Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Biomedica

Tesi di Laurea Magistrale

Identificazione di midollo osseo attivo da immagini TC: machine learning e analisi wavelet

Relatrice:

Prof.ssa Gabriella Balestra

Correlatrice:

Prof.ssa Samanta Rosati

Candidato:

Daniele Scaffidi Gennarino

Anno accademico 2020/2021

A chi mi è stato accanto

Abstract

Il midollo osseo riveste un ruolo centrale nello sviluppo del piano di trattamento radioterapico a cui sono sottoposti i soggetti affetti da tumori della parte bassa dell'addome. Esso si distingue in due parti: midollo attivo o rosso (RM), responsabile del processo di ematopoiesi, e midollo giallo (YM), sede di conservazione dei lipidi da parte degli adipociti. Circa il 50 % del midollo attivo si colloca tra le pelvi e la spina lombare ed è un tessuto particolarmente radiosensibile. Si è notato che quando il trattamento radioterapico coinvolge il midollo osseo attivo si va incontro a un fenomeno di emato-tossicità che porta ad essere più esposti a contrarre infezioni di gravità anche molto elevata per cui il trattamento chemio-radioterapico deve essere interrotto oppure la sua regolarità alterata con una complessiva perdita di efficacia della terapia. È evidente, dunque, come il midollo rosso debba essere distinto da quello giallo in modo da poter sviluppare un piano di trattamento radioterapico che eviti il più possibile la sua irradiazione.
Lo stato dell'arte prevede, per la rilevazione del midollo attivo, la possibilità di impiegare due tecniche principali, la risonanza magnetica (MR) e la tomografia ad emissione di positroni (PET), oppure alcune tecniche ibride che coinvolgono PET e una metodica tra TC e MR. Tuttavia, queste tipologie di esami presentano diversi svantaggi, tra cui l'elevato costo e la difficoltà di accesso sul territorio.
L'idea su cui si basa questa tesi è l'impiego delle immagini di tomografia computerizzata (TC) per l'individuazione del midollo attivo. In questo modo viene risolto il problema del costo e della difficoltà di accesso, senza la somministrazione di alcuna ulteriore dose di radiazione al paziente, in quanto la TC è già un esame standard nel workflow ospedaliero che precede lo sviluppo del piano di trattamento radioterapico. In questo lavoro, si è scelto di seguire la strada della radiomica andando a ricavare informazioni dai pixel delle immagini e implementando tecniche di machine learning.
La popolazione analizzata è stata quella di 50 pazienti in cura presso l'oncologia dell'Ospedale Molinette di Torino, i quali sono stati sottoposti sia a PET che a TC. La prima è stata impiegata come riferimento, mentre la seconda ha permesso la suddivisione del midollo osseo in tre regioni anatomiche: midollo osseo lombosacrale (LSBM), midollo osseo iliaco (IBM) e porzione bassa del midollo osseo pelvico (LPBM). I primi 40 pazienti sono stati impiegati per la realizzazione di un construction set costituito da 36 feature statistiche del 1° e del 2° ordine, mentre gli ultimi 10 sono stati inseriti in un validation set.
In una prima fase del lavoro, si è svolto un confronto, per ogni regione anatomica, tra le performance ricavate da tre tipologie di classificatori diversi (DT, KNN e NN) a partire da un set di dati di training ricavato mediante estrazione proporzionale e clustering attraverso reti SOM e le performance relative agli stessi classificatori addestrati sulla base di un training set estratto in maniera random. Quest'ultimo era stato il risultato di uno studio condotto in precedenza. In tutti i casi l'ottimizzazione dei parametri di ciascun metodo di classificazione e la feature selection è stata eseguita mediante algoritmi genetici.
In seguito, si è rivolta l'attenzione alla struttura LPBM, in quanto è risultata affetta dalle maggiori difficoltà di classificazione. Dopo una prima analisi delle sue slice in termini di RM presente, si è deciso di effettuare una diversa estrazione delle slice da inserire all'interno del construction set per questa struttura. Inoltre, è stato impiegato un nuovo metodo di estrazione del training set basato sulla variabilità intra-cluster per cercare di aumentare la rappresentatività dello stesso dell'intero construction set. Successivamente, è stata effettuata un'analisi dei misclassificati comuni tra diverse

tipologie di classificatori impiegati, è stato implementato un metodo di classificazione basato sul majority voting di quest'ultimi e sono state valutate le sue prestazioni.
In ultimo, sono stati rivalutati i descrittori impiegati all'interno di questo studio e si è esplorato il campo delle feature di ordine superiore, prendendo in considerazione, in particolare, quelle derivanti dalla decomposizione wavelet. In tale ottica, sono stati ricavati un nuovo construction set e un nuovo validation set per la struttura LPBM comprendenti 70 feature derivanti dall'uso delle wavelet, sono state allenate alcune tipologie di classificatori e i risultati sono stati messi a confronto con quelli ottenuti a partire dai descrittori del 1° e del 2° ordine.
Il lavoro svolto ha evidenziato le maggiori capacità delle feature derivanti dalla decomposizione wavelet di risolvere il problema di individuazione del midollo attivo sulle immagini di TC della struttura LPBM rispetto alle tradizionali feature statistiche del 1° e del 2° ordine.

Indice

1 Introduzione

1.1 Midollo osseo

Il midollo osseo è un tessuto connettivo altamente differenziato che svolge diverse funzioni all'interno dell'individuo, tra cui quello fondamentale dell'ematopoiesi, ossia di produzione delle cellule del sangue.

Esso si concentra nei canali delle ossa lunga e nella parte centrale delle ossa piatte e si distingue principalmente in due tipologie [1]:

- Midollo osseo rosso, detto anche attivo o ematopoietico (Red Marrow, RM): si trova prevalentemente nello scheletro assiale, contiene una vasta rete di vasi sanguigni ed è costituito per il 40% da tessuto adiposo e per il 60% da cellule ematopoietiche. Queste ultime sono responsabili del processo di ematopoiesi, il cui risultato è la produzione di cellule ematiche appartenenti a due sottofamiglie: cellule linfoidi (cellule B, T e NK) e cellule mieloidi (monociti, macrofagi, neutrofili, basofili, eosinofili, eritrociti e megacariociti).
- Midollo osseo giallo (Yellow Marrow, YM): è prevalente nello scheletro appendicolare, presenta una percentuale di tessuto adiposo nettamente superiore a quello rosso e pari a circa il 95% e costituisce la sede di conservazione dei lipidi da parte degli adipociti.

Nel corso della vita questi due tipi di midollo sono presenti in percentuale diversa nell'uomo. Più nel dettaglio, alla nascita tutto il midollo è virtualmente rosso e con l'avanzare dell'età esso viene convertito gradualmente in midollo giallo, perdendo quindi la capacità di produrre cellule ematiche. In talune occasioni, quali emorragie di particolare gravità, è stata tuttavia individuata la capacità del midollo giallo di riconvertirsi in midollo rosso, per sopperire all'elevata richiesta di nuove cellule del sangue.

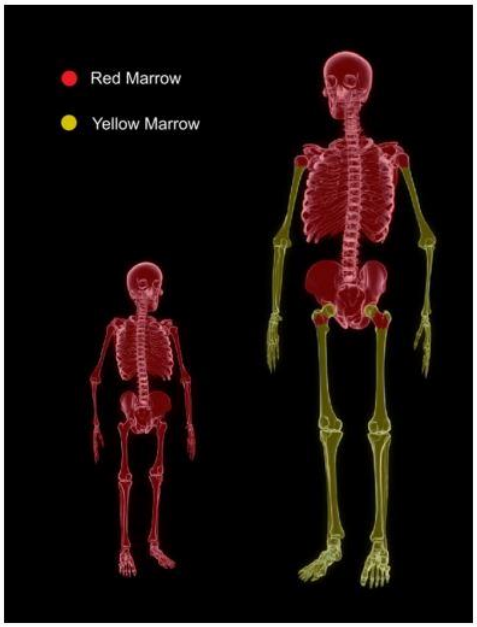


Figura 1 Distribuzione del midollo rosso e giallo negli scheletri assiale e appendicolare in età infantile e adulta. Fonte: [2]

1.2 Radioterapia e midollo attivo

Quando si parla di trattamento dei tumori le tecniche maggiormente adoperate sono quelle della chemioterapia e della radioterapia, spesso combinate, che consistono in diversi cicli di rilascio di farmaco e, rispettivamente, di radiazioni nella regione di interesse.

In quest'ottica, bisogna osservare che le cellule ematopoietiche del midollo osseo attivo presentano una elevata radiosensibilità, per cui durante un trattamento chemio-radioterapico si può andare in contro ad una tossicità ematologica. Quest'ultima si può manifestare in maniere diverse come ad esempio sotto forma di anemia, piastrinopenia o leucopenia. Si è notato che quando il trattamento radioterapico coinvolge il midollo osseo attivo uno degli sviluppi più probabili di emato-tossicità è la neutropenia, ossia un tipo particolare di leucopenia caratterizzato da una diminuzione dei neutrofili al di sotto di valori di soglia considerati sicuri [3]. L'effetto è, quindi, quello di essere sensibilmente più esposti a contrarre infezioni di gravità anche molto elevata per cui il trattamento chemio-radioterapico viene interrotto oppure la regolarità dello stesso viene alterata con una conseguente perdita di efficacia della terapia.

Dal momento che oltre il 50% di midollo osseo attivo si trova nella regione compresa tra la pelvi e la spina lombare, appare chiaro che, nei casi di trattamento di tumori della parte bassa dell'addome come ad esempio quello anale, l'individuazione del midollo rosso riveste un'importanza fondamentale. Il piano di trattamento radioterapico deve, infatti, essere altamente focalizzato e ridurre la dose di radiazione che investe il midollo attivo.

1.3 Individuazione del midollo attivo

Attualmente sono disponibili diverse tecniche per l'individuazione del midollo attivo. Di seguito vengono discusse quelle principali.

1.3.1 Biopsia

Consiste nel prelievo di un campione, in genere cilindrico, di midollo osseo direttamente dal paziente, esso viene quindi analizzato in laboratorio da parte del medico per determinarne la natura. Si tratta sicuramente del metodo più accurato per riconoscere e individuare il midollo attivo in quanto è possibile una osservazione diretta dello stesso. Tuttavia, è un esame invasivo e l'informazione che si ricava è relativa unicamente al campione estratto, quindi non è pensabile il suo utilizzo nell'ambito della pianificazione di un trattamento di radioterapia [4].

1.3.2 MRI

La risonanza magnetica è una tecnica che trova impiego nell'individuazione del midollo rosso in quanto si basa sulla differenza di quest'ultimo da quello giallo in termini di concentrazioni di acqua e grasso.

Nei lipidi i protoni si trovano legati ai gruppi metilene CH_2 di molecole relativamente pesanti che consentono un tempo di rilassamento spin-reticolo (T_1) molto breve e un segnale di elevata intensità per i tessuti adiposi in immagini di MR pesate in T_1. Al contrario, il tempo di rilassamento spin-spin (T_2) risulta relativamente elevato e ciò comporta un segnale di bassa intensità nelle immagini pesate

in T_2 [1]. Per l'acqua il comportamento è inverso, per cui in un'immagine pesata in T_1 apparirà relativamente scura, mentre avrà un colore più chiaro in un'immagine pesata in T_2.

Lo studio di Andreychenko et al [5] ha mostrato come sia possibile realizzare un sistema di individuazione del midollo attivo proprio a partire dalle immagini pesate in T_1 e T_2. In particolare, ciò che viene fatto è calcolare un parametro denominato fat fraction (FF) e definito come:

$$FF = \frac{Fat\ Image}{Fat\ Image + Water\ Image}$$

dove Fat Image e Water Image rappresentano, rispettivamente, l'immagine pesata in T_1 e quella pesata in T_2. A questo punto ciascun pixel viene classificato come appartenente o meno alla regione del midollo attivo sulla base del suo valore di intensità P, infatti:

$$se \qquad S_{min} < P < S_{max} \qquad allora\ P \in RM$$

dove S_{min} e S_{max} sono delle soglie il cui valore può cambiare in base alla specifica sequenza MR utilizzata.

L'MRI offre sicuramente il vantaggio di essere una tecnica ad elevata sensibilità e risoluzione spaziale che utilizza un campo elettromagnetico non ionizzante il quale non presenta effetti dannosi per il paziente. Tuttavia, si tratta di un esame costoso che richiede dei tempi di acquisizione relativamente lunghi e che non è sempre di facile accesso sul territorio.

Inoltre, è doveroso precisare come il metodo che è stato sviluppato per il riconoscimento del midollo attivo sia soggetto ad errori qualora si presentino dei tessuti che mostrano un contenuto di FF molto simile a quello del midollo attivo, ma che in realtà non svolgono alcuna funzione ematopoietica. In tal senso sono state previste delle operazioni di post-processing che contribuiscono a ridurre le aree erroneamente individuate, ma che non risolvono a pieno il problema, il quale resta quindi una causa di imprecisione. In generale, è stato anche visto che il calcolo della FF consente prevalentemente l'identificazione della zona di midollo attivo più estesa, perdendo quindi le regioni più piccole. La segmentazione della sola area più estesa è però, in linea di principio, sufficiente allo scopo di ottimizzare il piano di trattamento radioterapico per ridurre la tossicità ematologica [5].

1.3.3 PET

La tomografia a emissione di positroni è una tecnica della medicina nucleare che consente di mappare dei processi fisiologici all'interno dell'organismo. Questa si inserisce tra le metodiche utilizzabili per l'individuazione del midollo attivo dal momento che, in relazione al radiotracciante impiegato, sono possibili diversi sistemi di captazione delle cellule del midollo osseo rosso. I principali meccanismi adottati sono [1]:

- Captazione da parte del sistema reticolo endoteliale (RES) del midollo attivo: un esempio è quello del colloide zolfo Tc-99m, utilizzato prevalentemente in scintigrafia per la valutazione del midollo osseo. Il problema di questo radiotracciante è quello di non essere individuato solo dal RES ma anche dal fegato e dalla milza, con una conseguente riduzione dello stesso nella zona di interesse.

- Captazione da parte delle cellule eritropoietiche: viene utilizzato il citrato di Fe-52 che non viene assorbito da fegato e milza, ma ha il non trascurabile difetto di avere un costo particolarmente elevato.
- Captazione da parte delle cellule granulopoietiche: si basano sul labeling di globuli bianchi e anticorpi monoclonali, per esempio ottenuti dal Tc-99m. Sono difficili da produrre e vengono anche essi rilevati dal fegato e dalla milza.

In generale, la tecnica oggi più largamente impiegata è quella che fa uso del radiotracciante 2-fluoro-2-deossi-glucosio (FDG), un analogo del glucosio che, come tale, viene individuato maggiormente dalle cellule ematopoietiche contenute nel midollo attivo di quanto invece non accada in quello giallo. Questo radiofarmaco viene captato dai trasportatori di glucosio e condotto all'interno delle cellule dove la fosforilazione tramite esochinasi ne blocca la possibilità di attraversare ulteriormente la membrana delle cellule e venire rilasciato. A questo punto, dal momento che esso presenta un radionuclide F-18 sostituito al tipico gruppo ossidrile in posizione C-2 del glucosio, non può essere metabolizzato dalle cellule (glicolisi). È necessario attendere il decadimento del F-18 con rilascio conseguente di positroni affinché il radiofarmaco ottenga il gruppo 2-idrossile necessario alla glicolisi. Appare quindi chiaro come l'uso dell'FDG consenta una individuazione del midollo attivo a partire dal suo livello di attività metabolica e non dalla sua struttura morfologica come avveniva nei metodi precedentemente descritti.

Diversi studi [6] [7] hanno mostrato come sia possibile da immagini FDG-PET effettuare una segmentazione delle aree di midollo attivo tramite il calcolo di un parametro denominato Standardized Uptake Value (SUV), definito come:

$$SUV = \frac{[attività_{assorbita}]}{\frac{attività_{iniettata}}{massa_{paziente}}}$$

dove $attività_{iniettata}$ rappresenta la quantità di radiotracciante iniettata al paziente e $[attività_{assorbita}]$ la concentrazione di radiotracciante assorbito in un punto. Esistono diverse versioni alternative di questo indicatore, come ad esempio:

- SUV_{max}: pari al valore massimo dei SUV calcolati per ciascun pixel;
- SUV_{mean}: dato dal valor medio dei SUV calcolati per ciascun voxel all'interno di un dato volume di interesse (VOI);
- SUV_{peak}: pari al valore medio dei SUV in un volume sferico di diametro pari a 1.2 cm, rispetto al SUV_{mean} è un parametro meno sensibile al rumore statistico tipico delle immagini PET;
- SUV_{LBM}: in cui il SUV non viene normalizzato rispetto all'intera massa del paziente, ma rispetto alla sola massa magra dal momento che il tessuto adiposo non è metabolicamente attivo al pari di altri tessuti; è ritenuto avere una maggiore stabilità e una minore variabilità tra i pazienti rispetto al SUV;
- SUV_{BSA}: in cui il SUV non viene normalizzato rispetto alla massa del paziente ma rispetto alla sua superficie corporea; è anch'esso considerato preferibile rispetto al SUV_{mean} [8].

Una volta calcolato il SUV, ciò che viene fatto è un'operazione di thresholding per cui vengono identificati come midollo osseo attivo solamente quei punti il cui valore di SUV si trova al di sopra di una certa soglia come quella del SUV_{mean} [7].

Appare doveroso precisare che però la PET non è una tecnica esente da problemi. Anzitutto, si tratta di un esame che fornisce informazioni su tutto il corpo e che, quindi, non può essere limitato a singole aree di interesse. Ciò si ripercuote in un effetto di rumore di fondo che rende difficile lo sviluppo di sistemi automatici di segmentazione. Inoltre, è noto che la PET non è caratterizzata da una elevata risoluzione spaziale, principalmente da attribuirsi ai fenomeni di assorbimento e scattering. Attualmente la massima risoluzione ottenibile è stata misurata pari a 4-5 mm, in base agli attuali standard di valutazione delle prestazioni. Tuttavia, nelle immagini diagnostiche dell'uomo la risoluzione vera identificata si aggira al momento intorno agli 8-10 mm. Questa limitazione in termini di risoluzione spaziale è la causa di quello che viene definito Partial Volume Effect (PVE), per il quale delle lesioni di dimensione ridotta in un tessuto appaiono più grandi ma di minore intensità e contrasto. Per lesioni con dimensione minore di 2-3 volte la risoluzione spaziale della PET non è quindi possibile quantificare accuratamente la concentrazione del tracciante utilizzato [6].

In merito proprio al tracciante impiegato, bisogna sottolineare come l'esame FDG-PET sia caratterizzato da un problema di natura logistica-economica. Il radiofarmaco 2-fluoro-2-deossi-glucosio ha, infatti, un tempo di emivita piuttosto breve (circa 2 h) e pertanto solamente quei centri collocati nei pressi di ciclotroni o che possiedono denaro sufficiente per l'acquisto di uno di essi possono fare uso di una tecnica FDG-PET. Se si considera inoltre che di per sé la PET è anche un esame più costoso rispetto agli altri metodi di imaging, ciò che ne deriva è, come nel caso dell'MRI, una bassa disponibilità nel territorio.

Infine, la metodica dell'FDG-PET è prevista solamente come opzionale dalle linee guida internazionali sulle fasi del processo diagnostico e, di conseguenza, non tutti i pazienti vengono sottoposti ad essa [9].

Di seguito vengono riportate alcune tecnologie ibride col fine di sopperire ad alcuni dei problemi dei metodi sopra descritti.

1.3.3.1 PET/TC

Si tratta di un esame che fonde insieme l'uso di una PET con l'uso di una tomografia computerizzata (TC), al fine di aumentare la risoluzione spaziale della PET e di aggiungere un'informazione morfologica oltre a quella funzionale.

Sono stati sviluppati diversi sistemi di segmentazione automatica del midollo attivo tramite l'uso di questa tecnologia [10] [11] [12]. Fondamentalmente, lo schema operativo è il seguente:

- Acquisizione dell'immagine TC;
- Segmentazione delle ossa: a partire dalla TC con metodi come l'algoritmo "atlas-based" [10], che sovrappone all'immagine in esame delle maschere di segmenti ossei preregistrati soggetti a continua deformazione sino alla corretta segmentazione delle regioni scheletriche di interesse, oppure per mezzo di contorni attivi od ancora operazioni di thresholding;
- Acquisizione dell'immagine PET;
- Moltiplicazione delle due immagini: al fine di ottenere le sole aree scheletriche con la mappa dell'attività metabolica;
- Segmentazione del midollo attivo e quantificazione della sua attività: sulla base, ad esempio, del parametro SUV.

Con questa tecnica permangono, tuttavia, i problemi di costo di cui discusso in precedenza per la PET e si aggiunge anche un ulteriore inconveniente legato all'alto livello di radiazioni a cui viene sottoposto un paziente per il fatto che, oltre alla PET, anche la TC è un esame che richiede elevate dosi di radiazioni.

1.3.3.2 PET/MRI

Questa tecnica ibrida vuole andare ad aumentare, come la precedente, la risoluzione spaziale della PET [13]. In questo caso, però, ciò che cambia è che invece della segmentazione delle ossa si opera una segmentazione diretta del midollo osseo attivo a partire dall'immagine di MRI. La maschera così ottenuta può essere moltiplicata per l'immagine PET fornendo come risultato la mappa della sua attività metabolica.

Rispetto al metodo precedente il contenuto in termini di radiazioni è ridotto, ma il costo di questi sistemi è così elevato da aver limitato fortemente il loro utilizzo per quanto riguarda la pianificazione della radioterapia.

1.3 Radiomica

Nel vasto campo della medicina, la radiomica si pone l'obiettivo di analizzare le immagini mediche con lo scopo di estrarre informazioni di tipo quantitativo per mezzo di più o meno sofisticati algoritmi matematici [14].

L'idea è che la semplice visualizzazione da parte dell'operatore di un'immagine ottenuta tramite TC, MRI, PET o altra tecnica di imaging medicale limiti il processo di interpretazione delle stesse ad una analisi puramente qualitativa. I sistemi basati sulla radiomica vedono invece le immagini come delle matrici di pixel, tra i quali sono sepolti precise informazioni riguardanti la morfologia delle strutture in esame.

L'analisi che viene condotta è definita texture analysis e consiste nell'estrazione di relazioni, denominate feature, tra i pixel delle immagini e che possono essere di varia natura (feature di forma, feature statistiche del primo ordine, del secondo ordine e di ordini successivi).

La radiomica può quindi essere impiegata per lo sviluppo di sistemi di supporto decisionale in campo medico e il miglioramento di modelli predittivi, come nei casi degli algoritmi di segmentazione automatica delle immagini.

1.4 Obiettivo

Nelle immagini di tomografia computerizzata il midollo osseo attivo appare indistinguibile da quello giallo sia dal punto di vista dell'analisi visiva che da una prima indagine sui valori dei pixel che costituiscono il midollo. Tuttavia, l'utilizzo della TC per l'individuazione di midollo rosso costituirebbe un enorme vantaggio in termini economici e di accessibilità rispetto all'impiego delle metodiche di MRI e PET attualmente in uso. I soggetti che devono cominciare trattamenti di tipo oncologico sono stati generalmente già sottoposti ad un esame di TC, dato che quest'ultimo rappresenta un esame standard nel workflow ospedaliero che precede la pianificazione radioterapica.

In tale ottica, l'obiettivo di questo lavoro è quello di andare ad esplorare la possibilità di realizzare un sistema di segmentazione del midollo attivo a partire da immagini di tomografia computerizzata

sfruttando la radiomica e, in particolare, tecniche di machine learning. Dopo aver estratto un certo numero di feature dalle immagini in analisi, verranno quindi implementati dei metodi di classificazione e ne verranno valutate le prestazioni sul problema in esame.

2 Materiali e Metodi

2.1 Popolazione

Le immagini utilizzate per questo studio provengono da un gruppo di 50 pazienti in cura presso il dipartimento oncologico dell'Ospedale Molinette di Torino e affetti da carcinomi a cellule transizionali e/o squamose. Ciascun soggetto è stato sottoposto a due esami:

- Tomografia computerizzata: è stato utilizzato uno scanner CT Philips "BigBore" che presenta un pixel spacing di 0.93 mm e una distanza tra le slice di 3 mm e i pazienti sono stati disposti in posizione supina con dei supporti per ginocchia e caviglia;
- Tomografia a emissione di positroni: è stata impiegata una Philips Gemini PET/CT Tomography.

Le immagini provenienti dalla PET sono state utilizzate per ricavare le machere di riferimento del midollo osseo attivo. A tale scopo è stata effettuata una segmentazione semi-automatica tramite il calcolo del parametro SUV descritto in precedenza, il quale è stato normalizzato rispetto a quello calcolato nel fegato e nelle pelvi di ciascun paziente. Per andare a individuare il midollo attivo è stata effettuata un'operazione di thresholding prendendo come soglia il SUV_{mean} del paziente in esame e considerando come appartenenti a RM tutti quei pixel aventi valore di SUV superiore a tale soglia.

Per quanto riguarda invece le immagini di TC, esse sono state sottoposte ad un radioterapista oncologico, il quale ha suddiviso il midollo osseo in tre regioni distinte:

1) Midollo osseo lombosacrale (LSBM): include l'area tra il bordo somatico superiore della quinta vertebra lombare (L5) fino alla parte inferiore del coccige;
2) Porzione bassa del midollo osseo pelvico (LPBM): comprende il pube bilaterale, l'ischio, l'acetabolo e il femore prossimale;
3) Midollo osseo iliaco (IBM): include l'area tra le creste iliache e il bordo superiore.

In una prima fase di pre-processing, si è provveduto alla rimozione del tavolino medico e dei tessuti non ossei, in seguito è stata effettuata una rimozione dell'osso corticale, in quanto, come noto, esso non contiene midollo osseo. A tale scopo, per ciascuna slice di ogni paziente si è eseguito un

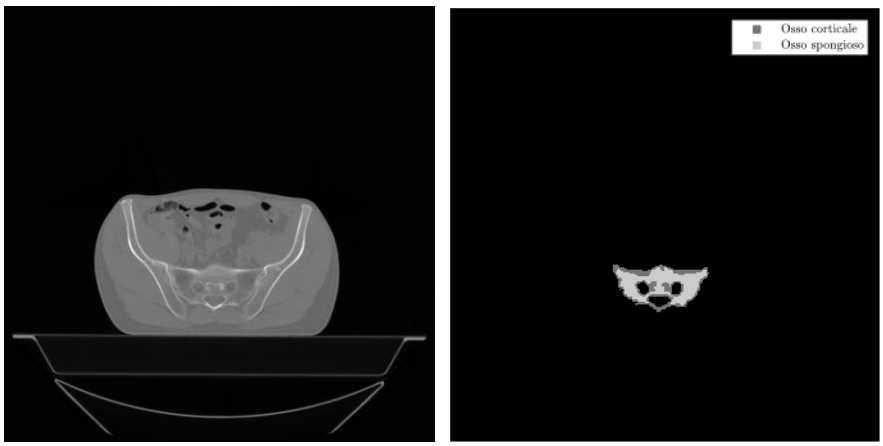


Figura 2: A sinistra: immagine TC del paziente 1 slice -93.5; A destra: maschera dopo rimozione di tavolino medico e tessuti non ossei con isolamento della struttura LSBM, si evidenziano i due cluster di osso corticale e spongioso.

k-means clustering in base all'intensità di ogni pixel andando a distinguere due cluster: il primo, ad intensità maggiore, racchiude i pixel facenti parte dell'osso corticale a densità maggiore, mentre il secondo comprende i pixel dell'osso spongioso. In tal modo dalle immagini TC è stato possibile ottenere le maschere binarie del solo midollo osseo (figura 2).

2.2 Feature extraction e organizzazione del dataset

A partire dalle immagini TC dei 50 pazienti a disposizione, è stata effettuata la texture analysis su ciascuna slice analizzata. Sono state utilizzate a tal fine delle regioni ottagonali di dimensione 5x5 pixel, denominate ROI, fatte muovere su tutta l'area dell'osso spongioso di un pixel alla volta sia in direzione orizzontale che verticale cosicché le ROI risultassero sempre parzialmente sovrapposte. Da ognuna di queste regioni sono stati estratti 36 descrittori così suddivisi:

- 4 feature statistiche del primo ordine: media, deviazione standard, skewness e kurtosis dei pixel in esame;
- 32 feature statistiche del secondo ordine: in particolare 22 di queste ultime provenienti dalla Gray Level Co-occurrence Matrix (GLCM), 5 estratte dalla Gray Level Dipendence Matrix (GLDM) e le ultime 5 derivanti dalla Gray Level Run Length Matrix (GLRLM).

Alla fine ciò che si è ottenuto è una matrice di dimensione n x 41 (con n numero di ROI) dove le prime due colonne rappresentano, rispettivamente, l'ID del paziente e il numero della slice da cui proviene la ROI per la quale sono state calcolate le feature, la terza e la quarta colonna contengono le coordinate x e y del pixel centrale della ROI, si trovano poi i valori delle 36 feature e infine, nell'ultima colonna, si trova un valore che indica l'appartenenza di quella ROI a RM, YM o al bordo tra i due (0 = YM, 1 = RM, 2 = bordo). Di seguito è possibile vedere un esempio della matrice ricavata:

ID paziente	Slice	X	Y	Feature 1	Feature 2	…	Feature 35	Feature 36	Classe
1	-21.5	159	322	1.17e+03	20.73	…	19.52	0.976	0
1	-24.5	228	327	1.21e+03	40.65	…	19.52	0.976	1
1	-54.5	297	355	1.17e+03	48.41	…	20.25	0.988	2
2	-153	185	247	1.30e+03	67.30	…	20.26	0.988	0
3	-148	147	180	1.19e+03	42.44	…	20.26	0.988	1

Il dataset è stato, quindi, suddiviso in construction set e validation set, dove il primo conteneva solamente ROI appartenenti ai primi 40 pazienti mentre il secondo quelle provenienti dai restanti 10 pazienti.

Nel dettaglio, per l'organizzazione del training set e del test set, impiegati rispettivamente per l'allenamento e la verifica dei classificatori costruiti si è proceduto come segue. A partire dai primi 40 pazienti, sono state estratte 5 slice contenenti almeno 25 ROI di RM e altrettante di YM e sono state sottoposte a clusterizzazione mediante una rete SOM (Self Organizing Map) di dimensione 12 e vicinato 1, la quale ha suddiviso le ROI in un determinato numero di cluster.

La rete SOM è un tipo particolare di rete neurale utilizzata nell'ambito dell'apprendimento non supervisionato. Essa è caratterizzata dalla formazione di una mappa topografica solitamente 2D (talvolta 3D) di un certo numero di neuroni. Nel momento in cui la rete riceve in ingresso un input, i neuroni entrano in competizione tra loro per determinare quale presenta le caratteristiche più idonee

al riconoscimento dell'input, questo neurone viene definito "vincitore". A questo punto i pesi del neurone vincitore vengono modificati e, assieme a questi, anche quelli dei suoi neuroni vicini. Con questa tecnica neuroni vicini tra loro sono rappresentativi di elementi simili tra loro, mentre neuroni distanti tra loro di elementi diversi tra loro.

Da ognuno dei cluster derivati dalla rete SOM sono state estratte, in maniera proporzionale alla numerosità complessiva del cluster, un numero di ROI tale da ottenere un training set composto da un totale di circa 10000 ROI di cui metà appartenenti a RM e metà a YM. Tutti gli elementi non inclusi nel training set sono stati inseriti nel test set. L'insieme di training set e test set costituisce il construction set. Dal momento che le strutture in esame sono tre (LSBM, IBM e LPBM), sono stati ricavati tre construction set distinti.

Successivamente, è stata effettuata una normalizzazione delle feature utilizzando la tecnica del min-max scaling e riportando quindi l'intervallo di ciascuna feature tra 0 e 1. Se x rappresenta una colonna di una feature della matrice ottenuta, l'operazione che viene effettuata è:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

dove x_{norm} rappresenta la colonna normalizzata. Questa normalizzazione è stata eseguita su tutte le ROI delle tre strutture separatamente, andando ad utilizzare come valori di minimo e massimo quelli provenienti da ciascuno dei tre construction set.

2.3 Metodi di classificazione

Di seguito viene riportata una descrizione dei classificatori utilizzati in questo studio.

2.3.1 Decision tree

Si tratta di un classificatore supervisionato basato su una struttura ad albero in cui, a partire da un nodo radice, si sviluppano una serie di ramificazioni terminanti con nodi finali detti di foglia (figura 3). Ciascun nodo interno dell'albero esegue un test su una variabile e ogni ramo rappresenta quindi il risultato di un test effettuato. Alla fine, i nodi di foglia rappresentano le classi di destinazione.

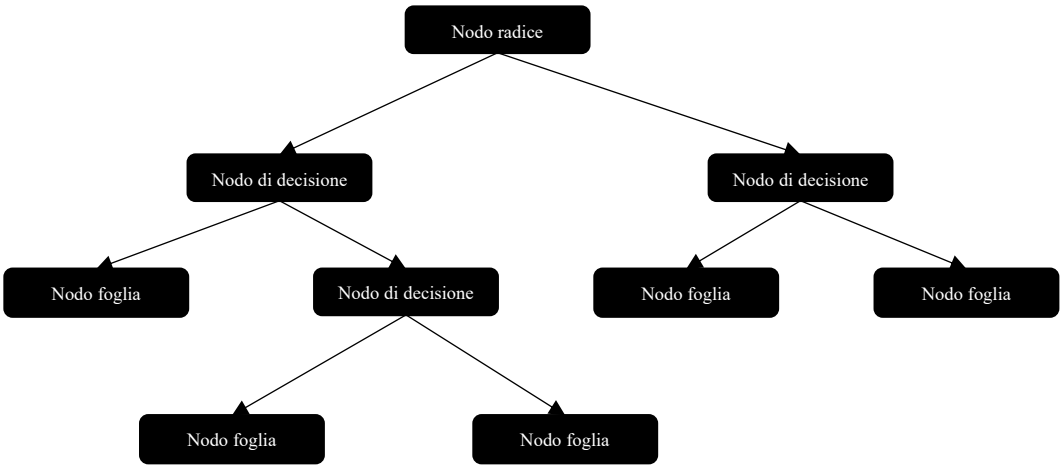


Figura 3 Schema costitutivo di un decision tree

Durante la fase di training del classificatore si procede in modo iterativo per ricercare la migliore regola di separazione possibile (splitting rule), ossia quella regola che consente di suddividere il

training set in partizioni che siano il più pure possibile, dove per pura si intende una partizione contenente elementi appartenenti ad un'unica classe. Una volta trovata la regola migliore questa viene associata ad un nodo interno dell'albero e si procede così via sino a quando tutte le partizioni sono pure o sino a quando non ci sono più feature utili per ricavare splitting rule. A questo punto si sono ottenuti i nodi finali e a questi viene associata la classe di appartenenza tramite majority voting degli elementi contenuti nella sua partizione.

2.3.2 K-nearest neighbors

È un classificatore supervisionato che associa ciascun elemento ad una classe sulla base della sua similarità con gli elementi del training set. La misura di similarità può essere di vario tipo, ma in questo lavoro è stata utilizzata la distanza euclidea. Ciò che viene fatto è, quindi, calcolare la distanza dell'elemento da classificare da ognuno degli elementi del training set, disporre le distanze in ordine crescente, selezionare le prime K distanze ordinate e assegnare l'elemento alla classe più rappresentata tra le K selezionate. L'unico parametro per il quale è necessaria una ottimizzazione è pertanto il numero K di distanze da considerare. A tale scopo è buona misura inizializzare il valore di K come:

$$K_{in} = \sqrt{N}$$

dove N rappresenta il numero di elementi che costituiscono il training set.

