, indicano un basso grado di circolarità e quindi configurazioni più irregolari.

2) Il secondo è il rapporto tra asse maggiore e asse minore, utilizzato come indicatore del livello di elongazione del cluster. È stato ricavato secondo la seguente formula

$$rapporto\_assi = \frac{asse\ maggiore}{asse\ minore}$$

Gli assi principali sono stati stimati mediante analisi delle componenti principali (PCA) applicata ai punti del contorno. In questo modo è stato possibile stimare l'asse maggiore e l'asse minore di ciascun gruppo e calcolarne il rapporto. Tale rapporto assume valori maggiori o uguali ad 1: valori prossimi a 1 indicano una forma più vicina a quella circolare, mentre valori più elevati riflettono una crescente differenza tra le due direzioni principali del cluster, evidenziando una distribuzione spaziale progressivamente meno uniforme e orientata lungo una direzione prevalente.

Poiché i descrittori sono stati calcolati sia sul contorno convesso sia su quello concavo, per ciascun cluster sono stati ottenuti due insiemi di misure geometriche complementari, consentendo di descrivere sia la forma globale sia eventuali irregolarità locali.

Nei casi di ambiguità, caratterizzati da configurazioni geometriche quasi collineari o da una estensione spaziale estremamente limitata (inferiore a 2 pixel), alcuni contorni non risultano definibili in modo affidabile. In tali situazioni, le corrispondenti misure geometriche non sono state considerate.

L'implementazione dei descrittori geometrici è stata realizzata utilizzando le seguenti librerie di Python: SciPy, per il calcolo dell'inviluppo convesso mediante ConvexHull, la libreria alphashape per la determinazione del contorno concavo, Shapely per la gestione delle geometrie e scikit-learn per l'analisi delle componenti principali (PCA).

Nelle figure da 24 a 31 sono riportati esempi rappresentativi dei cluster ottenuti mediante i diversi approcci di identificazione e clustering, evidenziando i contorni convessi e concavi e gli assi principali stimati per ciascun gruppo.

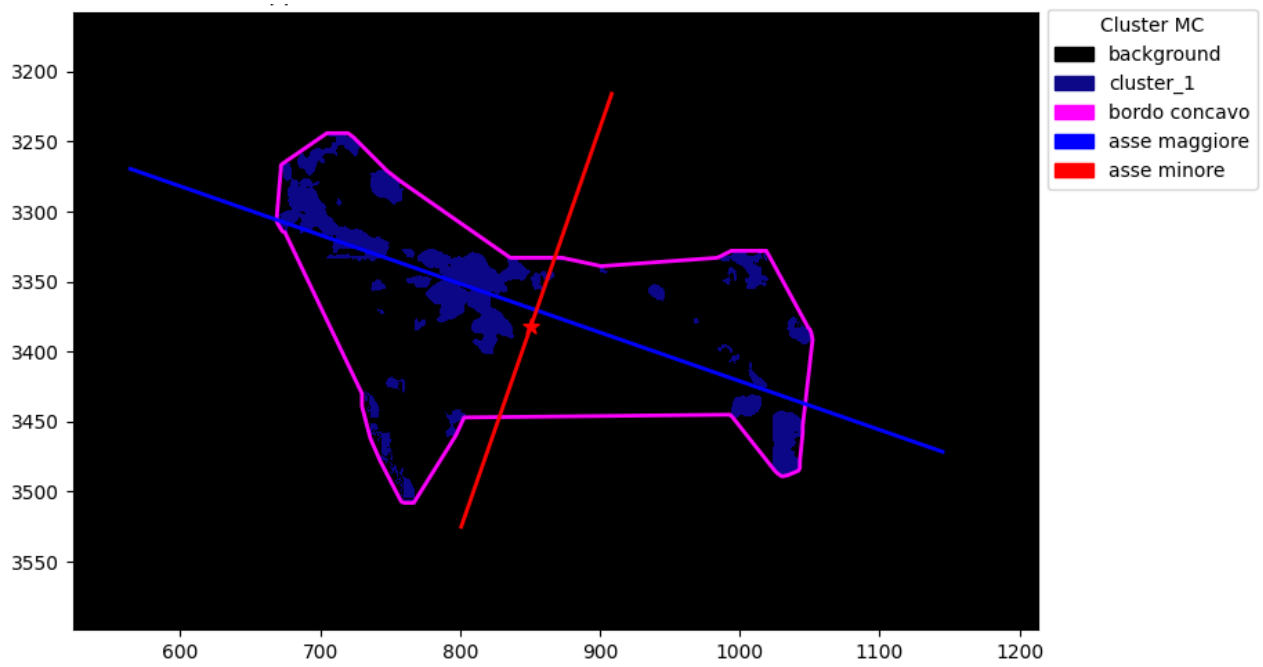

Figura 24: Cluster ottenuti utilizzando il metodo DBSCAN sulle MC identificate con il metodo delle Componenti Connesse, con rappresentazione del contorno concavo e degli assi principali.

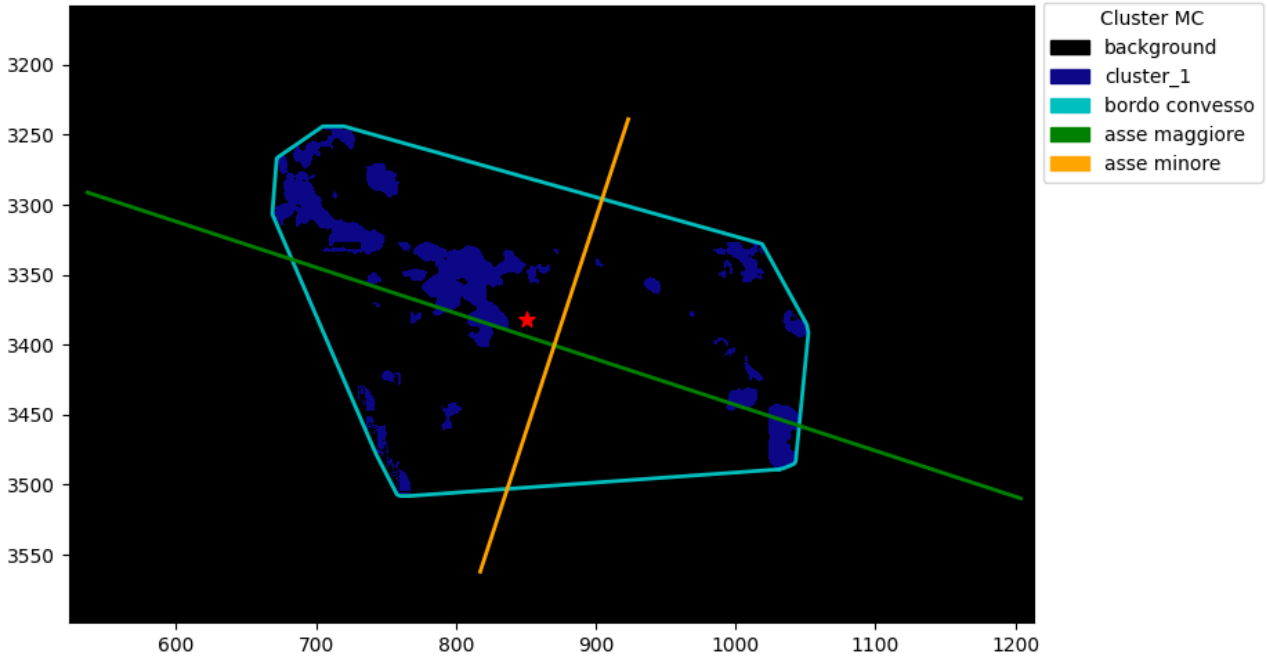

Figura 25: Cluster ottenuti utilizzando il metodo DBSCAN sulle MC identificate con il metodo delle Componenti Connesse, con rappresentazione del contorno convesso e degli assi principali.

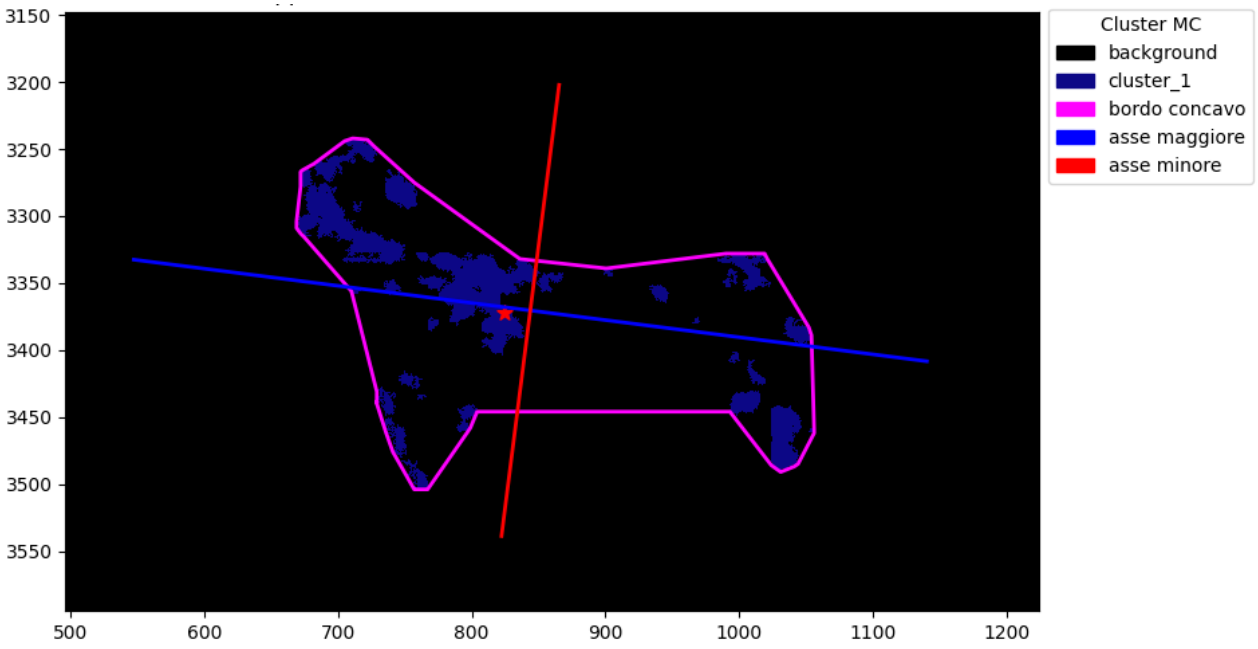

Figura 26: Cluster ottenuti utilizzando il metodo DBSCAN sulle MC identificate con il metodo Watershed, con rappresentazione del contorno concavo e degli assi principali.

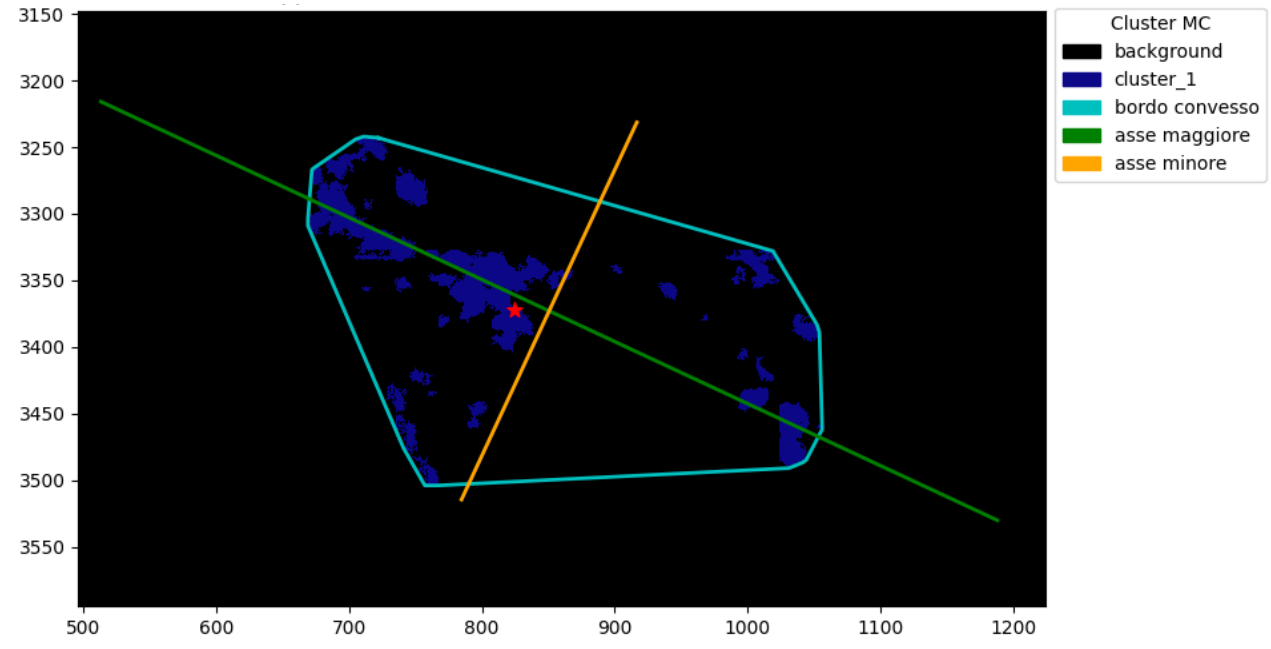

Figura 27: Cluster ottenuti utilizzando il metodo DBSCAN sulle MC identificate con il metodo Watershed, con rappresentazione del contorno convesso e degli assi principali.

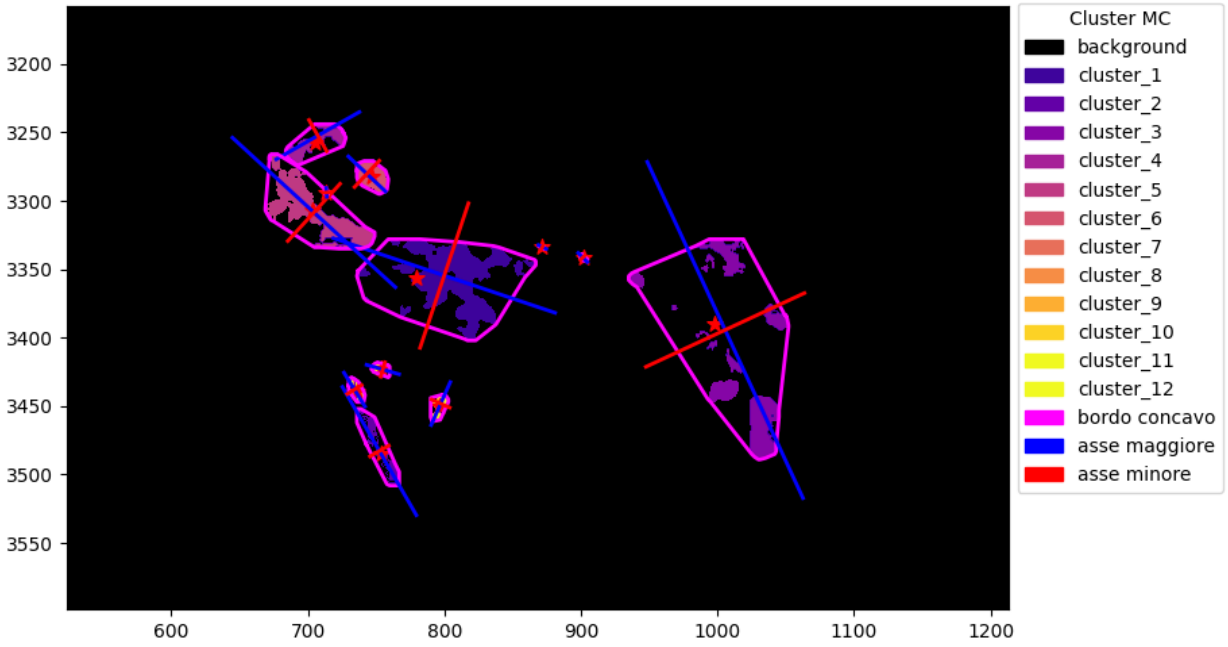

Figura 28: Cluster ottenuti utilizzando il metodo OPTICS sulle MC identificate con il metodo delle Componenti Connesse, con rappresentazione del contorno concavo e degli assi principali.