2.3.3 Multilayer perceptron Neural Network

È una rete neurale di tipo feedforward particolarmente utilizzata nei problemi di classificazione e costituita da una struttura a più layer come segue (figura 4):

- Layer di input: rappresentato da un numero di neuroni pari al numero di feature in ingresso alla rete;
- Layer nascosti: sono in numero variabile in funzione del problema trattato e il numero di neuroni di ogni layer nascosto può anch'esso essere diverso;
- Layer di output: possiede un numero di neuroni che è funzione del numero di classi del problema in esame.

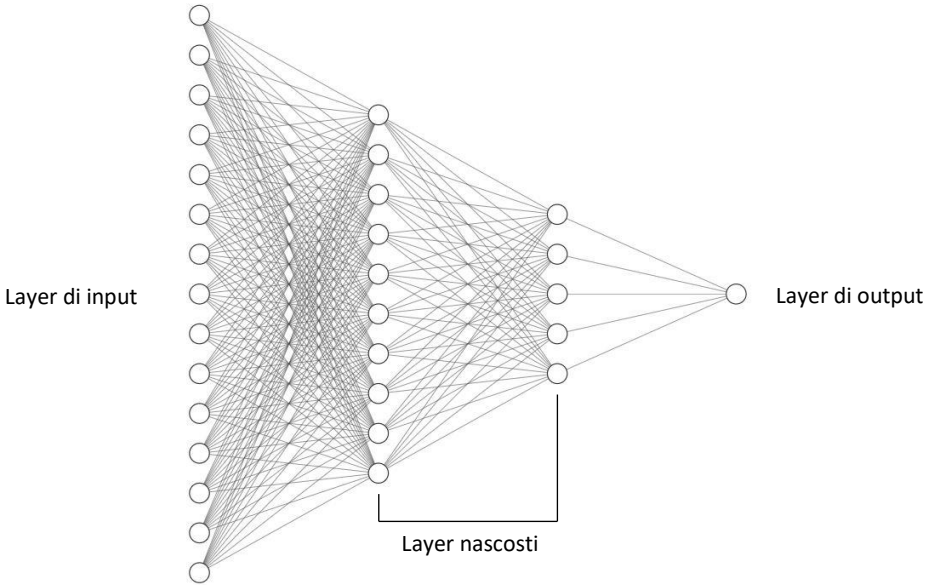


Figura 4 Schema costitutivo di esempio di una rete neurale multilayer perceptron. A sinistra un layer di input con 15 neuroni, al centro 2 layer nascosti con 10 e 5 neuroni rispettivamente, a destra un layer di output con un solo neurone.

I neuroni di un layer sono organizzati in modo tale da essere ognuno connesso a tutti i neuroni dei layer adiacenti. Ciascun neurone riceve input dall'ambiente o da altri neuroni per mezzo di connessioni dotate ognuna di un determinato peso, detto peso sinaptico. Viene effettuata una somma pesata degli input per mezzo di una particolare funzione di attivazione e l'output così generato viene inviato ad altri neuroni o risulta essere l'output della rete. Nella fase di apprendimento ciò che succede è che l'output generato dalla rete viene confrontato con l'output desiderato e viene calcolato l'errore tra i due, il quale viene riportato indietro per modificare i pesi sinaptici delle connessioni al fine di diminuire l'errore al passo successivo.

Per questo lavoro le reti costruite possiedono sempre un layer di input con un numero di neuroni pari a quello derivante da una feature selection effettuata e un layer di output con un solo neurone e una funzione di attivazione lineare che restituisce in output un valore tra 0 e 1. Il valore restituito dal neurone di output viene utilizzato per determinare la classe di appartenenza dell'elemento analizzato come segue:

$$se\ Output\ < 0.5\ \rightarrow Output\ \in YM$$

$$se\ Output\ \geq 0.5\ \rightarrow Output\ \in RM$$

Per quanto concerne il numero di layer nascosti e di neuroni dei layer nascosti, essi dipendono strettamente dal problema in analisi per cui sono stati sottoposti ad una ottimizzazione descritta nel paragrafo seguente.

2.4 Algoritmi genetici

Il metodo di ottimizzazione impiegato in questo studio è un metodo wrapper, il quale per definizione dipende dall'algoritmo di apprendimento utilizzato, unito ad un approccio di tipo randomizzato: si tratta di un algoritmo genetico (GA). Esso è un metodo di evolutionary computing che si basa sull'evoluzione di un insieme di soluzioni iniziali ricombinate tra loro e trasformate casualmente con lo scopo di andare ad individuare una soluzione "ottima".

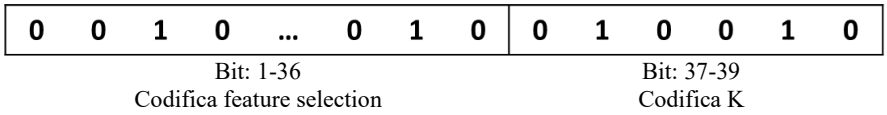
In prima istanza è necessario codificare la soluzione sotto forma di un vettore binario tramite l'uso di qualche tecnica. Fatto ciò, gli step principali dell'algoritmo sono i seguenti:

- Generazione di una popolazione iniziale: viene generata in maniera random una popolazione di soluzioni iniziali che siano tutte quante ammissibili, ossia che soddisfino gli eventuali vincoli del problema in esame;
- Calcolo della fitness della popolazione iniziale: la fitness è un valore numerico che viene attribuito a ognuna delle soluzioni della popolazione sulla base della sua bontà rispetto al problema di ottimizzazione affrontato;
- Selezione dei genitori: dalla popolazione iniziale viene scelto un sottoinsieme di soluzioni che diverranno i genitori della successiva generazione di soluzioni;
- Applicazione degli operatori genetici: gli operatori genetici permettono l'evoluzione delle soluzioni effettuando una ricombinazione e una trasformazione delle stesse. Possiamo distinguere due operatori: il primo è la mutazione che consiste nel complemento di uno o più bit della soluzione in esame e definito da una certa probabilità P_m, il secondo è il crossover, il quale consiste nel tagliare due soluzioni in uno o più punti e nello scambiarne tra loro le sottostringhe ed è anche esso definito da una certa probabilità P_c;

- Valutazione della nuova generazione: ogni soluzione della nuova generazione viene valutata tramite il calcolo della fitness e si ripete la selezione dei genitori della generazione seguente;
- Condizione di stop: l'algoritmo agisce iterativamente sino a che una determinata condizione di stop viene raggiunta.

In questo caso il GA è stato adoperato sia per effettuare feature selection che per andare a determinare e ottimizzare i parametri dei classificatori testati. Pertanto ogni soluzione è stata codificata in due parti: la prima è rappresentata da un vettore binario di 36 bit dove ogni bit corrisponde a una delle feature estratte e dove il valore 1 corrisponde alla selezione di quella specifica feature come facente parte di quelle utilizzate per la classificazione, mentre il valore 0 rappresenta la mancata selezione; la seconda parte della soluzione è costituita da un vettore binario di lunghezza variabile in funzione di ciascun classificatore, che codifica la scelta dei parametri dello stesso. Più nel dettaglio:

- Decision tree: data la sua organizzazione, il decision tree non presenta dei parametri da ottimizzare, per cui l'esecuzione del GA avverrà solo per la feature selection e ciascuna soluzione sarà costituita solo dalla prima parte descritta, ovvero da un vettore binario di 36 bit.
- KNN: a partire dal valore inizializzato di K_{in} si predispone una soluzione del GA che presenta la consueta prima parte di feature selection a 36 bit e una seconda parte rappresentata da un vettore binario di 6 bit il quale determina il numero di valori da esplorare intorno a K_{in}.



  In particolare, il valore di K di cui si ricerca l'ottimizzazione è definito come:

$$K = K_{in} + Bit_{dec} - 2^5$$

  dove Bit_{dec} è il valore decimale corrispondente al numero binario ottenuto dai 6 bit finali della soluzione dell'algoritmo genetico.

- NN: all'interno dell'algoritmo genetico si è predisposta la seconda parte della soluzione come un vettore di 3 bit, il cui valore in decimale, incrementato di un'unità, determina il numero di layer nascosti da implementare nella rete per quella soluzione (da 1 a 8 layer possibili). Il numero di neuroni di ogni layer nascosto è stato organizzato in modo tale che il primo layer nascosto abbia un numero di neuroni pari a quello del layer di input, mentre i successivi abbiano una struttura piramidale per cui ciascun layer nascosto conta un numero di neuroni pari alla metà di quelli del layer che lo precede (nel caso di un numero di neuroni dispari la metà viene approssimata per eccesso). Tutti i neuroni dei layer nascosti sono caratterizzati da una funzione di attivazione a sigmoide.

Per quanto riguarda il calcolo della fitness è stata adoperata la seguente funzione:

$$fitness = 1 - accuracy + 0.3 \cdot |sensitivity - specificity|$$

dove i parametri di accuracy, sensitivity e specificity sono quelli ricavati dalla valutazione delle performance di ciascun classificatore allenato sul training set e verificato sul test set con un sottocampionamento di quest'ultimo di un fattore pari a cinque. Data la struttura della funzione,

appare chiaro come un valore di fitness sia tanto migliore quanto più sia basso. Una funzione di questo tipo tende a favorire soluzioni ad elevata accuratezza e, parzialmente, soluzioni che abbiano valori di sensibilità e specificità simili. L'intento è quello di evitare una generale tendenza dei classificatori a sovra-segmentare o sotto-segmentare il midollo attivo.

Andando ad esaminare i parametri dell'algoritmo genetico sono stati scelti:

- Popolazione iniziale costituita da 500 individui con il vincolo di possedere almeno 2 feature selezionate; ogni ripetizione dell'algoritmo parte sempre dalla stessa popolazione iniziale;
- Probabilità di crossover $P_c = 1$ e probabilità di mutazione $P_m = 0.5$: valori così elevati sono scelti a causa dell'elevato numero di soluzioni possibili onde aumentare la variabilità delle soluzioni tra le varie generazioni ed evitare la convergenza a soluzioni apparentemente ottimizzate, per contro le soluzioni ottenute non saranno molto ottimizzate;
- Numero di genitori pari al 70% del numero di individui della popolazione iniziale (350 genitori): il metodo di selezione è quello della roulette, per cui ogni soluzione ha una probabilità di essere scelta inversamente proporzionale al suo valore di fitness;
- Numero di iterazioni dell'algoritmo genetico pari a 100, con l'aggiunta di una condizione di stop su 30 iterazioni consecutive: se non si osservano miglioramenti nella funzione di fitness per più di 30 iterazioni consecutive la ripetizione del GA viene interrotta e si passa alla successiva;
- Numero di ripetizioni pari a 5: il GA viene ripetuto 5 volte in quanto, sebbene si utilizzi la stessa popolazione iniziale, la natura casuale dell'algoritmo non assicura che si giunga sempre alla stessa soluzione.

L'algoritmo genetico è stato effettuato separatamente per ciascuna delle tre strutture in cui è stato diviso il midollo osseo, per cui l'apprendimento di ogni classificatore avviene una struttura alla volta.

3 Risultati GA

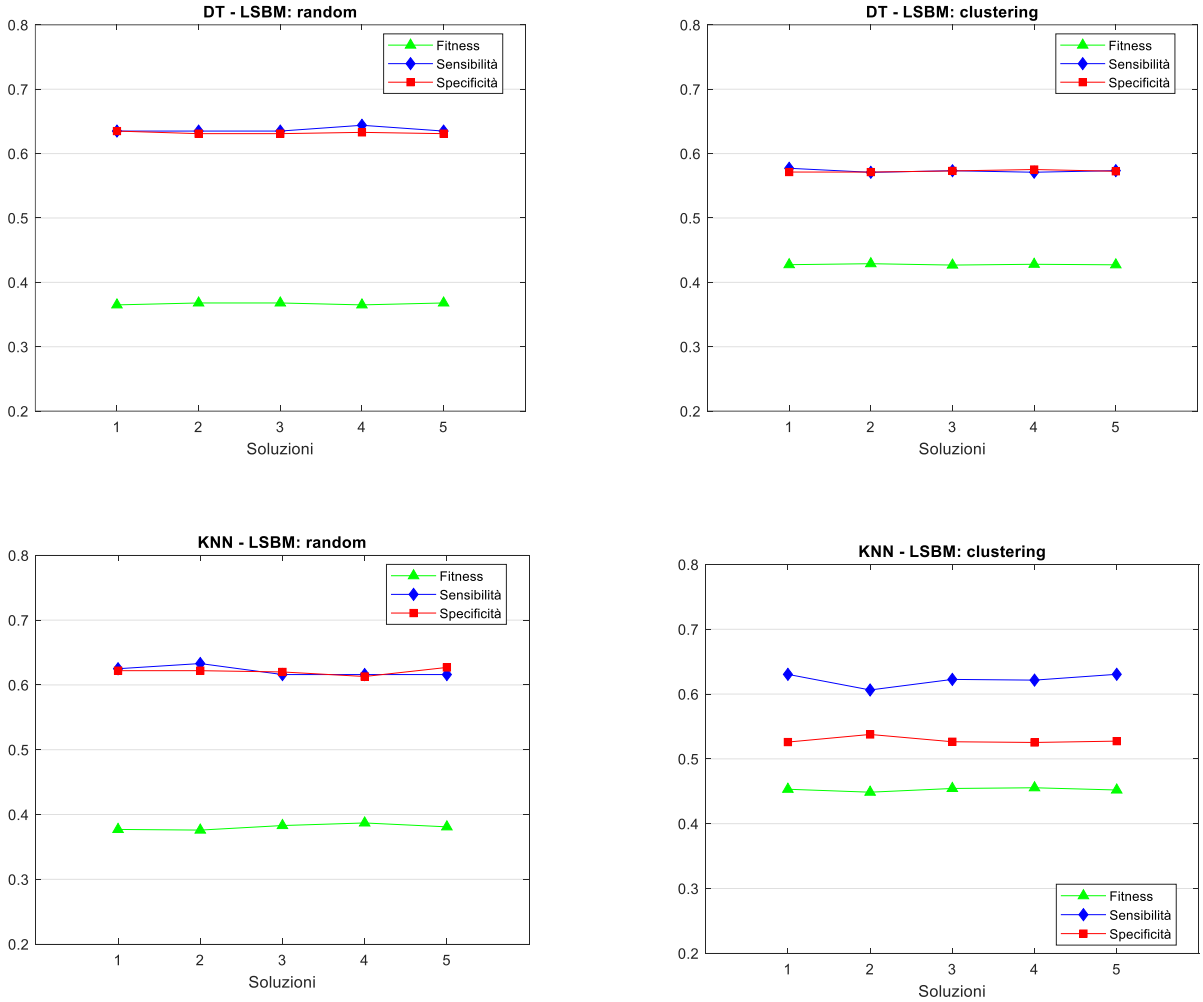
3.1 Training set: random vs. clustering

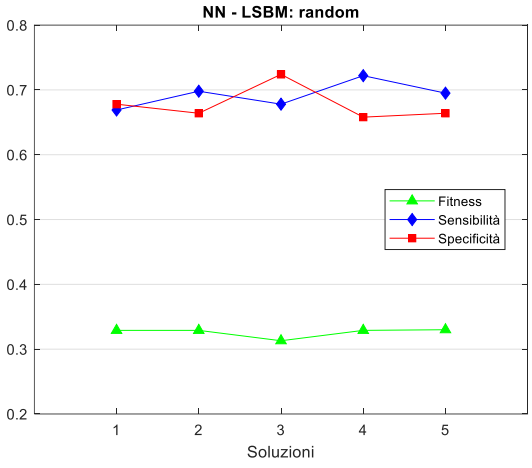
Nel corso di uno studio precedente [15], effettuato sulla struttura LSBM, è stato notato che il training set ricavato mediante clusterizzazione con rete SOM portava ad un allenamento di reti di deep learning migliore rispetto ad un generico training set estratto in maniera random.

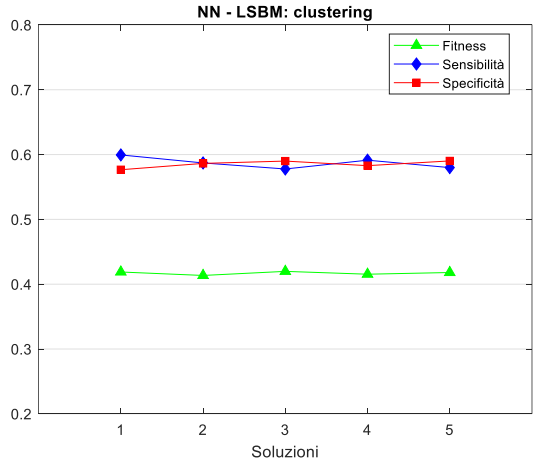
Per tale ragione, si è deciso di confrontare i risultati di questo nuovo lavoro su Machine learning con quelli precedentemente ottenuti a partire da un training set random [16], così da verificare se anche con tali metodi la scelta della clusterizzazione conducesse a performance migliori.

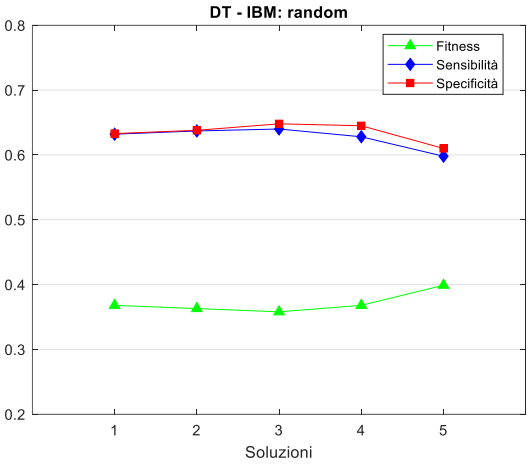
Da ciascuna delle 5 ripetizioni del GA è stata ottenuta una soluzione caratterizzata da un dato valore di fitness, di sensibilità e di specificità.

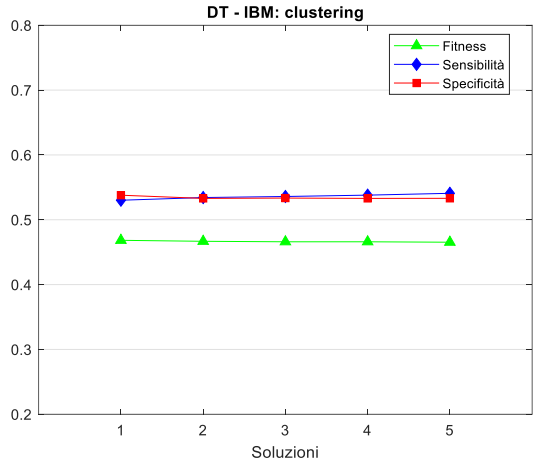
Di seguito il confronto tra i due metodi di ottenimento del training set, per ogni struttura e ogni classificatore (figura 5).

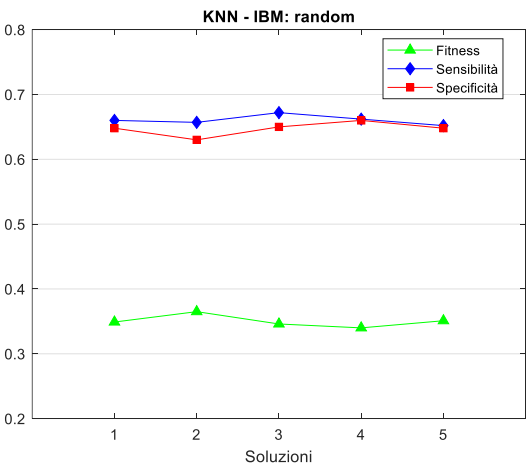


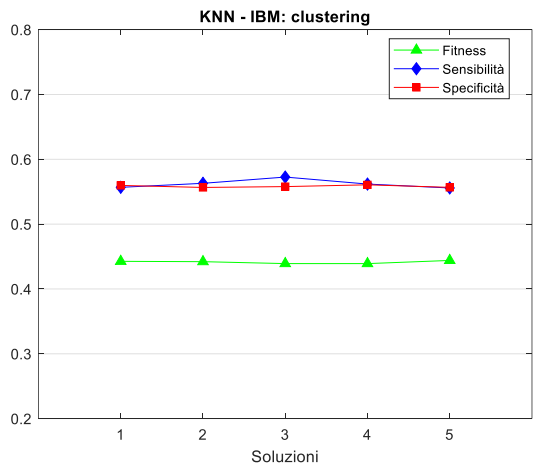

NN - LSBM: random
Fitness
Sensibilità
Specificità
Soluzioni

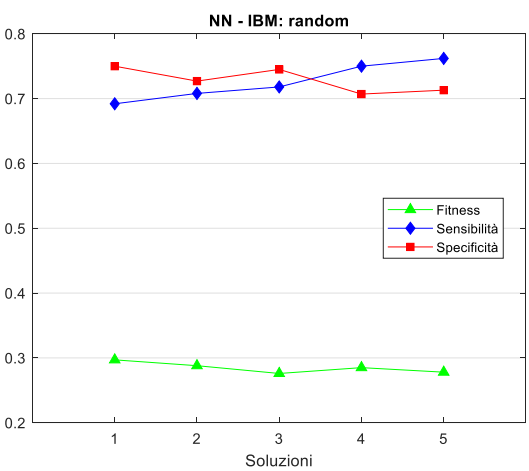

NN - LSBM: clustering
Fitness
Sensibilità
Specificità
Soluzioni

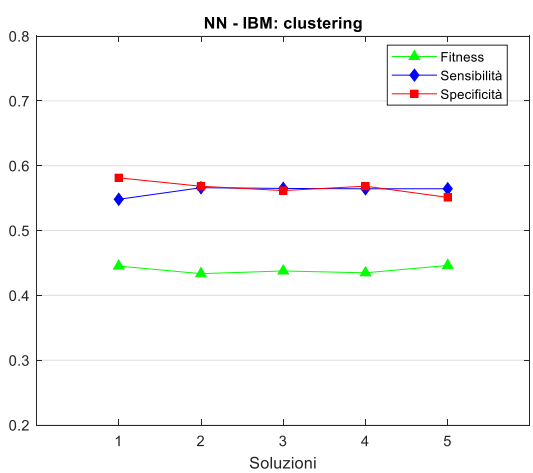

DT - IBM: random
Fitness
Sensibilità
Specificità
Soluzioni


DT - IBM: clustering
Fitness
Sensibilità
Specificità
Soluzioni


KNN - IBM: random
Fitness
Sensibilità
Specificità
Soluzioni


KNN - IBM: clustering
Fitness
Sensibilità
Specificità
Soluzioni


NN - IBM: random
Fitness
Sensibilità
Specificità
Soluzioni


NN - IBM: clustering
Fitness
Sensibilità
Specificità
Soluzioni

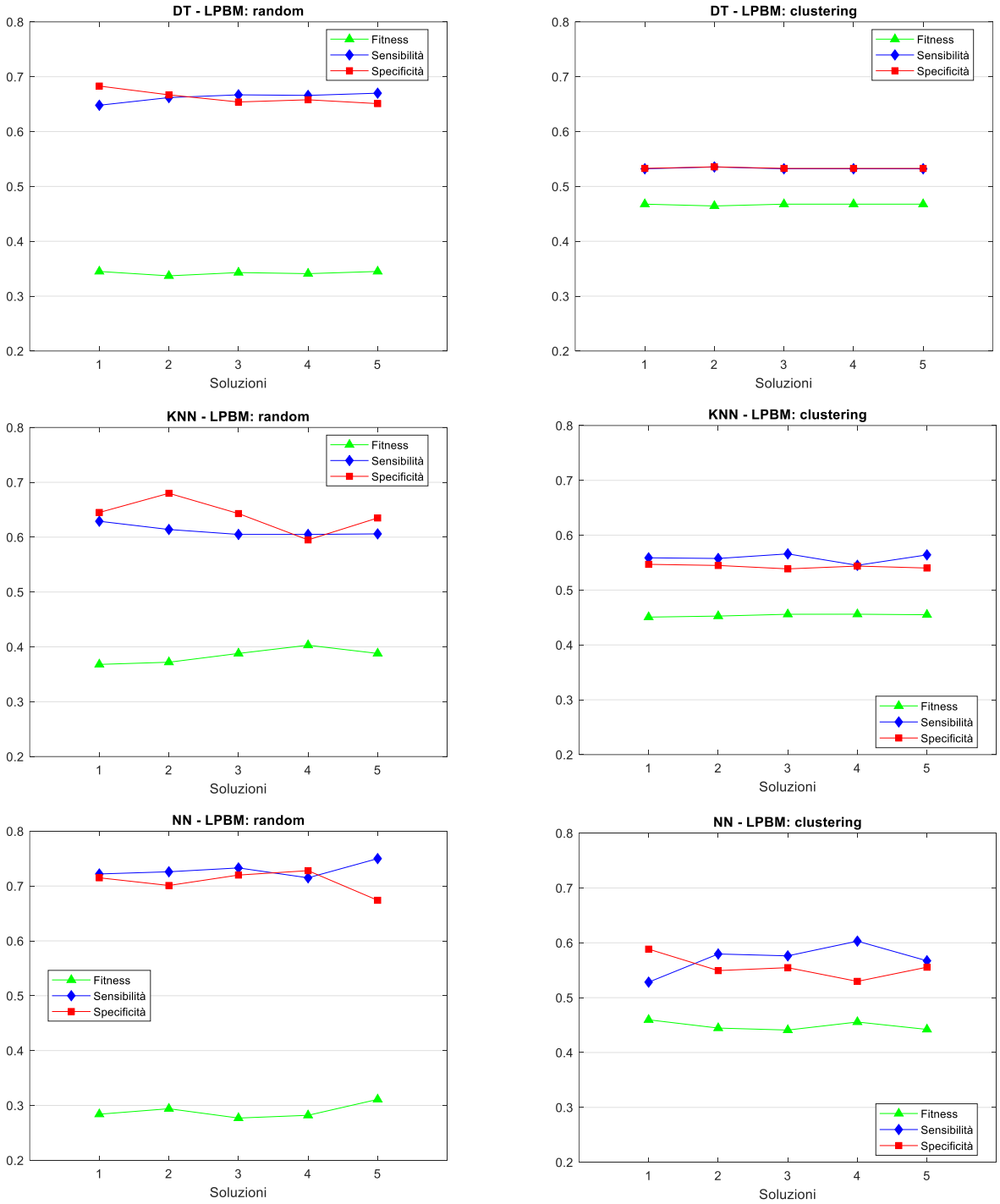


Figura 5 Risultati dell'algoritmo genetico per ciascuna struttura e per ogni classificatore. In ascissa sono riportate le cinque soluzioni ottenute e in ordinata i valori di fitness, sensibilità e specificità, rispettivamente in verde, blu e rosso.

I risultati in termini di fitness, sensibilità e specificità appaiono mediamente peggiori per tutte le strutture per il training set con clustering rispetto a quello random in tutti i classificatori analizzati. In quest'ultimo caso, abbiamo che, in funzione della struttura e del classificatore, i valori di fitness si aggirano intorno allo 0.25-0.4 con valori di sensibilità e specificità sempre superiori allo 0.6 e che, in taluni casi, superano anche lo 0.7. Con il metodo di ottenimento via clustering, invece, la fitness non scende al di sotto dello 0.4 per nessuna combinazione di struttura e classificatore. Tuttavia, è fondamentale considerare che la dimensione del test set utilizzato nello studio con training set random era notevolmente minore rispetto a quello attualmente qui utilizzato, per cui i

valori di fitness, sensibilità e specificità potrebbero essere più elevati semplicemente in virtù di tale ragione. Di conseguenza, si sono considerate non pienamente attendibili i risultati ottenuti a tale livello dalla sola valutazione sulle ROI e si è proceduto con l'ottenimento delle maschere di segmentazione dell'RM seguendo esattamente le stesse operazioni di voting e di post-processing dello studio svolto con il training set random. Per il voting sono state considerate le tre soluzioni a fitness minore, a sensibilità maggiore e a specificità maggiore.
Vengono ora riportati, per ogni classificatore, alcuni esempi di maschere di segmentazione prima di effettuare il post processing e una valutazione sui primi 20 pazienti delle performance ottenute a seguito del post processing, in modo da poter confrontare i risultati dei due metodi di ottenimento del training set.

- Classificatore DT:

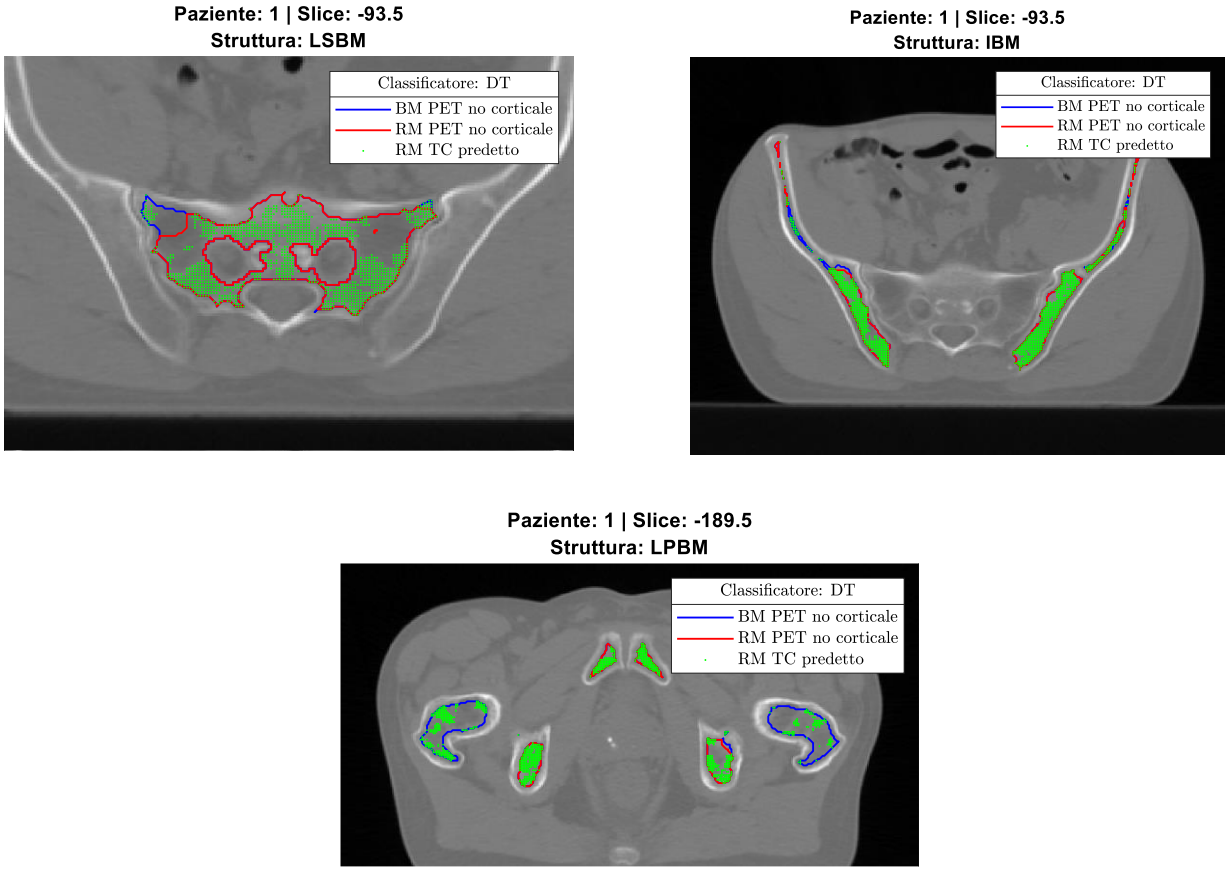


Figura 6 Maschere di segmentazione di LSBM, IBM e LPBM tramite classificatore DT. In verde la classificazione ottenuta, in blu il contorno della maschera del midollo osseo proveniente dalla PET con esclusione dell'osso corticale, i rosso il contorno della maschera del midollo attivo proveniente dalla PET con esclusione dell'osso corticale.

Per la struttura LSBM le performance appaiono molto simili tra i due training set random e con clusterizzazione nel caso del classificatore DT. Tuttavia, nel caso di clustering le performance della recall sembrano soggette a una variabilità inter-paziente leggermente minore (figura 7).

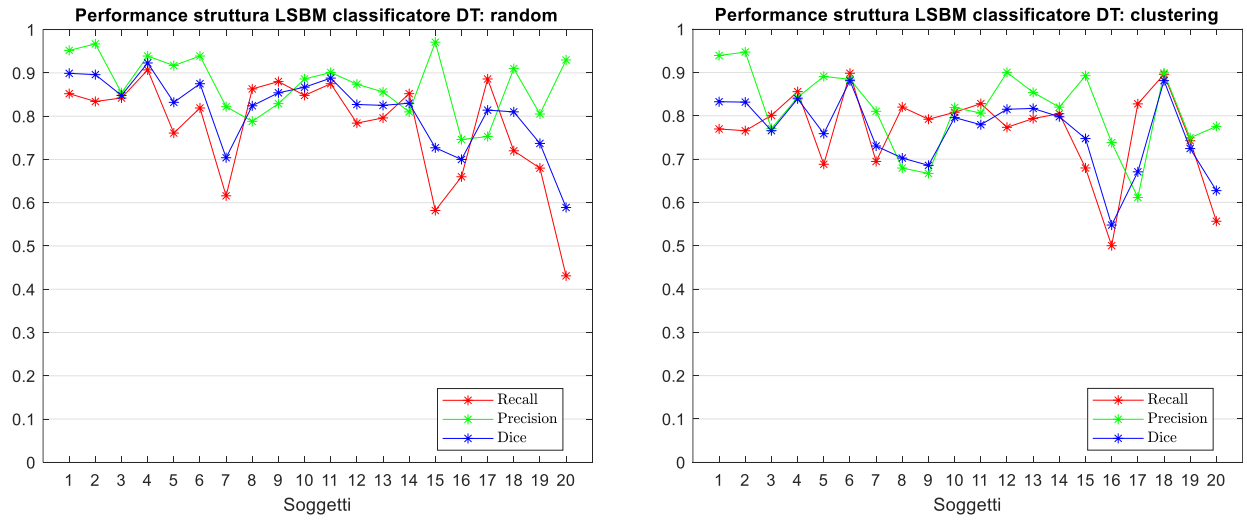


Figura 7 Performance struttura LSBM e classificatore DT, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Per la struttura IBM le performance appaiono leggermente più basse nel caso di training set con clusterizzazione. Anche qui, però, nel caso di clustering le performance della recall sembrano soggette a una variabilità inter-paziente di poco minore (figura 8).