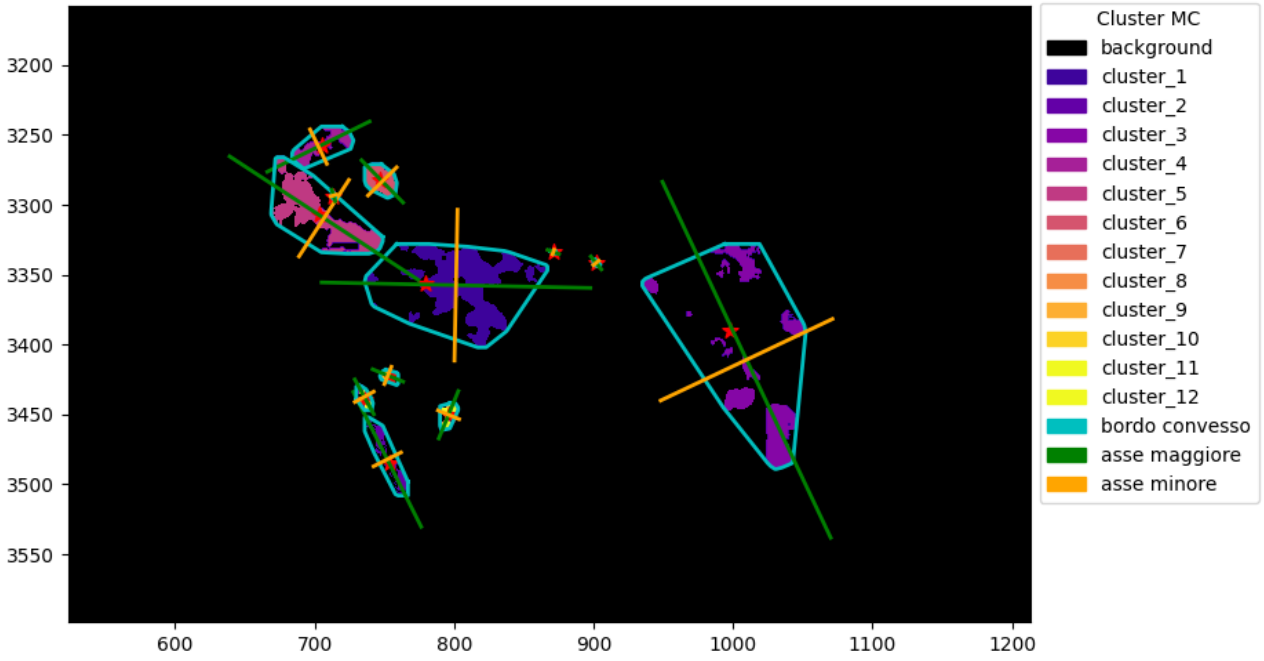

Figura 29: Cluster ottenuti utilizzando il metodo OPTICS sulle MC identificate con il metodo delle Componenti Connesse, con rappresentazione del contorno convesso e degli assi principali.

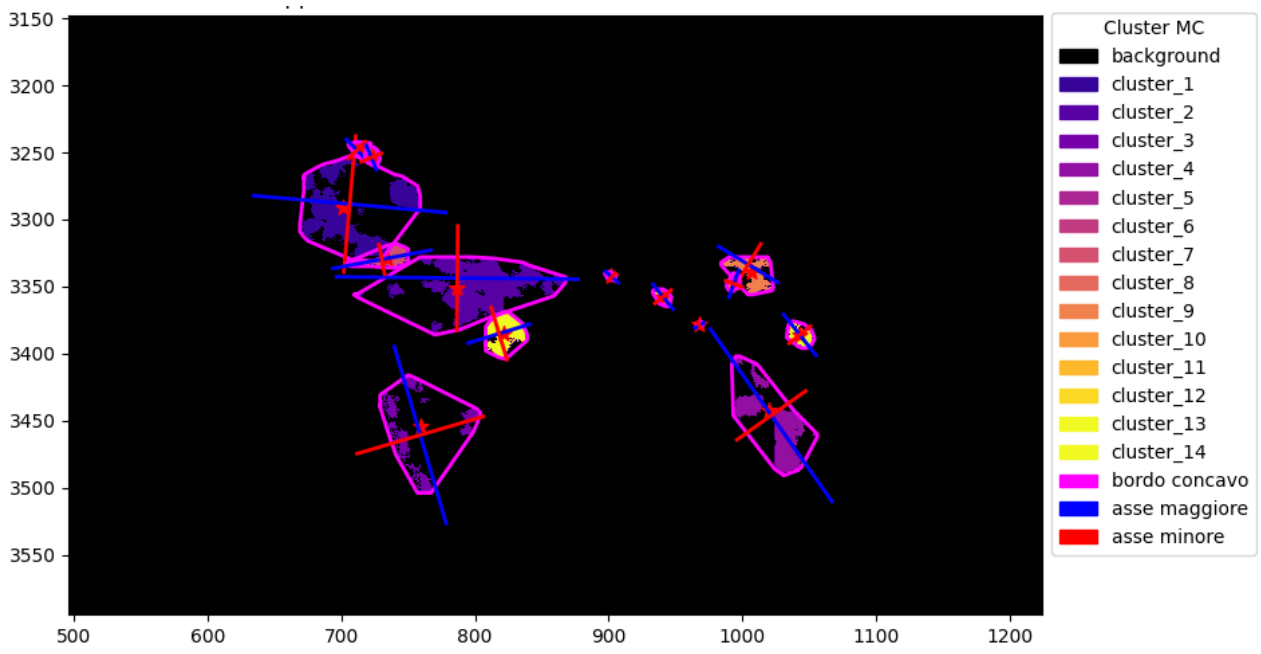

Figura 30: Cluster ottenuti utilizzando il metodo OPTICS sulle MC identificate con il metodo Watershed, con rappresentazione del contorno concavo e degli assi principali.

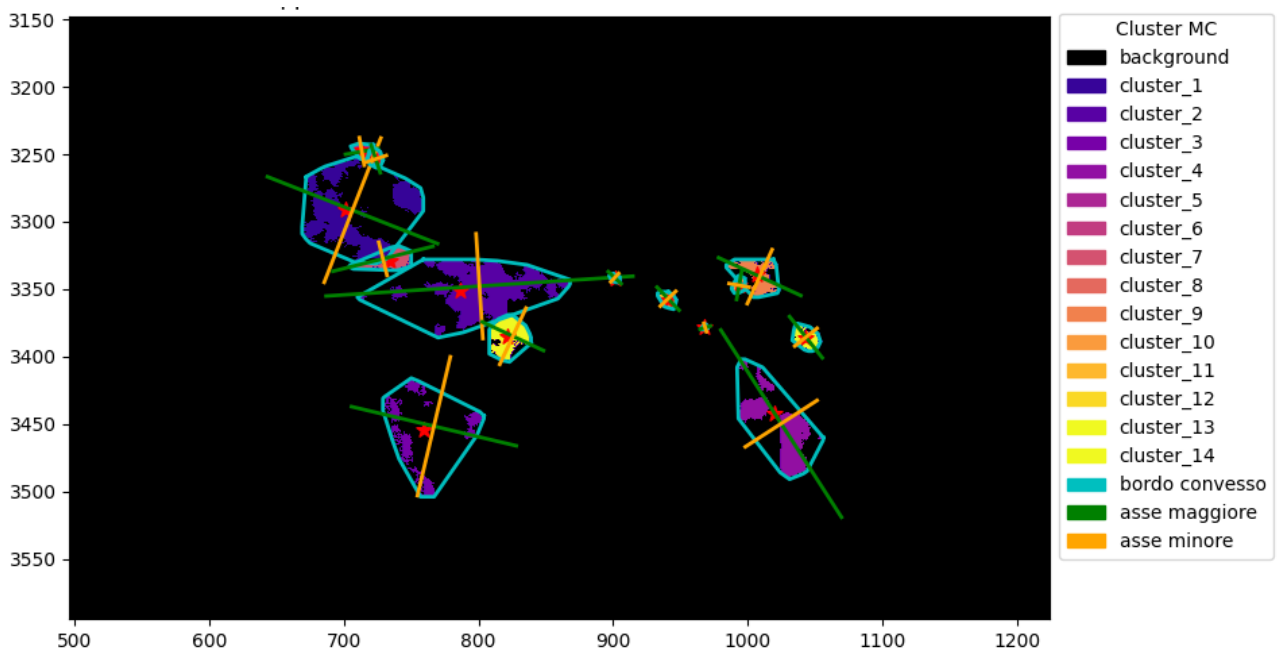

Figura 31: Cluster ottenuti utilizzando il metodo OPTICS sulle MC identificate con il metodo Watershed, con rappresentazione del contorno convesso e degli assi principali.

3.5 Classificazione

Il terzo step della pipeline consiste nell'utilizzare algoritmi di ML per affrontare un problema di classificazione binaria e distinguere le ROI in benigne o maligne. A partire dalle feature estratte nelle fasi precedenti, è stata costruita una matrice dei dati su cui sono stati applicati i modelli di classificazione.

Il processo è stato articolato nelle seguenti fasi:

1) Costruzione della matrice dei dati
2) Pre-processing dei dati
3) Selezione dei modelli di classificazione
4) Definizione della pipeline di training
5) Ottimizzazione degli iperparametri mediante cross-validazione.

Le fasi sopra elencate sono descritte in dettaglio nei paragrafi successivi.

3.5.1 Costruzione della matrice dei dati

La matrice dei dati è stata costruita rappresentando ciascuna ROI mediante le feature radiomiche e i descrittori geometrici estratti dalle MC contenute al suo interno. In particolare, come descritto in precedenza, le feature calcolate a livello di singola MC e di cluster sono state aggregate al fine di ottenere una rappresentazione univoca per ciascuna ROI.

Per ciascuna ROI sono disponibili informazioni relative alla classe (benigna o maligna), al file della mammografia di origine e alla maschera di riferimento. Si osserva, inoltre, che più ROI possono essere associate alla stessa mammografia e che tali ROI possono anche appartenere a classi differenti. In particolare, nelle tabelle 3-4-5 è indicato il numero di dati disponibili inizialmente rispettivamente nel dataset totale, nel training set e nel test set.

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
202	398	1251	1090 (87.1%)	161 (12.9%)

Tabella 3: Numero di dati disponibili inizialmente, in termini di pazienti, immagini, e numero di ROI con distinzione tra casi benigni e maligni nel Dataset totale

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
153	318	1004	876 (87.3%)	128 (12.7%)

Tabella 4: Numero di dati disponibili inizialmente, in termini di pazienti, immagini, e numero di ROI, con distinzione tra casi benigni e maligni nel Training set (80%)

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
49	80	247	214 (86.7%)	33 (13.3%)

Tabella 5: Numero di dati disponibili inizialmente, in termini di pazienti, immagini, e numero di ROI, con distinzione tra casi benigni e maligni nel Test set (20%)

I 5 fold utilizzati durante la procedura di cross-validazione per la stima degli iperparametri sono stati generati in maniera casuale su immagini appartenenti al training set con stratificazione in funzione della classe, facendo in modo che tutte le ROI appartenenti a uno stesso paziente fossero presenti in un solo fold.

Le classi considerate nel problema di classificazione sono:

1) 1: maligno
2) 0: benigno

L'analisi della distribuzione delle classi evidenzia un forte sbilanciamento del dataset, con 1090 ROI benigne (87.1%) e 161 ROI maligne (12.9%). In particolare, il training set presenta 876 ROI benigne (87.3%) e 128 ROI maligne (12.7%), invece nel test set abbiamo 214 ROI benigne

(86.7%) e 33 ROI maligne (13.3%). Inoltre, i fold generati a partire dalle immagini incluse nel training set presentano la distribuzione delle classi mostrata in Figura 32.

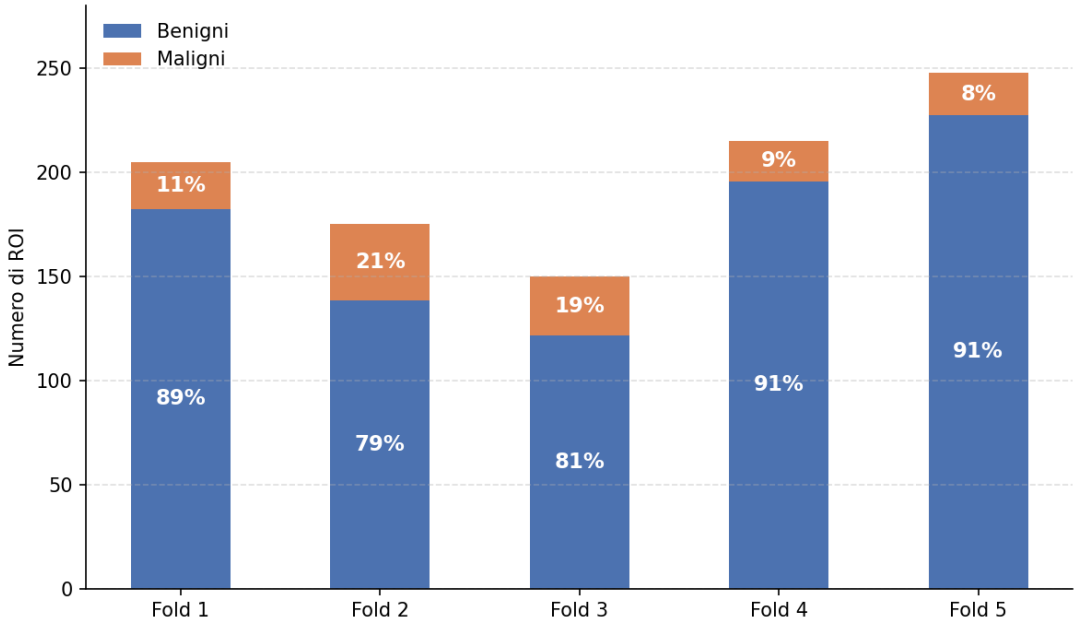

Figura 32: Distribuzione delle classi nei fold ricavati sul training set.

3.5.1.1 Configurazioni delle feature

Per la costruzione della matrice dei dati sono state considerate diverse configurazioni delle feature. In particolare, sono stati analizzati sia insiemi di feature singoli sia combinazioni di essi.

Nel caso delle feature radiomiche, sono stati considerati:

- Un insieme di 108 feature di base, che riassumono aspetti inerenti a quelli considerati durante la valutazione clinica;
- L'insieme completo delle feature radiomiche pari a 408

Nel caso dei descrittori geometrici, sono stati considerati:

- 8 descrittori geometrici derivati da contorni concavi, che permettono una rappresentazione più aderente alla distribuzione spaziale delle MC;
- 8 descrittori geometrici derivati da contorni convessi, che restituiscono una rappresentazione più regolare e globale della forma.

A partire da queste configurazioni di base, sono state infine considerate tutte le possibili combinazioni tra i diversi insiemi di feature come illustrato nelle figure 33 e 34.

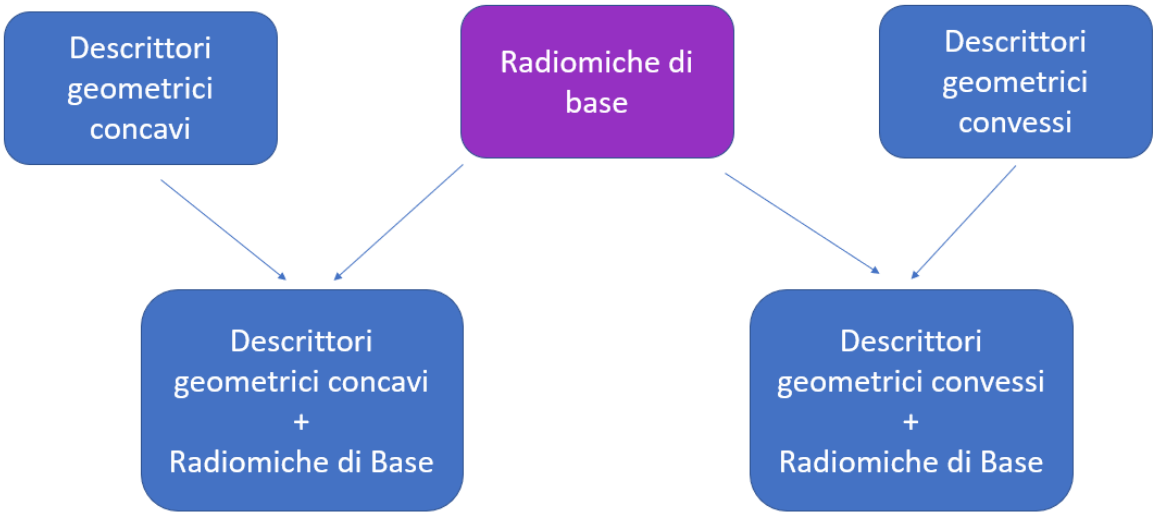

Figura 33: Combinazioni di descrittori geometrici e feature radiomiche di base.

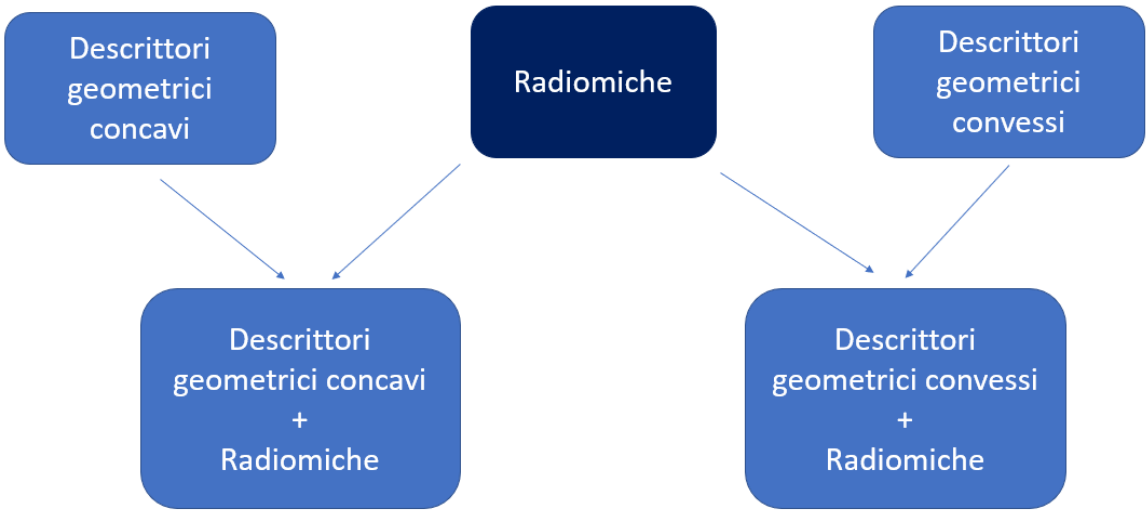

Figura 34: Combinazioni di descrittori geometrici e feature radiomiche.

Le configurazioni di feature illustrate nelle figure 33-34 sono state considerate per ciascuna delle quattro combinazioni definite in precedenza, ottenute combinando i due metodi di identificazione delle MC (Componenti Connesse e Watershed) con i due algoritmi di clustering (DBSCAN e OPTICS).

Il numero di feature varia in funzione della configurazione considerata. Inoltre, come verrà descritto nel paragrafo successivo, la fase di pre-processing comporta una riduzione del numero di feature, dovuta alla rimozione di quelle caratterizzate da un elevata incidenza di valori indefiniti (NaN). Nella tabella 6 è riportato il numero di feature iniziali e il numero di feature dopo tale fase, considerando i singoli insiemi di feature. Le combinazioni tra i diversi insiemi non sono riportate esplicitamente, in quanto il numero totale di feature può essere ricavato dalla somma dei contributi delle singole configurazioni.

Configurazione	Feature iniziali	Feature dopo pre-processing
Radiomiche di base	108	81
Radiomiche	408	306
Descrittori concavi	8	6
Descrittori convessi	8	6

Tabella 6: Confronto tra il numero di feature iniziali e il numero di feature residue dopo la fase di pre-processing.

3.5.2 Pre-processing

Una volta definita la matrice dei dati, è stato eseguito uno step preliminare di pre-processing finalizzato a migliorare la qualità dei dati utilizzati per l'addestramento dei classificatori. Il pre-processing è stato articolato in due fasi principali:

1) Selezione delle feature
2) Pulizia delle ROI

Nel primo step è stata analizzata la presenza di valori mancanti nelle feature estratte. Si è osservato che alcune caratteristiche, in particolare diverse feature relative alla deviazione standard delle misure calcolate a livello di singola MC o di singolo cluster, presentavano una percentuale elevata di dati mancanti. Per questo motivo, tutte le feature con una percentuale di valori mancanti superiore al 30% sono state rimosse dalla matrice dei dati, in quanto considerate non affidabili.

Nel secondo step, è stata analizzata la distribuzione dei valori mancanti a livello di ROI. In particolare, sono state escluse tutte le ROI che presentavano dati mancanti per più del 50% delle feature. Tale soglia è stata scelta in quanto rappresentativa di una condizione in cui la maggior

parte dell'informazione risulta non disponibile, rendendo la ROI poco affidabile ai fini della classificazione. Nel training in esame, tale criterio ha di fatto selezionato ROI in cui tutte le feature risultavano mancanti, riconducibili verosimilmente a casi in cui le MC presenti erano troppo piccole per consentire un'estrazione affidabile delle caratteristiche, rendendo non informativa la descrizione della ROI.

A seguito di questa fase di pre-processing, il training è passato da 1004 a 948 ROI. La distribuzione finale delle classi è pari a 823 ROI benigne (87%) e 125 ROI maligne (13%). È stato inoltre verificato che la procedura di pulizia non alterasse in modo sostanziale la distribuzione delle ROI nei fold predefiniti. Complessivamente, la rimozione delle ROI ha comportato una riduzione limitata del training, con la perdita di 53 ROI benigne e 3 ROI maligne corrispondenti complessivamente a circa il 5% del totale, mantenendo sostanzialmente invariata la distribuzione delle classi e la ripartizione nei fold.

Poiché da uno stesso paziente possono essere ricavate diverse immagine mammografiche che si distinguono per la proiezione (ML, CC, OBL) e una stessa mammografia può contribuire con più ROI, è stata inoltre valutata la riduzione nel dataset totale, nel training set e nel test set anche a livello di paziente e di immagine.

Nel dataset totale (Tabella 7) si è osservato che il numero complessivo di immagini è passato da 398 a 395 e il numero complessivo di pazienti da 202 a 201. In particolare, la rimozione delle ROI ha interessato 47 immagini, determinando la perdita completa di 3 di esse. Invece dal punto di vista dei soggetti nel dataset set la rimozione delle ROI ha interessato 34 soggetti, di cui 1 in possesso di una sola ROI appartenente alla classe benigna è stato perso completamente.

Nel training set (Tabella 8) si è osservato che il numero complessivo di immagini è passato da 318 a 316 e il numero complessivo di pazienti da 153 a 152. In particolare, la rimozione delle ROI ha interessato 39 immagini, determinando la perdita completa di 2 di esse. Invece dal punto di vista dei soggetti nel training set la rimozione delle ROI ha interessato 28 soggetti, di cui 1 in possesso di una sola ROI appartenente alla classe benigna è stato perso completamente.

Nel test set (Tabella 9) si è osservato che il numero complessivo di immagini è passato da 80 a 79 mentre il numero complessivo di pazienti è rimasto invariato e pari a 49. In particolare, la rimozione delle ROI ha interessato 8 immagini, determinando la perdita completa di 1 di esse.

Invece dal punto di vista dei pazienti nel test set la rimozione delle ROI ha interessato 6 soggetti, ma senza la perdita completa di nessuno di essi.

I risultati descritti fanno riferimento a una specifica configurazione di feature e a una specifica combinazione di metodi di identificazione e clustering delle MC, ma risultano coerenti anche per le altre configurazioni considerate. Infatti, anche se il numero e la tipologia di feature possano variare tra le combinazioni, la procedura di pre-processing applicata rimane invariata e produce effetti analoghi in termini di selezione delle feature e delle ROI.

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
201	395	1183	1025 (86.6%)	158 (13.4%)

Tabella 7: Numero di dati disponibili dopo il controllo qualità, in termini di pazienti, immagini, e numero di ROI, con distinzione tra casi benigni e maligni nel Dataset totale.

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
152	316	948	823 (86.9%)	125 (13.1%)

Tabella 8: Numero di dati disponibili dopo il controllo qualità, in termini di pazienti, immagini, e numero di ROI, con distinzione tra casi benigni e maligni nel Training set (80%).

		ROI		
Pazienti	Immagini	Totale	Benigne	Maligne
49	79	235	202 (86.0%)	33 (14.0%)

Tabella 9: Numero di dati disponibili dopo il controllo qualità, in termini di pazienti, immagini, e numero di ROI, con distinzione tra casi benigni e maligni nel Test set (20%).

3.5.3 Modelli di classificazione

Per affrontare il problema di classificazione binaria delle ROI sono stati considerati tre diversi algoritmi di machine learning: RF, SVM e XGBoost. In particolare, il modello RF è stato considerato per la sua capacità di gestire dataset caratterizzati da feature eterogenee e per la sua robustezza rispetto al rumore e all'overfitting. La SVM è stata invece inclusa in quanto particolarmente adatta a problemi di classificazione in spazi ad alta dimensionalità, come quelli derivanti dall'estrazione di un elevato numero di feature radiomiche. Infine, XGBoost è stato selezionato per l'efficacia nel modellizzare relazioni non lineari complesse tra le feature.

L'implementazione dei classificatori è stata realizzata in ambiente Python utilizzando le librerie scikit-learn per Random Forest e SVM e la libreria XGBoost per il modello boosting.

3.5.4 Pipeline e training

Per l'addestramento dei classificatori e la stima degli iperparametri è stata implementata in Python una pipeline che integra le operazioni di pre-processing dei dati e i modelli di classificazione.

In primo luogo, le feature sono state normalizzate mediante RobustScaler, che effettua una standardizzazione basata sulla mediana e sull'intervallo interquartile (IQR), risultando meno sensibile alla presenza di outlier.

In secondo luogo, tenendo conto dello sbilanciamento tra le classi, nei modelli è stata adottata una strategia di bilanciamento dei pesi. In particolare, per RF e SVM è stato utilizzato il parametro class_weight = "balanced". Questa opzione consente di pesare automaticamente le classi in modo inversamente proporzionale alla loro frequenza nel training set, attribuendo un peso maggiore alla classe minoritaria. In questo modo, gli errori commessi sulla classe maligna risultano maggiormente penalizzati durante l'addestramento, migliorando la capacità del modello di identificare i casi più rari.

Per il modello XGBoost, invece, è stato introdotto il parametro scale_pos_weight definito come rapporto tra il numero di ROI benigne e il numero di ROI maligne presenti nel training set.

Questa scelta consente di compensare lo sbilanciamento tra le classi, attribuendo un peso maggiore alla classe minoritaria durante l'ottimizzazione del modello.

Tutte queste operazioni di pre-processing sono state integrate all'interno della pipeline di training, in modo da essere applicate esclusivamente sui dati di training all'interno di ciascun fold di cross-validazione e successivamente ai corrispondenti dati di validazione. Questo approccio consente di prevenire fenomeni di data leakage, evitando che informazioni provenienti dai dati di validazione influenzino il processo di addestramento e a garantire una stima più affidabile delle prestazioni del modello.

Una volta definita la pipeline, è stata effettuata la ricerca degli iperparametri mediante una procedura di grid search, applicata separatamente a ciascun modello.

La pipeline di training è stata implementata utilizzando gli strumenti messi a disposizione dalla libreria scikit-learn.

3.5.5 Ottimizzazione degli iperparametri mediante cross-validazione

La definizione degli iperparametri è stata effettuata mediante una procedura di cross-validazione (CV) a 5 fold. In particolare, il training set è stato suddiviso in 5 sottoinsiemi: a ogni iterazione, il modello è stato addestrato su 4 fold e validato sul fold rimanente.

La ricerca degli iperparametri ottimali è stata implementata mediante GridSearchCV, applicata all'intera pipeline di training. In questo modo, per ciascuna combinazione di iperparametri, tutte le operazioni di pre-processing e la fase di classificazione vengono eseguite all'interno di ogni iterazione della cross-validazione.

In questo contesto, la metrica utilizzata per la valutazione delle prestazioni è stata la PR-AUC, ritenuta più informativa rispetto alla ROC-AUC in questo contesto, in quanto maggiormente sensibile alla capacità del modello di identificare correttamente la classe minoritaria.

Per ciascun modello è stata inizialmente definita una griglia di iperparametri, costruita a partire da una configurazione preliminare e successivamente mantenuta invariata per tutte le combinazioni di feature e metodi di clustering, al fine di garantire coerenza nel confronto tra le diverse configurazioni.