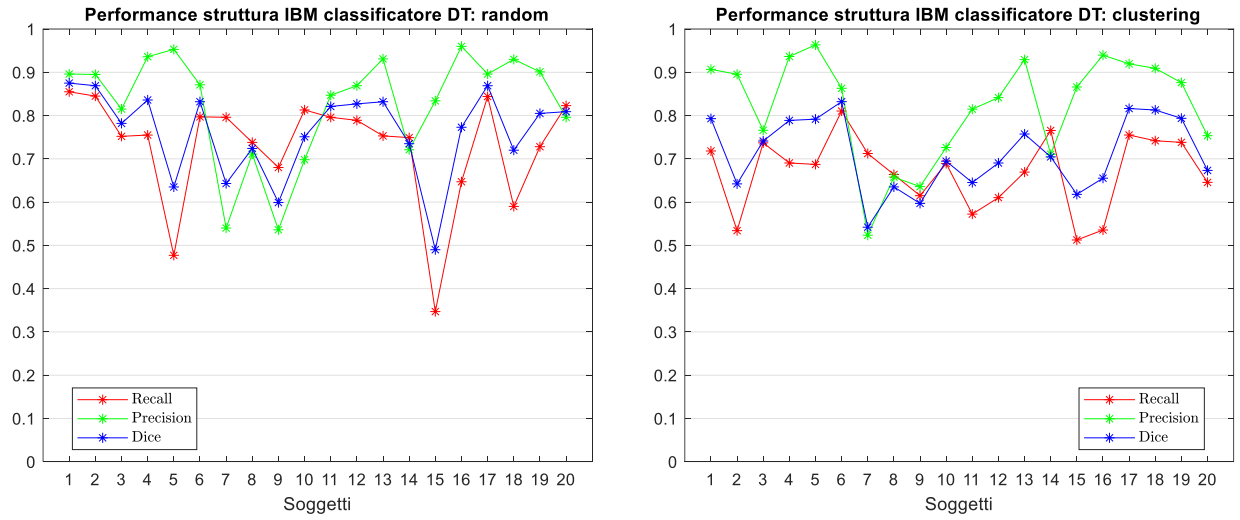


Figura 8 Performance struttura IBM e classificatore DT, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Come per la struttura precedente, anche per l'LPBM è possibile vedere come le performance appaiono più basse nel caso di training set con clustering, ma la variabilità inter-paziente della recall sia leggermente minore (figura 9).

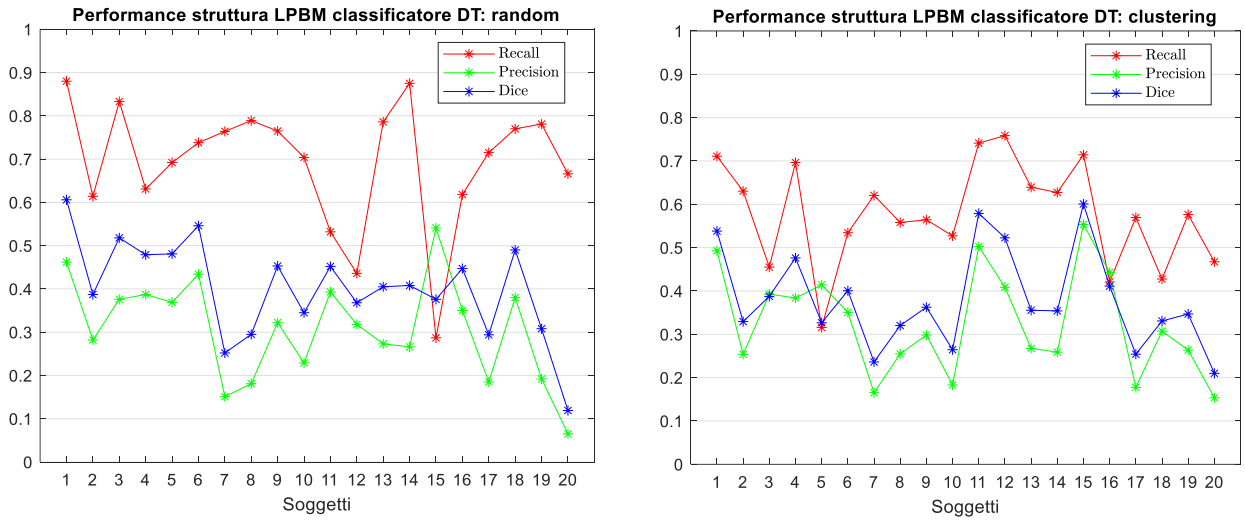


Figura 9 Performance struttura LPBM e classificatore DT, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

- Classificatore KNN:

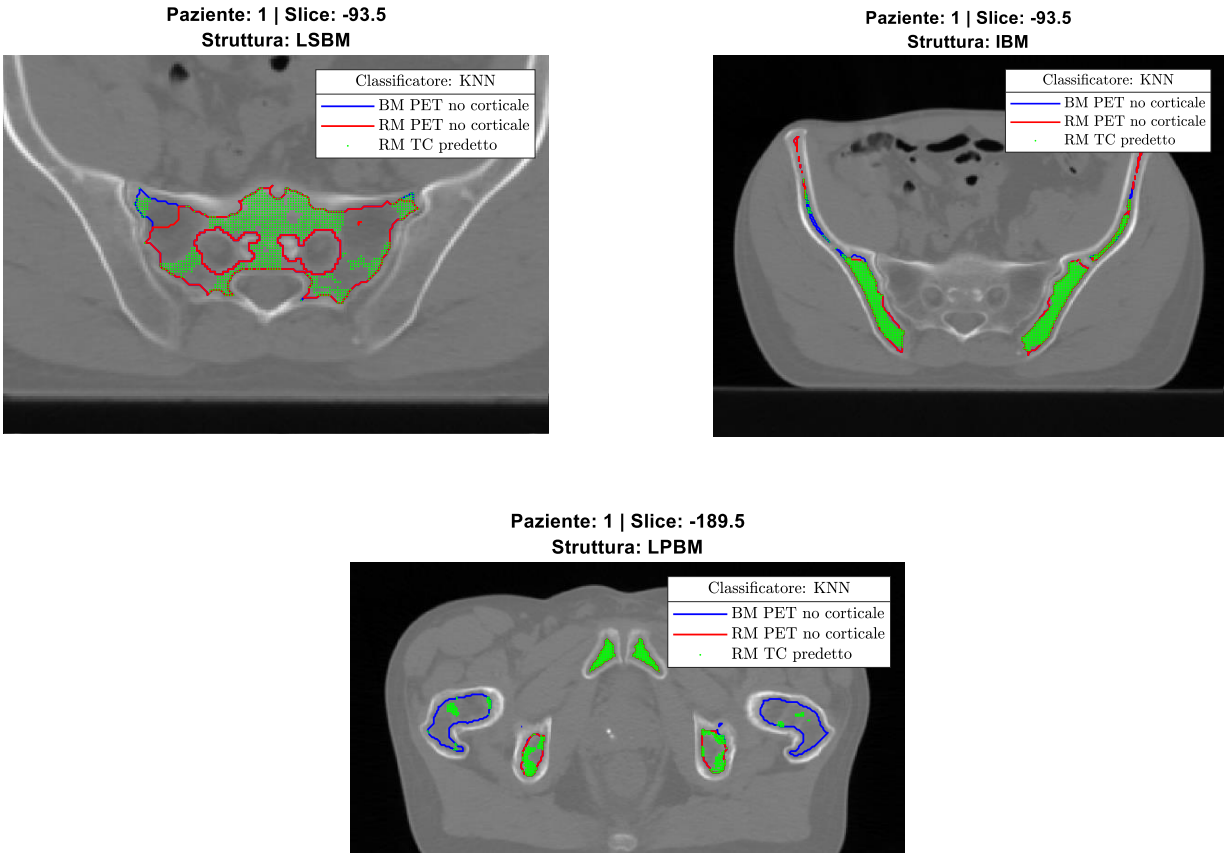


Figura 10 Maschere di segmentazione di LSBM, IBM e LPBM tramite classificatore KNN. In verde la classificazione ottenuta, in blu il contorno della maschera del midollo osseo proveniente dalla PET con esclusione dell'osso corticale, i rosso il contorno della maschera del midollo attivo proveniente dalla PET con esclusione dell'osso corticale.

Passando al caso del classificatore KNN, si può notare come per la struttura LSBM le performance appaiano leggermente migliori nel caso di training set con clusterizzazione. Inoltre, in tal caso, le performance sembrano soggette a una variabilità inter-paziente minore (figura 11).

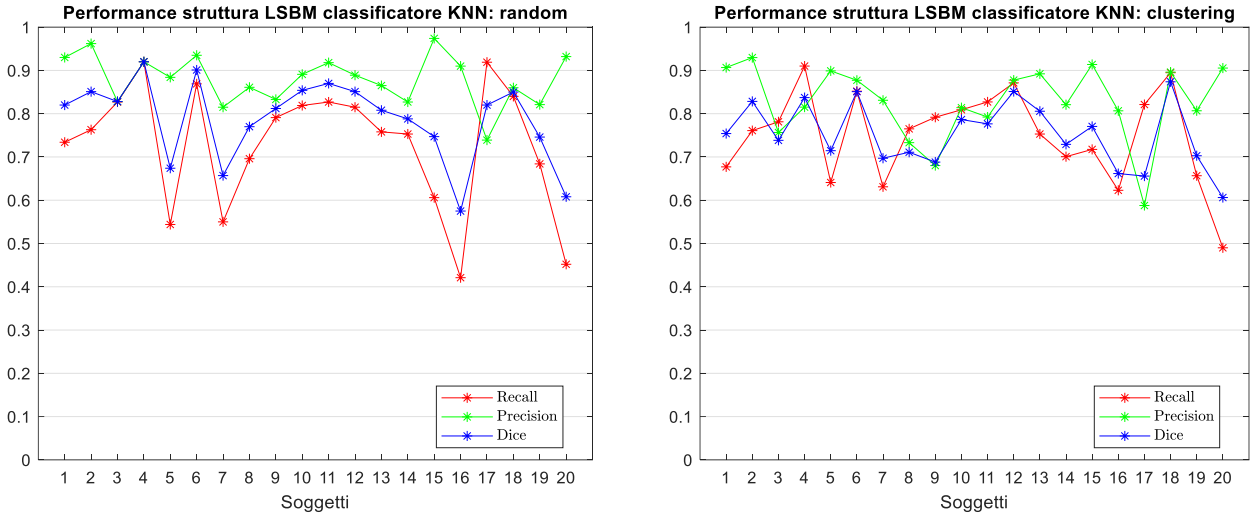


Figura 11 Performance struttura LSBM e classificatore KNN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Per la struttura IBM non si apprezzano grosse differenze tra i due metodi di ottenimento del training set (figura 12).

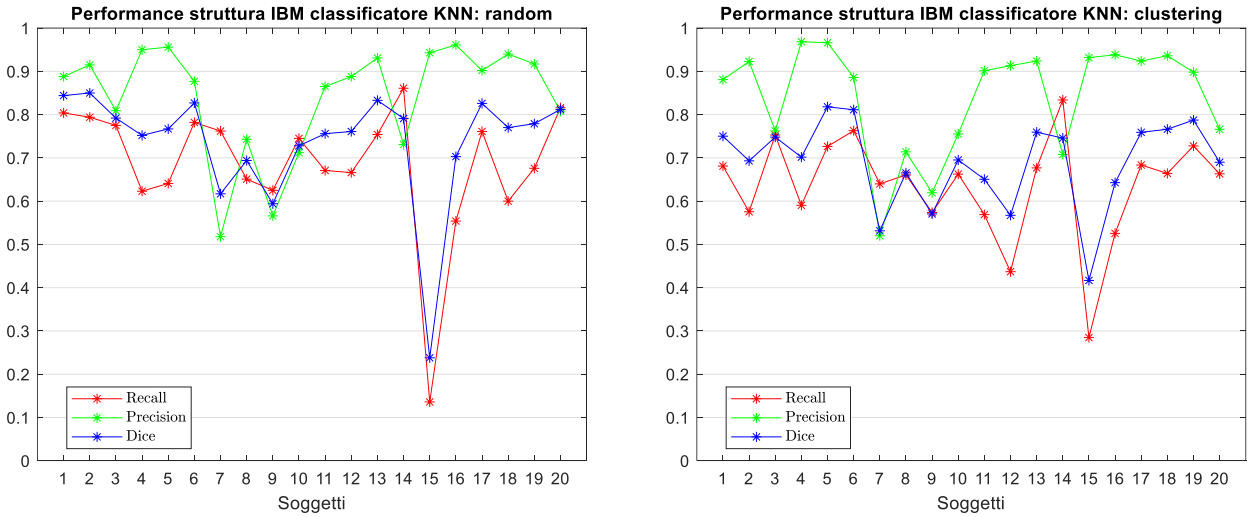


Figura 12 Performance struttura IBM e classificatore KNN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Lo stesso si può dire per la struttura LPBM (figura 13).

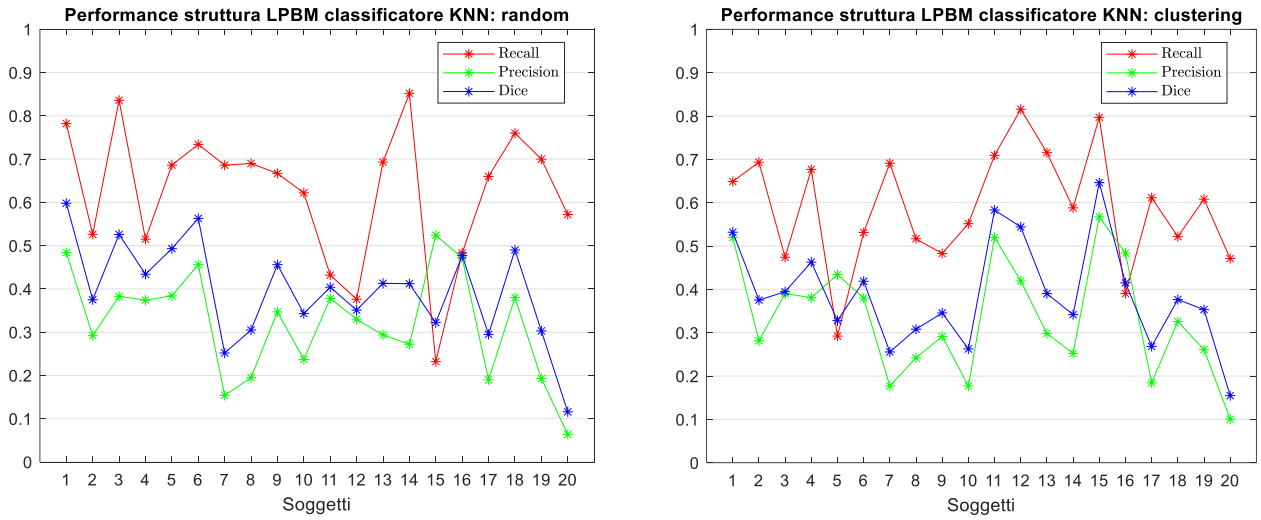


Figura 13 Performance struttura LPBM e classificatore KNN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

- Classificatore NN:

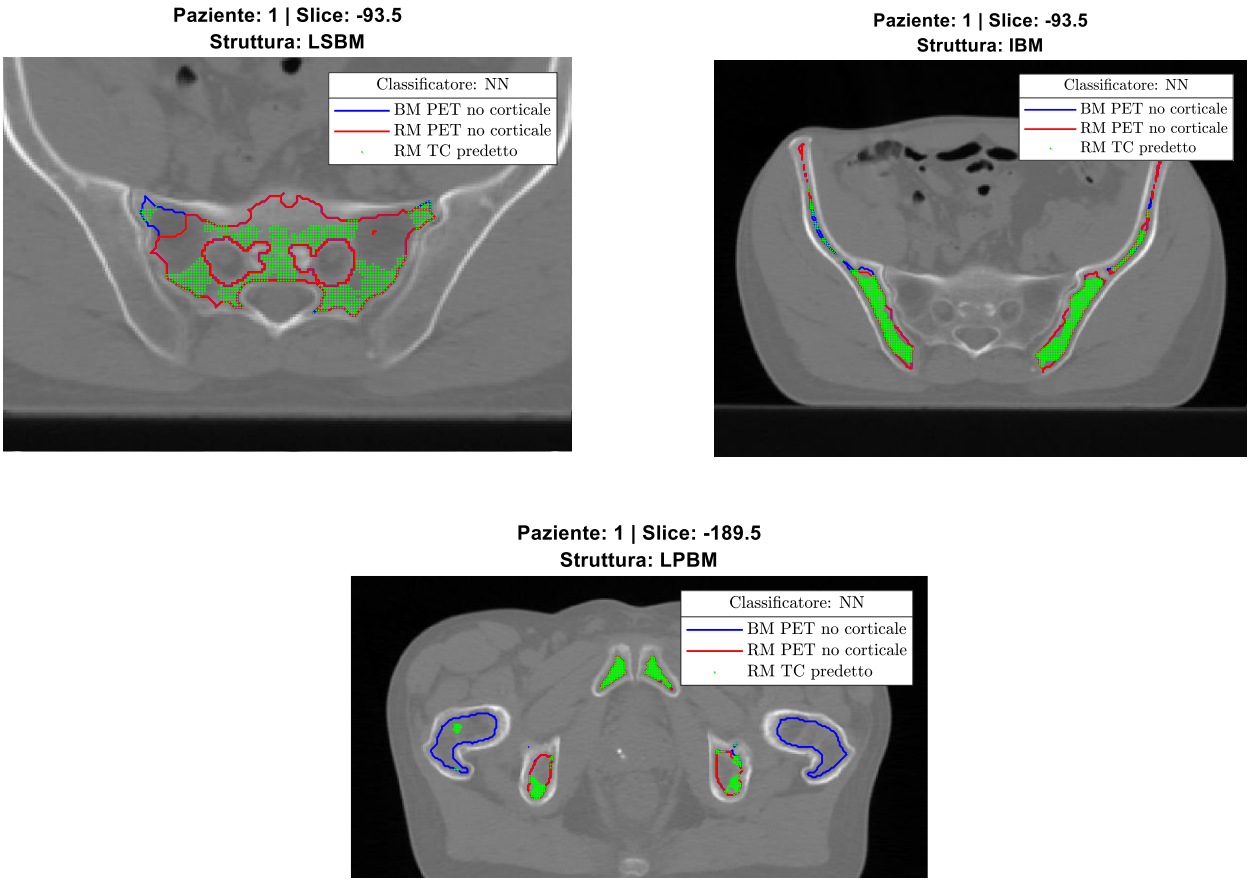


Figura 14 Maschere di segmentazione di LSBM, IBM e LPBM tramite classificatore NN. In verde la classificazione ottenuta, in blu il contorno della maschera del midollo osseo proveniente dalla PET con esclusione dell'osso corticale, i rosso il contorno della maschera del midollo attivo proveniente dalla PET con esclusione dell'osso corticale.

Analogamente al caso del classificatore KNN, anche per la rete neurale le performance appaiono mediamente migliori nel caso di training set con clustering per la struttura LSBM e le performance in termini di recall mostrano una variabilità inter-paziente minore (figura 15).

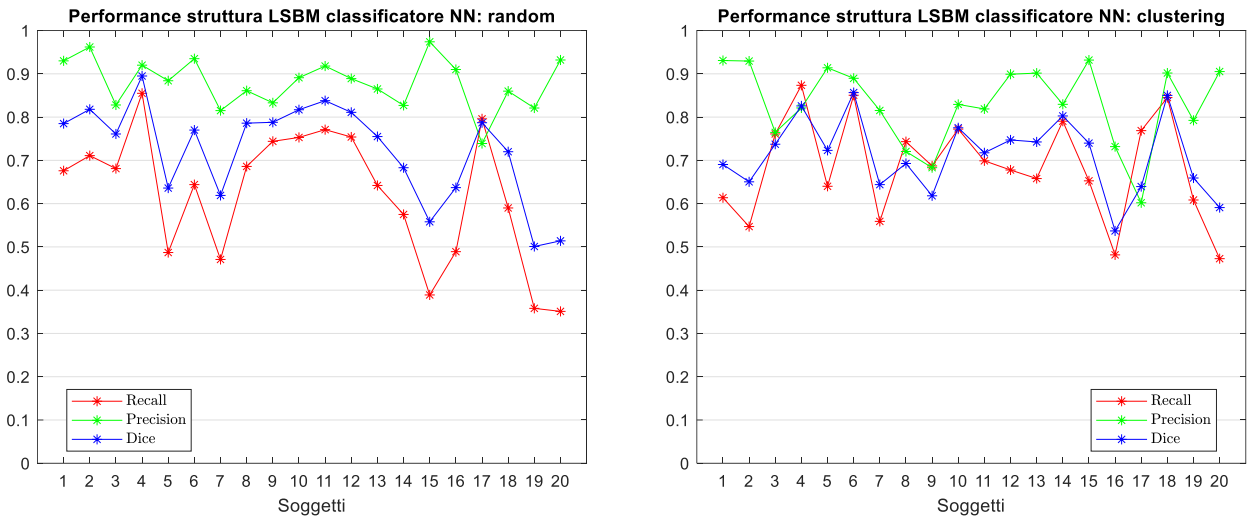


Figura 15 Performance struttura LSBM e classificatore NN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Per la struttura IBM si ha invece che le performance sono molto simili tra i due training set, sebbene la recall presenti una variabilità inter-paziente leggermente minore (figura 16).

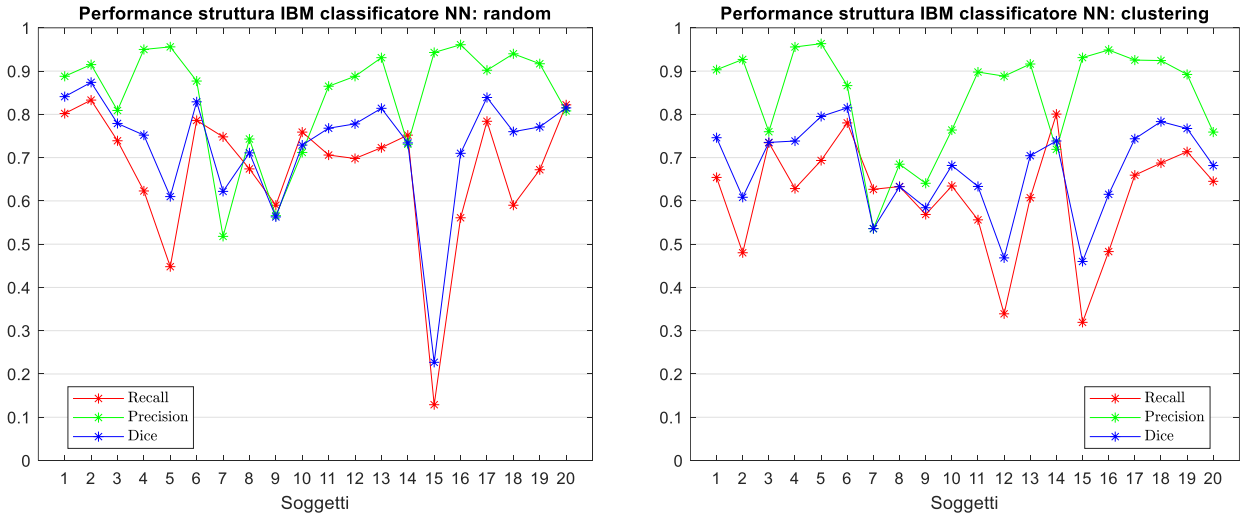


Figura 16 Performance struttura IBM e classificatore NN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Infine, nel caso della struttura LPBM non si evidenziano grosse differenze tra i due metodi di ottenimento del training set (figura 17).

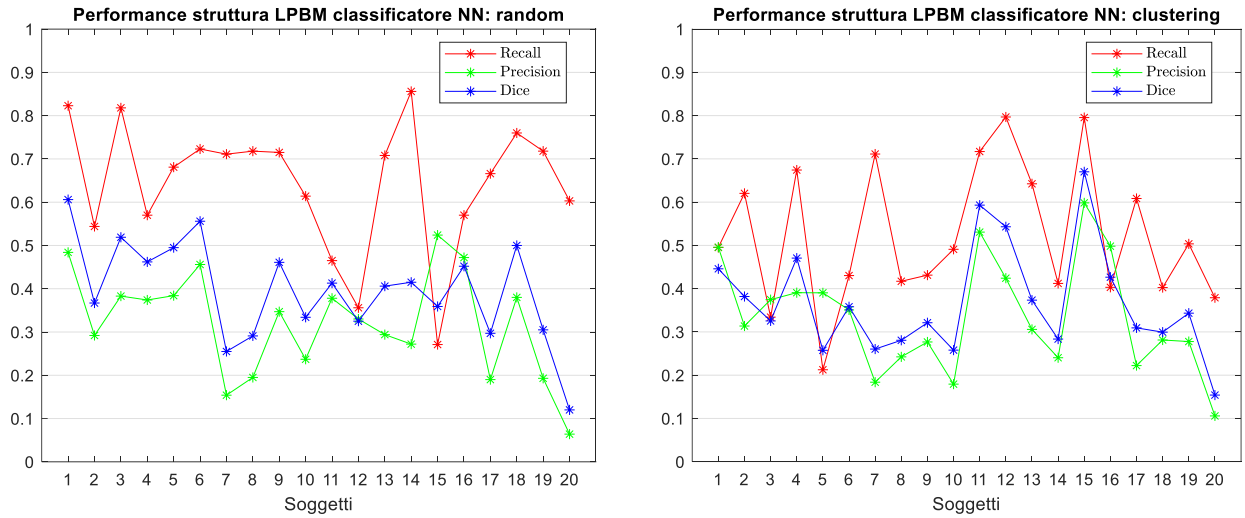


Figura 17 Performance struttura LPBM e classificatore NN, a sinistra con training set random, a destra con clustering. In rosso è riportata la recall, in verde la precision e in blu la dice.

Nella tabella sotto riportata, sono sintetizzate le considerazioni sulle performance e la loro variabilità per i classificatori allenati a partire da training set random e da training set con clustering.

Training set random		Training set SOM	
Performance migliori	**Variabilità minore**	**Performance migliori**	**Variabilità minore**
IBM-DT		LSBM-KNN	LSBM-DT
LPBM-DT		LSBM-NN	IBM-DT
			LPBM-DT
			LSBM-KNN
			LSBM-NN
			IBM-NN

I casi struttura-classificatore in cui le performance erano sostanzialmente molto simili tra le due prove non sono stati considerati in tabella. Sembra chiaro che, a fronte di performance leggermente migliori da parte del DT sulle strutture IBM e LPBM con training set random, nella maggioranza dei casi rimanenti i classificatori portano a performance migliori e/o meno variabili nel caso di allenamento mediante training set ottenuto tramite clustering.

Si sono, quindi, analizzate da un lato le performance sui primi 40 pazienti, ossia quelli le cui slice erano in parte presenti nel construction set, e sui restanti 10 pazienti, cioè quelli del validation set.

Di seguito i risultati divisi per ciascuna tipologia di classificatore.

- Classificatore DT:

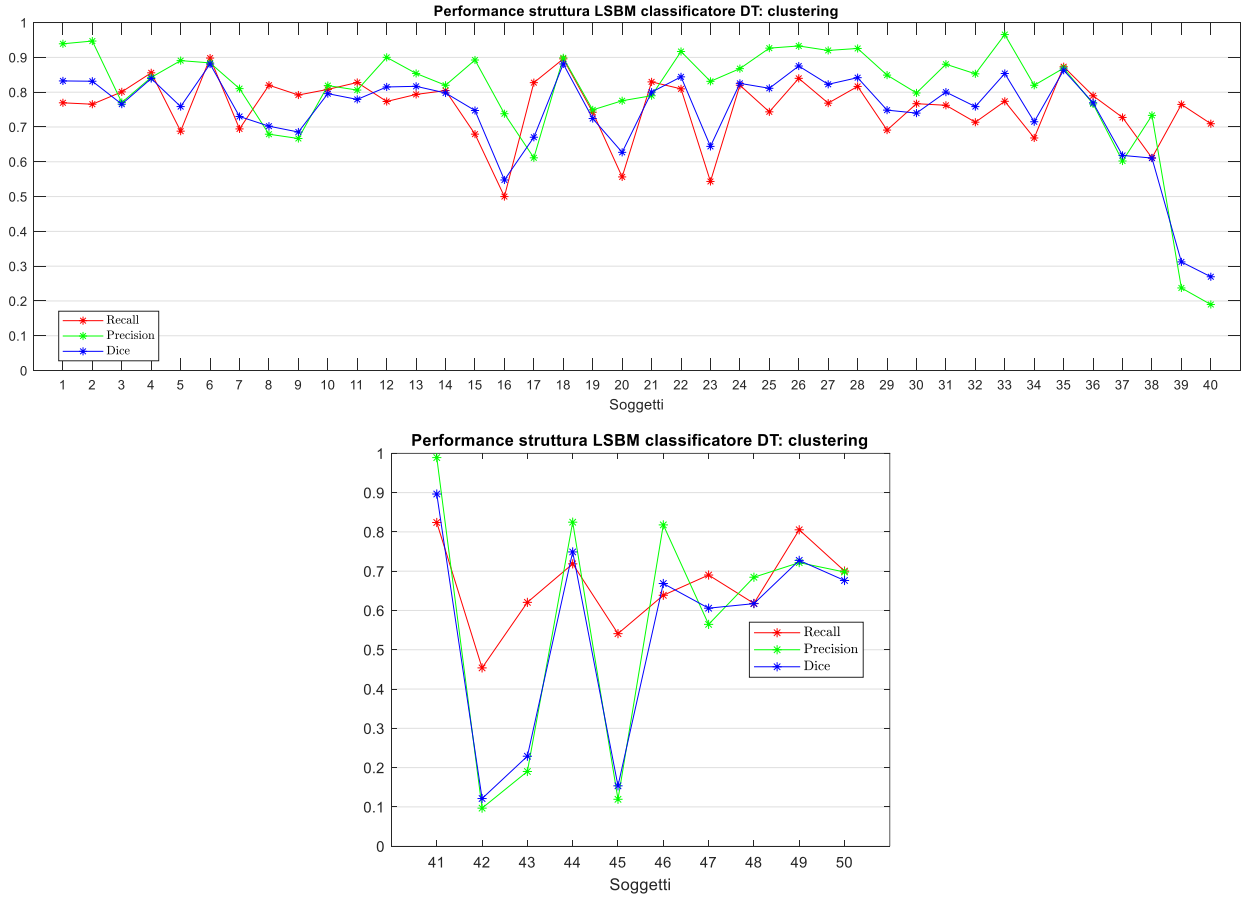
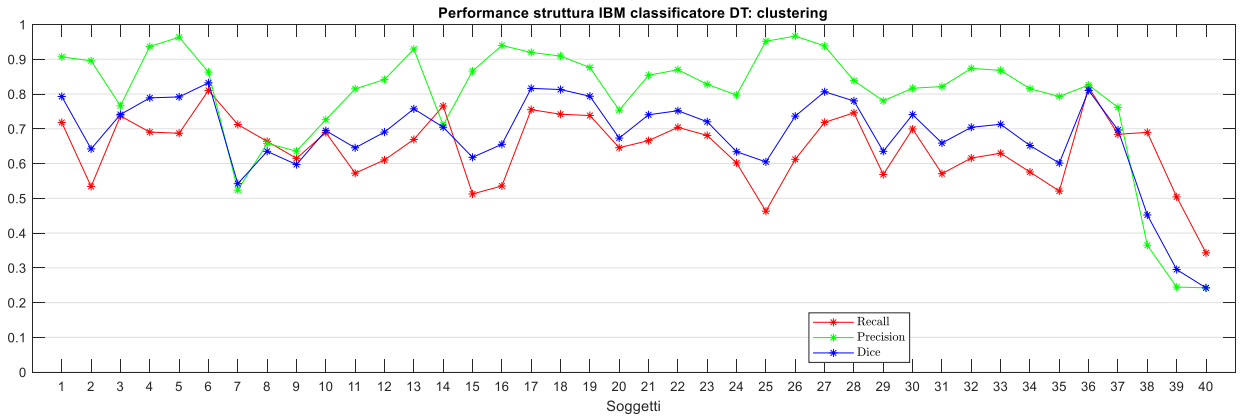


Figura 18 Performance classificatore DT e struttura LSBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Nel caso della struttura LSBM, per il classificatore DT, le performance in termini di recall, precision e dice non mostrano una variabilità particolare tra i diversi soggetti del construction set. Ciò non è però vero per i pazienti 39 e 40 con performance in termini di precision decisamente inferiori.

Nel validation set le performance sono in linea col construction set, tranne che per i pazienti 42, 43 e 45 che mostrano una chiara diminuzione, particolarmente evidente nella precision (figura 18).



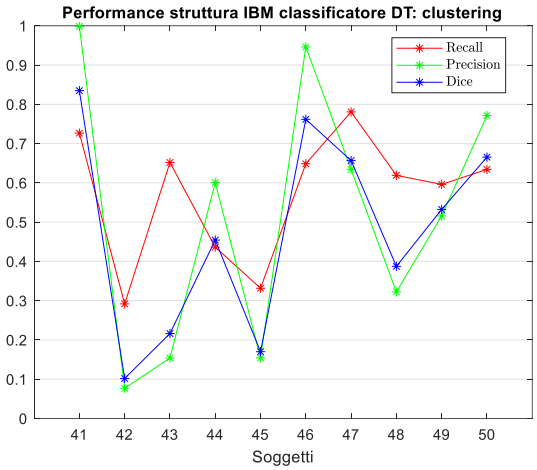


Figura 19 Performance classificatore DT e struttura IBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Per la struttura IBM si ha anche qui che le performance in termini di recall, precision e dice non mostrano una variabilità inter-paziente particolarmente elevata nel construction set. Tuttavia, ciò non vale per i pazienti 38, 39 e 40 con performance in termini di precision chiaramente inferiori.

Per quanto riguarda il validation set si riscontra che le performance, soprattutto di precision, hanno una diminuzione molto elevata soprattutto per i soggetti 42, 43, 45 e 48 (figura 19).

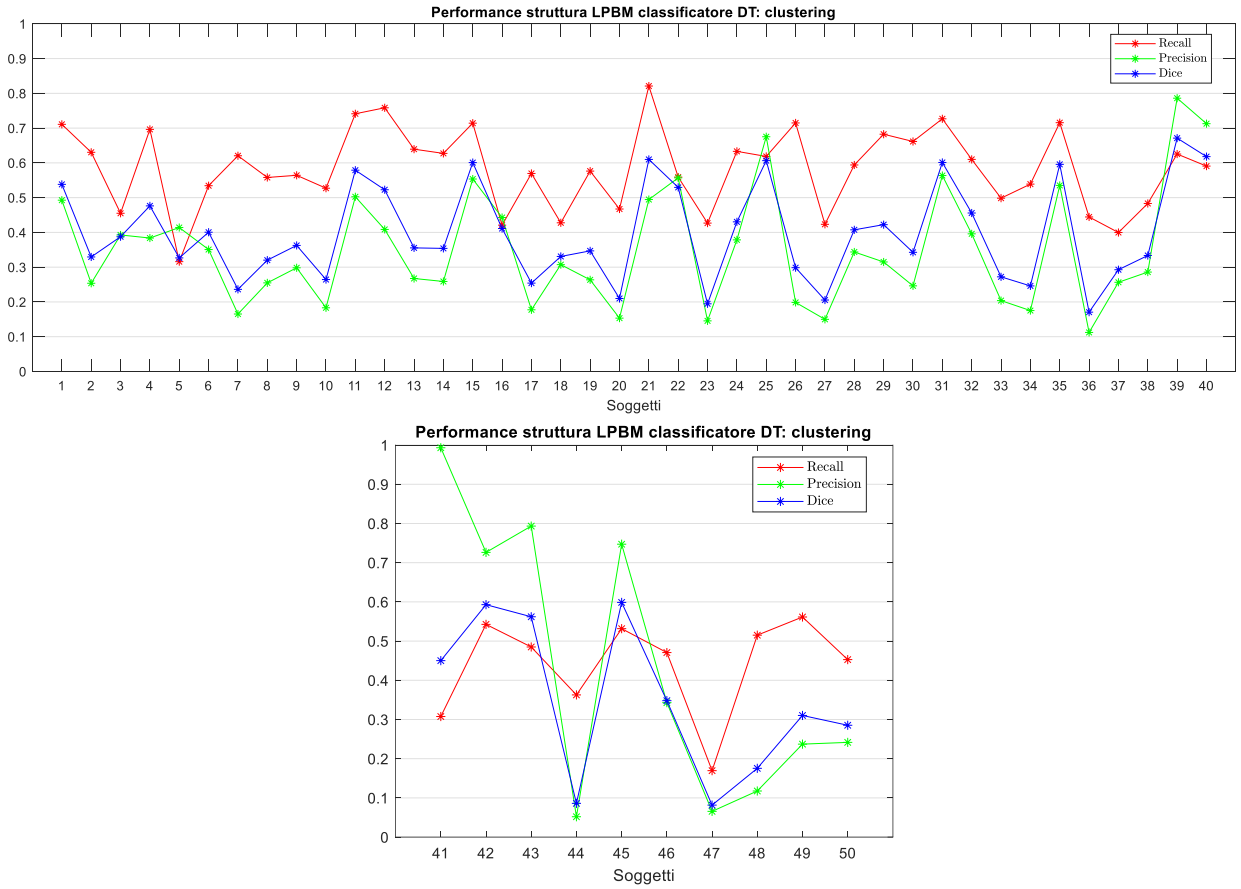


Figura 20 Performance classificatore DT e struttura LPBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

In ultimo, per la struttura LPBM è possibile notare ancora una volta come le performance in termini di recall, precision e dice non mostrino una grossa variabilità tra i soggetti del construction set sebbene chiaramente inferiori alle precedenti strutture. In tal caso, per i pazienti 25, 39 e 40 si ha un miglioramento delle performance in termini di precision.

Nel validation set le performance hanno una diminuzione molto elevata soprattutto per i soggetti 44 e 47, mentre mostrano valori alti per i soggetti 42, 43 e 45, in particolar modo sulla precision (figura 20).

- Classificatore KNN:

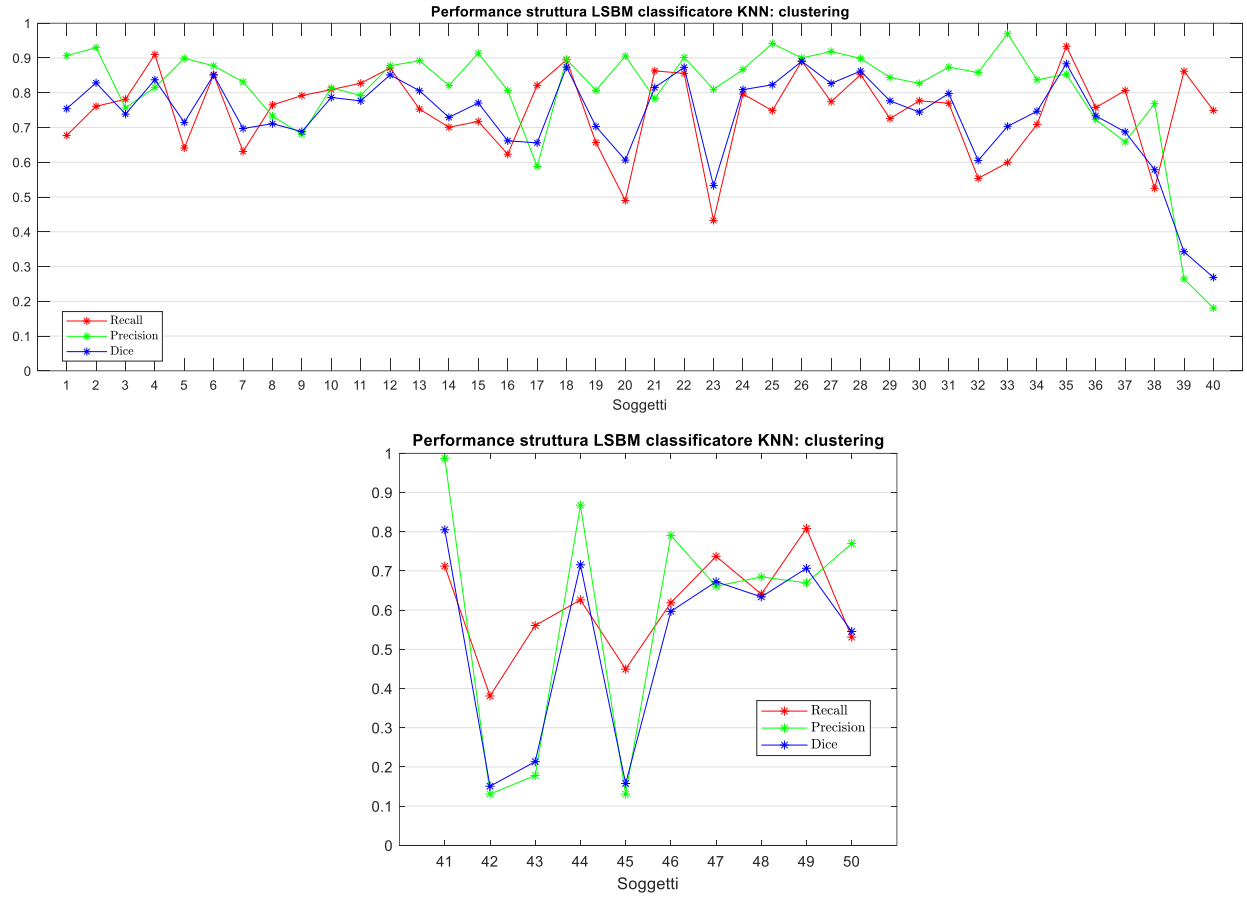


Figura 21 Performance classificatore KNN e struttura LSBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Nel caso della struttura LSBM si può vedere che le performance in termini di recall, precision e dice non sono molto variabili tra i diversi soggetti del construction set. Al pari del DT per questa struttura, nei pazienti 39 e 40 le performance in termini di precision sono decisamente inferiori.

Per il validation set le performance sono più basse del construction set e per i pazienti 42, 43 e 45 la diminuzione è sensibilmente più netta (figura 21).

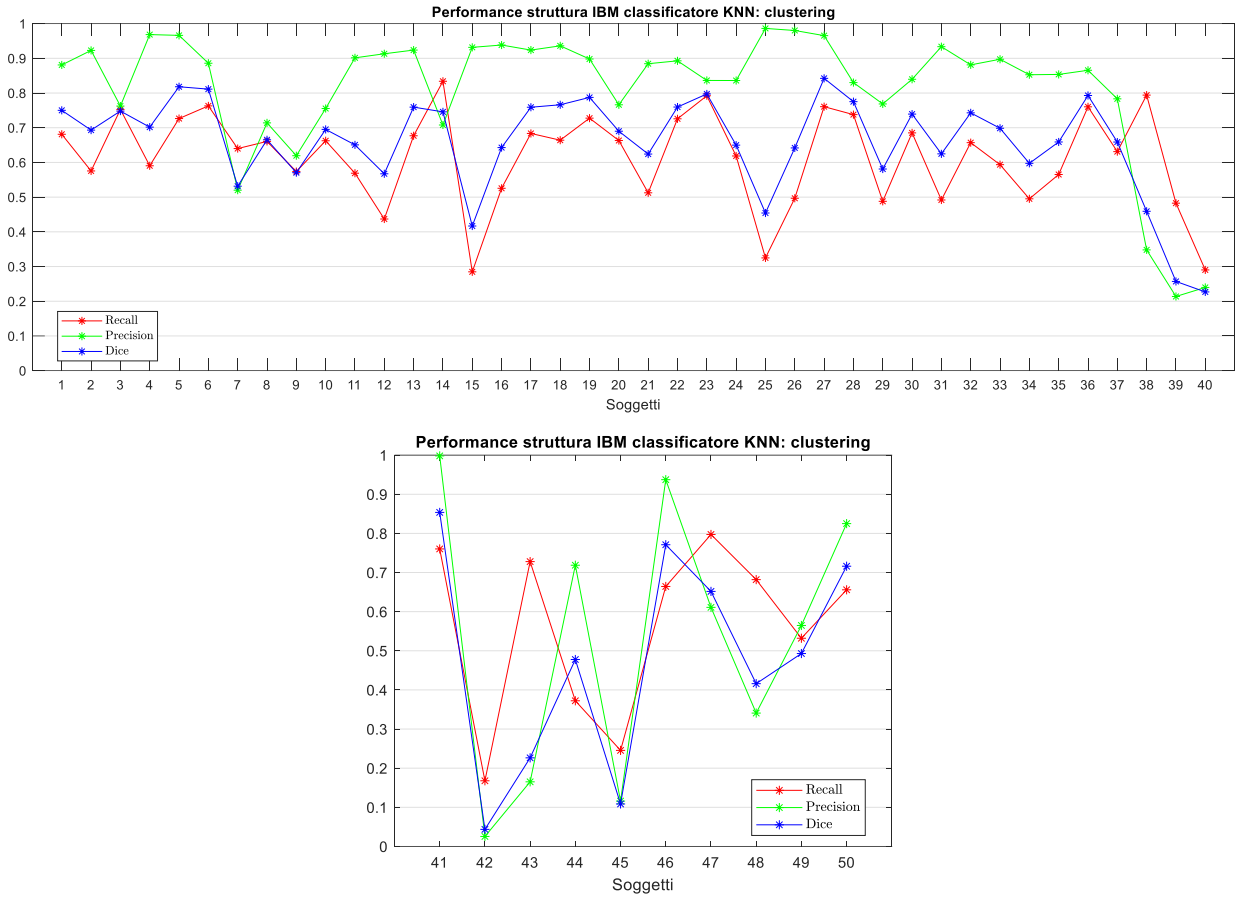
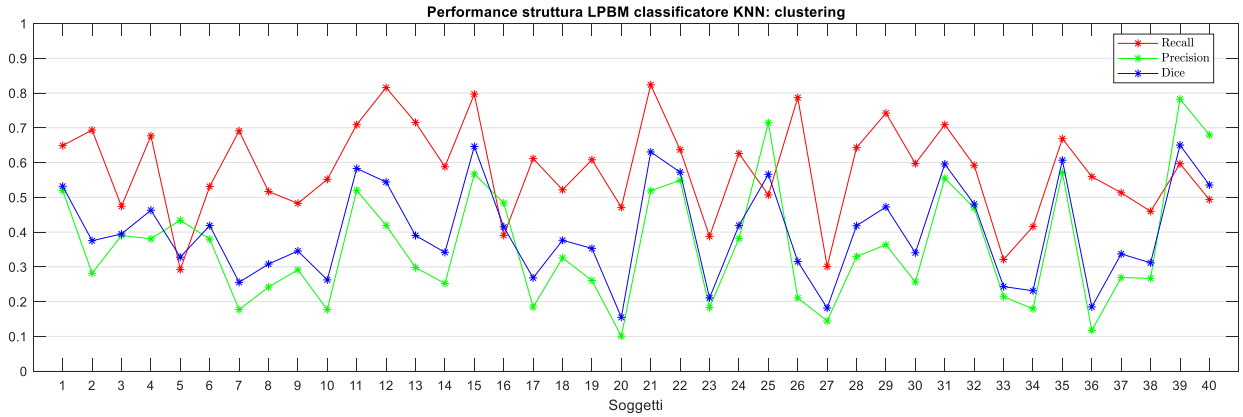


Figura 22 Performance classificatore KNN e struttura IBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

In merito all'IBM, le performance di recall e dice sono inferiori rispetto al caso del DT e mostrano una variabilità inter-paziente più elevata nel construction set. Nuovamente, i pazienti 38, 39 e 40 hanno precision inferiore.

Nel validation set si apprezza una diminuzione delle performance più marcata soprattutto per i soggetti 42, 43, 45 e 48 (figura 22).



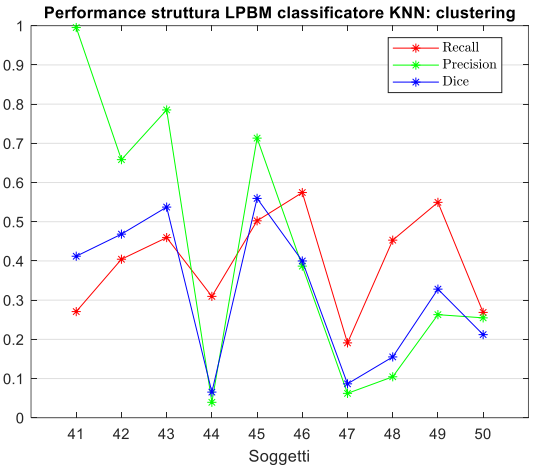


Figura 23 Performance classificatore KNN e struttura LPBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Relativamente alla struttura LPBM, recall, precision e dice appaiono molto simili al caso con classificatore DT per questa struttura sebbene presentino una variabilità leggermente maggiore tra i soggetti del construction set. Per i pazienti 25, 39 e 40 le performance di precision sono migliori.

Infine, si osservano performance di precision più basse soprattutto per i soggetti 44 e 47 del validation set, mentre valori elevati per i soggetti 42, 43 e 45 (figura 23).

- Classificatore NN:

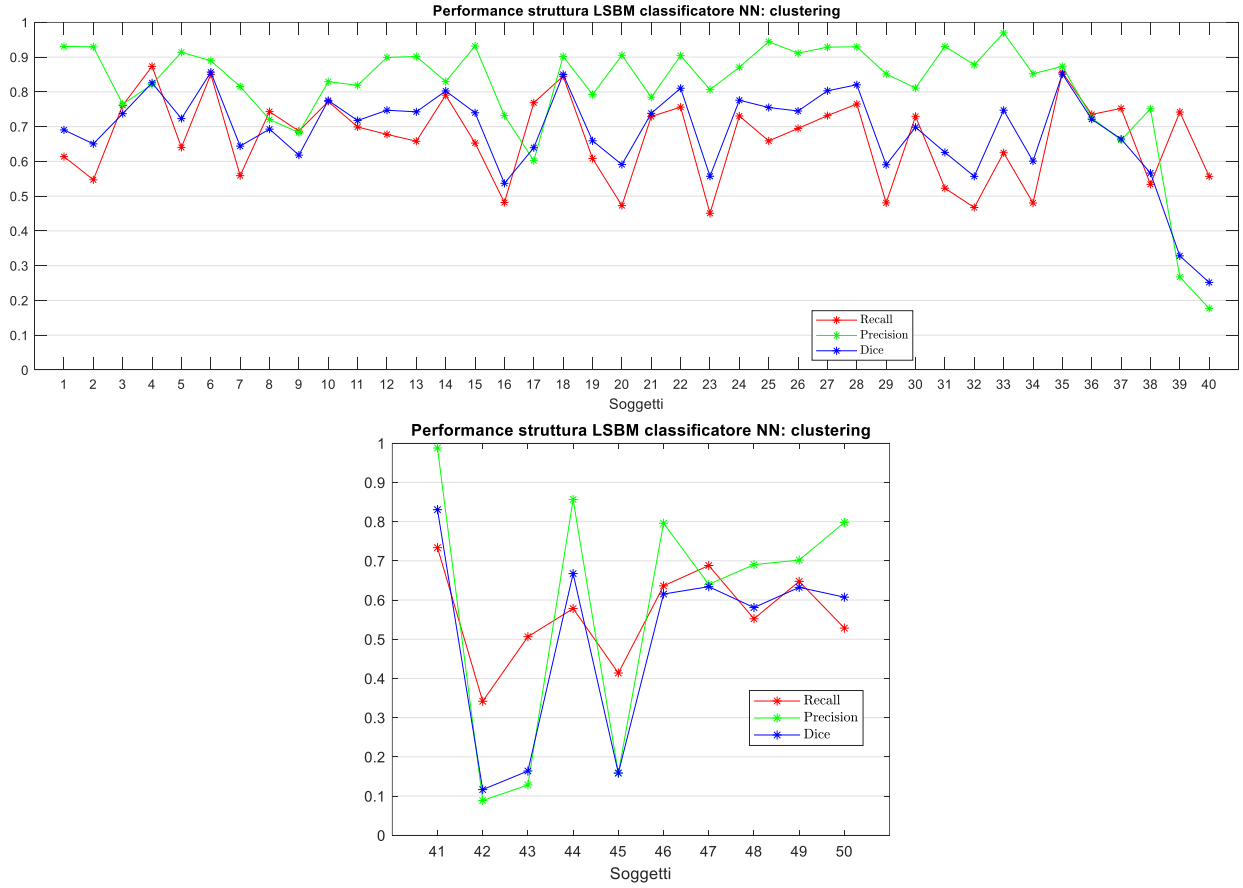


Figura 24 Performance classificatore NN e struttura LSBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Il classificatore NN non mostra, sulla struttura LSBM, performance di recall, precision e dice con variabilità inter-paziente particolarmente evidente sul construction set. Ciò non è però vero per i pazienti 39 e 40, i quali presentano una precision decisamente inferiore.

Nel validation set le performance sono leggermente più basse del construction set e per i pazienti 42, 43 e 45 la diminuzione sulla precision è chiaramente più netta (figura 24).

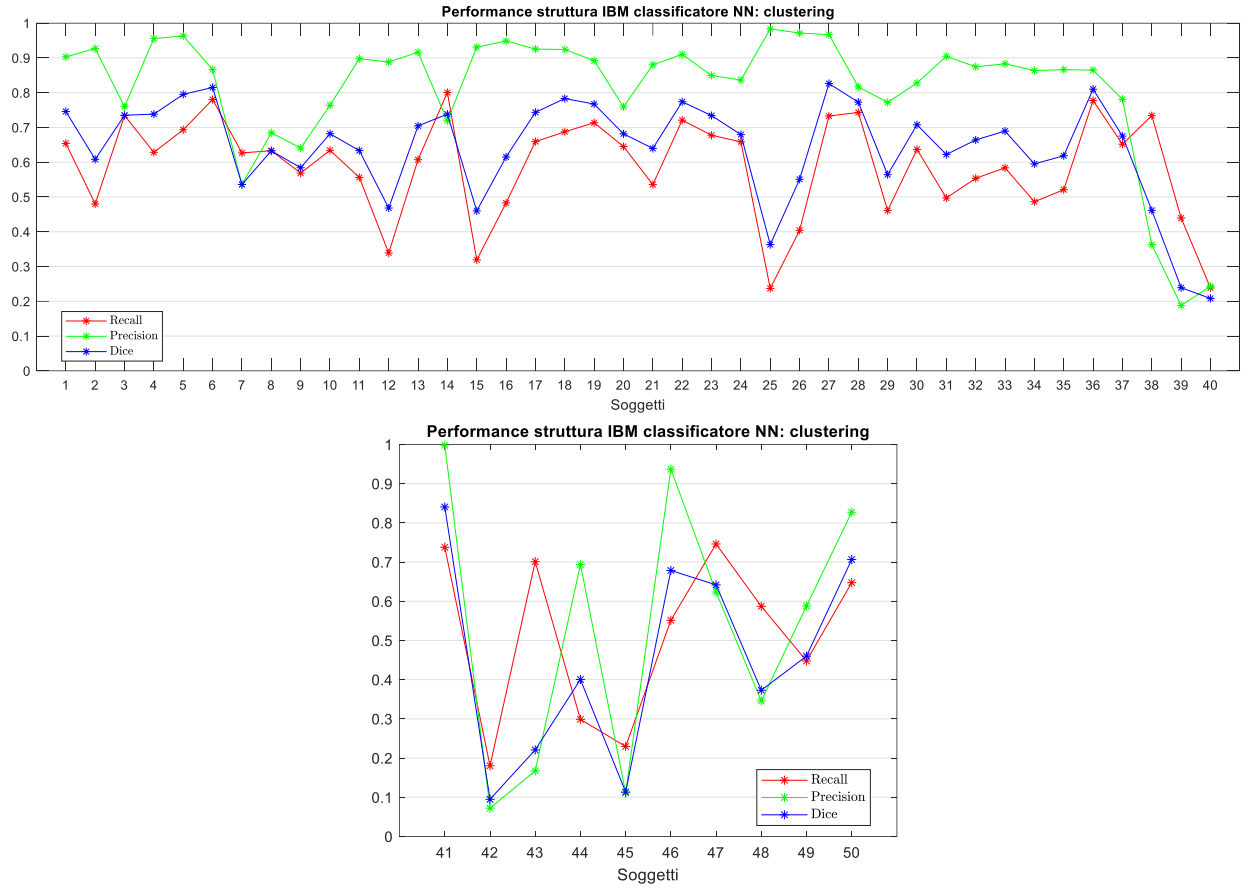
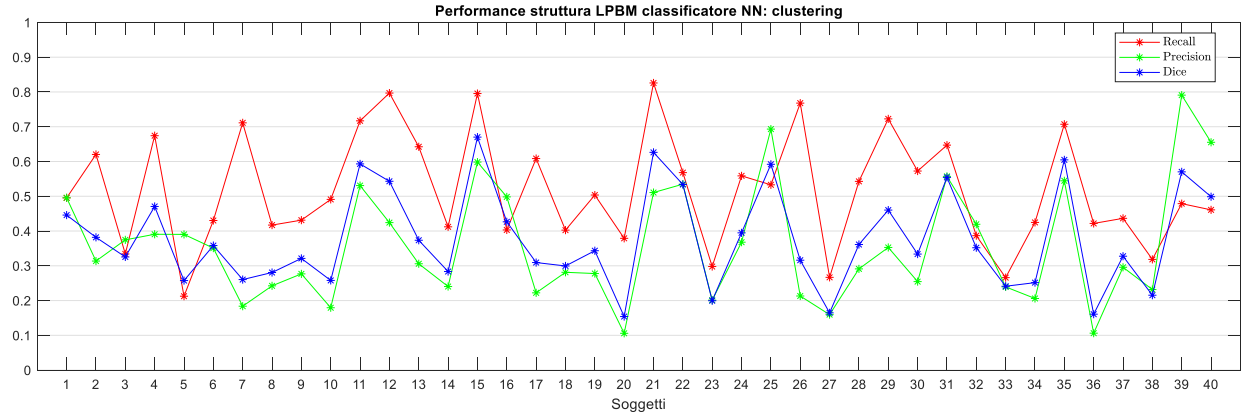


Figura 25 Performance classificatore NN e struttura IBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

Si nota come per la struttura IBM le performance in termini di recall e dice siano inferiori rispetto al caso del DT e molto simili a quelle del KNN con però una variabilità maggiore tra i diversi soggetti del construction set. Anche in tal caso i pazienti 38, 39 e 40 hanno performance di precision inferiori.

Le performance di precision per i soggetti 42, 43, 45 e 48 del validation set sono decisamente più basse (figura 25).



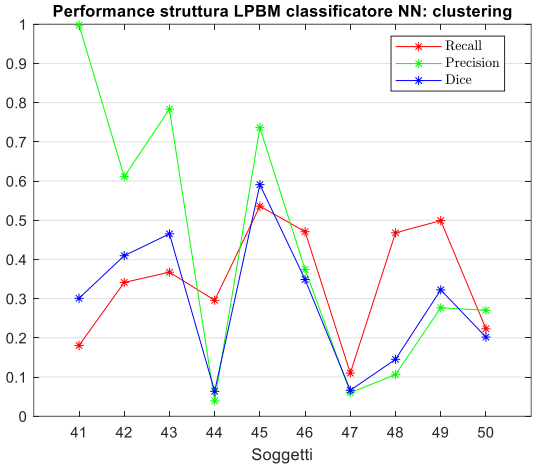


Figura 26 Performance classificatore NN e struttura LPBM. In alto le performance relative ai primi 40 pazienti, le cui slice erano state parzialmente inserite nel construction set, in basso i pazienti 41-50 che rappresentano il validation set.

In ultimo, per quanto concerne la struttura LPBM si può dire che le performance dimostrate dalla NN sono molto simili a quelle del DT e del KNN sebbene con una variabilità inter-paziente più elevata sul construction set. Da notare che, ancora una volta, i pazienti 25, 39 e 40 hanno precision più elevata.

I soggetti 44 e 47 del validation set presentano performance di precision inferiori, mentre i soggetti 42, 43 e 45 sono caratterizzati da valori più elevati (figura 26).

3.2 Analisi dei pazienti anomali

I risultati precedentemente riportati mostrano che, in funzione della struttura, alcuni pazienti presentano delle performance diverse dal comportamento medio sugli altri. Ciò succede a prescindere dal tipo di classificatore utilizzato ed è riassunto nella tabella sottostante.

LSBM		IBM		LPBM	
↓	↑	↓	↑	↓	↑
39		38	41	44	25
40		39	46	47	39
42		40	50		40
43		42			42
45		43			43
		45			45
		48			

↓	Performance peggiori
↑	Performance migliori

Nel lavoro svolto in precedenza con training set random su classificatori di machine learning erano state analizzate tutte e tre le strutture di midollo attivo LSBM, LPBM e IBM ma il dataset conteneva solamente i primi 25 pazienti, quindi non si hanno informazioni sui pazienti problematici prima evidenziati. Tuttavia, era stato riscontrato che la struttura LPBM mostrava i risultati peggiori ed era stata indagato il rapporto tra pixel appartenenti a RM e pixel appartenenti a BM per le 3 diverse strutture, evidenziando come proprio la struttura LPBM avesse tale rapporto più basso.

È nato, quindi, il dubbio che per i pazienti con performance più basse rilevati nei casi sopracitati ci potesse essere un rapporto tra pixel di RM e pixel di BM molto basso nella struttura di interesse.

Per rispondere a questo interrogativo, si è proceduto a calcolare il rapporto $\frac{RM}{BM}$ per ciascuna struttura a partire dalle immagini PET di riferimento (figura 27).

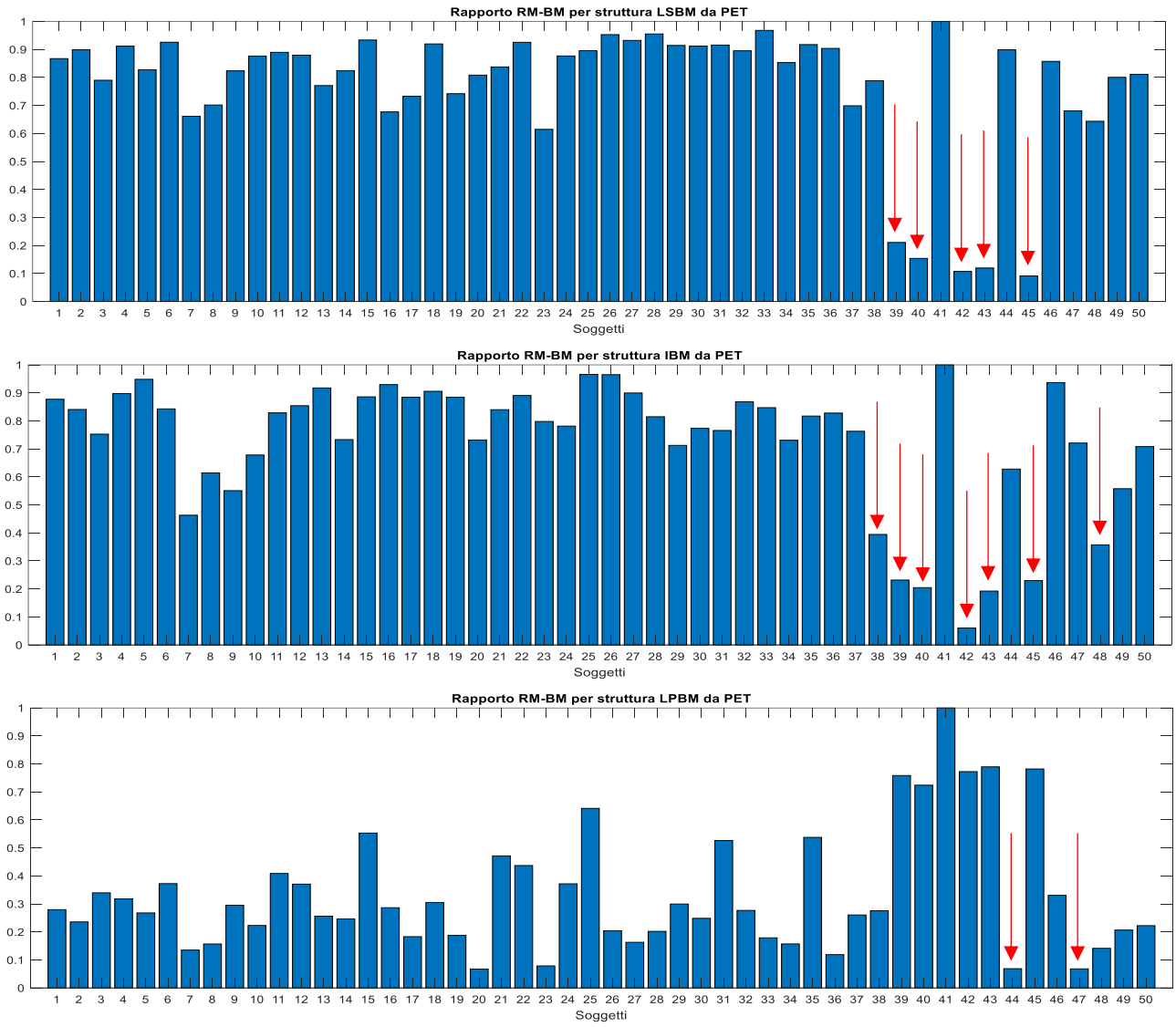


Figura 27 Grafici a barre rappresentanti, per ciascun paziente, il rapporto tra il numero di pixel appartenenti a RM e quello dei pixel appartenenti a YM. In alto la struttura LSBM, al centro la struttura IBM e in basso la struttura LPBM. Le frecce rosse indicano quei pazienti con performance molto più basse rispetto agli altri nella propria struttura.

Si può notare che effettivamente i soggetti per i quali le performance di precision sono in genere più basse sono quelli a rapporto minore, mentre i pazienti che presentano precision più elevata sono quelli a rapporto più alto per la loro struttura.

Dal momento che la precision indica la percentuale di corretto RM individuato sul totale di tessuto riconosciuto come RM, se un paziente avesse un elevato rapporto $\frac{RM}{BM}$ e la classificazione dei pixel avvenisse con una certa tendenza X a segmentare, allora sarebbe più probabile che classificando un pixel come RM questo appartenga effettivamente a RM e quindi a dire che non c'è stata grossa sovra-segmentazione, viceversa se il rapporto è basso questa probabilità sarebbe inferiore e la precision calerebbe concludendo che c'è stata grossa sovra-segmentazione.

Ciò che è importante notare è quindi che nella struttura LPBM i pazienti ad elevato rapporto $\frac{RM}{BM}$ non hanno performance migliori rispetto a quelli a basso rapporto, la loro recall è del tutto simile, mentre la precision, per quanto descritto, non è un parametro affidabile.

Andando a tirare le somme di quanto mostrato in questi due paragrafi, possiamo dire che dei 3 classificatori testati, il DT sembra essere quello che fornisce in generale performance leggermente migliori su tutte e tre le strutture.
Le diminuzioni delle performance in termini di recall per le strutture LSBM e IBM nei pazienti a rapporto $\frac{RM}{BM}$ basso del validation set sono giustificate dal fatto che il training di base avviene su pazienti diversi e, inoltre, questi ultimi hanno tutti rapporto $\frac{RM}{BM}$ alto.
Le performance sulla struttura IBM sono inferiori a quelle della struttura LSBM sebbene il construction set delle due non sembri così diverso in termini di rapporto $\frac{RM}{BM}$ (figura 28). È possibile che la differenza derivi dal fatto che il clustering eseguito con rete SOM di dimensione 12 e vicinato 1 non sia particolarmente adatto a questa struttura.

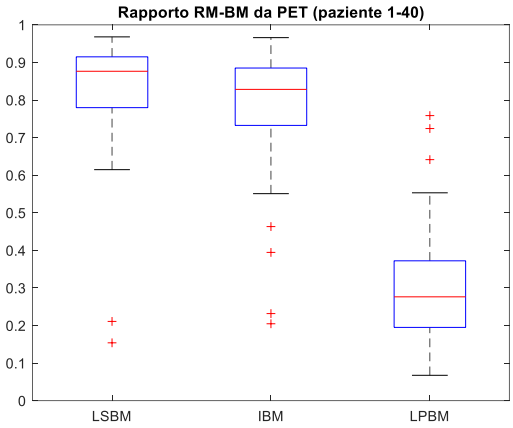


Figura 28 Boxplot del rapporto tra il numero di pixel appartenenti a RM e quello dei pixel appartenenti a YM per le tre strutture sul construction set.

Le performance sulla struttura LPBM sono sempre le più basse. Sebbene il training venga fatto su strutture con rapporto $\frac{RM}{BM}$ basso, le performance sono scarse sia che un paziente abbia rapporto elevato che basso. Infatti, come già sottolineato, anche quando il rapporto è alto, la recall non aumenta particolarmente. Anche qui è possibile che il clustering eseguito con rete SOM di dimensione 12 e vicinato 1 non sia particolarmente adatto alla struttura, però apparirebbe più plausibile che la difficoltà di classificazione derivi da una mancanza di rappresentatività del construction set. A tal proposito un'altra considerazione da fare è relativa al vincolo imposto di prendere come slice per il construction set solamente 5 slice in maniera random per ciascun paziente che presentano almeno 25 ROI di RM e altrettante di YM, in quanto è possibile che, con l'estrazione random, la struttura LPBM sia sottorappresentata rispetto alle altre strutture, a causa magari di un numero di slice valide maggiore. Sulla stessa filosofia di pensiero, il fatto di prendere 10000 ROI per tutte e tre le strutture potrebbe non essere equo, se ad esempio la struttura LPBM presentasse un numero molto maggiore di ROI.

3.3 Slice e ROI del construction set

Al fine di fare luce meglio su quest'ultima ipotesi, sono state calcolate per i primi 40 pazienti, ossia quelli facenti parte del construction set, il numero di slice totali, il numero di slice valide (quelle che presentano almeno 25 ROI di RM e altrettante di YM) e il numero di slice da cui sono state estratte ROI in maniera random per la realizzazione del construction set. Tutto questo per ciascuna delle tre strutture.
Inoltre, sempre per le tre strutture, è stato calcolato il rapporto tra il numero di pixel presenti nel training set e il numero di pixel totali della struttura (figura 29).

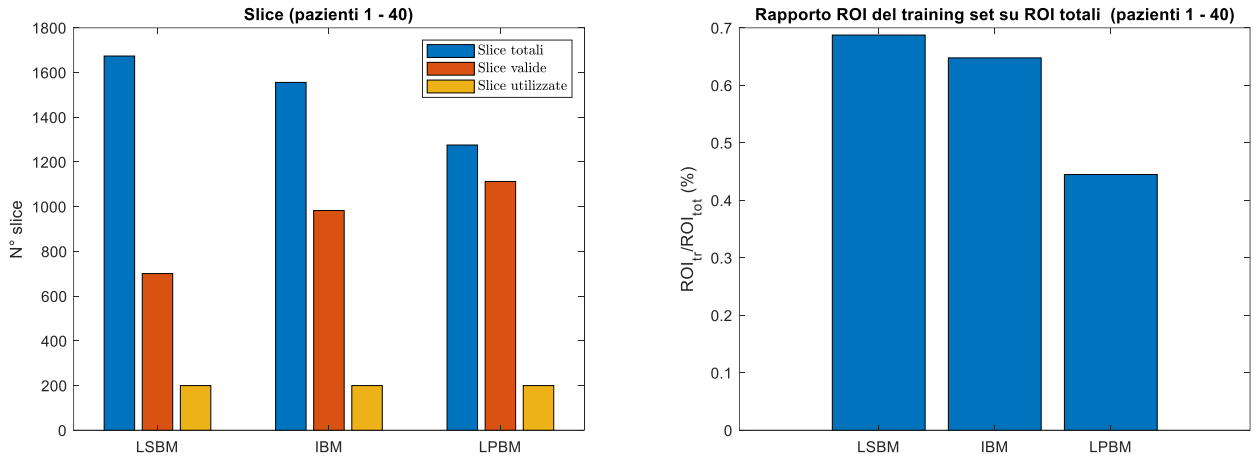


Figura 29 A sinistra grafico a barre riportante in blu il numero di slice totali, in arancione il numero di slice con almeno 25 ROI di RM e altrettante di YM e in giallo il numero di slice incluse nel construction set, per ciascuna delle tre strutture. A destra il rapporto percentuale tra il numero di ROI incluse nel training set e il numero totale delle ROI presenti in ognuna delle strutture.