Per ciascuna combinazione di iperparametri, la procedura di cross-validazione restituisce un insieme di valori della metrica considerata, uno per ciascun fold. A partire da tali valori, vengono quindi calcolati la mediana e l'intervallo interquartile (IQR), al fine di fornire una stima robusta della tendenza centrale e della variabilità delle prestazioni.

Le metriche così ottenute sono state infine utilizzate per confrontare in modo consistente le prestazioni dei modelli tra le diverse configurazioni di feature e le differenti combinazioni di metodi di identificazione e clustering delle MC.

3.5.5.1 Ottimizzazione degli iperparametri RF

Per il modello RF è stata definita una griglia iniziale di iperparametri comprendente:

1) **n_estimators:** numero di alberi appresi, che contribuisce a migliorare la stabilità del modello. Un numero maggiore di alberi riduce la varianza ma aumenta il costo computazionale.
2) **max depth**: profondità massima degli alberi decisionali, che controlla la complessità del modello, limitando il numero massimo di split consecutivi lungo ciascun ramo.
3) **min_samples_leaf**: numero minimo di osservazioni per foglia di ogni albero, che evita suddivisioni su pochi dati e contribuisce a ridurre il rischio di overfitting.
4) **max_features**: numero di feature considerate a ogni split in ogni albero. Limitando casualmente le feature disponibili, si ottengono alberi tra loro diversi, riducendo la correlazione e migliorando la capacità di generalizzazione.

I valori considerati sono stati:

5) n_estimators: [150, 300, 450, 600]
6) max_depth: [None, 10, 20]
7) min_samples_leaf: [1, 2, 5]
8) max_features: ["sqrt", "log2"]

Il criterio utilizzato per verificare la conformità della griglia al problema in esame si è basato su un approccio empirico applicato su una configurazione di riferimento, al fine di verificare la robustezza rispetto a variazioni locali dei parametri. La prima ottimizzazione ha individuato come configurazione ottimali i valori: (600, 10, 5 e "log2"), con prestazioni in termini di PR-AUC sui fold pari a:

- media: 0.5615 ± 0.193
- mediana: 0.5658, IQR: 0.17

È stata quindi effettuata un'analisi di sensibilità degli iperparametri, verificando come variazioni locali dei parametri influenzassero la loro selezione e le prestazioni del modello.

1) Variando **n_estimators** nell'intervallo [600, 800, 1000], il modello continuava a selezionare il valore 600, indicando una buona robustezza del parametro.
2) Variando **min_samples_leaf** nell'intervallo [5, 7, 10], veniva selezionato il valore 10, ma senza miglioramenti significativi delle prestazioni e con un aumento della variabilità tra i fold.

In assenza di benefici sostanziali e considerando la maggiore stabilità della configurazione iniziale, si è deciso di mantenere la griglia di partenza, utilizzata per tutte le configurazioni successive.

3.5.5.2 Ottimizzazione degli iperparametri SVM

Per il modello Support Vector Machine è stata utilizzata una funzione kernel di tipo **RBF** (Radial Basis Function), scelta per la sua capacità di modellare relazioni non lineari tra le feature in spazi ad alta dimensionalità, come nel caso delle feature radiomiche. Il kernel RBF consente infatti di trasformare i dati in modo non lineare, rendendo possibile la separazione tra le classi anche in presenza di relazioni complesse tra le feature.

La griglia di iperparametri considerata include:

1) **C:** parametro di regolarizzazione, che controlla il compromesso tra ampiezza del margine e accuratezza sul training set.
2) **gamma:** parametro del kernel che definisce l'influenza dei singoli punti. Valori elevati considerano solo i punti molto vicini, producendo frontiere più complesse, mentre valori più bassi portano a decision boundaries più lisce e globali.

I valori esplorati sono stati:

1) **C:** [0.1, 1, 10, 100]
2) **gamma:** ["auto"," scale"]

Nel primo caso, il valore di gamma viene impostato automaticamente come l'inverso del numero di feature, senza tenere conto della distribuzione dei dati. Nel secondo caso, invece, gamma viene calcolato considerando anche la varianza delle feature, in modo da riflettere meglio la distribuzione dei dati.

In particolare, la configurazione iniziale ha selezionato i parametri C = 0.1 e gamma = "auto" con prestazioni in termini di PR-AUC sui fold pari a:

- media 0.6 ± 0.146
- mediana: 0.66, IQR: 0.188

Poiché il valore ottimale di C si trovava al limite inferiore della griglia, sono state effettuate ulteriori prove estendendo l'intervallo verso valori più piccoli (0.01 e 0.05). Tuttavia, in entrambi i casi la configurazione selezionata e le prestazioni sono rimaste invariate.

Ciò ha confermato la robustezza della griglia iniziale, che è stata quindi mantenuta per tutte le configurazioni successive.

3.5.5.3 Ottimizzazione degli iperparametri del modello XGBoost

Per il modello XGBoost è stata definita una griglia iniziale comprendente i principali parametri che regolano la complessità e la capacità di generalizzazione del modello:

1) **n_estimators:** numero di alberi costruiti. Valori maggiori migliorano la capacità di apprendimento ma aumentano il rischio di overfitting.
2) **learning_rate:** parametro che controlla il contributo di ciascun albero al modello finale.
3) **max_depth:** profondità degli alberi, che regola la complessità del modello limitando il numero di split consecutivi lungo ciascun ramo
4) **min_child_weight:** numero minimo di campioni richiesto in un nodo. Valori più elevati riducono il rischio di overfitting.
5) **subsample:** frazione di campioni utilizzati per ciascun albero, che introduce casualità nel processo di apprendimento e migliora la generalizzazione del modello.

I valori considerati sono stati:

1) n_estimators: [200, 400, 600]
2) learning_rate: [0.03, 0.05, 0.1]

3) max depth: [3, 4, 5]
4) min_child_weight: [1, 3, 5]
5) subsample: [0.8, 1.0]

La configurazione ottimale iniziale è risultata: (200, 0.03, 3, 5, 1.0) con prestazioni in termini di PR-AUC sui fold pari a:

- media: 0.56 ± 0.147
- mediana: 0.63, IQR: 0.205.

Successivamente, è stata valutata l'introduzione di ulteriori parametri per verificare possibili miglioramenti:

1) **colsample_bytree**: frazione di feature utilizzate per la costruzione di ciascun albero, usato per valutare l'effetto del campionamento delle feature introducendo variabilità e potenzialmente migliorare la capacità di generalizzazione, con valori [0.6, 0.8, 1.0]
2) **reg_alpha e reg_lambda:** parametri di regolarizzazione che contribuiscono a ridurre l'overfitting, con valori rispettivamente di [0, 0.5] e [1, 5].

Tuttavia, tali estensioni non hanno prodotto miglioramenti significativi in termini di PR-AUC. Pertanto, si è deciso di mantenere la griglia iniziale, ritenuta adeguata in termini di prestazioni e complessità.

3.5.6 Variazione della soglia decisionale

Nella fase iniziale, la classificazione delle ROI è stata effettuata utilizzando la soglia decisionale standard pari a 0.5 sulle probabilità della classe positiva. Tuttavia, data la natura fortemente sbilanciata del dataset, tale soglia potrebbe non essere ottimale in termini di compromesso tra precisione e recall.

In particolare, l'analisi preliminare delle prestazioni ha evidenziato come alcuni modelli presentassero uno sbilanciamento tra precision e recall, risultando in alcuni casi troppo conservativi (valore elevato di precision, valore basso di recall) o, viceversa, troppo permissivi (valore elevato di recall, valore basso di precision). Per questo motivo, è stata introdotta una procedura di ottimizzazione della soglia decisionale.

A tal fine, per ciascun modello sono state considerate le probabilità predette sui fold di validazione (out of fold predictions) raccolte e utilizzate per stimare la soglia ottimale. In particolare, i valori di probabilità sono stati ordinati in ordine crescente e ciascuno di essi è stato considerato come possibile soglia decisionale. Per ogni soglia candidato, è stata calcolata la metrica F_β con il parametro β che consente di modulare l'importanza relativa tra precisione e recall. Il valore della metrica F_β è compreso tra 0 e 1, dove valori prossimi a 1 indicano un miglior compromesso tra precisione e recall, mentre valori prossimi a 0 indicano prestazioni scarse. Per ciascun valore di beta, è stata selezionata la soglia che massimizza il valore della metrica F_β. In particolare, sono stati considerati tre casi:

1) $\boldsymbol{\beta = 1}$: corrispondente alla metrica F1-score, che assegna uguale importanza a precisione e recall.
2) $\boldsymbol{\beta = 2}$: che attribuisce maggiore peso al recall, penalizzando maggiormente i falsi negativi (FN).
3) $\boldsymbol{\beta = 0.5}$: che attribuisce maggiore peso alla precision, penalizzando maggiormente i falsi positivi (FP).

In questo modo è stato possibile analizzare il comportamento dei modelli in scenari differenti, privilegiando alternativamente la capacità di individuare i casi maligni (riduzione dei FN) oppure la riduzione dei falsi allarmi (riduzione dei FP).

Le soglie ottimali così ottenute sono state successivamente applicate al set di test, al fine di valutare l'impatto della variazione della soglia decisionale sulle prestazioni finali dei modelli.

3.5.7 Selezione del modello finale

La selezione del modello finale è stata effettuata sulla base delle prestazioni ottenute sul training set mediante cross-validazione, al fine di evitare che il test set influenzasse la scelta del classificatore. In particolare, i modelli sono stati confrontati considerando come metrica principale la PR-AUC mediana sui fold di validazione, affiancata dalla misura di dispersione intervallo interquartile, in modo da valutare sia la performance media sia la stabilità del modello.

Poiché, nei problemi con classi sbilanciate, la soglia decisionale può influenzare il compromesso tra precisione e recall, sono state analizzate le prestazioni dei modelli anche al variare della soglia. Tuttavia, la scelta finale non si è basata unicamente su questo aspetto, ma

ha considerato complessivamente il comportamento del modello. In particolare, è stata selezionata la combinazione di classificatore e configurazione delle feature che mostrava buone capacità discriminative e prestazioni stabili sul training set, verificando che tali proprietà si mantenessero ragionevolmente consistenti anche al variare della soglia decisionale, su dati di test, al fine di valutarne la capacità di generalizzazione.

3.5.8 Explainability: Metodo SHAP

Una volta costruito il modello di classificazione e valutato sul test set, è stata effettuata un'analisi di explainability, con l'obiettivo di interpretare il contributo delle singole feature alle decisioni del modello e verificare la coerenza dei risultati con il significato clinico delle variabili considerate.

A tal fine è stato utilizzato il metodo **SHAP**, che consente di quantificare il contributo di ciascuna feature alla predizione del modello. In particolare, i valori SHAP permettono non solo di identificare le variabili maggiormente influenti, ma anche di valutare la direzione del loro contributo alla classificazione, evidenziando se determinati valori della feature contribuiscano ad aumentare o ridurre la probabilità di classificazione nella classe maligna.

L'analisi è stata effettuata sui dati appartenenti al test set dopo l'applicazione degli stessi step di pre-processing utilizzati nella pipeline di training. Ciò ha consentito di interpretare il contributo delle feature maggiormente influenti nelle decisioni del modello.

L'analisi di explainability è stata implementata utilizzando la libreria Python SHAP.

CAPITOLO 4 – RISULTATI

In questo capitolo vengono illustrati i risultati sperimentali ottenuti. In particolare, sono state valutate diverse configurazioni di feature, includendo sia descrittori geometrici (concavi e convessi), sia feature radiomiche complete e di base, nonché le loro combinazioni. Tuttavia, per chiarezza espositiva, nel seguito vengono riportati i risultati relativi alle configurazioni più rappresentative, selezionate sulla base delle prestazioni ottenute e della loro rilevanza ai fini interpretativi.

In particolare, l'analisi si concentra su tre configurazioni principali: i descrittori geometrici basati su contorni concavi, le feature radiomiche di base e la loro combinazione. Tale scelta consente di valutare il contributo delle informazioni geometriche e radiomiche sia singolarmente sia in maniera integrata, al fine di verificare se la loro combinazione permetta una caratterizzazione più completa delle ROI.

Le configurazioni selezionate sono state analizzate considerando tutte le combinazioni tra i metodi di identificazione delle MC (Componenti Connesse e Watershed) e gli algoritmi di clustering (DBSCAN e OPTICS), in modo da valutare la robustezza dei risultati rispetto alle diverse strategie adottate nella pipeline.

I risultati sono presentati distinguendo tra le prestazioni ottenute in fase di training mediante cross-validazione e quelle osservate sul test set.

4.1 Risultati derivanti dall'analisi del training set

In questa sezione vengono analizzate le prestazioni dei modelli in fase di training, valutate mediante una procedura di cross-validazione a 5 fold. Per ciascuna configurazione di feature, le prestazioni sono state riassunte in termini di mediana e intervallo interquartile (IQR) della metrica PR-AUC, così da fornire una stima robusta sia della tendenza centrale sia della variabilità tra i fold.

Per garantire una rappresentazione chiara e coerente dei risultati, le configurazioni di feature sono state analizzate distinguendo tra feature indipendenti e dipendenti dall'analisi di clustering. In particolare, le feature radiomiche di base (Figura 36), che non richiedendo l'applicazione degli algoritmi di clustering, sono state riportate separatamente rispetto alle configurazioni che includono descrittori geometrici (Figura 35). Queste ultime sono state

invece valutate considerando tutte le combinazioni tra metodi di identificazione delle MC (Componenti Connesse e Watershed) e algoritmi di clustering (DBSCAN e OPTICS).

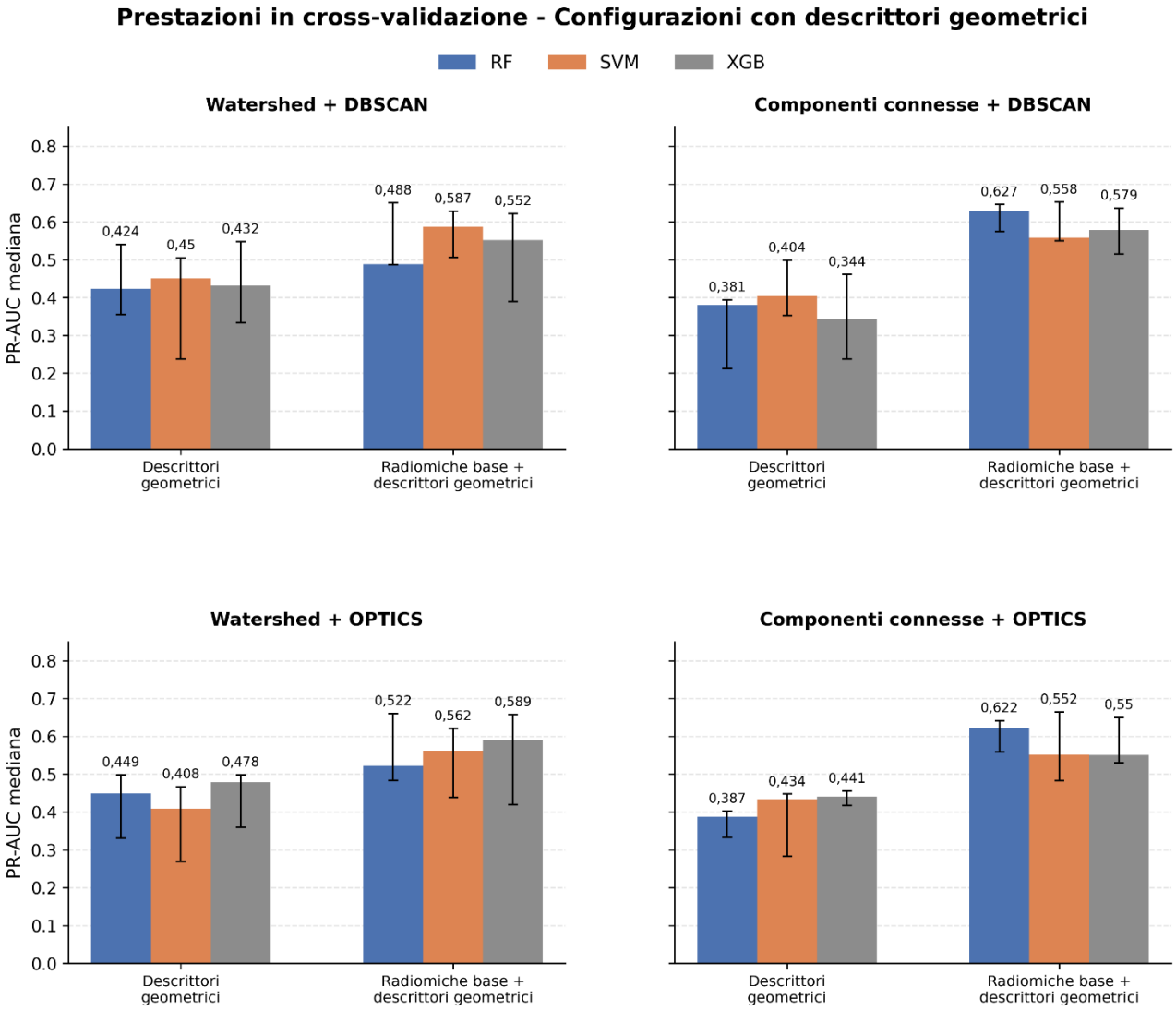

Figura 35: Risultati sul training set per la configurazione di feature relativa ai soli descrittori concavi e ai descrittori geometrici di concavità combinati con le feature radiomiche di base. Le colonne colorate mostrano il valore mediano della PR-AUC per ciascun classificatore, mentre le barre d'errore rappresentano la variabilità interquartile (Q1-Q3).

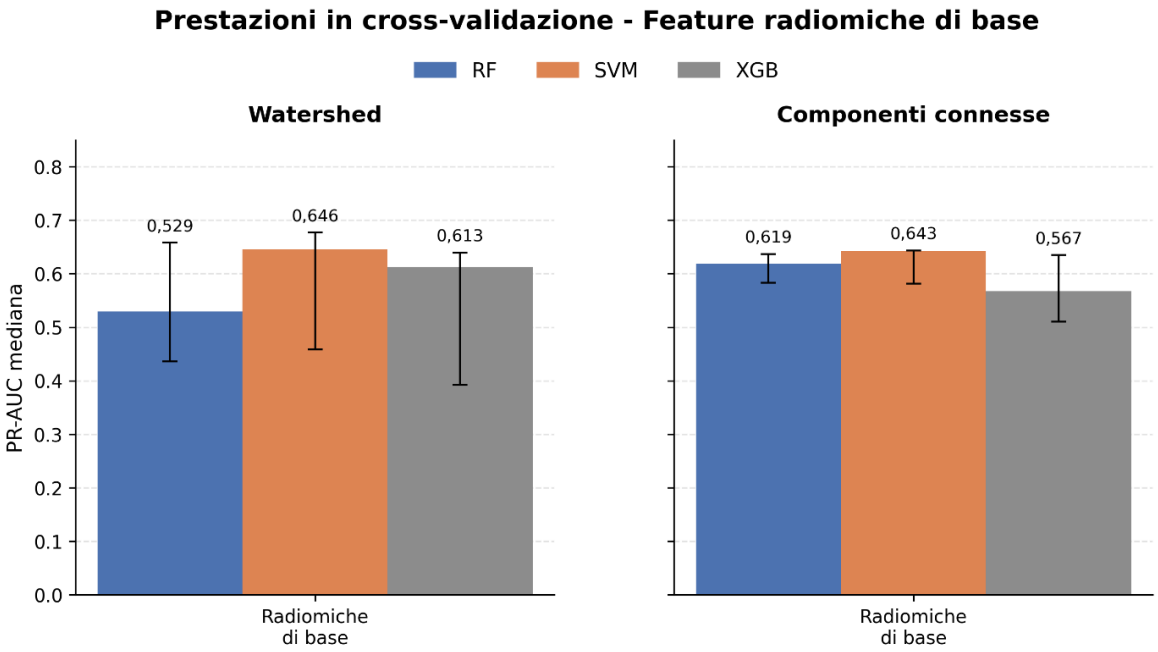

Figura 36: Risultati sul training set per la configurazione di feature relativa alle sole feature radiomiche di base. Le colonne colorate mostrano il valore mediano della PR-AUC per ciascun classificatore, mentre le barre d'errore rappresentano la variabilità interquartile (Q1-Q3).

4.1.1 Analisi dei risultati derivanti dall'analisi del training set

Le figure 35-36 mostrano le prestazioni in termini di PR-AUC mediana e IQR per le diverse configurazioni di feature considerate. In particolare, viene riportato il confronto tra i tre classificatori (RF, SVM e XGBoost) per ciascuna combinazione di metodo di identificazione delle MC e algoritmo di clustering.

Dall'analisi dei risultati emerge innanzitutto come i descrittori geometrici basati su contorni concavi, se considerati singolarmente, presentino prestazioni significativamente inferiori rispetto alle feature radiomiche di base, indipendentemente dal classificatore utilizzato e dalla combinazione tra metodo di identificazione delle MC e algoritmo di clustering. In particolare, i valori di PR-AUC ottenuti con le sole feature concave risultano sistematicamente più bassi, evidenziando come l'informazione geometrica, se utilizzata singolarmente, non sia sufficiente a discriminare in modo efficace tra ROI benigne e maligne.

Un confronto più significativo riguarda invece l'integrazione tra feature radiomiche di base e descrittori geometrici di concavità rispetto all'utilizzo delle sole feature radiomiche. In questo

caso, l'aggiunta delle informazioni geometriche non porta a un miglioramento sistematico delle prestazioni, ma mostra risultati generalmente comparabili e, in alcune configurazioni, leggermente inferiori. Ad esempio, nel caso Watershed, il classificatore SVM passa da circa 0.646 con le sole radiomiche di base a circa 0.587 nella configurazione radiomiche base + concavi, mentre per XGBoost si osserva una riduzione da circa 0.613 a 0.552. Analogamente, nel caso Componenti Connesse, SVM passa da circa 0.643 a 0.558 e XGBoost da circa 0.567 a circa 0.55.

Questi risultati suggeriscono che le feature radiomiche di base forniscono già una rappresentazione informativa delle ROI e che l'aggiunta dei descrittori geometrici di concavità non introduce un contributo discriminativo rilevante, risultando in alcuni casi ridondante o potenzialmente fonte di rumore. Tuttavia, il fatto che le prestazioni rimangano complessivamente comparabili indica che le informazioni geometriche possono comunque contribuire a descrivere aspetti complementari della distribuzione delle MC, pur senza determinare un miglioramento significativo delle performance in fase di training.

Infine, si osserva come l'influenza del metodo di clustering (DBSCAN vs OPTICS) sia generalmente contenuta, mentre il metodo di identificazione delle MC (Componenti Connesse vs Watershed) può determinare variazioni più evidenti nelle prestazioni, in particolare per alcune combinazioni di feature e classificatori, suggerendo come la fase di identificazione delle MC abbia un impatto più rilevante rispetto alla successiva analisi di clustering.

4.2 Risultati derivanti dall'analisi del test set

In questa sezione vengono analizzate le prestazioni dei modelli sui dati di test, al fine di valutare la capacità di generalizzazione delle diverse configurazioni di feature considerate. In continuità con quanto osservato in fase di training, l'analisi si concentra sulle configurazioni più rappresentative, ovvero le feature radiomiche di base e la loro combinazione con i descrittori geometrici di concavità. Le configurazioni basate esclusivamente su descrittori geometrici non vengono riportate, in quanto hanno mostrato prestazioni significativamente inferiori già in fase di training.

In continuità con quanto definito nell'analisi del training set, le configurazioni che includono descrittori geometrici sono state valutate considerando tutte le combinazioni tra metodi di identificazione delle MC (Componenti Connesse e Watershed) e algoritmi di clustering

(DBSCAN e OPTICS). I risultati sono stati ottenuti utilizzando i tre classificatori considerati, ovvero RF, SVM e XGBoost.

4.2.1 Analisi del test set

Per la valutazione finale dei modelli è stato utilizzato un test set indipendente, composto da 247 ROI totali, di cui 214 benigne (87%) e 33 maligne (13%). Il test set, quindi, presenta una distribuzione delle classi fortemente sbilanciata, analoga a quella osservata nel training set.

Sul test set è stata applicata la stessa procedura di pre-processing definita sul training set, al fine di garantire coerenza tra le fasi di addestramento e valutazione. In particolare, in primo luogo sono state rimosse le feature precedentemente identificate come non affidabili sul training. In secondo luogo, sono state escluse le ROI caratterizzate da un'elevata percentuale di valori mancanti (ROI con informazione non disponibile e che nel caso specifico avevano Nan in tutte le feature). A seguito della pulizia si ottengono i seguenti risultati:

- da 247 a 235 ROI → rimosse 12 ROI benigne (5%)

Le trasformazioni di pre-processing sono state applicate ai dati di test utilizzando esclusivamente i parametri stimati sul training set. In particolare, la normalizzazione mediante RobustScaler è stata applicata utilizzando mediana e IQR stimati sul training set. Questo approccio garantisce una valutazione corretta delle prestazioni del modello.

4.2.2 Metriche di valutazione e criteri di analisi

Le prestazioni dei modelli sono state analizzate principalmente in termini di PR-AUC. A supporto di tale analisi, sono state inoltre considerate ulteriori metriche di valutazione, tra cui: accuratezza, specificità, NPV, precision e recall, in modo di fornire una valutazione completa del comportamento dei modelli.

Per ciascuna configurazione, le prestazioni sono state valutate considerando diverse soglie decisionali applicate alle probabilità predette. In particolare, oltre alla soglia standard pari a 0.5, sono state considerate soglie ottimizzate in fase di training mediante metriche della famiglia F_β (con $\beta = 2$, $\beta = 1$, $\beta = \frac{1}{2}$). È importante sottolineare che tali soglie non influenzano i valori di PR-AUC poiché questa metrica viene calcolata sulle probabilità predette. Di conseguenza, a parità di PR-AUC, la scelta della soglia consente di modulare il comportamento del classificatore in funzione delle esigenze applicative.

Per quantificare la variabilità delle prestazioni stimate sul test-set, sono stati inoltre calcolati gli intervalli di confidenza al 95% mediante tecnica bootstrap. In particolare, a partire dalle probabilità predette dai modelli sul test set, sono stati effettuati 2000 ricampionamenti con reinserimento, calcolando ad ogni iterazione il valore della PR-AUC. Gli intervalli di confidenza sono stati quindi ottenuti come percentili della distribuzione empirica delle metriche così ottenute. Sono rappresentati graficamente nelle Figure 37-38 tramite barre di errore e forniscono una misura della stabilità delle prestazioni dei modelli rispetto alla variabilità del campione.

Le figure 37-38 riportano il confronto in termini di PR-AUC tra le diverse configurazioni di feature e classificatori. Tale metrica consente di valutare in modo diretto la capacità discriminativa dei modelli, indipendentemente dalla soglia decisionale. In particolare, nella figura 37 vengono riportati i risultati sul test set nel caso delle feature radiomiche di base e nella figura 38 nel caso di feature radiomiche di base con descrittori geometrici di concavità.

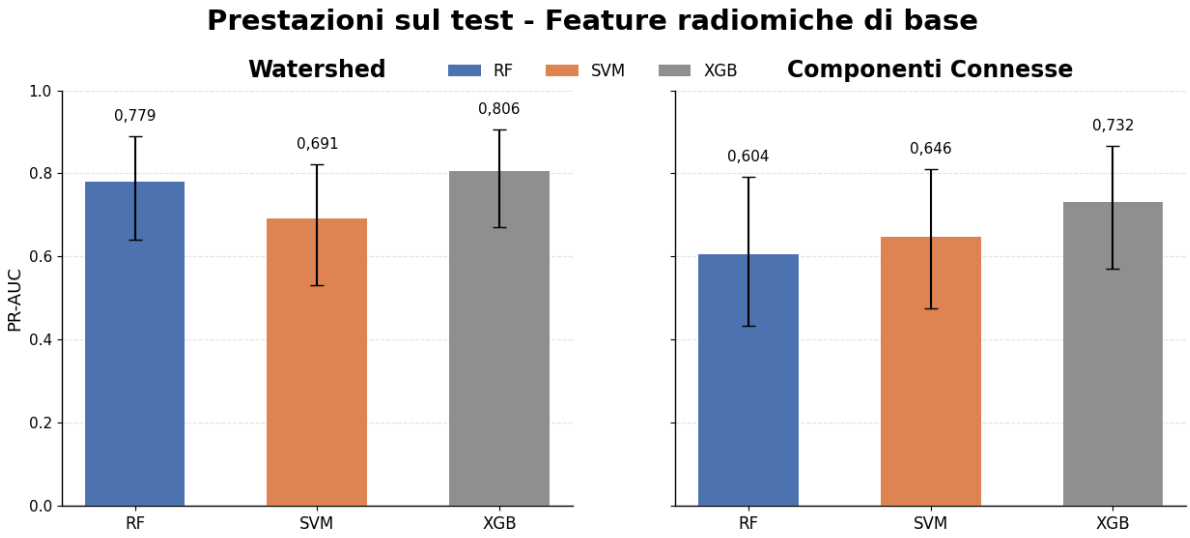

Figura 37: Prestazioni sul test set nel caso delle feature radiomiche di base. Le colonne colorate rappresentano il valore di PR-AUC per ciascun classificatore, mentre le barre d'errore indicano gli intervalli di confidenza al 95%, calcolati tramite 2000 ricampionamenti bootstrap.