Si può notare come effettivamente la struttura LPBM sia quella per la quale la scelta di un numero di slice da utilizzare come rappresentative della struttura in sé (perché usate nel training dei classificatori) può ricadere in un range più ampio di slice, nonostante addirittura il numero di slice totali della struttura LPBM sia minore rispetto alle altre due. Questo vuol dire che molte più slice in proporzione al totale costituiscono la totalità del RM per la struttura LPBM, quindi, in tal caso, è possibile che un training set estratto in maniera random non rappresenti a pieno l'LPBM e che ROI provenienti da altre slice possano portare un'informazione aggiuntiva (e magari necessaria) per la descrizione della struttura. Oltretutto, la struttura LPBM è quella che presenta il rapporto tra il numero di pixel presenti nel training set e il numero di pixel totali della struttura più basso, il che può costituire, in una certa misura, un ulteriore motivo di scarsa rappresentatività.

4 Struttura LPBM

4.1 Construction set stratificato

Nel construction set precedente per la struttura LPBM le slice erano state scelte in maniera random tra quelle che avevano almeno 25 ROI di RM e altrettante di YM. Come spiegato nel capitolo precedente, molte più slice in proporzione al totale costituiscono la totalità del RM per la struttura LPBM, per cui la selezione random delle slice potrebbe portare a non prendere in considerazione slice di parti della struttura significativamente diverse dal resto della stessa, con la conseguente perdita di informazione utile per la classificazione.

In questa parte del lavoro, si è scelto di mantenere sempre lo stesso numero di slice e di ROI per il training set, ma di estrarre le slice del construction set in maniera stratificata da ciascuno dei 40 pazienti. Quindi, dato un paziente e la sua struttura LPBM, si sono selezionate le slice con almeno 25 ROI di RM e altrettante di YM, sono state ordinate in senso decrescente e sono state selezionate 5 slice equidistanti (figura 30).

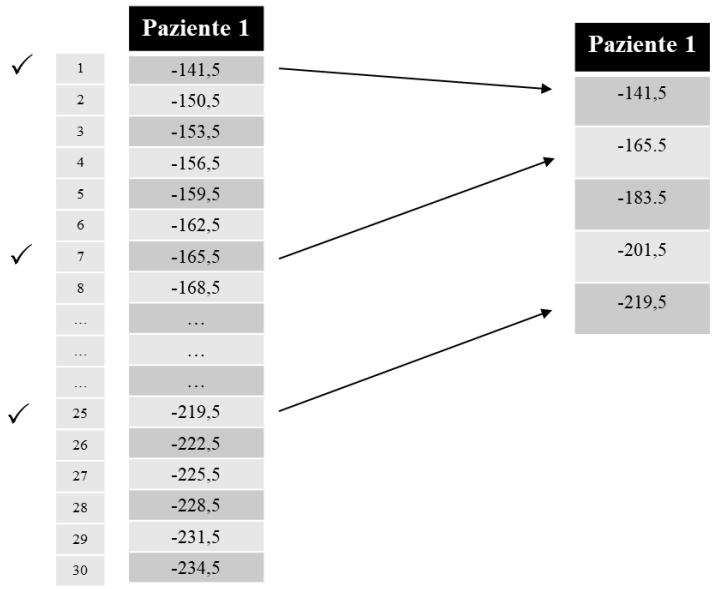


Figura 30 Esempio di selezione delle slice per il construction set stratificato della struttura LPBM per il paziente 1: a sinistra il numero totale di slice e l'indicazione di quali sono le slice con almeno 25 ROI di RM e YM, a destra l'indicazione delle slice selezionate in maniera stratificata.

A questo punto, si è proceduti a dividere il construction set nelle due classi RM e YM e ad effettuare una clusterizzazione mediante reti SOM di tutte le ROI contenute in ciascuna delle due classi. In particolare, per la clusterizzazione sono state effettuate 2 prove andando ad utilizzare:

- Reti SOM quadrate di dimensione 12, vicinato 1 e metrica di distanza euclidea sia per RM che per YM;
- Reti SOM quadrate di dimensione 10, vicinato 1 e metrica di distanza euclidea sia per RM che per YM.

Per ogni rete sono stati considerati tre tagli possibili del dendrogramma dei pesi sulla base della massima distanza euclidea tra i pesi dei neuroni.

La rete migliore per RM e YM è stata identificata come quella con il taglio che assicurava un numero ragionevole di cluster, che presentava bolle di neuroni adiacenti per ciascun suo cluster e che mostrava la varianza tra i pesi minore.

Di seguito sono riportati i dendrogrammi dei pesi per le due architetture di reti e per le due classi e la rappresentazione delle bolle derivanti dai cluster determinati da ciascun taglio analizzato.

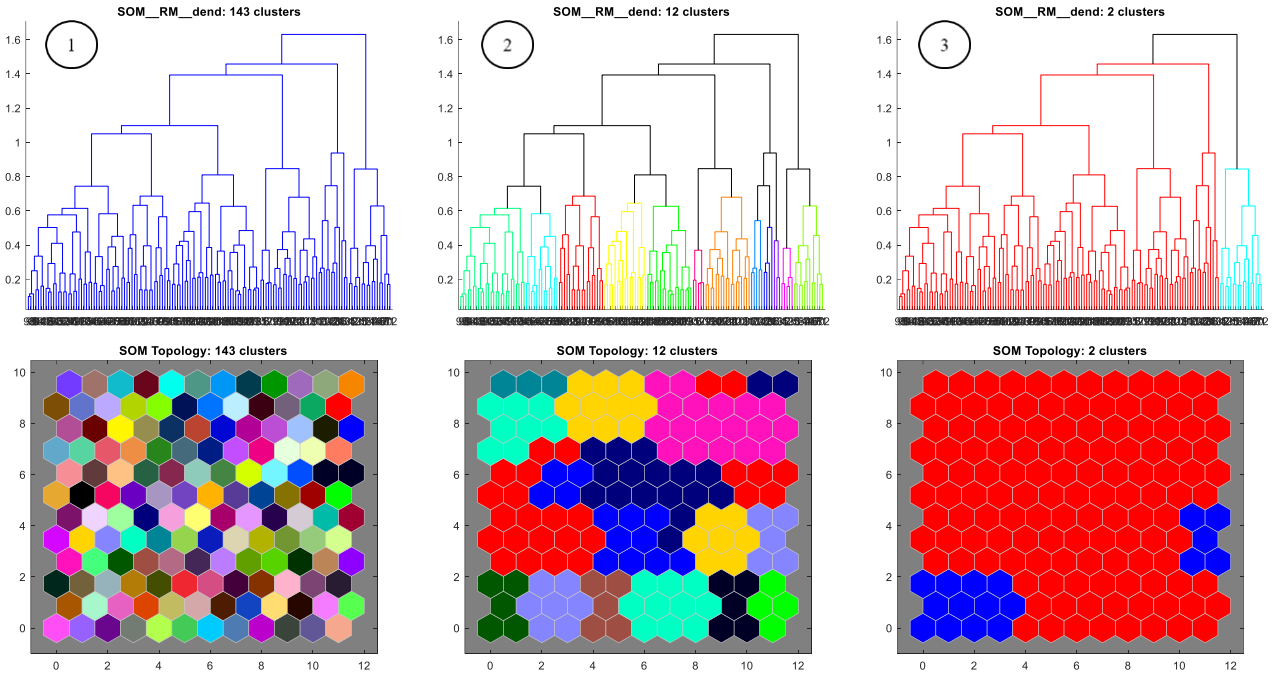


Figura 31 Dendrogrammi e topologia della rete SOM di dimensione 12 classe RM con indicazione delle bolle: viene riportato inoltre il numero di cluster generati da ogni taglio.

I tagli 2 e 3 non rispettano la condizione sulle bolle, per cui non possono essere presi in considerazione. Il taglio 1 porta a un numero di cluster troppo elevato, quindi questa rete non viene scelta per la classe RM (figura 31).

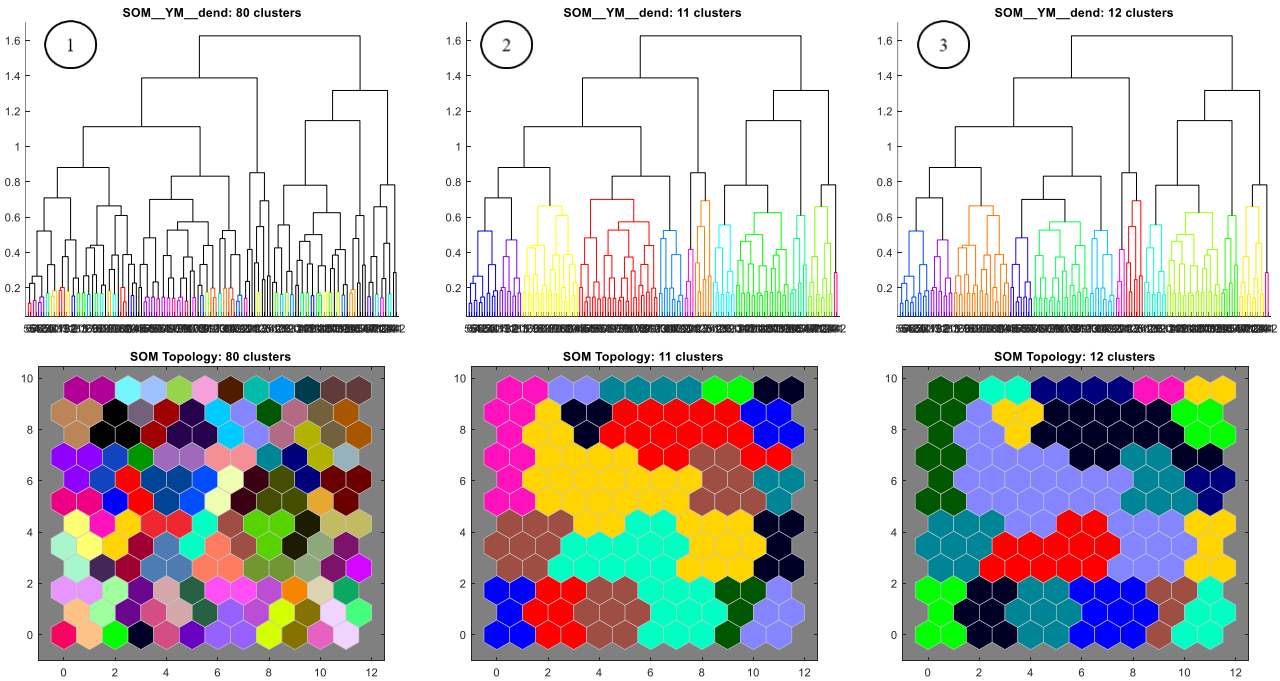


Figura 32 Dendrogrammi e topologia della rete SOM di dimensione 12 classe YM con indicazione delle bolle: viene riportato inoltre il numero di cluster generati da ogni taglio.

Anche per la classe YM i tagli 2 e 3 non rispettano la condizione sulle bolle e il taglio 1 porta a un numero di cluster molto elevato, quindi questa rete non viene selezionata per il clustering (figura 32).

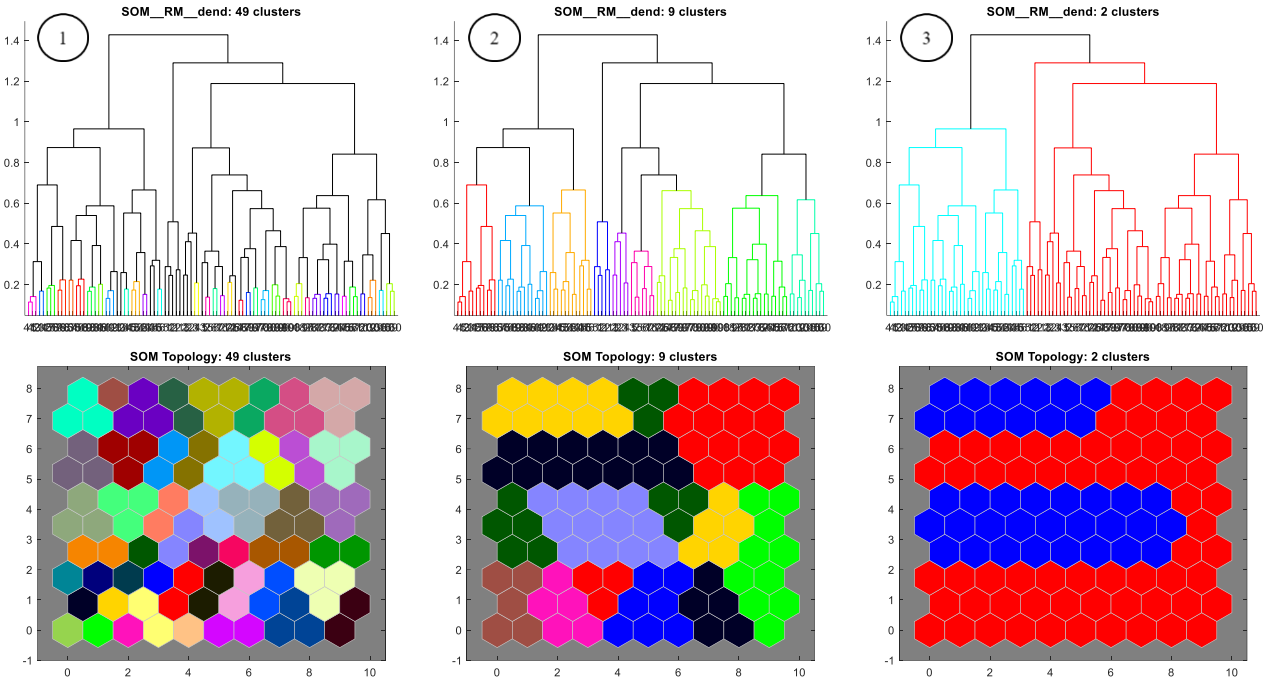


Figura 33 Dendrogrammi e topologia della rete SOM di dimensione 10 classe RM con indicazione delle bolle: viene riportato inoltre il numero di cluster generati da ogni taglio.

Per quanto riguarda la rete SOM di dimensione 10, essa presenta i tagli 2 e 3 che non rispettano la condizione sulle bolle, mentre il taglio 1 rispetta questa condizione e possiede un numero di cluster ragionevole. Pertanto, questa rete con il taglio 1 è stata scelta per il clustering delle ROI di RM della struttura LPBM (figura 33).

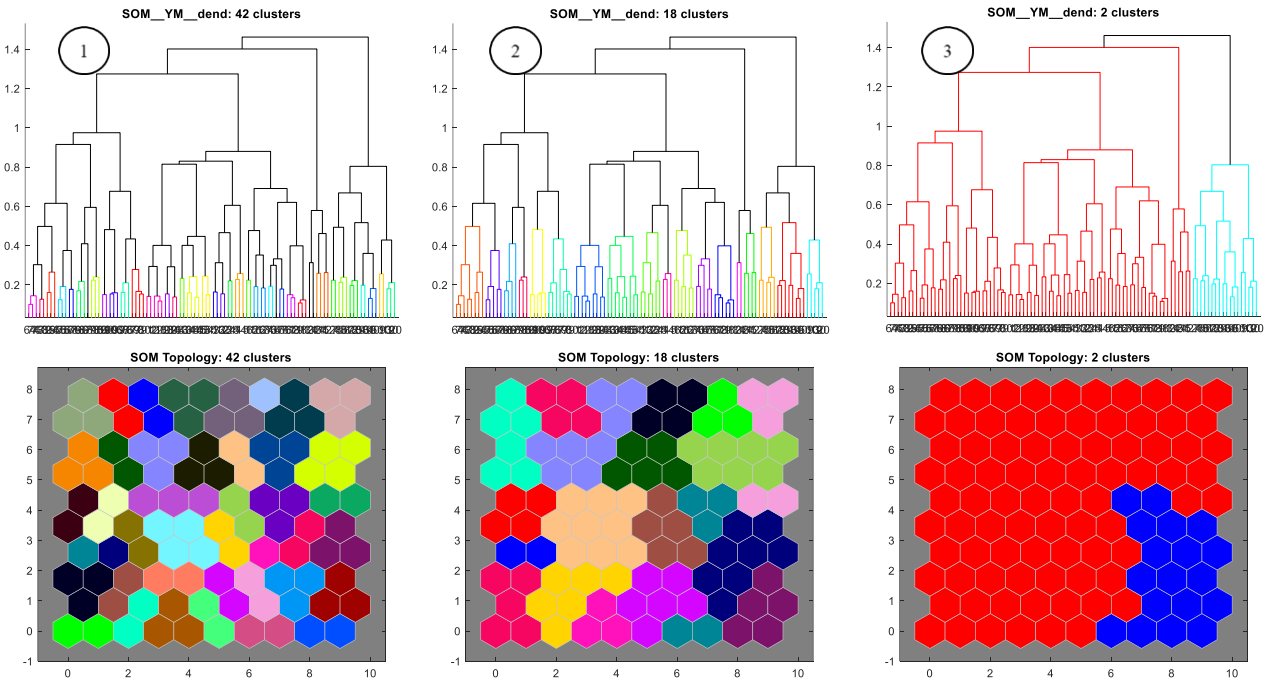


Figura 34 Dendrogrammi e topologia della rete SOM di dimensione 10 classe RM con indicazione delle bolle: viene riportato inoltre il numero di cluster generati da ogni taglio.

Passando in ultimo alla classe YM con rete SOM di dimensione 10, si nota che il taglio 2 non rispetta la condizione sulle bolle, per cui non può essere selezionato per il clustering. Al contrario, i tagli 1 e 3 rispettano tale condizione, tuttavia il taglio 3 forma solamente 2 cluster, pertanto non sembra la scelta migliore in virtù di una rete di dimensione 10. Inoltre, il taglio 1 mostra la variabilità più bassa, quindi è scelto insieme a questa rete per il clustering delle ROI di classe YM (figura 34).

Effettuata la clusterizzazione delle ROI separatamente per le due classi, si è ottenuto un training set bilanciato di circa 10000 ROI utilizzando l'estrazione proporzionale da ciascun cluster in funzione della numerosità degli elementi contenuti nello stesso. Le ROI non incluse nel training set sono state inserite nel test set.

A partire dal training set e dal test set ottenuti per mezzo di queste reti SOM, è stato eseguito lo stesso GA del construction set con estrazione random delle slice per il classificatore DT e sono state ottenute le maschere di segmentazione della struttura LPBM delle quali si sono poi valutate le performance.

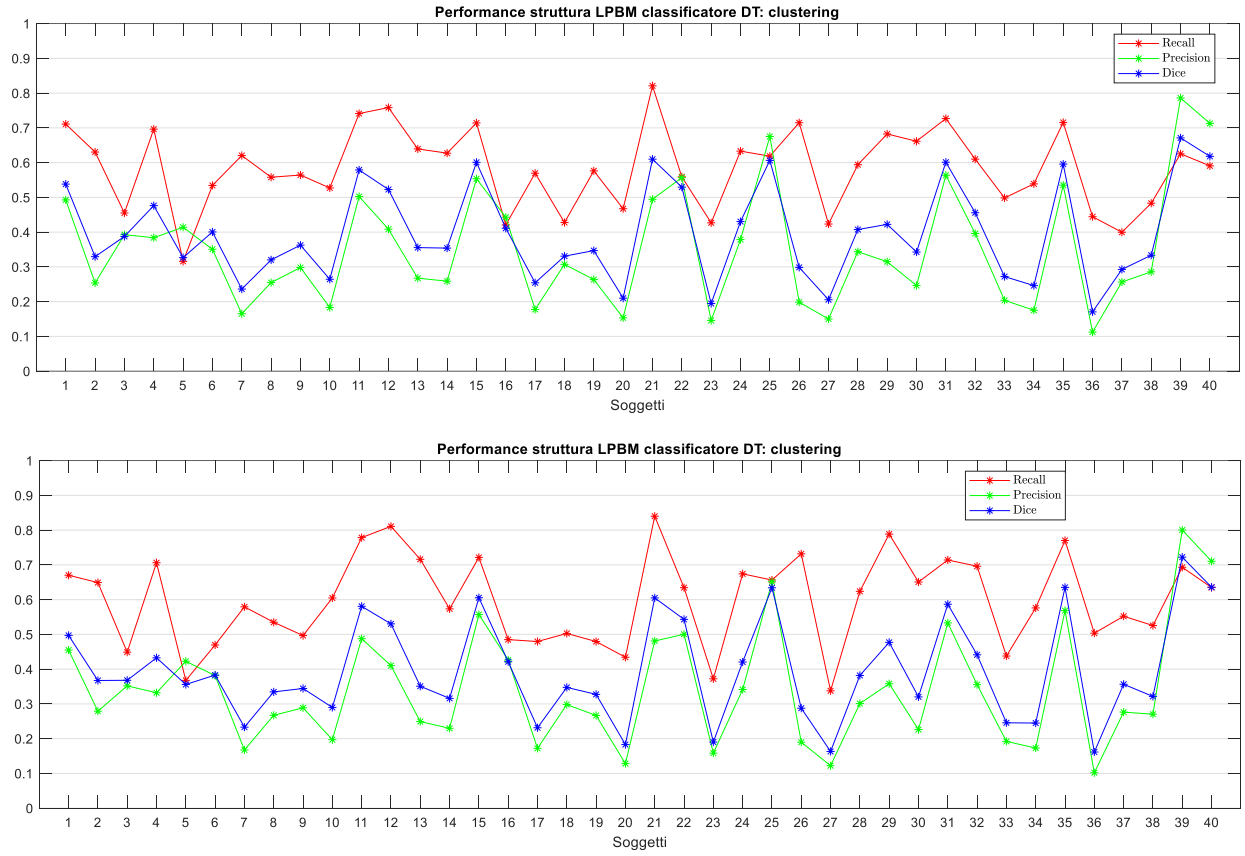


Figura 35 Performance classificatore DT e struttura LPBM. In alto le performance relative ai primi 40 pazienti, le cui slice, selezionate in maniera random, erano state parzialmente inserite nel construction set, in basso le performance relative ai primi 40 pazienti, le cui slice, selezionate in maniera stratificata, erano state parzialmente inserite nel construction set.

Dal confronto delle performance sui pazienti utilizzati per il construction set non appaiono particolari differenze tra i due metodi di selezione delle slice (figura 35). Pertanto, si è deciso di valutare le performance separando le slice utilizzate nel GA del classificatore da quelle restanti dei pazienti del construction set.

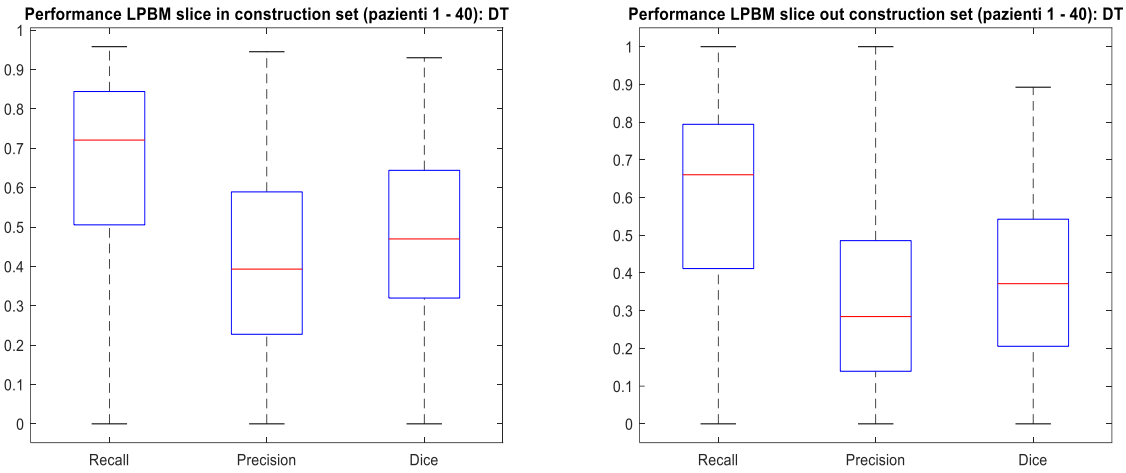


Figura 36 A sinistra le performance in termini di recall, precision e dice come boxplot sulle 200 slice presenti nel construction set, a destra gli stessi parametri di performance valutati sulle slice dei primi 40 pazienti che non erano state inserite nel construction set.

È possibile notare come le performance diminuiscano di alcuni punti percentuali per tutti e tre gli indicatori quando si passa alla classificazione delle slice non usate per il GA (figura 36). Questo comportamento naturalmente è del tutto ragionevole.
Purtroppo, però, anche le performance sulle slice del construction set non sono ottime. È chiaro, quindi, che il problema stia alla radice, infatti il GA, con i valori impostati, non riesce a trovare alcuna soluzione ottima per questo classificatore: performance migliori non possono di certo essere ottenute valutando tutta la struttura in sé.
Dal momento che le performance sono basse anche sulle slice usate nel GA, è opportuno andare a controllare le performance della soluzione a fitness minore trovata dal GA, per comprendere come questa sia stata scelta.
Di seguito i risultati sotto forma di confusion matrix per il classificatore DT sia con estrazione random delle slice che con il metodo stratificato.

Training set LSBM		Classe reale	
		RM	YM
Classe predetta	RM	4617	449
	YM	384	4554

Training set IBM		Classe reale	
		RM	YM
Classe predetta	RM	4581	439
	YM	418	4564

Training set LPBM (random)		Classe reale	
		RM	YM
Classe predetta	RM	4548	390
	YM	450	4613

Test set LSBM		Classe reale	
		RM	YM
Classe predetta	RM	247955	54854
	YM	184798	72862

Test set IBM		Classe reale	
		RM	YM
Classe predetta	RM	69613	18513
	YM	60538	19952

Test set LPBM (random)		Classe reale	
		RM	YM
Classe predetta	RM	42253	110314
	YM	37879	124006

Training set LPBM (stratificato)		Classe reale	
		RM	YM
Classe predetta	RM	4578	410
	YM	421	4589

Test set LPBM (stratificato)		Classe reale	
		RM	YM
Classe predetta	RM	43617	99599
	YM	39410	110004

Notiamo che le confusion matrix dei training set hanno sempre performance elevate (intorno al 90%) sia in termini di sensibilità che di specificità, mentre le confusion matrix dei test set hanno una diminuzione che porta entrambe le performance intorno al 50-55 %.

Lo stesso si presenta nel caso della confusion matrix sulle ROI dei pazienti 1-40 non incluse nel construction set e di seguito riportata.

ROI non constr LPBM (paz. 1-40)		Classe reale	
		RM	YM
Classe predetta	RM	204606	551299
	YM	194226	599202

Da queste evidenze sembra trasparire un chiaro problema di overfitting del classificatore DT.

Un aspetto interessante da andare ad analizzare è rappresentato dal numero di corretti classificati e di erroneamente classificati per ciascun cluster sia nel caso dell'RM che dell'YM per training set e test set (construction set stratificato) con classificatore DT in modo da verificare se il numero elevato di FP e FN nel test set sia da ricondurre a qualche cluster nello specifico.

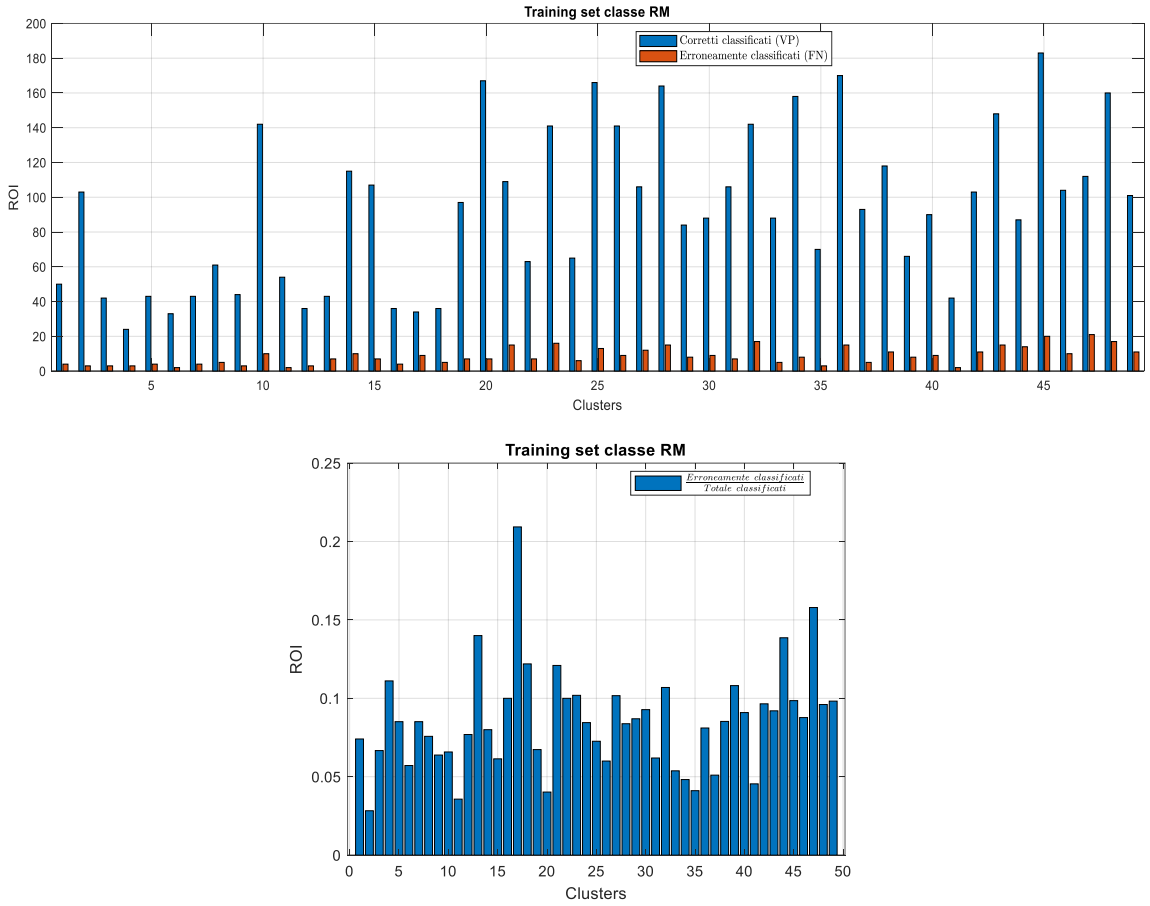


Figura 37 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun cluster della classe RM sul training set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun cluster della classe RM sul training set con classificatore DT

Nel caso della classe RM e del training set, il numero di FN, che di base era molto contenuto come evidente nelle confusion matrix, non sembra derivare particolarmente da nessun cluster, infatti si ha al massimo il 21 % di rapporto tra erroneamente classificati e ROI totali per il cluster 17 (figura 37).

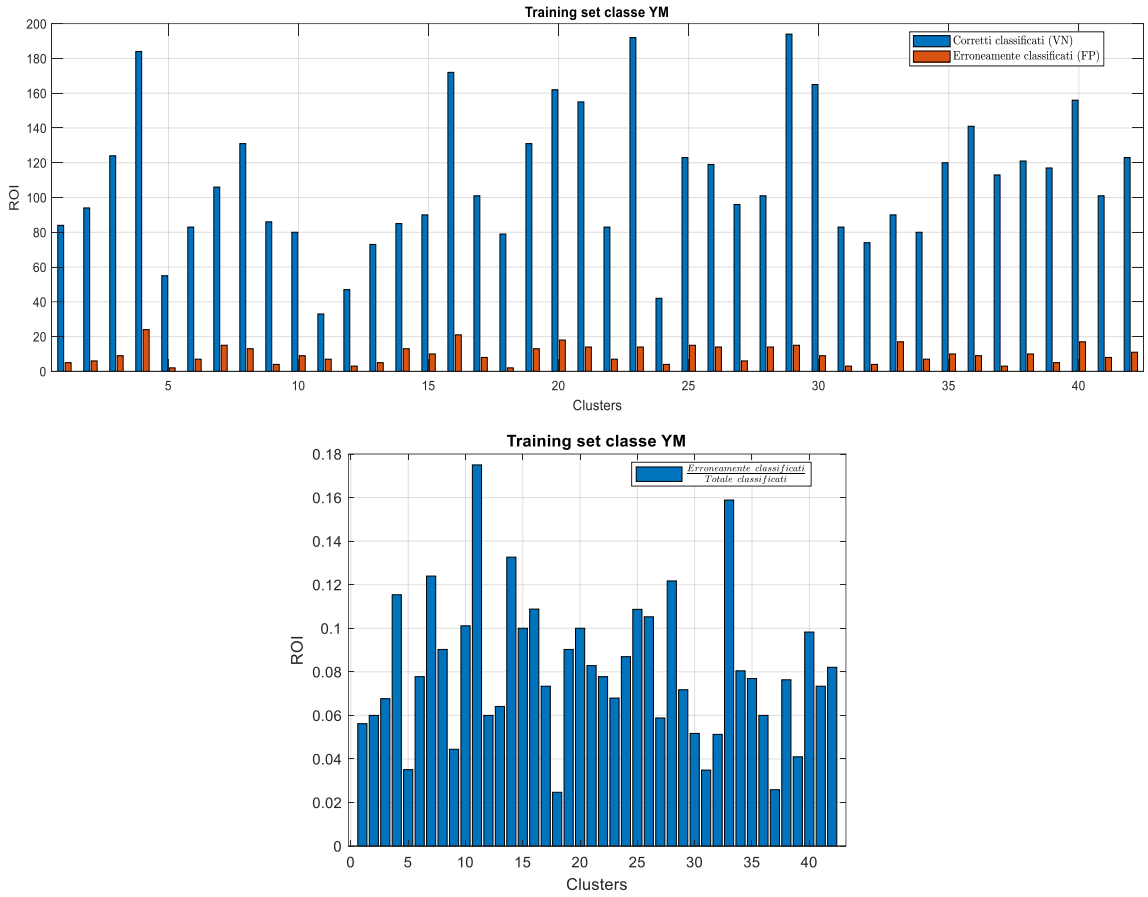
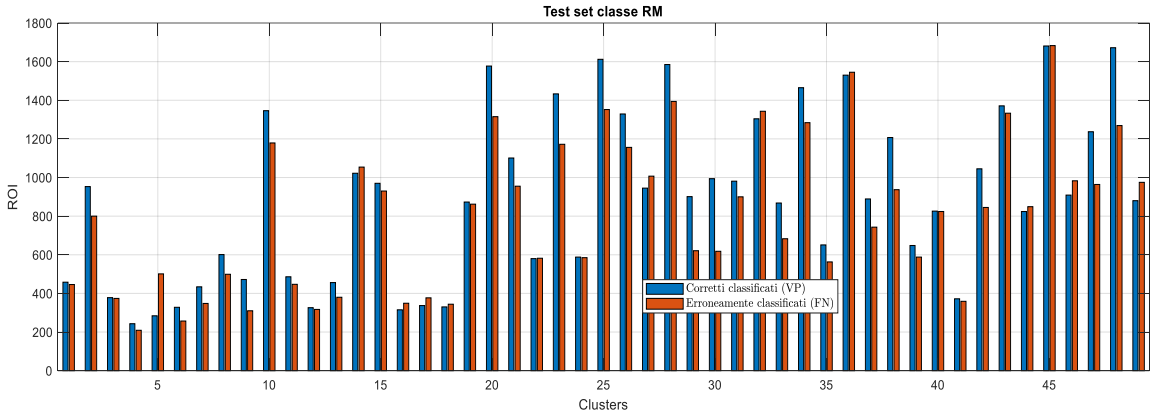


Figura 38 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun cluster della classe YM sul training set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun cluster della classe YM sul training set con classificatore DT

Le considerazioni precedentemente fatte per la classe RM valgono anche per la classe YM e training set, difatti la percentuale massima tra erroneamente classificati e ROI totali si ha per il cluster 11 e non supera il 17.5 % (figura 38).



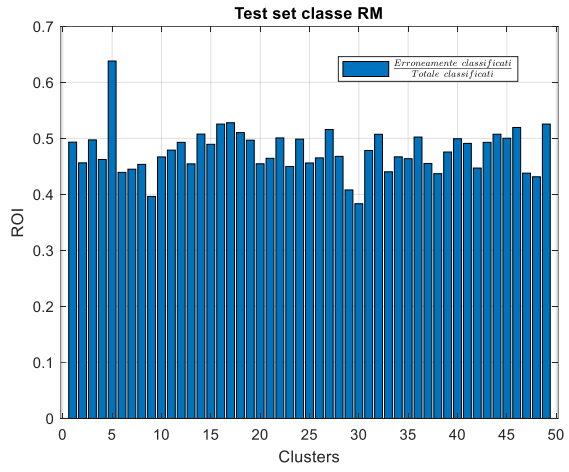


Figura 39 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun cluster della classe RM sul test set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun cluster della classe RM sul test set con classificatore DT

Sia per la classe RM che per la classe YM sul test set appare che il rapporto tra gli erroneamente classificati e il totale degli elementi di ciascun cluster sia sempre maggiore rispetto al caso del training set, tuttavia esso si aggira sempre intorno al 50 %. Pertanto, non è possibile attribuire il numero elevato di FP e FN a nessun cluster in particolare (figura 39 e 40).

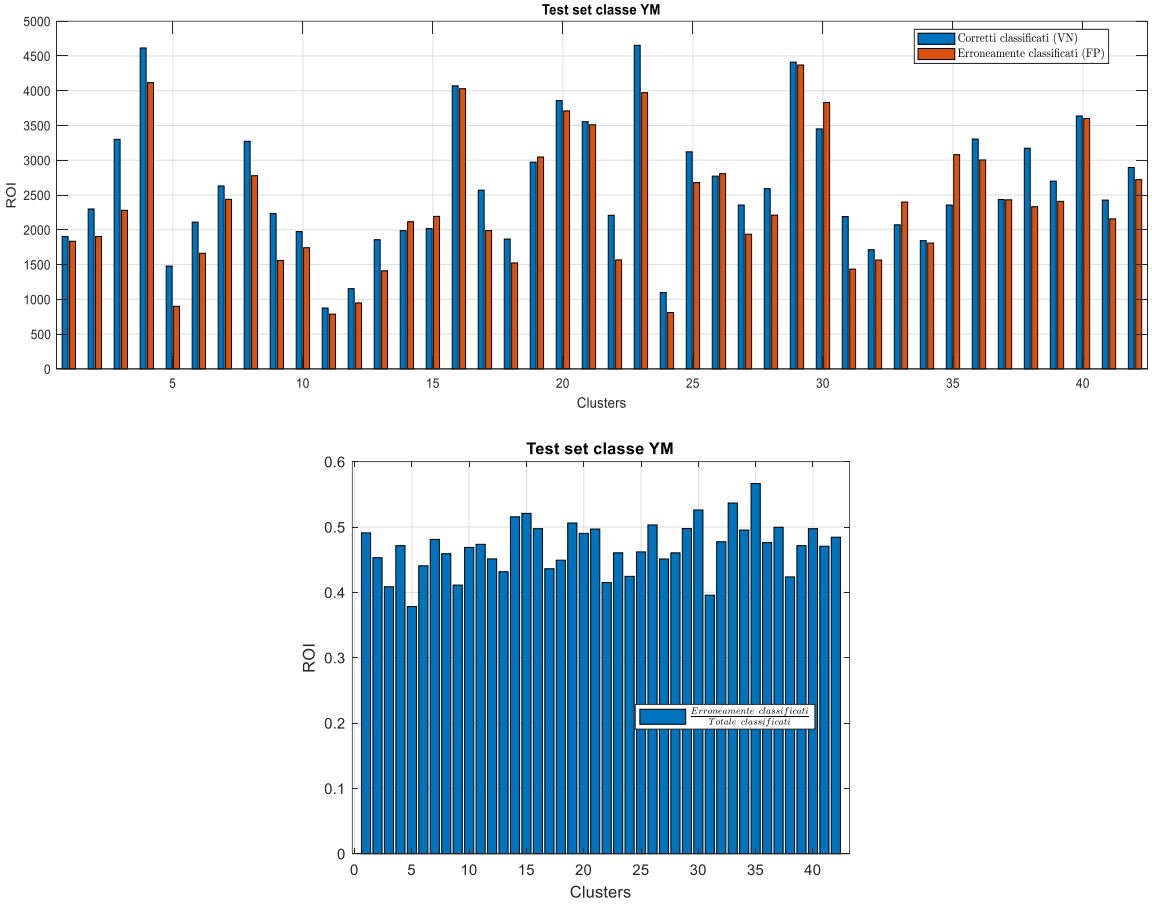


Figura 40 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun cluster della classe YM sul test set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun cluster della classe YM sul test set con classificatore DT

Per completezza, si è deciso di andare a controllare il numero di corretti classificati e di erroneamente classificati anche per ciascun paziente, invece che per ciascun cluster, sia nel caso dell'RM che dell'YM per training set e test set in modo da verificare se il numero elevato di FP e FN nel test set derivasse dalla difficoltà di classificazione di qualche paziente in particolare.

Di seguito vengono quindi riportati gli stessi bardiagram precedenti, ma con l'indicazione dei parametri calcolati suddivisi per paziente.

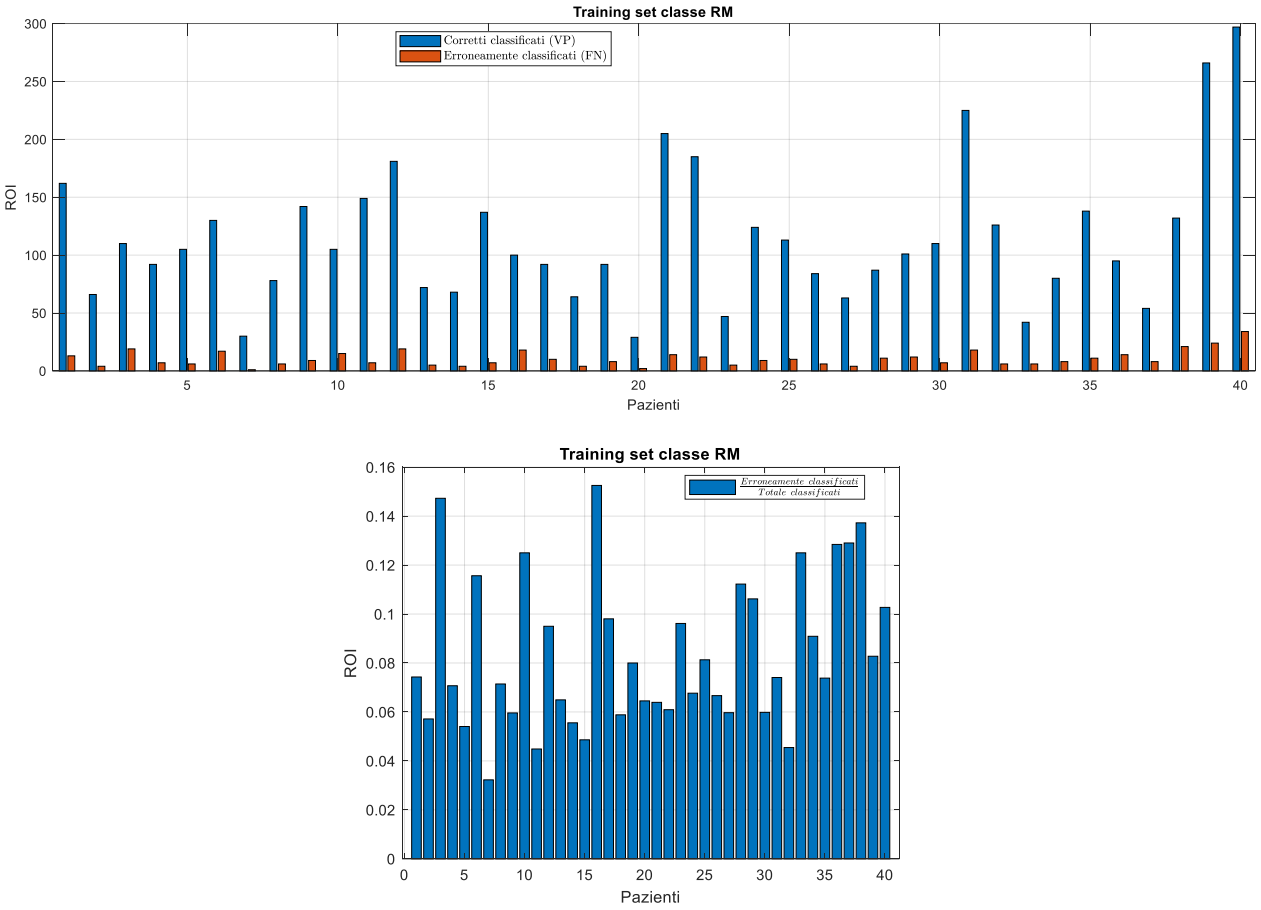
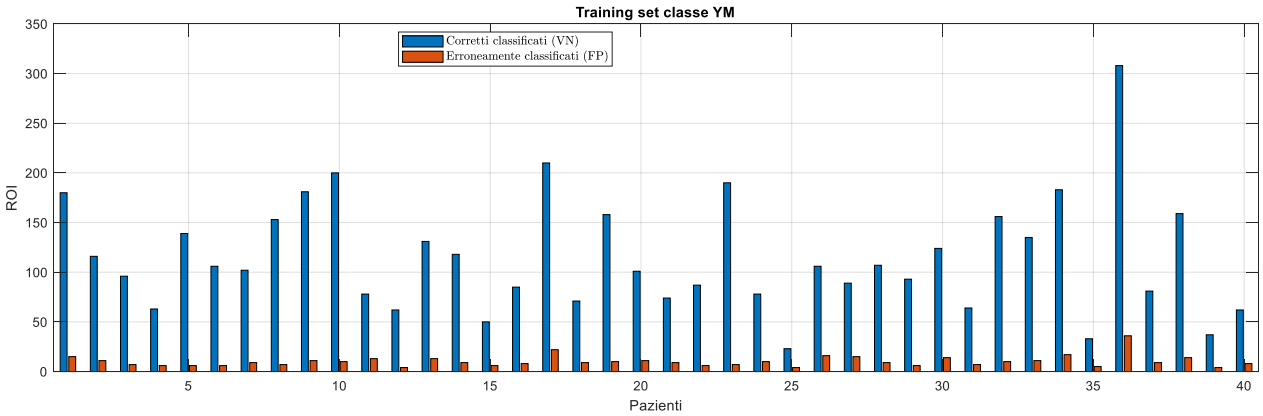


Figura 41 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun paziente della classe RM sul training set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun paziente della classe RM sul training set con classificatore DT

Andando a osservare la distribuzione del rapporto tra ROI erroneamente classificate e ROI totali contenute in ciascun paziente nel training set della classe RM, si nota che i valori sono molto contenuti e raggiungono un massimo del 15 % circa per il paziente 16 (figura 41).



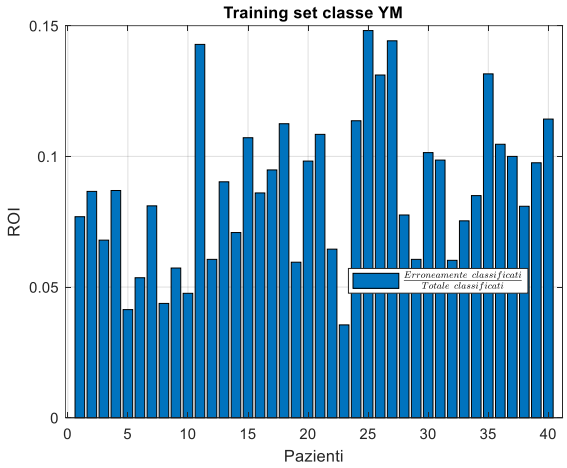


Figura 42 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun paziente della classe YM sul training set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun paziente della classe YM sul training set con classificatore DT

Lo stesso comportamento riguarda la classe YM sul training set, con un rapporto massimo pari al circa 15 % per il paziente 25 (figura 42).

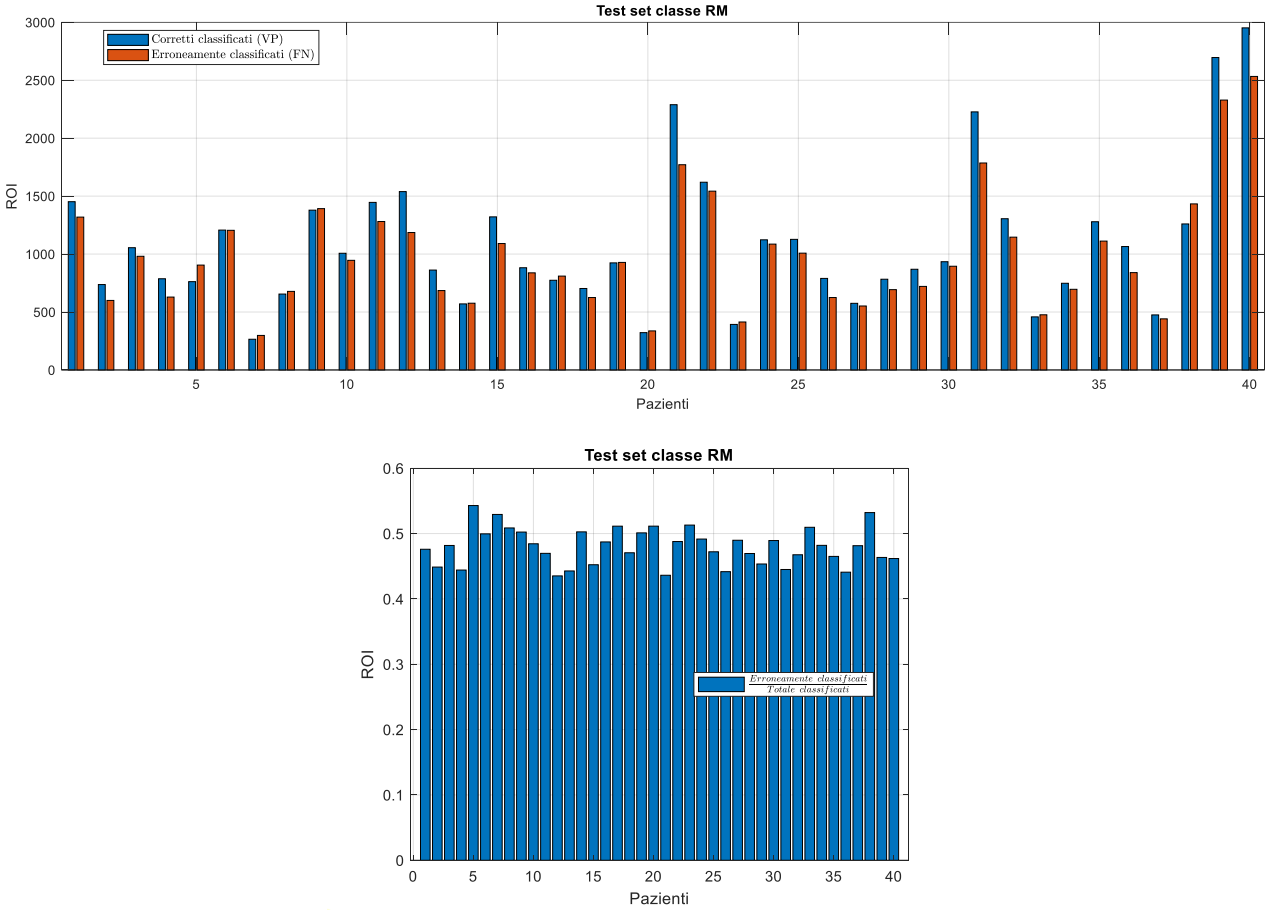


Figura 43 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun paziente della classe RM sul test set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun paziente della classe RM sul test set con classificatore DT

Relativamente all'analisi del test set, si nota che per tutti pazienti il rapporto tra ROI erroneamente classificate e totali si aggira sempre intorno al 50 % per entrambe le classi di midollo. Al pari, quindi, di quanto succedeva per la suddivisione in funzione del cluster, non si evidenzia un comportamento diverso di nessun paziente rispetto alla media degli altri e non è possibile assegnare

a nessun paziente nello specifico l'elevato numero di FP e FN che caratterizzano la classificazione sul test set (figure 43 e 44).

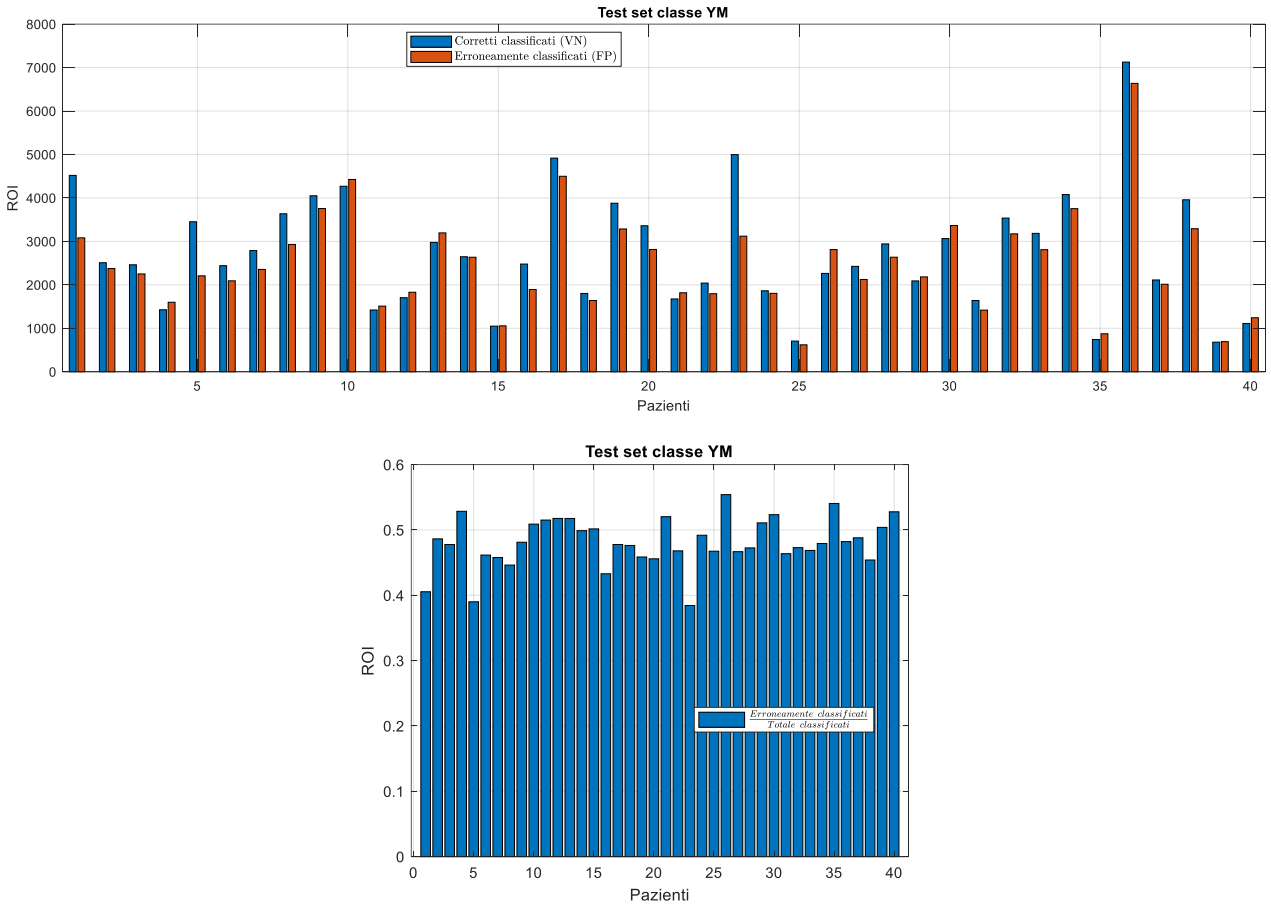


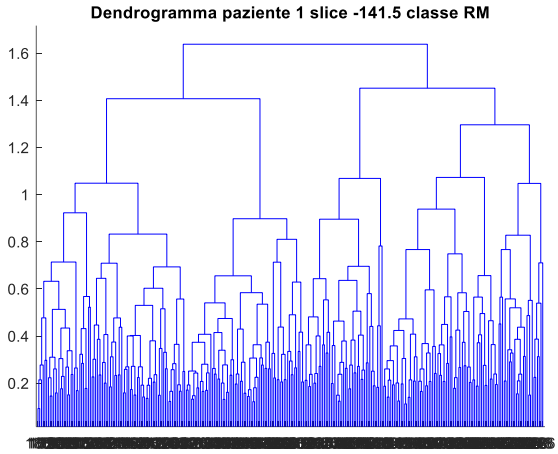
Figura 44 In alto bardiagram del numero di ROI correttamente classificate ed erroneamente classificate per ciascun paziente della classe YM sul test set con classificatore DT. In basso bardiagram del rapporto tra il numero di ROI erroneamente classificate rispetto al totale di ciascun paziente della classe YM sul test set con classificatore DT

4.2 Clustering con dendrogramma

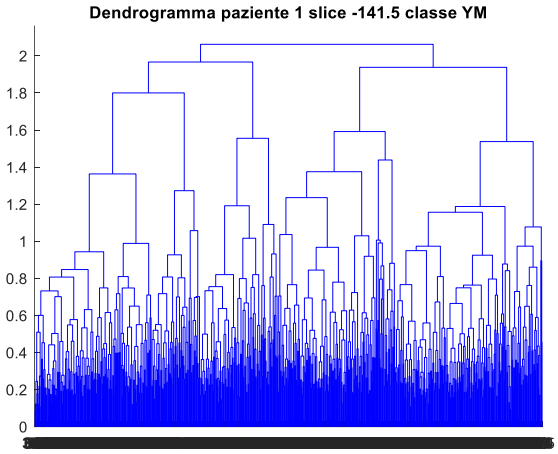
Dal momento che nessun classificatore fornisce risultati sufficientemente buoni, sorge spontaneo domandarsi se il problema sia qualcosa di comune a tutti, ossia, ancora una volta, il training set e la sua rappresentatività dell'intero construction set.

Per chiarire meglio questo aspetto, si sono osservati i dendrogrammi di ciascuna delle 5 slice presenti nei primi 5 pazienti del construction set e si è visto che tra loro sono molto diversi. In particolare, alcuni dendrogrammi sono caratterizzati da situazioni molto singolari in cui si uniscono cluster con numerosità molto diversa e comprendenti gruppi di pochi elementi ad elevata variabilità. Se questi dendrogrammi venissero sottoposti al taglio sulla base della distanza, come fatto in precedenza per le reti SOM, potrebbe venirsi a creare un numero di cluster tale per cui da un determinato cluster che contiene al suo interno una particolarità l'estrazione random delle ROI porta alla perdita di quella particolarità.

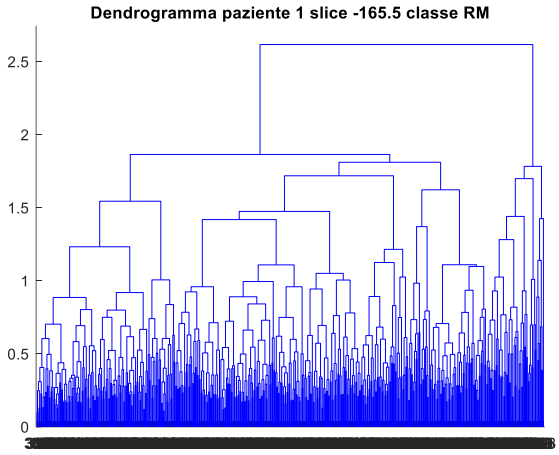
I dendrogrammi delle 5 slice del paziente 1 del construction set sono mostrati di seguito assieme all'indicazione del numero di cluster che verrebbero a formarsi usando il metodo del taglio sulla base della distanza.



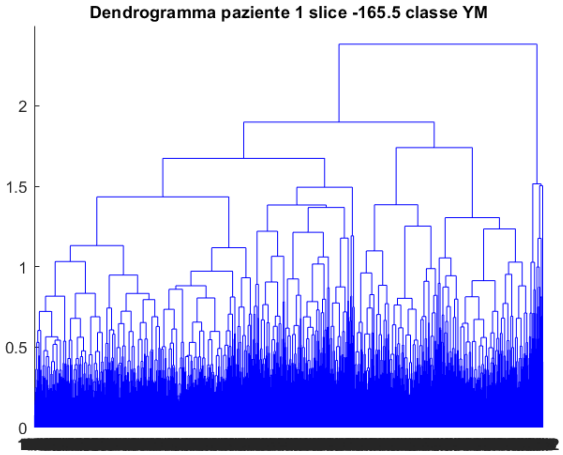
N° cluster: 33



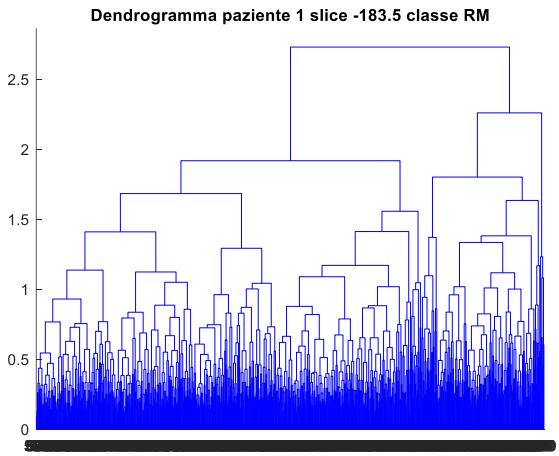
N° cluster: 3



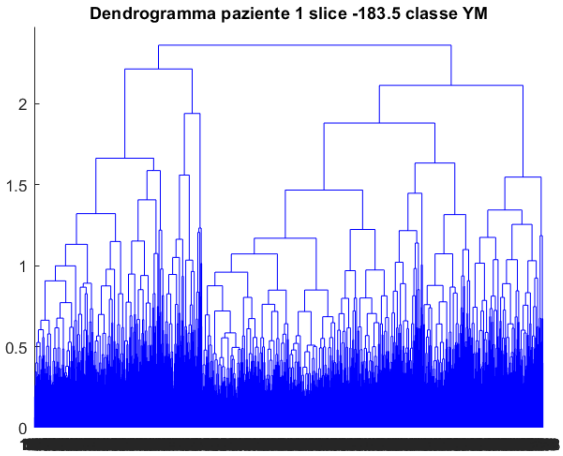
N° cluster: 42



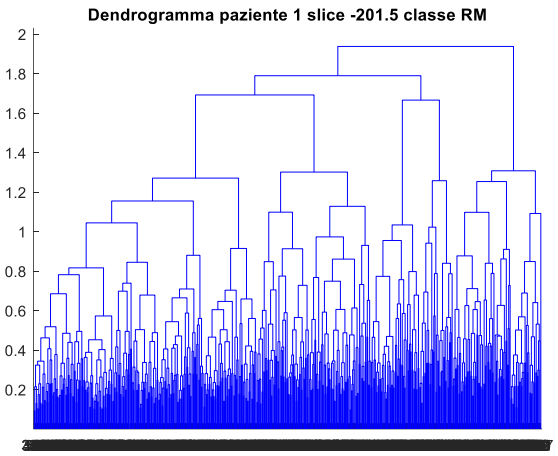
N° cluster: 53



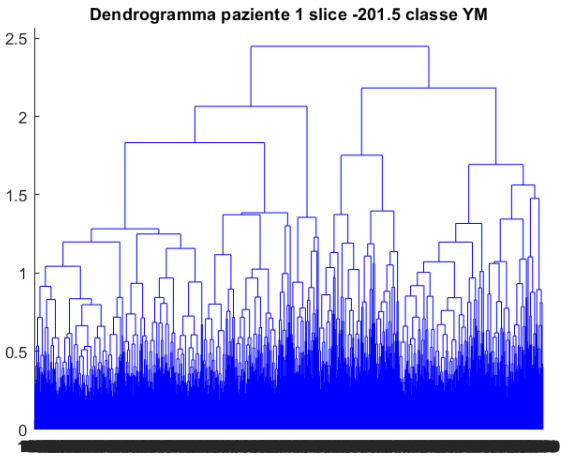
N° cluster: 32



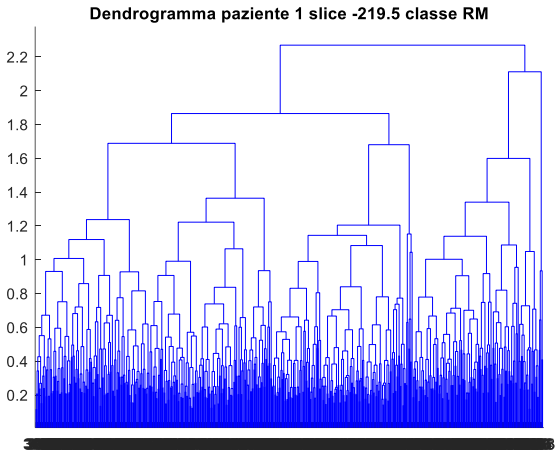
N° cluster: 82



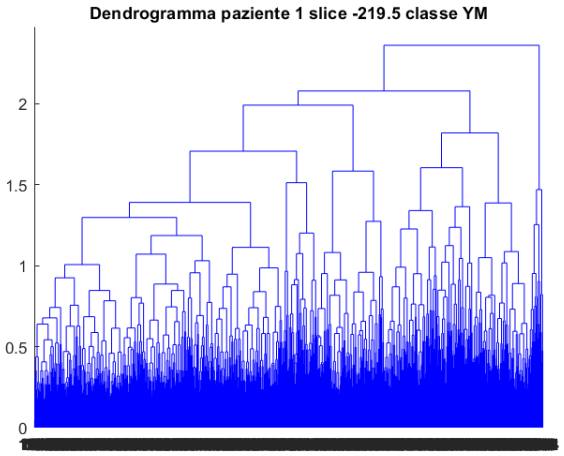
N° cluster: 34



N° cluster: 12



N° cluster: 41



N° cluster: 9

L'intenzione è, dunque, proprio quella di rivolgere l'attenzione a questi cluster più piccoli ad alta variabilità e di individuare quell'iterazione per cui l'unione con un cluster più piccolo porterebbe alla maggiore differenza di variabilità, aggiungendo di conseguenza dell'informazione con peso molto elevato.

Per procedere in tal senso, sono state selezionate alcune slice tra le 5 dei primi 5 pazienti del construction set e si è analizzata la differenza della distanza inter-cluster tra due iterazioni consecutive del dendrogramma, la variabilità intra-cluster del neo-cluster che si forma ad ogni iterazione e la differenza tra la variabilità del neo-cluster formato e quella del più numeroso cluster dei due da cui esso deriva a ciascuna iterazione.
Nel caso di quest'ultima differenza calcolata, per facilitare la lettura del grafico, è stata posta a zero la differenza qualora il neo-cluster che si forma, ad una determinata iterazione, derivi dall'unione di due cluster di cui anche solo uno contiene un singolo elemento.

Vengono quindi indicati i risultati relativi alla prima slice selezionata.

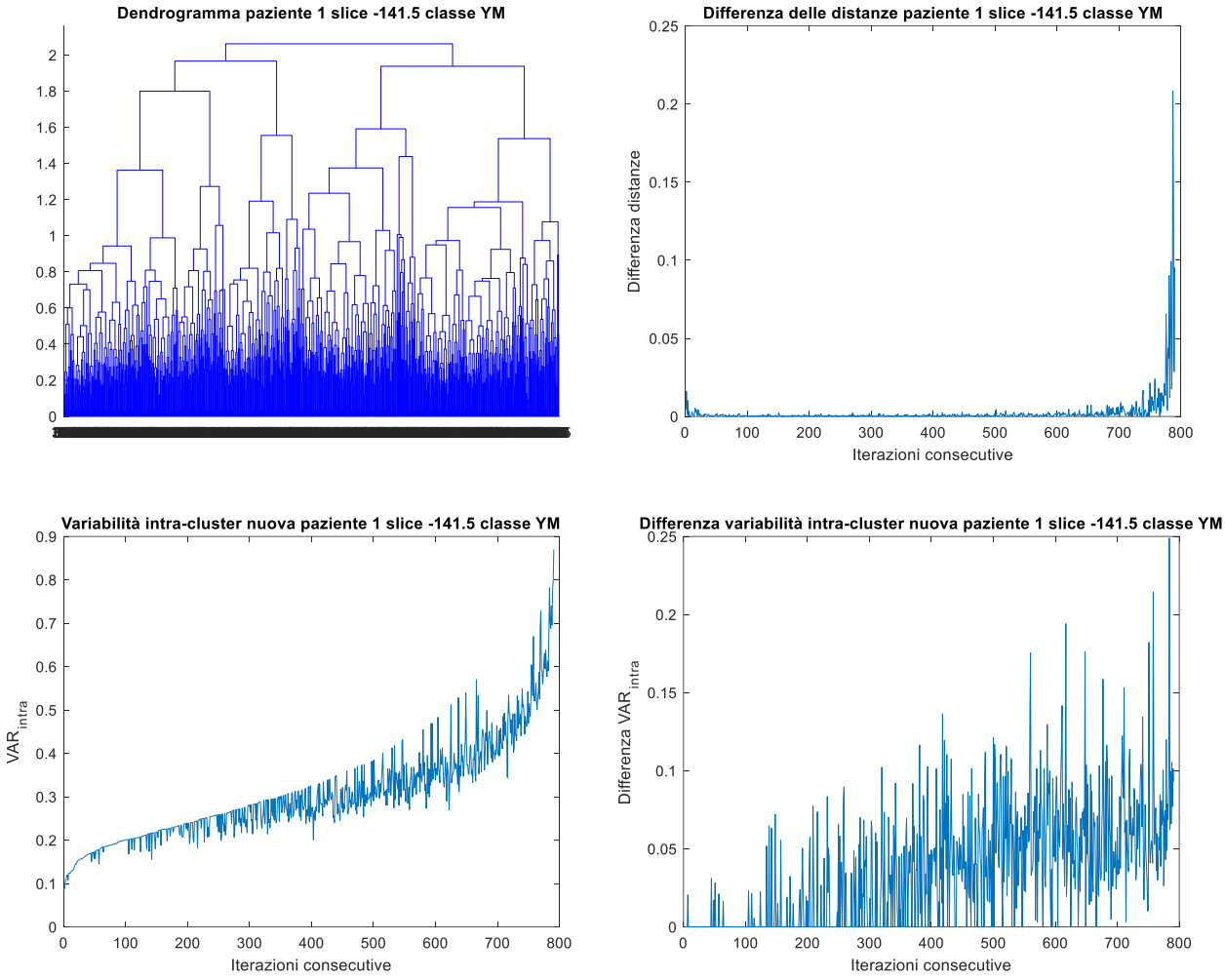


Figura 45 In alto a sinistra il dendrogramma della slice selezionata, in alto a destra il grafico della differenza delle distanze inter-cluster tra iterazioni consecutive del dendrogramma. In basso a sinistra la variabilità intra-cluster del neo-cluster che si viene a formare ad ogni iterazione, in basso a destra la differenza tra la variabilità intra-cluster del neo-cluster formato e la variabilità intra-cluster del più numeroso dei due cluster da cui esso deriva (valori posti a zero se uno dei cluster di origine contiene un solo elemento).