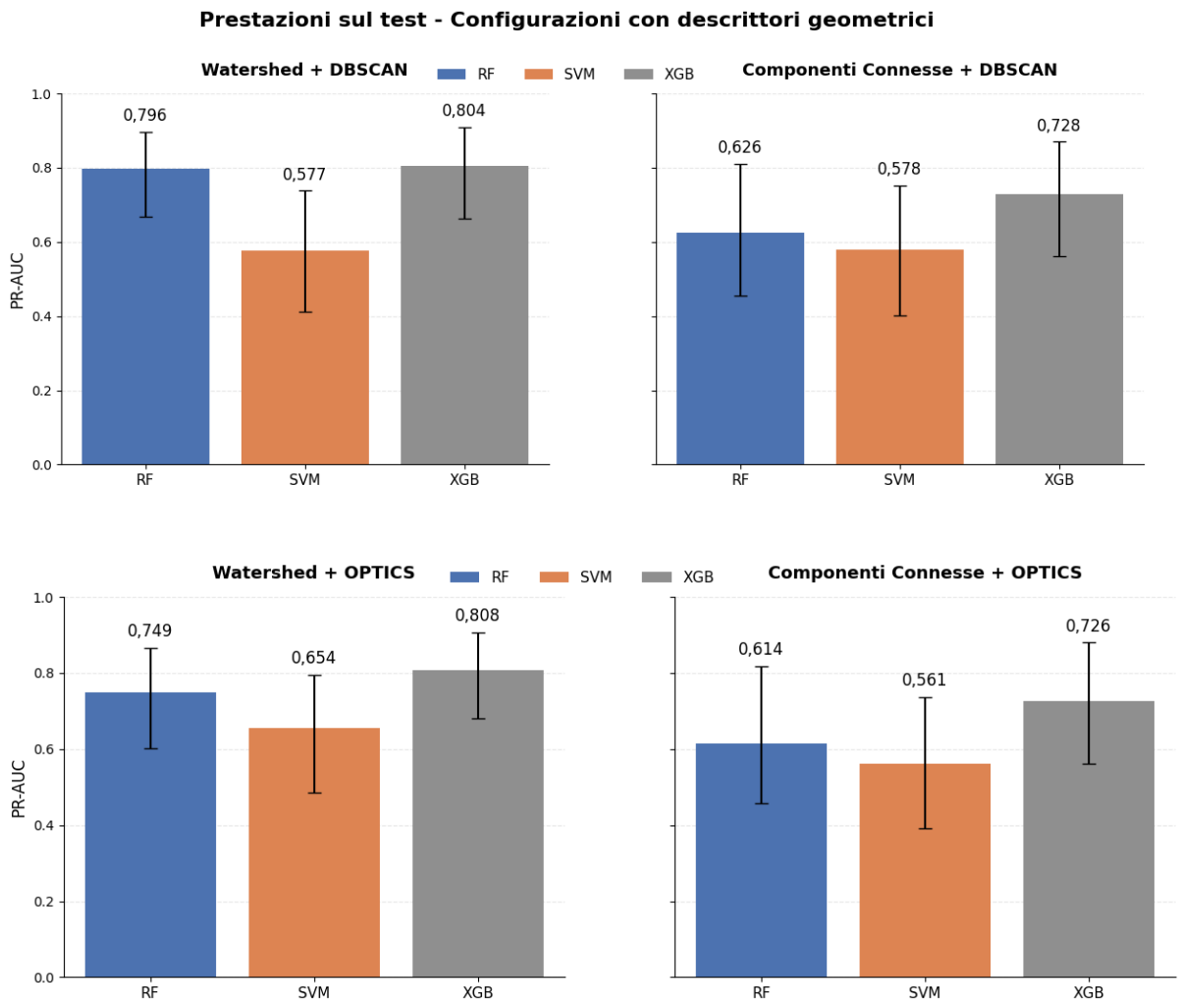

Figura 38: Prestazioni sul test set nel caso delle feature radiomiche di base combinate con i descrittori geometrici di concavità. Le colonne colorate rappresentano il valore di PR-AUC per ciascun classificatore, mentre le barre d'errore indicano gli intervalli di confidenza al 95%, calcolati tramite 2000 ricampionamenti bootstrap.

4.2.3 Confronto delle prestazioni tra modelli e configurazioni

Dall'analisi dei risultati emerge innanzitutto come le feature radiomiche di base confermino le buone prestazioni osservate in fase di training, mostrando un'elevata capacità di generalizzazione. In particolare, il classificatore XGBoost risulta il più performante, raggiungendo valori di PR-AUC prossimi a 0.80 nel caso in cui le ROI siano ottenute mediante il metodo Watershed, mentre prestazioni inferiori si osservano nel caso delle Componenti Connesse.

Il confronto tra i metodi di identificazione delle MC evidenzia infatti ancora come il metodo Watershed produca sistematicamente risultati migliori rispetto alle Componenti Connesse, suggerendo una maggiore accuratezza nell' identificazione delle MC e, conseguentemente, una migliore qualità delle feature estratte.

Per quanto riguarda le configurazioni che integrano descrittori geometrici di concavità, i risultati mostrano prestazioni complessivamente comparabili rispetto all'utilizzo delle sole feature radiomiche di base. In particolare, nel caso Watershed, il classificatore XGBoost mantiene valori di PR-AUC sostanzialmente invariati tra le due configurazioni, indicando che l'informazione geometrica aggiuntiva non introduce un miglioramento significativo delle prestazioni. Analogamente, nelle altre combinazioni considerate, le differenze risultano contenute e non sistematiche.

Questo comportamento suggerisce che le feature radiomiche di base forniscono già una rappresentazione sufficientemente informativa delle ROI, mentre i descrittori geometrici di concavità, pur contribuendo a descrivere ulteriori aspetti della morfologia delle MC, non apportano un contributo discriminativo rilevante in termini di performance predittiva.

Infine, il confronto tra gli algoritmi di clustering (DBSCAN e OPTICS) mostra differenze generalmente contenute, indicando come il loro impatto sia secondario rispetto alla scelta delle feature e del metodo di identificazione delle ROI.

4.2.4 Selezione del modello finale

Sulla base di tali risultati, è stata selezionata come configurazione di riferimento quella costituita dal classificatore XGBoost applicato alle feature radiomiche di base con identificazione delle ROI mediante Watershed, in quanto caratterizzata dalle migliori prestazioni complessive. La possibilità di selezionare diverse soglie decisionali consente inoltre di adattare il comportamento del modello alle specifiche esigenze applicative, privilegiando sensibilità o precisione a seconda del contesto.

Per questa configurazione, viene quindi presentata un'analisi più approfondita delle prestazioni al variare della soglia decisionale. In particolare, la Tabella 10 riporta i valori delle principali metriche di classificazione (accuratezza, precision, recall, specificità e NPV) ottenute utilizzando diverse soglie, tra cui quella standard (0.5) e quelle ottimizzate mediante metriche

F_β (con $\beta = 1$, $\beta = 2$, $\beta = \frac{1}{2}$). Per quantificare la robustezza delle stime puntuali e valutare il grado di incertezza associato alle metriche, i valori riportati sono accompagnati dai relativi intervalli di confidenza al 95%. Tali intervalli sono stati ricavanti mediante la tecnica bootstrap, effettuando 2000 ricampionamenti con reinserimento.

Soglia	Accuratezza	Precision	Recall	Specificità	NPV
0.5	0.855 [0.809-0.898]	0.491 [0.358-0.635]	0.848 [0.706-0.966]	0.856 [0.808-0.904]	0.972 [0.947-0.994]
0.586	0.855 [0.843-0.924]	0.562 [0.419-0.714]	0.818 [0.667-0.944]	0.896 [0.850-0.937]	0.968 [0.943-0.989]
0.381	0.804 [0.749-0.851]	0.411 [0.297-0.537]	0.909 [0.793-1.000]	0.787 [0.729-0.841]	0.981 [0.958-1.000]
0.791	0.928 [0.894-0.957]	0.786 [0.625-0.933]	0.667 [0.500-0.815]	0.97 [0.945-0.990]	0.947 [0.913-0.975]

Tabella 10: Metriche di classificazione del modello XGBoost sul set di test per diverse soglie decisionali (nel caso feature radiomiche di base e Watershed). Gli intervalli tra parentesi quadre rappresentano l'intervallo di confidenza al 95% calcolato tramite tecnica boostrap con 2000 iterazioni.

I risultati evidenziano come la variazione della soglia decisionale permetta di modulare efficacemente il comportamento del modello influenzando significativamente il bilanciamento tra precision e recall. In particolare, la soglia ottimizzata secondo F_β (con $\beta = 2$) consente di massimizzare il recall (0.909), risultando particolarmente adatta in contesti in cui è prioritario ridurre i falsi negativi. Al contrario, la soglia associata a F_β (con $\beta = 0.5$) privilegia la precision (0.786) e la specificità (0.97), riducendo i falsi positivi. La soglia basata su F_β (con $\beta = 1$)

rappresenta invece un compromesso equilibrato tra le due metriche. Complessivamente, l'elevato valore di NPV osservato in tutte le configurazioni suggerisce una buona affidabilità del modello nell'identificazione dei casi negativi. Questo risultato evidenzia come, a parità di capacità discriminativa del modello, la scelta della soglia consenta di adattarne il comportamento in funzione del contesto clinico di riferimento.

Nel complesso, i risultati indicano che la configurazione più efficace è rappresentata dall'utilizzo delle feature radiomiche di base in combinazione con il metodo Watershed e il classificatore XGBoost.

4.2.5 Analisi della calibrazione del modello

Alla luce dei risultati ottenuti in termini di capacità discriminativa, viene di seguito analizzata la calibrazione dei modelli nella configurazione basata su feature radiomiche di base e identificazione mediante Watershed. In questa fase, il confronto viene effettuato sui tre classificatori RF, SVM e XGBoost (Figura 39), al fine di valutare l'affidabilità delle probabilità predette, pur mantenendo XGBoost come modelli di riferimento sulla base delle prestazioni complessive precedentemente osservate.

La calibrazione del modello è stata analizzata mediante il calibration curve plot, che mette a confronto le probabilità predette dal modello con le frequenze osservate della classe positiva. In particolare, una calibrazione perfetta è rappresentata da una retta diagonale, mentre deviazioni da tale andamento indicano una sovrastima o sottostima delle probabilità predette.

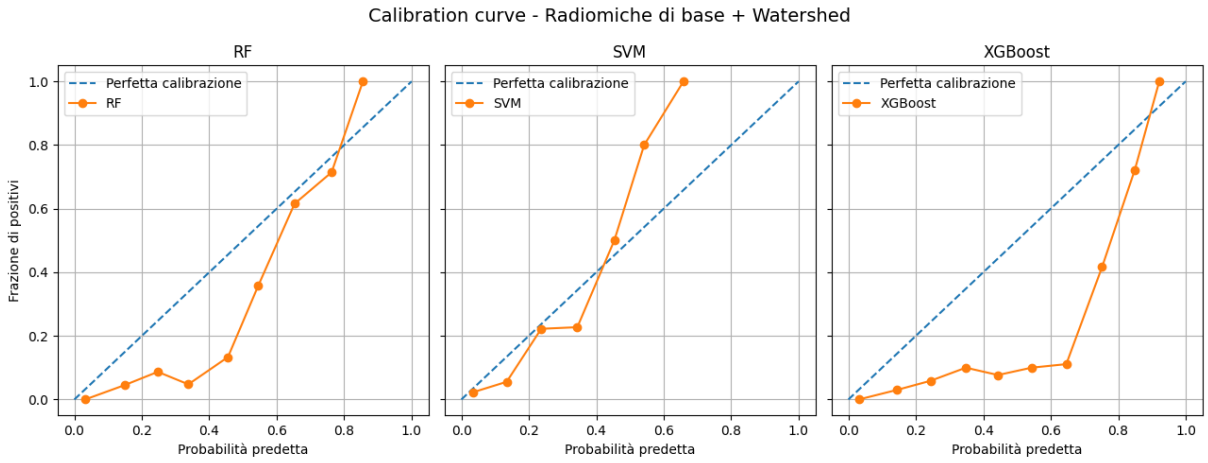

Figura 39: Calibration curve dei modelli RF, SVM e XGBoost nella configurazione basata su feature radiomiche di base e identificazione delle MC mediante Watershed.

Dall'analisi del calibration plot si osserva che nessuno dei tre classificatori risulta perfettamente calibrato sull'intero intervallo delle probabilità predette. In particolare, il confronto tra i tre classificatori non evidenzia differenze nette in termini di calibrazione, in quanto tutti i modelli presentano discostamenti dalla calibrazione ideale in diverse regioni dell'intervallo delle probabilità, suggerendo che la corrispondenza tra probabilità stimate e frequenze osservate non sia ottimale.

Alla luce di questi risultati, la calibrazione non rappresenta un criterio discriminante decisivo nella scelta del modello finale. La selezione del classificatore di riferimento resta quindi guidata dalle prestazioni complessive in termini di capacità discriminativa, per le quali XGBoost è risultato il modello più performante nella configurazione basata su feature radiomiche di base e identificazione mediante Watershed.

Nel complesso, questi risultati indicano che, anche se il modello XGBoost sia efficace nel distinguere tra le classi (come evidenziato dai valori di PR-AUC), le probabilità predette non possono essere interpretate direttamente come stime perfettamente affidabili della probabilità reale senza un'eventuale fase di ricalibrazione. Tuttavia, la buona capacità discriminativa osservata suggerisce che, pur in presenza di una calibrazione non ottimale, il modello mantenga un comportamento complessivamente coerente ai fini della classificazione, soprattutto nei contesti in cui l'obiettivo principale è la distinzione tra le classi piuttosto che l'interpretazione diretta delle probabilità.

4.2.6 Risultati Explainability

A partire da queste considerazioni, l'analisi viene quindi approfondita sul modello XGBoost, selezionato come configurazione finale. In particolare, viene condotto uno studio di interpretabilità, considerando sia la configurazione basata sulle sole feature radiomiche di base sia quella che include anche i descrittori geometrici di concavità, al fine di valutare il contributo delle diverse tipologie di feature nelle decisioni del modello.

Per analizzare l'interpretabilità del modello selezionato in modo da comprendere il contributo delle diverse features alla predizione finale è stato utilizzato il metodo SHAP che consente di quantificare l'impatto di ciascuna variabile sull'output del modello, fornendo una misura locale e globale dell'importanza delle feature.

Le figure 40-41 riportano i grafici di sintesi (SHAP summary plot) ottenuti per il classificatore XGBoost nelle due configurazioni considerate: feature radiomiche di base e combinazione di feature radiomiche di base con descrittori geometrici di concavità. In tali grafici, le feature sono ordinate in base alla loro importanza media. Per ciascuna feature, ogni punto rappresenta un'osservazione del dataset: il colore dei punti rappresenta il valore della feature (blu per valori bassi, rosso per valori alti) mentre la posizione sull'asse orizzontale rappresenta il valore SHAP, ossia l'entità e la direzione del contributo cha la feature ha avuto alla predizione del modello. In particolare, un valore SHAP positivo indica che la feature contribuisce ad incrementare la probabilità di assegnazione alla classe target (maligno), mentre un valore SHAP negativo indica un contributo che spinge la predizione verso la classe opposta (benigno).