Si può vedere come il grafico della differenza della variabilità intra-cluster possieda un picco massimo che rappresenta l'iterazione di unione tra due cluster di numerosità diversa dove il più piccolo dei due porta con sé una elevata variabilità che si ripercuote su quella del neo-cluster formato (figura 45). Tale picco si trova, inoltre, in corrispondenza di una iterazione che porta a una differenza di distanza inter-cluster piuttosto elevata. Difatti, le ultime iterazioni del dendrogramma sono quelle che conducono a una differenza di distanza con velocità di crescita in genere progressivamente maggiore quasi ad ogni iterazione. Questa iterazione risulta quindi piuttosto adeguata a un eventuale taglio del dendrogramma.
L'effetto viene qui riportato in termini di numero di cluster formatosi e di relativa numerosità di ciascuno di essi (figura 46).

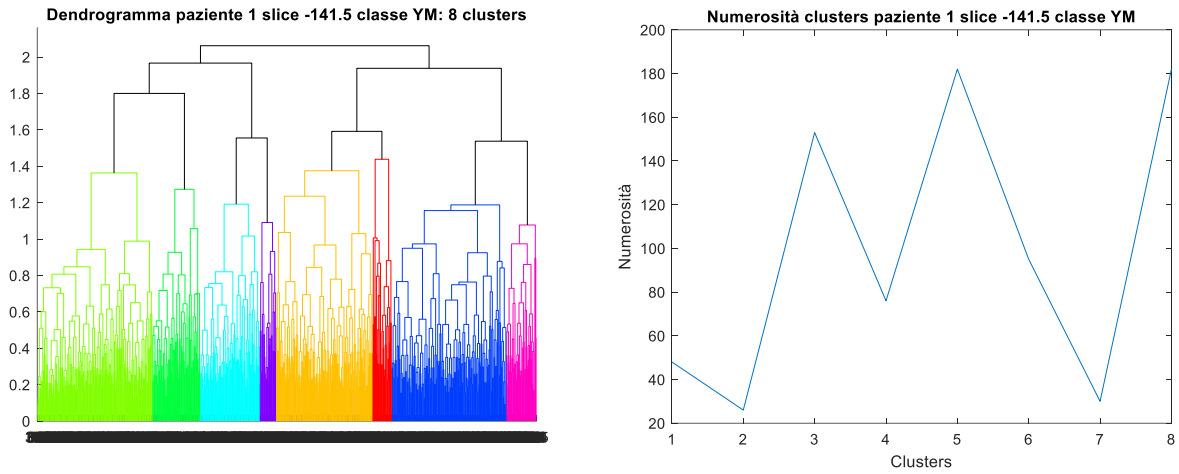


Figura 46 A sinistra dendrogramma della slice selezionata con indicazione del numero di cluster formatisi con il taglio individuato e la loro collocazione sul dendrogramma. A destra numerosità di ciascun cluster individuato dal taglio.

Volendo, pertanto, formalizzare e riassumere le fasi che portano alla determinazione di questo nuovo metodo di taglio basato sulla variabilità intra-cluster, si può dire che esso consta dei seguenti step:

- identificazione, per ciascuna iterazione del dendrogramma, del cluster a numerosità maggiore tra i due che si uniscono (la cui variabilità intra-cluster quindi ha il maggiore peso tra le due);
- calcolo della variabilità intra-cluster dello stesso;
- calcolo della variabilità intra-cluster del neo-cluster formato a ciascuna iterazione;
- calcolo della differenza per ogni iterazione tra le due variabilità sopra citate;
- identificazione dell'iterazione la cui differenza di variabilità è massima (ossia quell'iterazione per cui l'unione con un cluster più piccolo ha portato alla maggiore differenza di variabilità, aggiungendo quindi dell'informazione di grande rilevanza);
- taglio del dendrogramma in corrispondenza di quell'iterazione (al di sotto del suo nodo).

4.3 Generazione di nuovi training set

Utilizzando il nuovo metodo di taglio sulla base della variabilità intra-cluster si è proceduto all'ottenimento di un nuovo training set per la struttura LPBM a partire dal construction set stratificato.
Più nel dettaglio, per ciascun paziente sono state estratte 3 ROI da ogni cluster ottenuto tagliando i dendrogrammi di ognuna delle sue 5 slice presenti nel construction set con il taglio in base alla variabilità e laddove un cluster contasse meno di 3 ROI sono state selezionate tutte quelle presenti al suo interno. Questa operazione è stata svolta separatamente sia per la classe RM che per la classe YM.
I risultati di questa estrazione, in termini di numero di ROI estratte, sono riassunti nella tabella seguente.

	RM	YM	BM
Soggetto 1	1060	488	1548
Soggetto 2	641	560	1201
Soggetto 3	903	1104	2007
Soggetto 4	557	719	1276
Soggetto 5	611	842	1453
Soggetto 6	784	980	1764
Soggetto 7	320	1410	1730
Soggetto 8	155	267	422
Soggetto 9	523	1024	1547
Soggetto 10	546	1748	2294
Soggetto 11	1082	514	1596
Soggetto 12	831	894	1725
Soggetto 13	546	1431	1977
Soggetto 14	377	933	1310
Soggetto 15	697	746	1443
Soggetto 16	474	1000	1474
Soggetto 17	569	2562	3131
Soggetto 18	547	993	1540
Soggetto 19	443	2020	2463
Soggetto 20	221	1029	1250
Soggetto 21	869	940	1809
Soggetto 22	871	1257	2128
Soggetto 23	287	901	1188
Soggetto 24	758	1135	1893
Soggetto 25	640	861	1501
Soggetto 26	574	1359	1933
Soggetto 27	443	1663	2106
Soggetto 28	447	1088	1535
Soggetto 29	680	2218	2898
Soggetto 30	674	2078	2752
Soggetto 31	1003	1125	2128
Soggetto 32	978	1351	2329
Soggetto 33	361	564	925
Soggetto 34	289	1022	1311
Soggetto 35	871	473	1344
Soggetto 36	869	2312	3181
Soggetto 37	473	938	1411
Soggetto 38	1033	787	1820
Soggetto 39	1493	583	2076
Soggetto 40	1848	455	2303
TOTALE	**27348**	**44374**	**71722**

A questo punto, con le ROI così individuate sono stati realizzati due ulteriori dendrogrammi, uno per ogni classe, che sono stati tagliati ancora una volta con il nuovo metodo di taglio (figura 46). Da ciascun cluster ottenuto, sono state estratte in maniera proporzionale alla sua numerosità un certo numero di ROI, in modo da costruire un training set che presentasse circa 5000 ROI di ognuna delle due classi. Gli elementi non inseriti nel training set sono andati a costituire il test set.
La numerosità finale dei due set è mostrata di seguito.

	RM	YM	BM
Training set	5017	5050	10067
Test set	22331	39324	61655

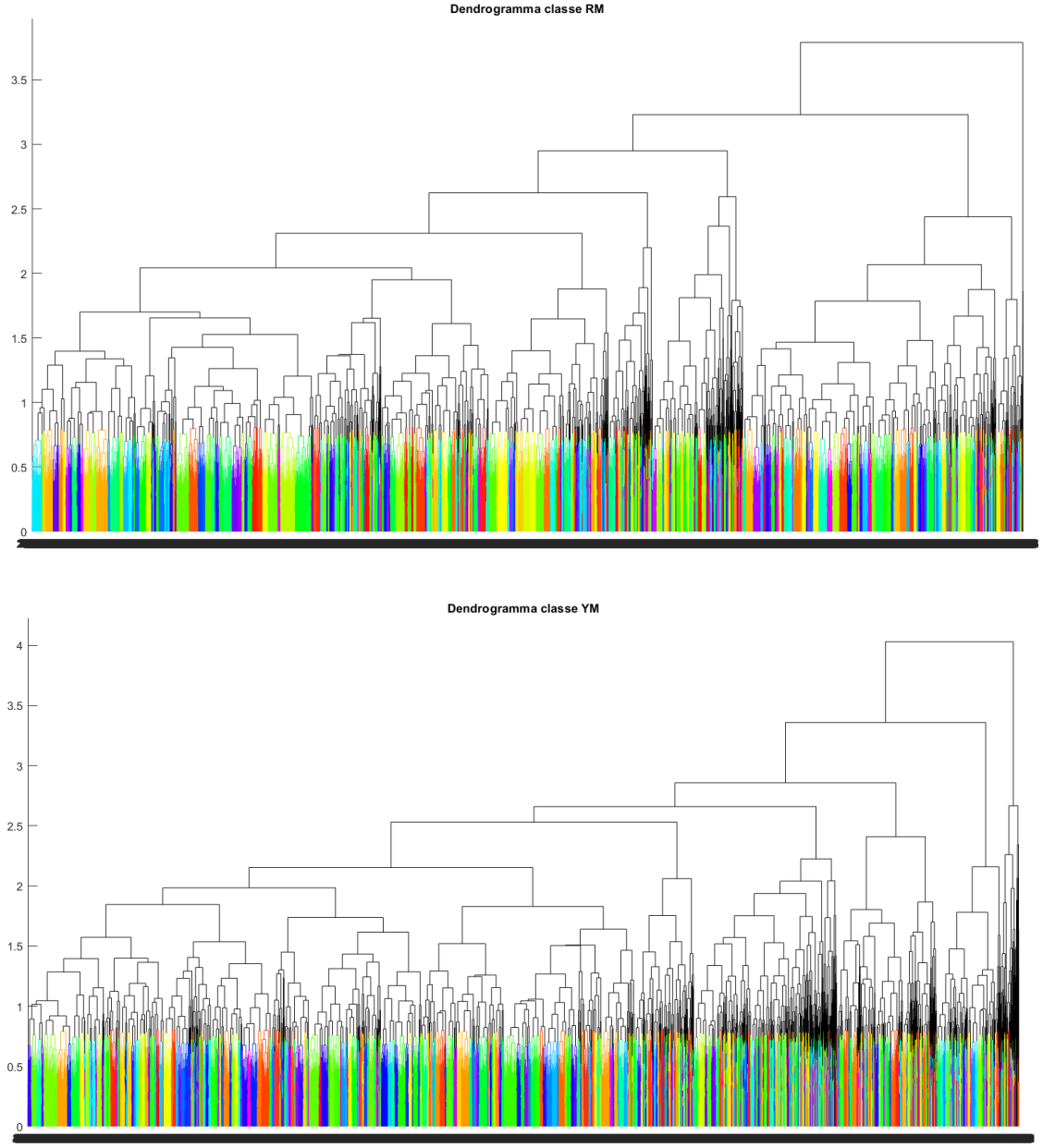
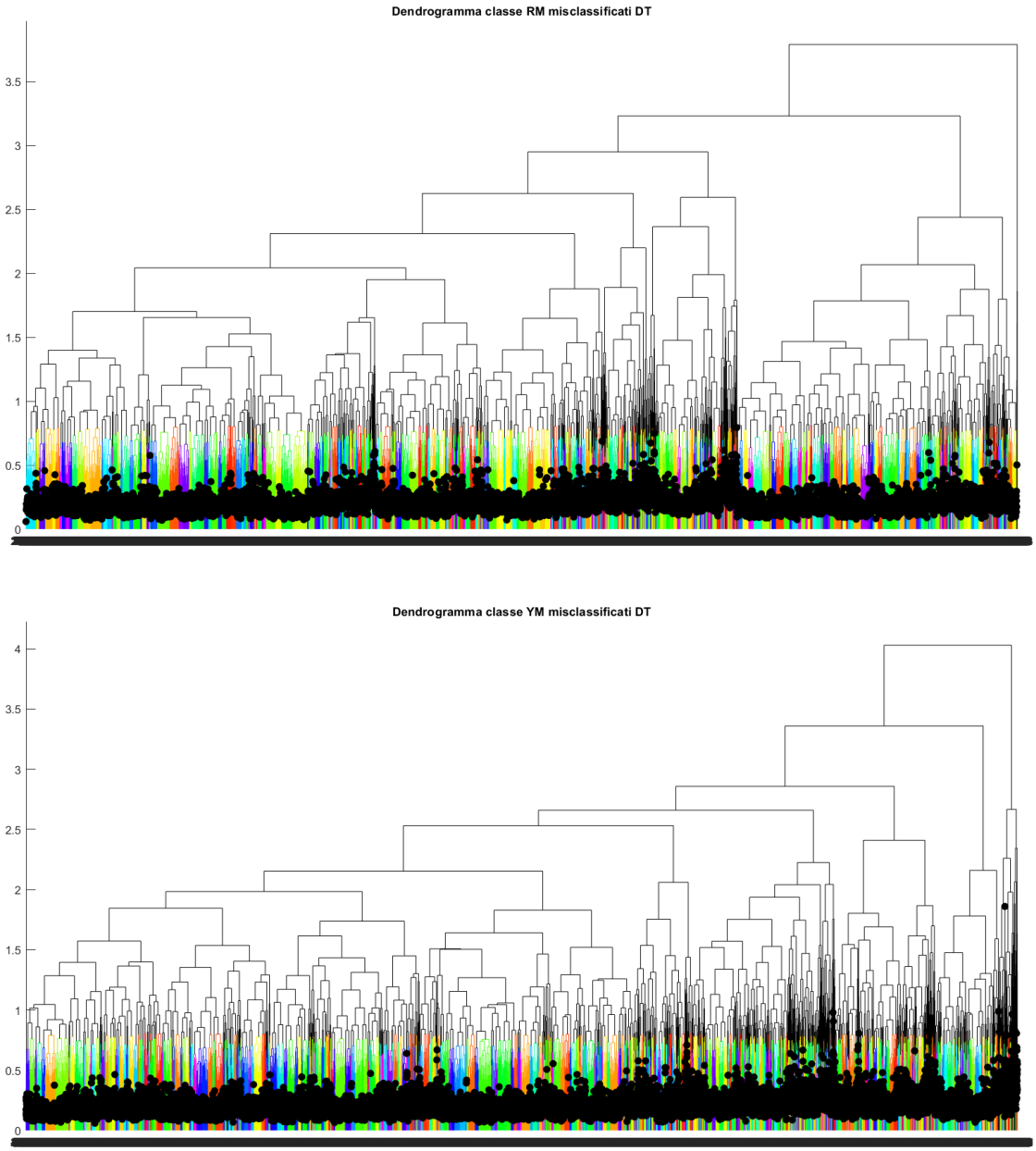


Figura 47 In alto dendrogramma delle ROI di RM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster. In basso dendrogramma delle ROI di YM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster.

Con questo nuovo training set sono stati allenati DT, KNN e NN, andando ad utilizzare i parametri ottenuti in precedenza con i GA (per il DT con i dati del construction set stratificato, mentre per KNN e NN con i dati del construction set random). I classificatori sono stati poi testati separatamente sul training set, sul test set (ossia le ROI non contenute nel training set ma selezionate nella fase di estrazione di 3 ROI per cluster) e sulle ROI restanti del construction set di partenza.

L'operazione di ottenimento del training set in maniera proporzionale è stata ripetuta 5 volte così come le prove sui classificatori, quindi, nel complesso, si sono ottenuti 5 training set, 5 test set e 5 modelli per ognuno dei classificatori.

Vengono riportate le confusion matrix derivanti dalla verifica dei classificatori e i dendrogrammi delle sue classi con l'indicazione di tutti gli elementi misclassificati nel training set e nel test set della prima ripetizione.



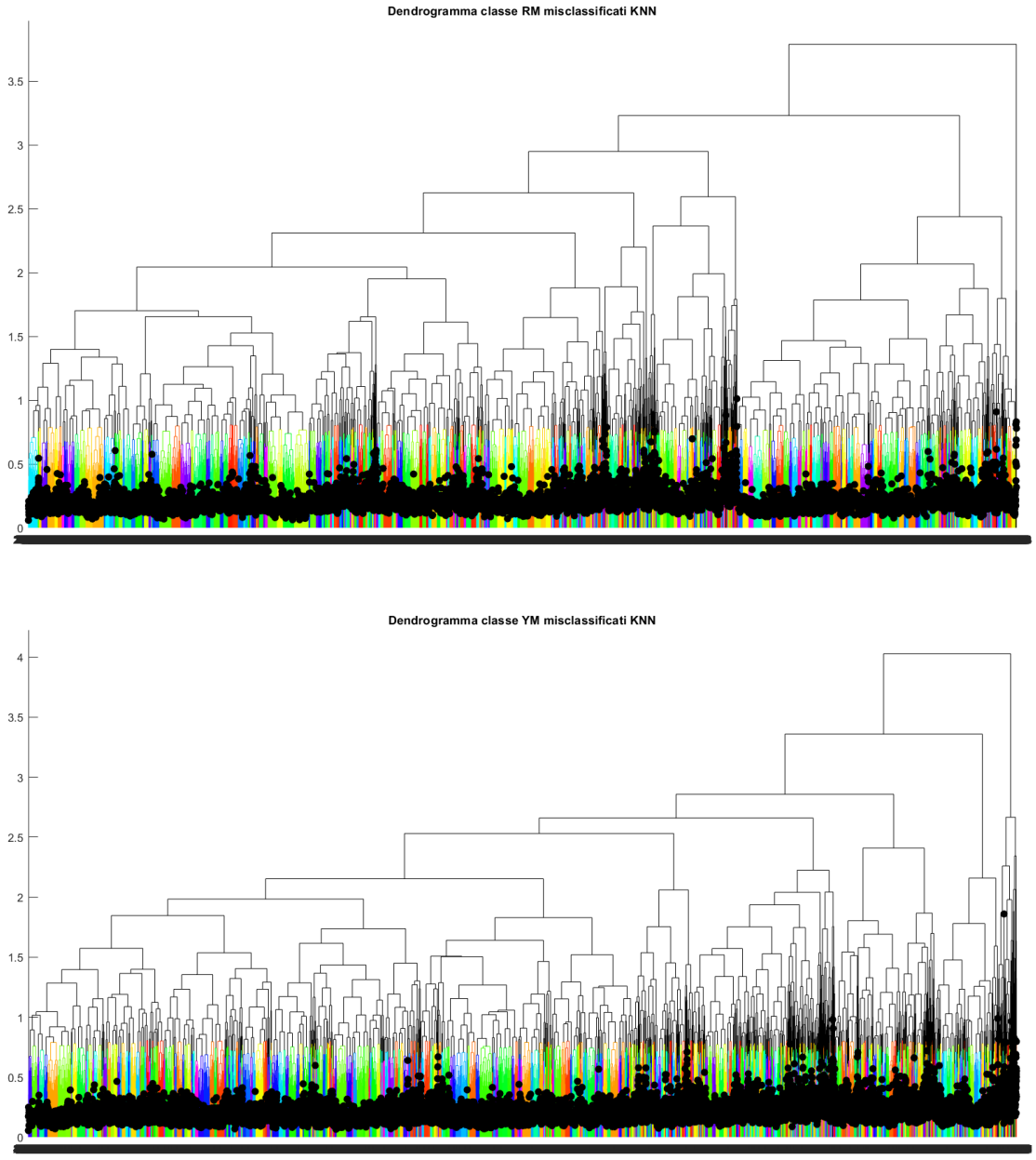
Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	4598	415
	YM	419	4635

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	11863	19028
	YM	10648	20296

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	32716	84581
	YM	27962	85647

Figura 48 In alto dendrogramma delle ROI di RM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster e indicazione dei misclassificati per il classificatore DT con marker circolare nero. Al centro dendrogramma delle ROI di YM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster cluster e indicazione dei misclassificati per il classificatore DT con marker circolare nero. In basso confusion matrix su training set, test set e ROI restanti del construction set per il classificatore DT. I dati sono relativi al training set 1 e al test set 1.

Come avveniva per i training set precedentemente realizzati, si nota un overfitting del classificatore DT, infatti sia la sensibilità che la specificità passano da valori oltre il 90 % per il training set a valori intorno al 50 – 55 % per il test set e le altre ROI del construction set (figura 48).



Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3105	2435
	YM	1912	2615

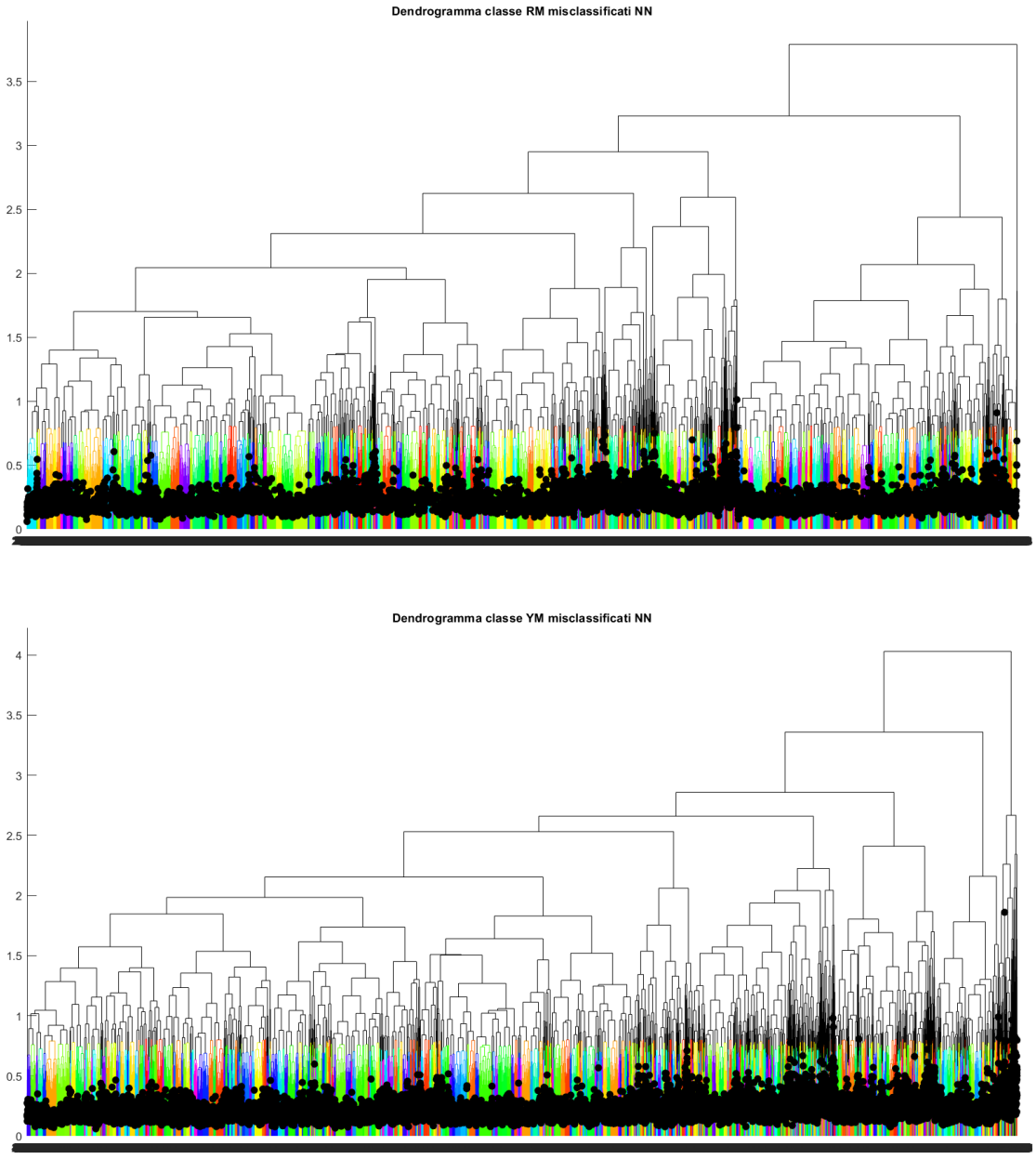
Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13415	20254
	YM	8916	19070

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	40303	97579
	YM	20375	72649

Figura 49 In alto dendrogramma delle ROI di RM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster e indicazione dei misclassificati per il classificatore KNN con marker circolare nero. Al centro dendrogramma delle ROI di YM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster cluster e indicazione dei misclassificati per il classificatore KNN con marker circolare nero. In basso confusion matrix su training set, test set e ROI restanti del construction set per il classificatore KNN. I dati sono relativi al training set 1 e al test set 1.

Nel caso dei classificatori KNN e NN, non si manifesta overfitting sul training set e si denota un miglioramento delle performance in termini di sensibilità sul test set e sulle ROI restanti del construction set. Tuttavia, le performance peggiorano sulla classe YM, infatti, rispetto ai casi visti in precedenza, la specificità è inferiore. Questo comportamento potrebbe, in parte, derivare dall'aver usato come feature e parametri per

questi classificatori quelli derivanti dal GA applicato al construction set ottenuto con selezione random delle slice (figura 49 e 50).



Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3268	2635
	YM	1749	2415

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	14138	21126
	YM	8193	18198

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	45329	105664
	YM	15349	64564

Figura 50 In alto dendrogramma delle ROI di RM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster e indicazione dei misclassificati per il classificatore NN con marker circolare nero. Al centro dendrogramma delle ROI di YM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster cluster e indicazione dei misclassificati per il classificatore NN con marker circolare nero. In basso confusion matrix su training set, test set e ROI restanti del construction set per il classificatore NN. I dati sono relativi al training set 1 e al test set 1.

Le confusion matrix relative ai training set e ai test set delle successive 4 ripetizioni sono indicati di seguito:

- Ripetizione 2:

Classificatore DT

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	4574	373
	YM	443	4677

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	11466	18604
	YM	10865	20720

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	31709	81674
	YM	28969	88554

Classificatore KNN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3123	2387
	YM	1894	2663

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13423	20080
	YM	8908	19244

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	39626	94755
	YM	21052	75473

Classificatore NN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3710	2993
	YM	1307	2057

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	16099	23957
	YM	6232	15367

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	48500	112997
	YM	12178	57231

- Ripetizione 3:

Classificatore DT

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	4548	391
	YM	469	4659

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	11360	18397
	YM	10971	20297

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	31629	80671
	YM	29049	89557

Classificatore KNN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3209	2433
	YM	1808	2617

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13819	20811
	YM	8512	18513

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	39408	95166
	YM	21270	75062

Classificatore NN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3150	2510
	YM	1867	2540

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13850	20128
	YM	8481	19196

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	41173	91808
	YM	19505	78420

- Ripetizione 4:

Classificatore DT

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	4609	397
	YM	408	4653

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	11709	18606
	YM	10622	20718

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	31792	82322
	YM	28886	87906

Classificatore KNN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3132	2426
	YM	1885	2624

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13221	20275
	YM	9110	19049

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	38404	92495
	YM	22274	77733

Classificatore NN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3013	2431
	YM	2004	2619

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13323	19254
	YM	9008	20070

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	40605	92476
	YM	20073	77752

- Ripetizione 5:

Classificatore DT

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	4619	382
	YM	398	4668

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	11603	18861
	YM	10728	20463

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	32295	83620
	YM	28383	86608

Classificatore KNN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3134	2415
	YM	1883	2635

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13361	20415
	YM	8970	18909

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	39528	96433
	YM	21150	73795

Classificatore NN

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	3174	2540
	YM	1843	2510

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	13809	20224
	YM	8522	19100

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	42630	98666
	YM	18048	71562

È possibile notare come le confusion matrix relative alla verifica dei classificatori si mantengano molto simili, in termini di numero di ROI correttamente ed erroneamente classificate, tra i 5 training set diversi che sono stati estratti.

Dato l'elevato numero di misclassificati in entrambe le classi, risulta difficile andare ad individuare visivamente quanto tra classificatori diversi i misclassificati si mantengano gli stessi.
A tale scopo, si è deciso di quantificare i misclassificati comuni ai tre classificatori nelle due classi calcolando l'indice di Sørensen - Dice per dati discreti, cioè:

$$ISD = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

dove $|X \cap Y|$ è la cardinalità dell'intersezione di due insiemi di misclassificati, ossia il numero di misclassificati comuni ai due insiemi, e $|X|$ e $|Y|$ sono le cardinalità dei due insiemi.

Sono indicati nelle tabelle sottostanti i risultati relativi a ciascuna ripetizione di training set e test set:

- Ripetizione 1:

Match misclassificati RM	DT	KNN	NN
DT	1	0.442	0.424
KNN	0.442	1	0.587
NN	0.424	0.587	1

Match misclassificati YM	DT	KNN	NN
DT	1	0.575	0.564
KNN	0.575	1	0.681
NN	0.564	0.681	1

- Ripetizione 2:

Match misclassificati RM	DT	KNN	NN
DT	1	0.455	0.392
KNN	0.455	1	0.548
NN	0.392	0.548	1

Match misclassificati YM	DT	KNN	NN
DT	1	0.585	0.536
KNN	0.585	1	0.672
NN	0.536	0.672	1

- Ripetizione 3:

Match misclassificati RM	DT	KNN	NN
DT	1	0.443	0.449
KNN	0.443	1	0.597
NN	0.449	0.597	1

Match misclassificati YM	DT	KNN	NN
DT	1	0.581	0.589
KNN	0.581	1	0.696
NN	0.589	0.696	1

- Ripetizione 4:

Match misclassificati RM	DT	KNN	NN
DT	1	0.451	0.444
KNN	0.451	1	0.602
NN	0.444	0.602	1

Match misclassificati YM	DT	KNN	NN
DT	1	0.581	0.597
KNN	0.581	1	0.693
NN	0.597	0.693	1

- Ripetizione 5:

Match misclassificati RM	DT	KNN	NN
DT	1	0.444	0.446
KNN	0.444	1	0.607
NN	0.446	0.607	1

Match misclassificati YM	DT	KNN	NN
DT	1	0.579	0.585
KNN	0.579	1	0.703
NN	0.585	0.703	1

Dall'analisi di queste tabelle si può notare come per la classe RM la quantità di misclassificati comuni tra due classificatori diversi sia tra il 39.2 % e il 60.7 %, mentre più elevata per la classe YM che presenta valori tra il 53.6 % e il 70.3 %. Appare anche che i valori più bassi di ISD si abbiano nel confronto tra il DT e KNN/NN, mentre i valori più alti nel confronto tra KNN e NN, denotando una maggiore somiglianza tra le classificazioni di questi ultimi due.

Lo stesso tipo di calcolo è stato condotto mantenendo sempre lo stesso classificatore, ma andando ad osservare i 5 training set diversi. Infatti, seppure le confusion matrix siano molto simili al variare della ripetizione di estrazione del training set e del test set, è probabile che i misclassificati non siano sempre gli stessi ed occorre verificare in che percentuale ce ne siano in comune.

Match RM DT	1	2	3	4	5
1	1	0.421	0.425	0.416	0.412
2	0.421	1	0.422	0.411	0.425
3	0.425	0.422	1	0.421	0.426
4	0.416	0.411	0.421	1	0.414
5	0.412	0.425	0.426	0.414	1

Match RM KNN	1	2	3	4	5
1	1	0.606	0.571	0.558	0.561
2	0.606	1	0.566	0.584	0.583
3	0.571	0.566	1	0.584	0.561
4	0.558	0.584	0.584	1	0.565
5	0.561	0.583	0.561	0.565	1

Match RM NN	1	2	3	4	5
1	1	0.642	0.682	0.635	0.691
2	0.642	1	0.697	0.645	0.694
3	0.682	0.697	1	0.740	0.770
4	0.635	0.645	0.740	1	0.744
5	0.691	0.694	0.770	0.744	1

Match YM DT	1	2	3	4	5
1	1	0.589	0.587	0.586	0.581
2	0.589	1	0.594	0.590	0.589
3	0.587	0.594	1	0.593	0.589
4	0.586	0.590	0.593	1	0.590
5	0.581	0.589	0.589	0.590	1

Match YM KNN	1	2	3	4	5
1	1	0.698	0.674	0.651	0.663
2	0.698	1	0.675	0.679	0.681
3	0.674	0.675	1	0.677	0.665
4	0.651	0.679	0.677	1	0.667
5	0.663	0.681	0.665	0.667	1

Match YM NN	1	2	3	4	5
1	1	0.743	0.764	0.726	0.770
2	0.743	1	0.782	0.750	0.782
3	0.764	0.782	1	0.805	0.829
4	0.726	0.750	0.805	1	0.806
5	0.770	0.782	0.829	0.806	1

Anche in tal caso si nota che la classe YM presenta generalmente dei valori di ISD più alti rispetto alla RM. L'aspetto interessante è che i misclassificati comuni siano in percentuale maggiore per i classificatori KNN (55 - 69 % circa) e NN (63 - 82 % circa) al variare del training set rispetto al DT (41 - 59 % circa). Ciò conferma la caratteristica del DT di essere un algoritmo di apprendimento ad elevata varianza, ossia a presentare la tendenza a cambiare significativamente al variare anche di poco dei suoi dati di training.

Dal momento che le percentuali di misclassificati comuni al variare dei 3 classificatori e dei 5 training set estratti non sono particolarmente elevate, si è deciso di provare a seguire la strada del voting.
Pertanto, sono state ottenute le classificazioni delle ROI presenti nel construction set stratificato ma non contenute né nel training set né nel test set andando ad effettuare majority voting in tre modi:

- Utilizzando DT, KNN e NN e tutti i training set: 15 classificatori
- Utilizzando DT e KNN e tutti i training set: 10 classificatori
- Utilizzando DT e NN e tutti i training set: 10 classificatori

La scelta di separare KNN e NN nei modi 2 e 3 deriva dal fatto che questi due classificatori presentano valori di ISD maggiori tra loro, per cui insieme potrebbero portare a una polarizzazione verso le loro classificazioni quando inseriti nel voting con il DT.
Di seguito sono riportate le confusion matrix delle classificazioni con voting. Nei modi 2 e 3, laddove il voting forniva un punteggio uguale per le due classi, le ROI sono state considerate come non classificate.

ROI restanti LPBM strat (Modo 1)		Classe reale	
		RM	YM
Classe predetta	RM	46923	107649
	YM	13755	62579

ROI restanti LPBM strat (Modo 2)		Classe reale	
		RM	YM
Classe predetta	RM	37049	85642
	YM	15073	61732
	NC	8556	22854

ROI restanti LPBM strat (Modo 3)		Classe reale	
		RM	YM
Classe predetta	RM	44215	99682
	YM	11332	55345
	NC	5131	15201

Appare evidente come il voting abbia condotto ad un generale miglioramento della sensibilità a discapito però della specificità. Questa non sembra pertanto la via più opportuna da seguire.

5 Feature di ordine superiore

Le analisi condotte sino ad adesso hanno portato risultati insoddisfacenti sulla struttura LPBM nonostante i diversi classificatori testati e i diversi metodi di ottenimento del training set. Pertanto, si è deciso di spostare l'attenzione sulle feature utilizzate.