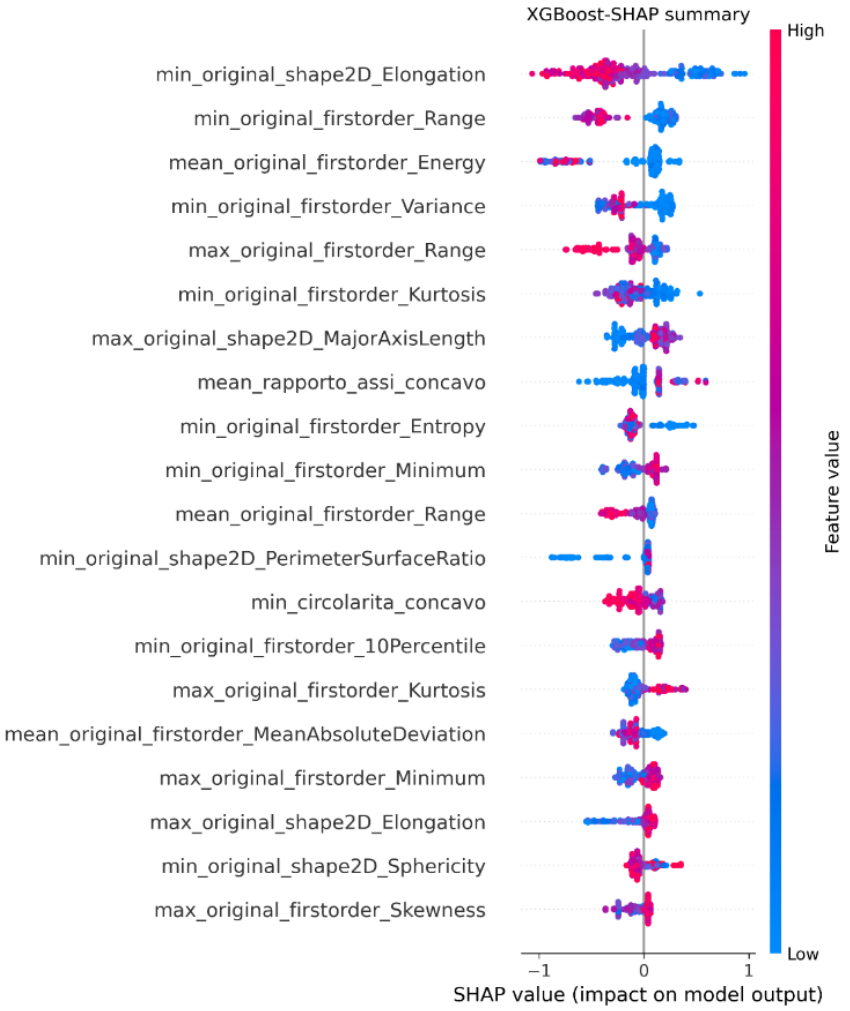

Figura 40: Analisi SHAP del modello XGBoost su feature radiomiche di base e descrittori geometrici di concavità (identificazione MC: Watershed, clustering: OPTICS). Il grafico riporta le 20 feature che maggiormente influenzano l'output del modello.

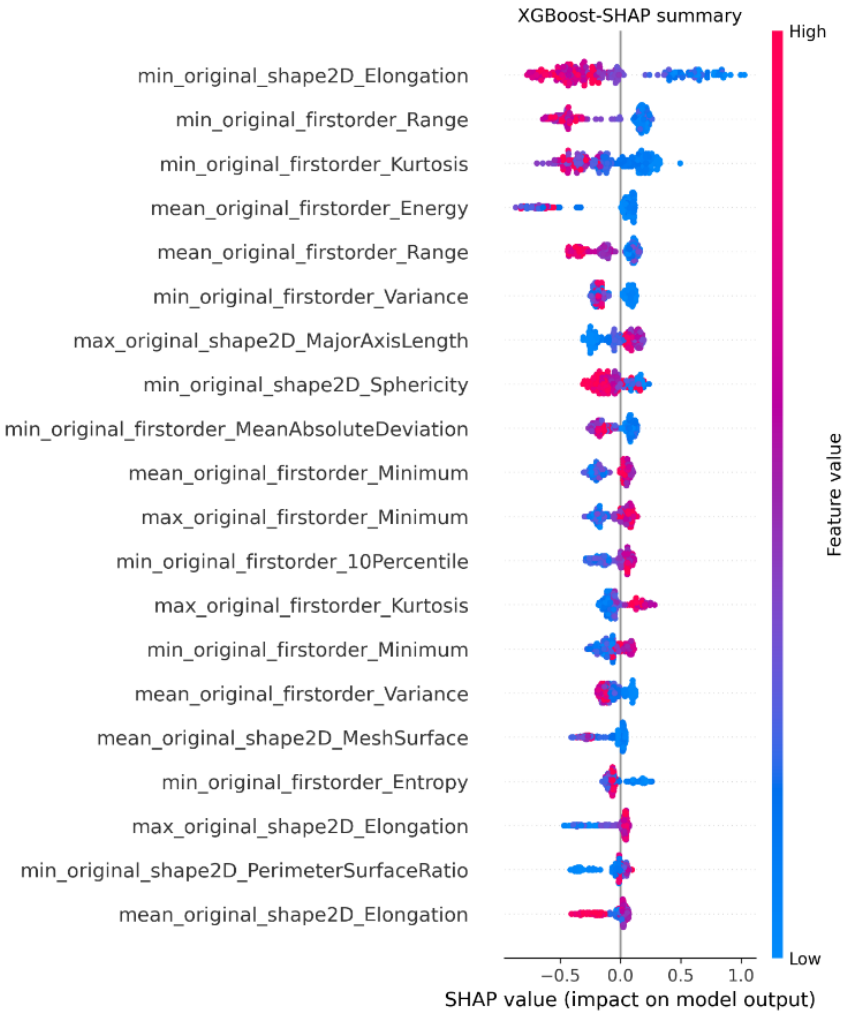

Figura 41: Analisi SHAP del modello XGBoost su feature radiomiche di base (identificazione MC: Watershed). Il grafico riporta le 20 feature che maggiormente influenzano l'output del modello.

Dall'analisi dei grafici SHAP emerge come le feature più rilevanti siano principalmente elongation, majorAxisLength, sphericity e perimeterSurfaceRatio che fanno parte delle feature di forma; e range, energy, eariance, kurtosis e entropy che fanno parte delle feature di primo ordine.

In entrambe le configurazioni (Figura 40 e 41), la feature con maggiore impatto risulta essere l'**elongazione (Elongation)**, suggerendo che la forma delle MC rappresenti un fattore discriminante rilevante per il modello. In PyRadiomics, l'elongation misura il grado di rotondità dell'oggetto definito come il rapporto tra asse minore e asse maggiore della regione. Valori prossimi ad 1 indicano una morfologia più compatta e circolare, mentre valori prossimi a 0

corrispondono a forme allungate e irregolari. L'analisi dei valori SHAP mostra che il modello distingue correttamente, associando valori più bassi di elongazione ad una maggiore probabilità di appartenenza alla classe maligna e valori più elevanti risultano più compatibili con la classe benigna. Tale comportamento risulta coerente con l'evidenza clinica secondo cui le MC maligne tendono a presentare morfologie meno compatte e più irregolari rispetto a quelle benigne.

Inoltre, si osserva che valori elevati o bassi delle diverse feature influenzano in modo differente la predizione, indicando che il modello sfrutta relazioni non lineari tra le caratteristiche delle ROI e la probabilità di appartenenza alla classe positiva.

Analizzando in dettaglio il modello SHAP in figura 40, si nota che tra le prime 20 feature più significative compaiono anche due descrittori geometrici di concavità riconducibili a:

1) **circolarità_concavo**: che assume valori tra 0 e 1, dove 1 corrisponde ad una forma perfettamente circolare e valori più bassi indicano forme progressivamente più irregolari. In accordo con quanto osservato sul grafico SHAP, valori maggiori sono associati a una maggiore probabilità di appartenere alla classe benigna, mentre valori minori risultano associati alla classe maligna.
2) **rapporto_assi_concavo**: definito come rapporto tra asse maggiore e minore, per cui valori maggiori indicano forme più allungate, mentre valori inferiori descrivono forme più compatte e rotondeggianti. Anche in questo caso, i risultati SHAP mostrano coerenza interpretativa, con valori più elevati associati a una maggiore probabilità di appartenere alla classe maligna.

Dal punto di vista clinico, i risultati ottenuti quindi risultano coerenti con la letteratura sulle MC. In particolare, le feature di forma, come elongazione e sfericità, sono rilevanti in quanto le MC maligne tendono a presentare morfologie irregolari e meno compatte, mentre quelle benigne risultano generalmente più regolari. Le feature radiomiche di primo ordine, come range, varianza ed entropia, riflettono l'eterogeneità dell'intensità all'interno della ROI. Una maggiore eterogeneità è spesso associata a tessuti patologici, in quanto può indicare una struttura più complessa e disorganizzata. La presenza di feature legate alla kurtosis suggerisce che la distribuzione dei valori di intensità all'interno della ROI contribuisce alla

discriminazione tra le classi, evidenziando differenze nella concentrazione dei valori intorno alla media.

Nel complesso, il modello sembra quindi basarsi su una combinazione di feature in linea con i criteri utilizzati nella pratica clinica per la valutazione delle MC e della loro relativa distribuzione, confermando coerenza tra le evidenze estratte automaticamente dal modello e le conoscenze cliniche consolidate.

CAPITOLO 5 – CONCLUSIONI

5.1 Discussione

Il presente lavoro di tesi ha avuto come obiettivo lo sviluppo e la validazione di un sistema automatico per la classificazione di MC, basato sull'integrazione di feature radiomiche e descrittori geometrici, con lo scopo di distinguere tra ROI benigne e maligne.

In linea con quanto delineato nel Capitolo 1, il lavoro si è inserito nel contesto dell'applicazione di tecniche di analisi automatica delle immagini a supporto della diagnosi precoce del tumore al seno, affrontando uno dei problemi più complessi dell'imaging mammografico, ovvero la caratterizzazione delle MC.

In primo luogo, si è partiti dalla definizione del dataset e delle ROI ottenute grazie alla pipeline DeepMica che ha consentito di avere a disposizione maschere di segmentazione standardizzate e di buona qualità. Successivamente il focus si è spostato sull'identificazione delle singole MC all'interno delle maschere. In questo contesto sono stati adottati due approcci: Componenti Connesse e Watershed. Il primo rappresenta una soluzione semplice e consolidata che tende a fondere MC molto vicine in un'unica regione. Il metodo Watershed, invece sfruttando la trasformata di distanza e l'uso di marcatori, si è dimostrato efficace nel separare strutture adiacenti, portando ad una identificazione più accurata delle singole MC. Questa differenza si è poi riflessa sulla qualità delle feature estratte e di conseguenza sulle prestazioni dei modelli di classificazione.

Una volta identificate le MC, è stata affrontata la fase di caratterizzazione quantitativa tramite due famiglie di descrittori: feature radiomiche e descrittori geometrici a livello di cluster. Le feature radiomiche, in particolare quelle di base (shape e first order) hanno rappresentato il nucleo informativo principale. Da un lato descrivono la morfologia delle singole MC (elongazione, sfericità, rapporto tra assi), dall'altro quantificano la distribuzione dei livelli di intensità nella ROI (range, varianza, entropia, energia, kurtosis), catturando la complessità interna del tessuto.

Parallelamente, l'analisi della distribuzione spaziale delle MC è stata affrontata tramite algoritmi di clustering basati sulla densità (DBSCAN e OPTICS), utilizzando le coordinate dei centroidi per individuare aggregazioni di MC. Da questi cluster sono stati derivati descrittori

geometrici (circolarità, rapporto tra assi principali) calcolati sia su contorni convessi sia su contorni con bordi concavi, allo scopo di quantificare la forma complessiva ed elongazione dei raggruppamenti, in analogia con i descrittori morfologici utilizzati nella pratica radiologica.

Una volta estratte le feature è stata effettuata una fase di classificazione. A partire dalla matrice delle feature, sono stati confrontati tre modelli supervisionati: RF, SVM e XGBoost su un problema di classificazione binaria fortemente sbilanciato (ossia in presenza di una prevalenza di ROI benigne). La pipeline di training ha integrato pre-processing (selezione e pulizia delle feature, normalizzazione robusta), strategie per il bilanciamento delle classi e un'ottimizzazione sistematica degli iperparametri tramite cross-validazione con Grid Search, utilizzando la PR-AUC come metrica principale, più informativa dell'accuratezza in presenza di sbilanciamento tra le classi.

Valutando i risultati ottenuti, emerge un quadro che attesta la validità della pipeline, confermandone l'effettiva utilità nel contesto applicativo considerato.

In primo luogo, è stato confermato come le feature radiomiche di base garantiscano performance solide e affidabili. Tra i modelli testati, XGBoost ha mostrato la maggiore robustezza, imponendosi come il classificatore più efficace sia in termini di capacità predittiva che di generalizzazione. Dal confronto tra le strategie di identificazione delle MC si è osservato, invece, che il metodo Watershed ha prodotto risultati mediamente superiori rispetto alle Componenti Connesse, validando l'ipotesi che una identificazione accurata sia il prerequisito fondamentale per un'estrazione delle feature fedele e di conseguenza per un'efficace classificazione. Infine, l'analisi relativa ai descrittori geometrici basati su cluster ha rivelato che questi, se utilizzati singolarmente, non raggiungono le prestazioni garantite dalle feature radiomiche. Anche la loro integrazione non ha comportato un incremento sistematico delle metriche di valutazione, pur consentendo di mantenere livelli di accuratezza complessivamente confrontabili con le configurazioni standard.

La gestione consapevole della soglia decisionale ha costituito un elemento importante nell'elaborazione dei risultati sperimentali. Considerando l'elevato sbilanciamento, la semplice soglia 0.5 sulle probabilità non è sufficiente per rappresentare al meglio il trade-off fra falsi positivi e falsi negativi. L'ottimizzazione della soglia attraverso metriche della famiglia F_β ha

permesso di esplorare scenari diversi: soglie più sensibili, che privilegiano il recall riducendo i falsi negativi, e soglie più stringenti, che aumentano la precision riducendo i falsi allarmi. Questo risultato mostra come il modello possa essere adattato alle esigenze operative di contesti clinici differenti (come, ad esempio, screening o approfondimento diagnostico).

Un ulteriore aspetto analizzato riguarda la calibrazione delle probabilità predette dal modello finale. I risultati mostrano come, pur in presenza di una buona capacità discriminativa, il modello non risulti perfettamente calibrato, evidenziando una tendenza alla sovrastima delle probabilità in alcune regioni dell'intervallo. Tuttavia, il comportamento osservato nelle regioni ad alta probabilità, dove si riscontra un miglior allineamento tra probabilità predette e frequenze osservate, suggerisce che il modello mantenga una buona affidabilità nelle predizioni più sicure. Questo risultato indica che, nel contesto applicativo considerato, il modello può essere efficacemente utilizzato per la classificazione, mentre l'interpretazione diretta delle probabilità potrebbe beneficiare di un'eventuale fase di ricalibrazione.

Infine, l'analisi di explainability tramite SHAP ha confermato che modello fonda le sue decisioni su feature morfologiche e di intensità coerenti con pratica clinica. Le feature di forma e le statistiche di primo ordine sono risultate determinanti. Anche se l'integrazione dei descrittori geometrici di cluster non abbia prodotto incrementi netti nelle metriche di classificazione, l'analisi SHAP ne ha validato l'importanza interpretativa, dimostrando che il sistema valuta correttamente anche la distribuzione spaziale delle MC. In conclusione, questo approccio rende il modello uno strumento trasparente e non più un sistema black-box. In particolare, le sue predizioni diventano spiegabili e coerenti con i criteri che il radiologo osserva per valutare le MC.

5.1.1 Conclusioni dello Studio

Nel complesso, il lavoro dimostra che una pipeline basata su segmentazione accurata (Watershed), feature radiomiche di base e un classificatore ad alberi boosting (XGBoost), opportunamente ottimizzato e analizzato in termini di soglia decisionale, è in grado di distinguere ROI con MC benigne e maligne con buone prestazioni e alto valore predittivo negativo. Le informazioni spaziali e geometriche derivate dalla distribuzione delle MC non migliorano in modo marcato le metriche globali, ma arricchiscono il quadro interpretativo e mantengono coerenza con i criteri clinici. Infine, l'uso di tecniche di explainability è

fondamentale per verificare che il modello si basi su pattern morfologici e di intensità che riflettono i criteri utilizzati nella pratica clinica.

Questi risultati supportano l'idea che sistemi di classificazione automatica delle MC, possano rappresentare un valido strumento di supporto alla decisione per il radiologo, con potenziale impatto sulla riduzione delle biopsie non necessarie e sul miglioramento della diagnosi precoce del tumore al seno.

5.2 Sviluppi Futuri

I risultati ottenuti suggeriscono diverse possibili direzioni di sviluppo.

In primo luogo, il ruolo dei descrittori geometrici merita ulteriori approfondimenti. Sebbene nel presente studio il loro contributo alle prestazioni sia risultato limitato, è plausibile che la rappresentazione adottata non riesca a catturare pienamente la complessità della distribuzione spaziale delle MC. Quindi, l'impiego di informazioni geometriche più sofisticate potrebbe risultare maggiormente discriminativo. In questo contesto, futuri sviluppi potrebbero includere l'utilizzo di descrittori topologici o basati su grafi, così come l'analisi delle relazioni tra MC attraverso modelli spaziali più avanzati.

Un ulteriore ambito di sviluppo riguarda l'integrazione di approcci multimodali, combinando le informazioni radiomiche con tecniche di deep learning e con dati clinici, quali età, storia clinica e densità mammaria. Questo tipo di integrazione potrebbe contribuire a una rappresentazione più completa del quadro diagnostico, migliorando le prestazioni complessive dei modelli.

Dal punto di vista dei dati, la disponibilità di dataset arricchiti da metadati completi, come quelli contenuti nei file DICOM, permetterebbe una caratterizzazione più accurata delle proprietà spaziali e di acquisizione, con potenziali benefici sulle prestazioni dei modelli.

Infine, sarà fondamentale estendere la validazione dei modelli su dataset multicentrici. Un'estensione a dataset provenienti da altri centri infatti permetterebbe di testare le capacità di generalizzazione dell'algoritmo rispetto a diverse popolazioni di pazienti, protocolli di acquisizione e strumentazioni radiologiche differenti, riducendo eventuali bias legati alla specificità del centro di origine.

BIBLIOGRAFIA

[1] J. Kim *et al.*, «Global patterns and trends in breast cancer incidence and mortality across 185 countries», *Nat Med*, vol. 31, fasc. 4, pp. 1154–1162, apr. 2025, doi: 10.1038/s41591-025-03502-3.

[2] B. Lauby-Secretan *et al.*, «Breast-Cancer Screening — Viewpoint of the IARC Working Group», *New England Journal of Medicine*, vol. 372, fasc. 24, pp. 2353–2358, giu. 2015, doi: 10.1056/NEJMsr1504363.

[3] S. H. Heywang-Köbrunner, A. Hacker, e S. Sedlacek, «Advantages and Disadvantages of Mammography Screening», *Breast Care (Basel)*, vol. 6, fasc. 3, pp. 199–207, giu. 2011, doi: 10.1159/000329005.

[4] M. El Khoury e B. Mesurolle, «Breast Mammographic Screening: The More Mammograms Read, the Better the Performance», *Can Assoc Radiol J*, vol. 73, fasc. 2, pp. 289–290, mag. 2022, doi: 10.1177/08465371211040699.

[5] L. Wang, «Mammography with deep learning for breast cancer detection», *Front. Oncol.*, vol. 14, p. 1281922, feb. 2024, doi: 10.3389/fonc.2024.1281922.

[6] D. Ribli, A. Horváth, Z. Unger, P. Pollner, e I. Csabai, «Detecting and classifying lesions in mammograms with Deep Learning», *Sci Rep*, vol. 8, fasc. 1, p. 4165, mar. 2018, doi: 10.1038/s41598-018-22437-z.

[7] R. Masud, M. Al-Rei, e C. Lokker, «Computer-Aided Detection for Breast Cancer Screening in Clinical Settings: Scoping Review», *JMIR Med Inform*, vol. 7, fasc. 3, p. e12660, lug. 2019, doi: 10.2196/12660.

[8] A. Logullo, K. Prigenzi, C. Nimir, A. Franco, e M. Campos, «Breast microcalcifications: Past, present and future (Review)», *Mol Clin Oncol*, vol. 16, fasc. 4, p. 81, feb. 2022, doi: 10.3892/mco.2022.2514.

[9] M. M.-S. Yao, H. Du, M. Hartman, W. P. Chan, e M. Feng, «End-to-End Calcification Distribution Pattern Recognition for Mammograms: An Interpretable Approach with GNN», *Diagnostics*, vol. 12, fasc. 6, p. 1376, giu. 2022, doi: 10.3390/diagnostics12061376.

[10] F. Prinzi, A. Orlando, S. Gaglio, e S. Vitabile, «Interpretable Radiomic Signature for Breast Microcalcification Detection and Classification», *J Digit Imaging. Inform. med.*, vol. 37, fasc. 3, pp. 1038–1053, feb. 2024, doi: 10.1007/s10278-024-01012-1.

[11] A. Papadopoulos, D. I. Fotiadis, e A. Likas, «Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector

machines», *Artificial Intelligence in Medicine*, vol. 34, fasc. 2, pp. 141–150, giu. 2005, doi: 10.1016/j.artmed.2004.10.001.

[12] Y. Gao, J. Lin, Y. Zhou, e R. Lin, «The application of traditional machine learning and deep learning techniques in mammography: a review», *Front. Oncol.*, vol. 13, p. 1213045, ago. 2023, doi: 10.3389/fonc.2023.1213045.

[13] M. Lei *et al.*, «Benchmarking Various Radiomic Toolkit Features While Applying the Image Biomarker Standardization Initiative toward Clinical Translation of Radiomic Analysis», *J Digit Imaging*, vol. 34, fasc. 5, pp. 1156–1170, ott. 2021, doi: 10.1007/s10278-021-00506-6.

[14] A. Zwanenburg *et al.*, «The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping», *Radiology*, vol. 295, fasc. 2, pp. 328–338, mag. 2020, doi: 10.1148/radiol.2020191145.

[15] R. Cattell, S. Chen, e C. Huang, «Robustness of radiomic features in magnetic resonance imaging: review and a phantom study», *Vis. Comput. Ind. Biomed. Art*, vol. 2, fasc. 1, p. 19, dic. 2019, doi: 10.1186/s42492-019-0025-6.

[16] S. Rizzo *et al.*, «Radiomics: the facts and the challenges of image analysis», *Eur Radiol Exp*, vol. 2, fasc. 1, p. 36, dic. 2018, doi: 10.1186/s41747-018-0068-z.

[17] M. Ciecholewski, «Microcalcification Segmentation from Mammograms: A Morphological Approach», *J Digit Imaging*, vol. 30, fasc. 2, pp. 172–184, apr. 2017, doi: 10.1007/s10278-016-9923-8.

[18] N. Alam, E. R. E. Denton, e R. Zwiggelaar, «Classification of Microcalcification Clusters in Digital Mammograms Using a Stack Generalization Based Classifier», *J. Imaging*, vol. 5, fasc. 9, p. 76, set. 2019, doi: 10.3390/jimaging5090076.

[19] M. Elter e A. Horsch, «CADx of mammographic masses and clustered microcalcifications: A review», *Medical Physics*, vol. 36, fasc. 6Part1, pp. 2052–2068, giu. 2009, doi: 10.1118/1.3121511.

[20] I. D. Mienye e Y. Sun, «A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects», *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[21] A. E. Ezugwu *et al.*, «A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects»,

Engineering Applications of Artificial Intelligence, vol. 110, p. 104743, apr. 2022, doi: 10.1016/j.engappai.2022.104743.

[22] M. Ester, H.-P. Kriegel, e X. Xu, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise».

[23] M. Ankerst, M. M. Breunig, H.-P. Kriegel, e J. Sander, «OPTICS: Ordering Points To Identify the Clustering Structure».

[24] E. Y. Boateng, J. Otoo, e D. A. Abaye, «Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review», *JDAIP*, vol. 08, fasc. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[25] S. A. K. Malalha, M. Burhanuddin, e D. N. B. Yunos, «Unveiling the Tapestry of Machine Learning: A Comparative Analysis of Support Vector Machines, Random Forests, and Neural Networks in Diverse Applications», vol. 45, fasc. 3, 2024.

[26] B. Schölkopf, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. in Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2002.

[27] A. Yaman e M. A. Cengiz, «The Effects of Kernel Functions and Optimal Hyperparameter Selection on Support Vector Machines», 2021.

[28] P. Probst, M. Wright, e A.-L. Boulesteix, «Hyperparameters and Tuning Strategies for Random Forest», *WIREs Data Min & Knowl*, vol. 9, fasc. 3, p. e1301, mag. 2019, doi: 10.1002/widm.1301.

[29] T. Chen e C. Guestrin, «XGBoost: A Scalable Tree Boosting System», in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[30] T. Saito e M. Rehmsmeier, «The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets», *PLoS ONE*, vol. 10, fasc. 3, p. e0118432, mar. 2015, doi: 10.1371/journal.pone.0118432.

[31] J. M. R. B. Mello *et al.*, «Breast cancer mammographic diagnosis performance in a public health institution: a retrospective cohort study», *Insights Imaging*, vol. 8, fasc. 6, pp. 581–588, dic. 2017, doi: 10.1007/s13244-017-0573-2.

[32] M. Assel e A. Vickers, «The F score ranks diagnostic tests and prediction models inconsistently with their clinical utility», *Diagn Progn Res*, vol. 9, fasc. 1, p. 30, dic. 2025, doi: 10.1186/s41512-025-00214-7.

[33] H. R. Arabnia, M. Joudaki, A. Bagheri, e H. R. Arabnia, «Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-score, H-measure, and AUC-based Metrics across Diverse Classifiers», 13 gennaio 2026, *Preprints*: 2025100958. doi: 10.20944/preprints202510.0958.v2.

[34] A. Gerbasi *et al.*, «DeepMiCa: Automatic segmentation and classification of breast MIcroCAlcifications from mammograms», *Computer Methods and Programs in Biomedicine*, vol. 235, p. 107483, giu. 2023, doi: 10.1016/j.cmpb.2023.107483.

[35] F. Bolelli, S. Allegretti, e C. Grana, «Connected Components Labeling on Bitonal Images», in *Image Analysis and Processing – ICIAP 2022*, vol. 13232, S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, e F. Tombari, A c. di, in Lecture Notes in Computer Science, vol. 13232. , Cham: Springer International Publishing, 2022, pp. 347–357. doi: 10.1007/978-3-031-06430-2_29.

[36] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, e Y. Chao, «The connected-component labeling problem: A review of state-of-the-art algorithms», *Pattern Recognition*, vol. 70, pp. 25–43, ott. 2017, doi: 10.1016/j.patcog.2017.04.018.

[37] R. D. Yapa e H. Koichi, «A connected component labeling algorithm for grayscale images and application of the algorithm on mammograms», in *Proceedings of the 2007 ACM symposium on Applied computing*, Seoul Korea: ACM, mar. 2007, pp. 146–152. doi: 10.1145/1244002.1244040.

[38] M. Orozco-Monteagudo, C. Mihai, H. Sahli, e A. Taboada-Crispi, «Combined Hierarchical Watershed Segmentation and SVM Classification for Pap Smear Cell Nucleus Extraction», vol. 16, fasc. 2, 2012.

[39] A. Tareef *et al.*, «Multi-Pass Fast Watershed for Accurate Segmentation of Overlapping Cervical Cells», *IEEE Trans. Med. Imaging*, vol. 37, fasc. 9, pp. 2044–2059, set. 2018, doi: 10.1109/TMI.2018.2815013.

[40] K. Hu, W. Yang, e X. Gao, «Microcalcification diagnosis in digital mammography using extreme learning machine based on hidden Markov tree model of dual-tree complex wavelet transform», *Expert Systems with Applications*, vol. 86, pp. 135–144, nov. 2017, doi: 10.1016/j.eswa.2017.05.062.

[41] C. Marasinou *et al.*, «Improving the Quantitative Analysis of Breast Microcalcifications: A Multiscale Approach», *J Digit Imaging*, vol. 36, fasc. 3, pp. 1016–1028, feb. 2023, doi: 10.1007/s10278-022-00751-3.

[42] B. Singh e M. Kaur, «An approach for classification of malignant and benign microcalcification clusters», *Sādhanā*, vol. 43, fasc. 3, p. 39, mar. 2018, doi: 10.1007/s12046-018-0805-2.

[43] L. Vivona, D. Cascio, R. Magro, F. Fauci, e G. Raso, «A fuzzy logic C-means clustering algorithm to enhance microcalcifications clusters in digital mammograms», in *2011 IEEE Nuclear Science Symposium Conference Record*, Valencia, Spain: IEEE, ott. 2011, pp. 3048–3050. doi: 10.1109/NSSMIC.2011.6152551.

[44] M. Alsheh Ali, M. Eriksson, K. Czene, P. Hall, e K. Humphreys, «Detection of potential microcalcification clusters using multivendor for-presentation digital mammograms for short-term breast cancer risk estimation», *Medical Physics*, vol. 46, fasc. 4, pp. 1938–1946, apr. 2019, doi: 10.1002/mp.13450.

[45] A. Fanizzi *et al.*, «A machine learning approach on multiscale texture analysis for breast microcalcification diagnosis», *BMC Bioinformatics*, vol. 21, fasc. S2, p. 91, mar. 2020, doi: 10.1186/s12859-020-3358-4.

[46] S. Ferrari, «Struttura topologica di una immagine», [Online]. Disponibile su: https://homes.di.unimi.it/ferrari/ElabImm2009_10/EI2009_10_04_struttura_immagine_doppio.pdf)