Come detto nei capitoli precedenti, il dataset attuale comprende feature statistiche del primo ordine e feature statistiche del secondo ordine (features di tessitura). L'interesse è quello di esplorare il campo delle feature statistiche di ordine superiore ed utilizzare quest'ultime per l'individuazione del midollo attivo.
Le principali classi di feature definite come "di ordine superiore" comprendono feature derivanti da:

- Wavelets decomposition
- Laplacian transforms
- Minkowski functionals
- Fractal dimensions

Nel corso di questo lavoro verrà analizzato nel dettaglio l'uso della prima di queste tecniche, ossia la decomposizione wavelet.

5.1 Scelta della trasformata wavelet

Nel campo dell'analisi wavelet, ci sono essenzialmente due tipologie di trasformate che possono essere adottate, la trasformata wavelet continua (CWT, continuous wavelet transform) e la trasformata wavelet discreta (DWT, discrete wavelet transform), ognuna delle quali presenta caratteristiche peculiari proprie che la rendono più idonea per determinate applicazioni piuttosto che per altre. Da una analisi della letteratura [17], risulta come la CWT sia più adeguata all'analisi tempo-frequenza, mentre la DWT trovi maggiore spazio nell'ambito della feature extraction. Essendo proprio la feature extraction il fine che ci si pone, la DWT è stata scelta per le successive elaborazioni e una sua descrizione viene fornita nel paragrafo successivo.

5.2 Trasformata wavelet discreta

Si consideri, per semplicità, il caso di un segnale monodimensionale $f(t)$, la sua trasformata wavelet discreta può essere scritta come:

$$W[f(t)] = W(s,\tau) = \int f(t)\, \psi^*_{s,\tau} dt$$

Essa rappresenta la decomposizione del segnale $f(t)$ tramite una serie di funzioni $\psi_{s,\tau}$ che sono versioni ritardate e scalate (con ritardo τ e fattore di scala s) della cosiddetta wavelet madre ψ e che prendono il nome di wavelet figlie:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\, \psi\left(\frac{t-\tau}{s}\right)$$

Così espresse, le wavelet figlie possono essere ritardate e scalate anche in maniera continua, tuttavia per definizione di trasformata discreta ciò non è possibile, per cui esse vengono sempre ritardate e scalate a passi discreti:

$$s = s_0^j \qquad \tau = k\tau_0 s_0^j$$

dove j e k sono numeri interi e il ritardo è funzione del fattore di scala. In genere, vengono scelti $s_0 = 2$ e $\tau_0 = 1$, cosicché si abbia un campionamento diadico sia sull'asse del tempo che su quello della frequenza (figura 51).

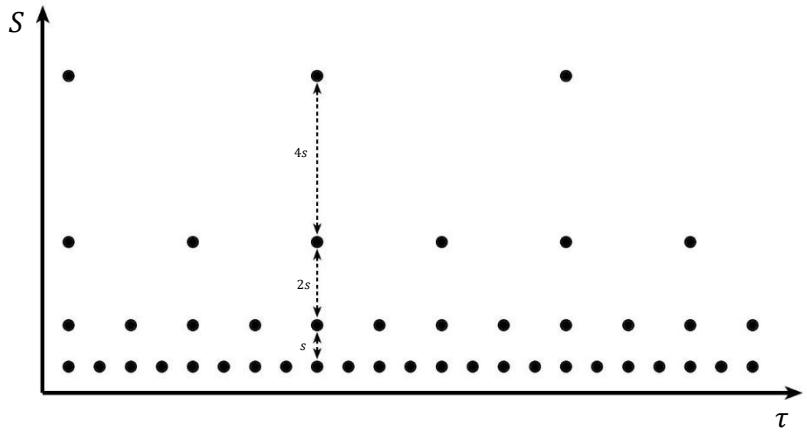


Figura 51 Campionamento diadico sull'asse tempo τ e sull'asse della frequenza s della wavelet figlia

L'espressione finale delle wavelet figlie risulta quindi:

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\,\psi\left(\frac{t - 2^j\,k}{2^j}\right)$$

All'aumentare del fattore di scala si dice che si passa al livello di decomposizione successivo, cioè la wavelet viene dilatata ("allungata" sull'asse tempi) e, come noto dalla teoria di Fourier, la dilatazione nel dominio del tempo si ripercuote in una compressione dello spettro nel dominio della frequenza. In aggiunta, data la dipendenza del ritardo dal fattore di scala, all'aumentare di quest'ultimo si ha anche uno shift di tutte le frequenze verso lo zero. A causa della scelta di s_0, ad ogni livello di decomposizione, lo spettro della wavelet figlia dimezza la sua banda e anche la sua frequenza centrale si dimezza (figura 52). Se si guarda alla wavelet madre da un diverso punto di vista, essa può essere considerata come un filtro passa banda [18] [19], per cui nella teoria dei filtri si direbbe che il suo fattore di merito Q, dato dal rapporto tra la frequenza centrale e la larghezza di banda, è costante.

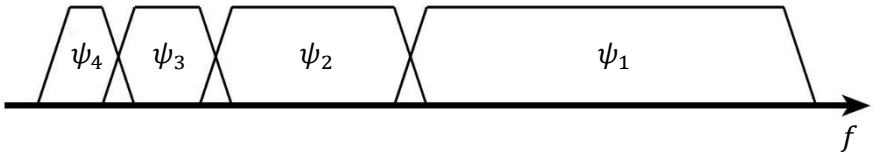


Figura 52 Dimezzamento della banda e della frequenza centrale della wavelet figlia al crescere del livello di decomposizione

Di base, ciò che viene fatto nel calcolo della trasformata wavelet è quindi scegliere un fattore di scala della wavelet, far scorrere la wavelet attraverso il segnale tramite l'aggiunta progressiva di un ritardo ed ogni volta moltiplicare la wavelet per il segnale in analisi. In questo modo, ad ogni passo temporale, cioè ad ogni ritardo scelto, corrisponde un coefficiente dato dal prodotto tra la wavelet e il segnale, il quale esprime "quanto" della wavelet è contenuto nel segnale per quel dato ritardo e quello specifico fattore di scala. In seguito, viene variato il fattore di scala e si ripete il procedimento.

Dal momento che, per ciascun incremento del fattore di scala, la banda della wavelet figlia si dimezza, appare evidente che per coprire l'intero spettro del segnale sarà necessario utilizzare un numero infinito di wavelet figlie. Nel caso della trasformata wavelet discreta questo problema è risolto mediante l'introduzione di una funzione di scaling detta wavelet padre φ, che può essere considerata come un filtro passa basso e che assolve quindi allo scopo di limitare il numero di wavelet necessarie per analizzare l'intero spettro del segnale in esame [19].
Nel complesso, si può dire quindi che una serie di wavelet scalate insieme con una funzione di scaling può essere vista come un banco di filtri. Al primo livello di decomposizione, lo spettro del segnale viene diviso in due parti uguali, una passa alto e una passa basso, dove in realtà la parte passa alto è un filtro passa banda in quando la larghezza di banda del segnale è limitata (spesso, dunque, si parla del filtro passa banda come di un passa alto). Successivamente, cioè al livello seguente, la parte passa basso viene ulteriormente suddivisa in due allo stesso modo della precedente e si procede in questo modo per tutti i livelli di decomposizione scelti. In pratica, risulta che il segnale ha attraversato una serie di filtri passa banda, di cui ognuno ha larghezza di banda pari alla metà del precedente, e un filtro passa basso (figura 53).

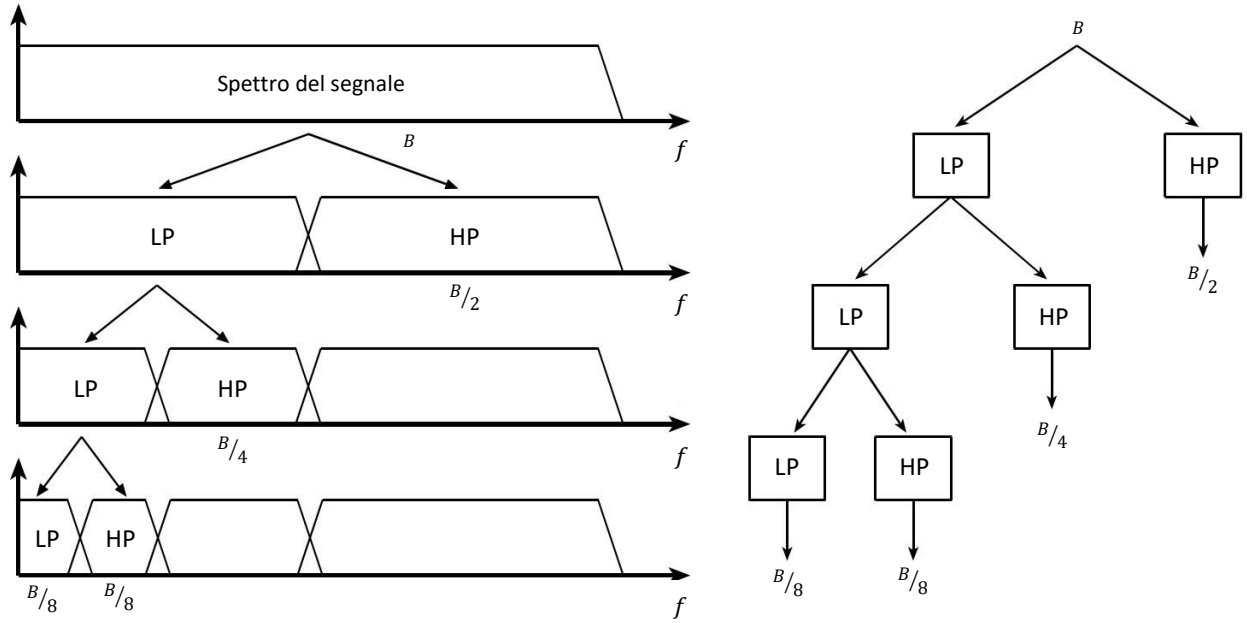


Figura 53 Decomposizione wavelet con banco di filtri: LP indica l'applicazione di un filtro passa basso e HP indica l'applicazione di un filtro passa alto

In questo modo, viene risolto anche un altro problema comune della trasformata wavelet, ossia quello di non possedere una soluzione analitica per la maggior parte delle funzioni, infatti non è necessario esprimere la wavelet in forma esplicita, ma semplicemente come filtro.

5.3 Decomposizione wavelet 2D

Il caso, mostrato nel paragrafo precedente, di trasformata wavelet applicata a un segnale monodimensionale può essere esteso facilmente al caso di un'immagine bidimensionale, si parla quindi di decomposizione wavelet 2D.

I passi fondamentali sono mostrati di seguito secondo quello che prende il nome di schema multirisoluzione di Mallat [20] (figura 54):

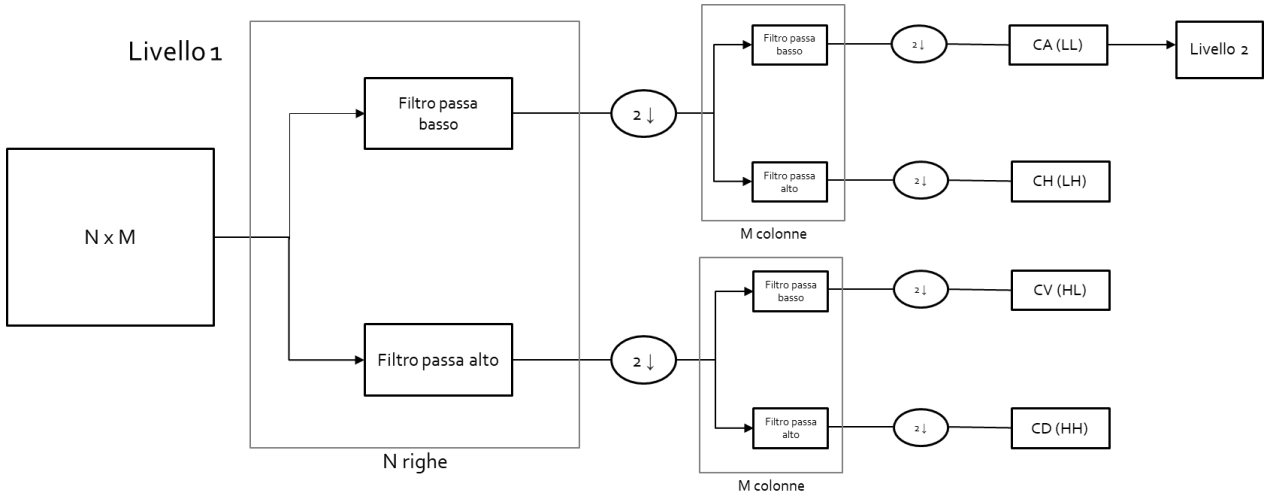


Figura 54 Schema multirisoluzione di Mallat

Essenzialmente l'immagine viene filtrata mediante un filtro passa basso e un filtro passa alto una riga alla volta, sottocampionata di un fattore 2, successivamente filtrata attraverso gli stessi filtri una colonna alla volta e nuovamente sottocampionata di un fattore 2 [21].
L'immagine viene così analizzata a più livelli, in modo da carpire i dettagli delle immagini (qui intesi come variazioni di intensità di toni di grigio tra pixel adiacenti) da un livello più basso (più fine) a uno progressivamente più alto (più grossolano). Il filtro passa basso consente di conservare le basse frequenze e rappresenta quindi un'approssimazione dell'immagine, mentre il filtro passa alto ha la funzione di conservare le alte frequenze e porta dunque con sé l'informazione relativa ai dettagli dell'immagine.

Nel complesso, dalla decomposizione wavelet 2 D si ottengono 4 bande:

- CA: banda dei coefficienti di approssimazione, risulta dall'applicazione di un filtro passa basso prima una riga alla volta e poi una colonna alla volta e per questo è anche detta banda LL (low pass-low pass filtering);
- CH: banda dei coefficienti di dettaglio orizzontali, risulta dall'applicazione di un filtro passa basso una riga alla volta e di uno passa alto una colonna alla volta e per questo è anche detta banda LH (low pass-high pass filtering), mostra le variazioni di intensità di tono di grigio al livello calcolato in senso orizzontale;
- CV: banda dei coefficienti di dettaglio verticali, risulta dall'applicazione di un filtro passa alto una riga alla volta e di uno passa basso una colonna alla volta e per questo è anche detta banda HL (high pass-low pass filtering), mostra le variazioni di intensità di tono di grigio al livello calcolato in senso verticale;
- CD: banda dei coefficienti di dettaglio diagonali, risulta dall'applicazione di un filtro passa alto una riga alla volta e di uno passa alto una colonna alla volta e per questo è anche detta banda HH (high pass-high pass filtering), mostra le variazioni di intensità di tono di grigio al livello calcolato in direzione diagonale.

Un esempio di decomposizione wavelet 2 D al livello 2 è mostrato di seguito insieme con una schematizzazione dello stesso a partire dall'immagine con le dimensioni di partenza (figura 55):

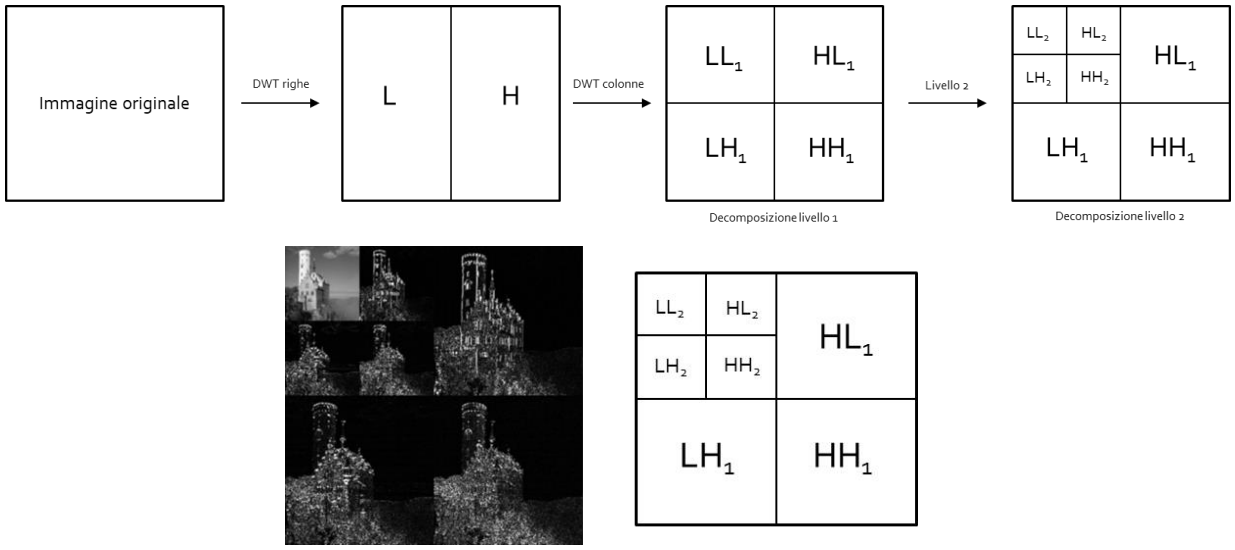


Figura 55 Schematizzazione della decomposizione wavelet 2 D al livello 2 ed esempio di applicazione a uno scatto del Castello di Lichtenstein

A partire dalle bande precedentemente descritte è possibile calcolare una serie di feature come quelle statistiche di primo e/o secondo ordine. In pratica, invece di utilizzare la matrice originale si calcolano le feature dopo aver effettuato la decomposizione.

5.4 Feature extraction

Allo stesso modo di quanto fatto per la texture analysis eseguita sulle immagini di partenza, anche in questo caso le immagini dopo decomposizione wavelet sono state suddivise in ROI ottagonali di 5x5 pixel, le quali sono state fatte traslare di un pixel alla volta sia in direzione orizzontale che verticale. Applicando la decomposizione wavelet 2 D come descritto nel paragrafo precedente, l'immagine presenta le sue dimensioni dimezzate progressivamente all'aumentare del livello analizzato, pertanto, per mantenere le dimensioni originali dopo ogni livello di decomposizione, si è deciso di adottare una variante della decomposizione wavelet denominata decomposizione wavelet non decimata (o stazionaria). In essa invece di sottocampionare l'immagine di un fattore 2 dopo ciascun livello di decomposizione per poi applicare gli stessi filtri adottati in precedenza, ciò che viene fatto è sovracampionare i filtri di un fattore 2 dopo ogni loro utilizzo, mentre l'immagine conserva le dimensioni originarie [21].
Per quanto riguarda il processo di feature extraction in seguito a decomposizione wavelet 2 D, è necessario trattare 3 tematiche principali:

- Scelta della tipologia di wavelet madre da impiegare;
- Scelta del numero di livelli di decomposizione da utilizzare;
- Scelta delle feature da estrarre da una o più bande di decomposizione.

Si è quindi svolta un'analisi della letteratura dalla quale è emerso che le tipologie di wavelet più largamente impiegate in problemi di radiomica come quello qui in analisi sono le wavelet ortogonali Haar, Daubechies 4 e Coiflet 1 [22] [23], mentre il numero di livelli di decomposizione più opportuno in problemi di classificazione su immagini TC appare essere pari a 5 [24]. Lo studio svolto evidenzia infatti come al sesto livello di decomposizione l'immagine perda la maggior parte dei suoi dettagli e discute, sulla base di risultati sperimentali, come l'accuratezza di classificazione sia inferiore al livello 4 rispetto che al 5 e come essa decresca a partire dal livello 6.

Per scegliere al meglio la tipologia di wavelet madre, si è deciso di effettuare anzitutto un'analisi visiva degli effetti di ciascuna wavelet su di una slice (-201.5) relativa al paziente 1, andando a testare le wavelet Haar, Daubechies 4 e Coiflet 1 su 5 livelli di decomposizione.

- Haar banda LL (coefficienti di approssimazione, figura 56):

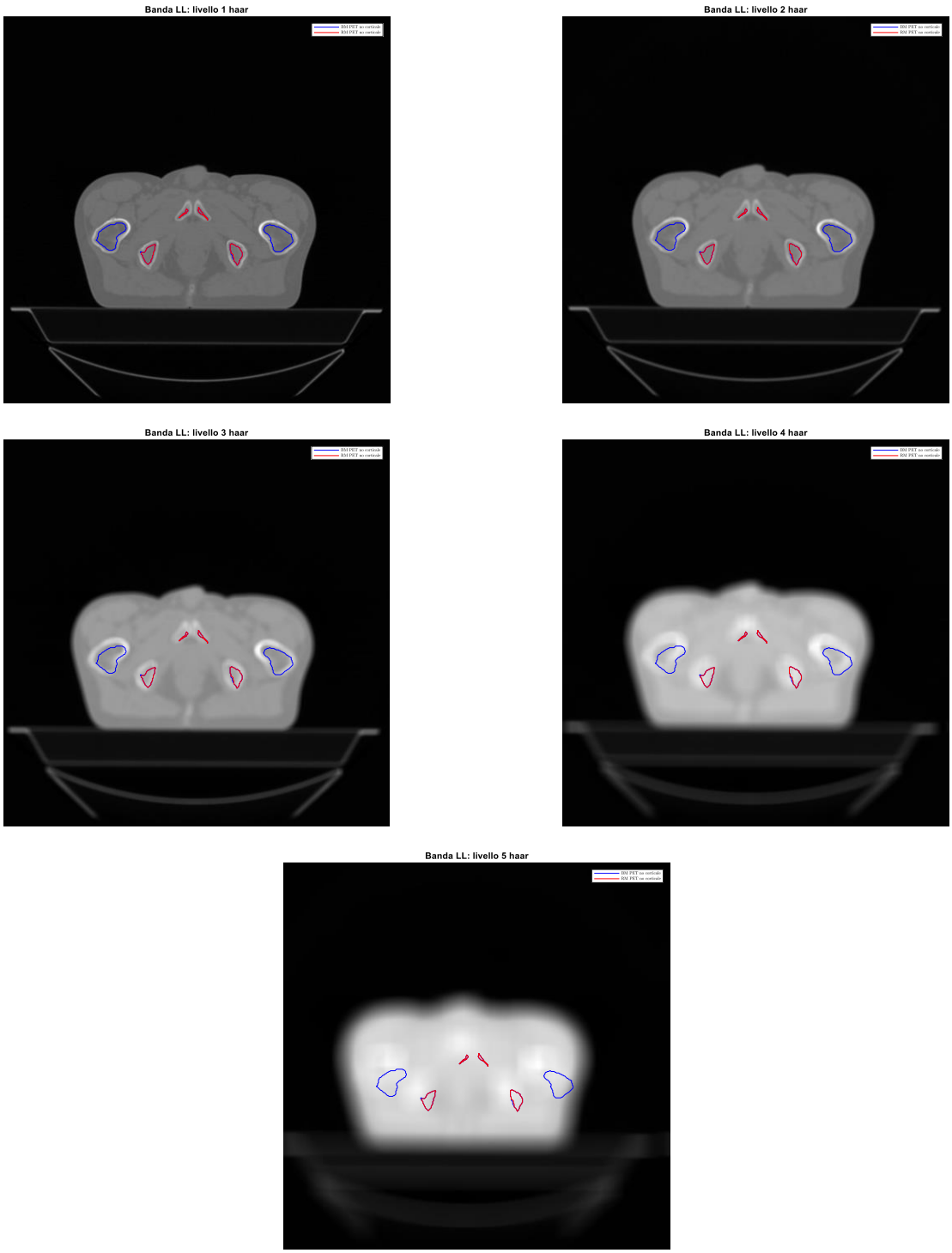


Figura 56 Bande LL dei livelli di decomposizione 1-5 con applicazione della wavelet madre Haar sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

- Haar banda HL (coefficienti di dettaglio verticali, figura 57):

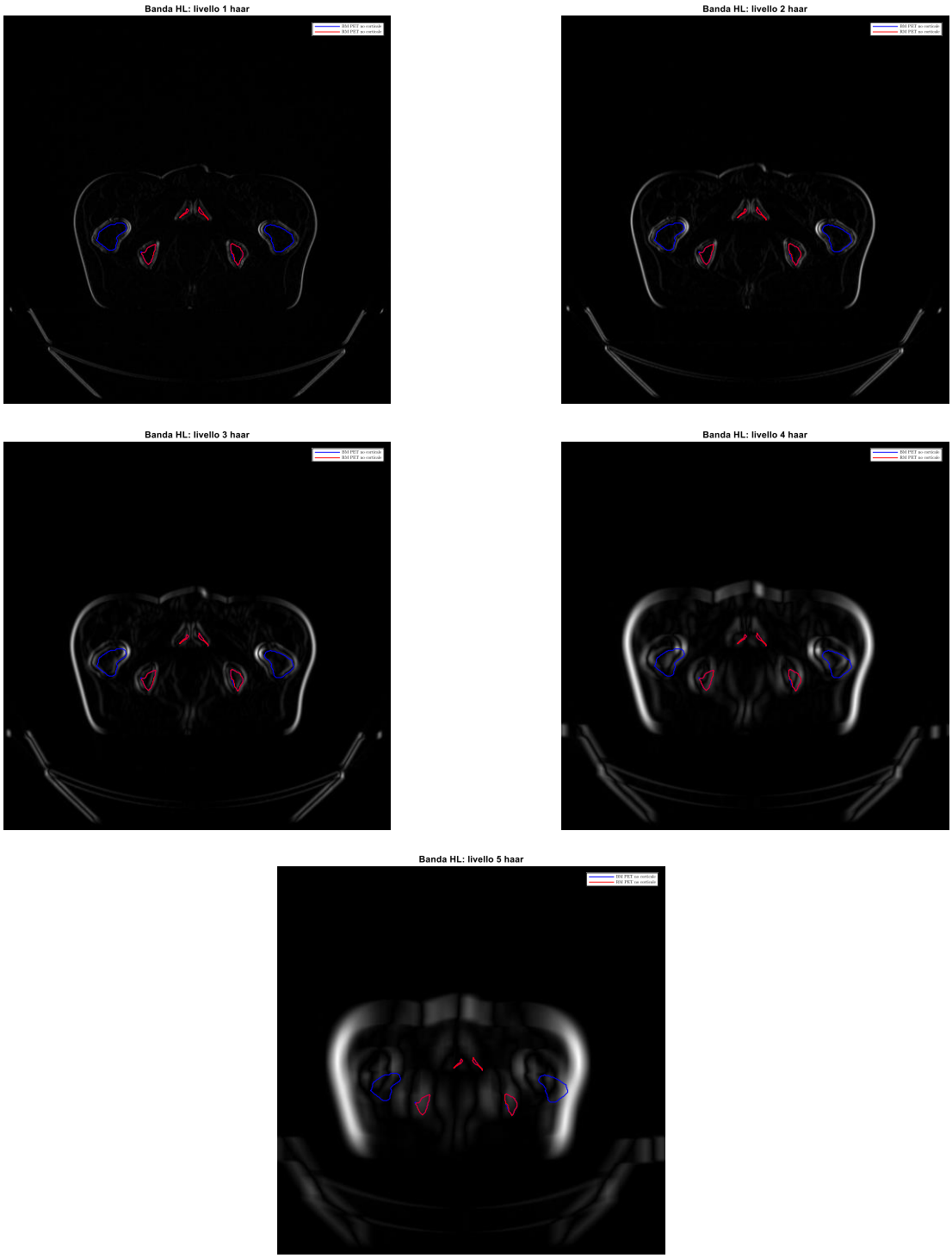


Figura 57 Bande HL dei livelli di decomposizione 1-5 con applicazione della wavelet madre Haar sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

- Haar banda LH (coefficienti di dettaglio orizzontali, figura 58):

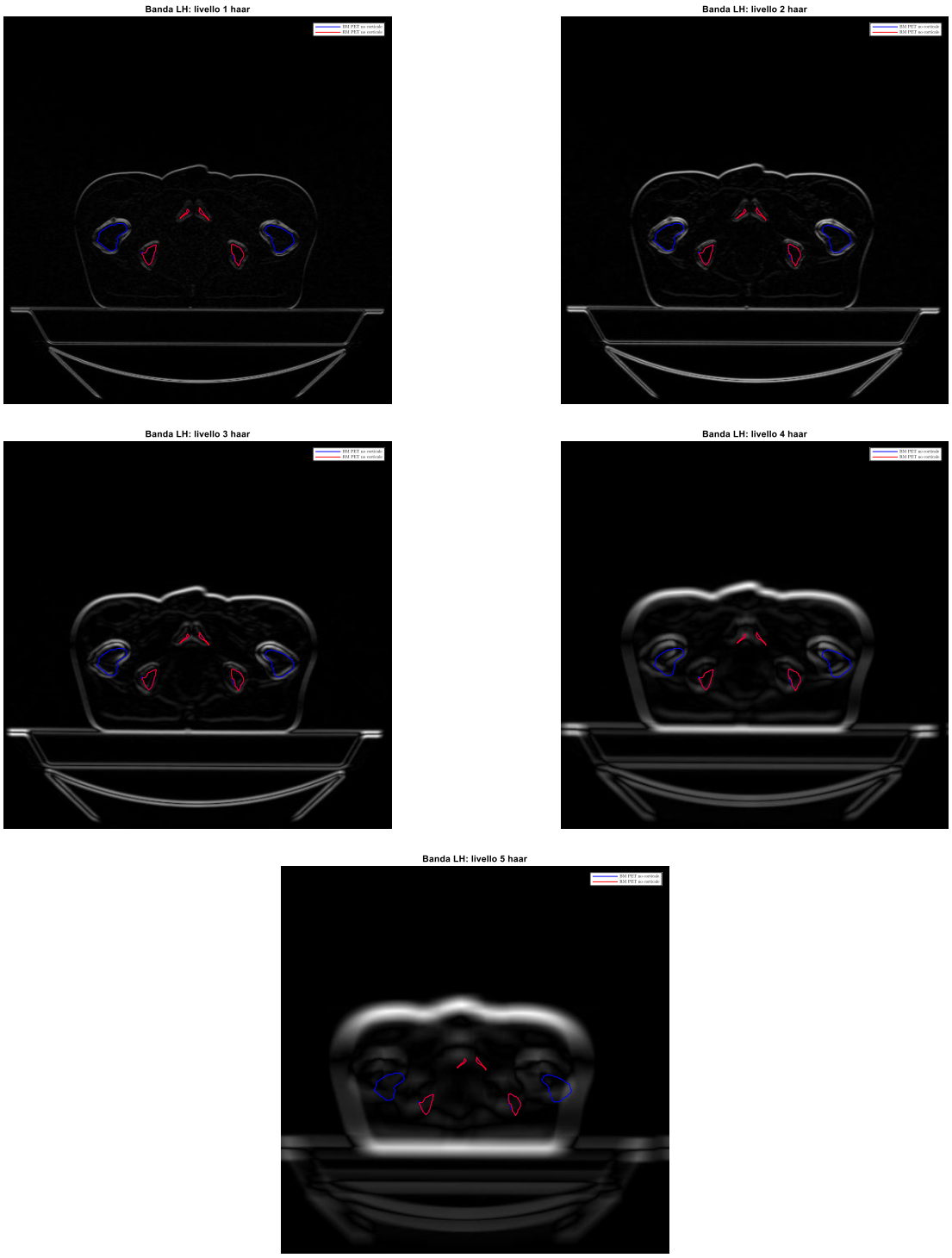


Figura 58 Bande LH dei livelli di decomposizione 1-5 con applicazione della wavelet madre Haar sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

- Haar banda HH (coefficienti di dettaglio diagonali, figura 59):

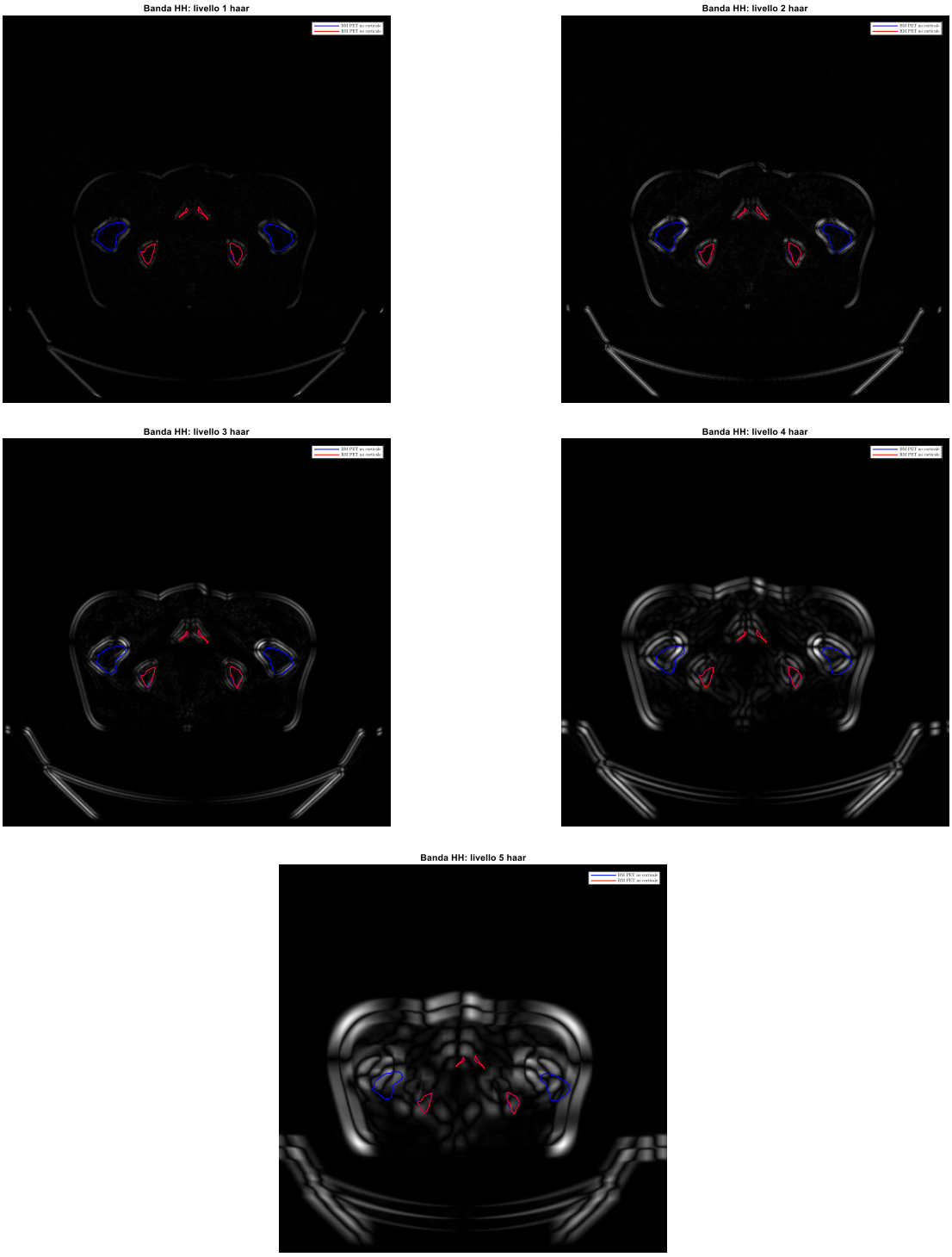


Figura 59 Bande HH dei livelli di decomposizione 1-5 con applicazione della wavelet madre Haar sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

Come si può vedere, le bande dei coefficienti di dettaglio mettono in evidenza solo i contorni dell'immagine su diversi livelli a seconda del livello di decomposizione utilizzato, pertanto non sono utili nel risolvere il problema di classificazione delle singole ROI. Vengono quindi omesse per le successive tipologie di wavelet madre.

- Daubechies 4 banda LL (coefficienti di approssimazione, figura 60):

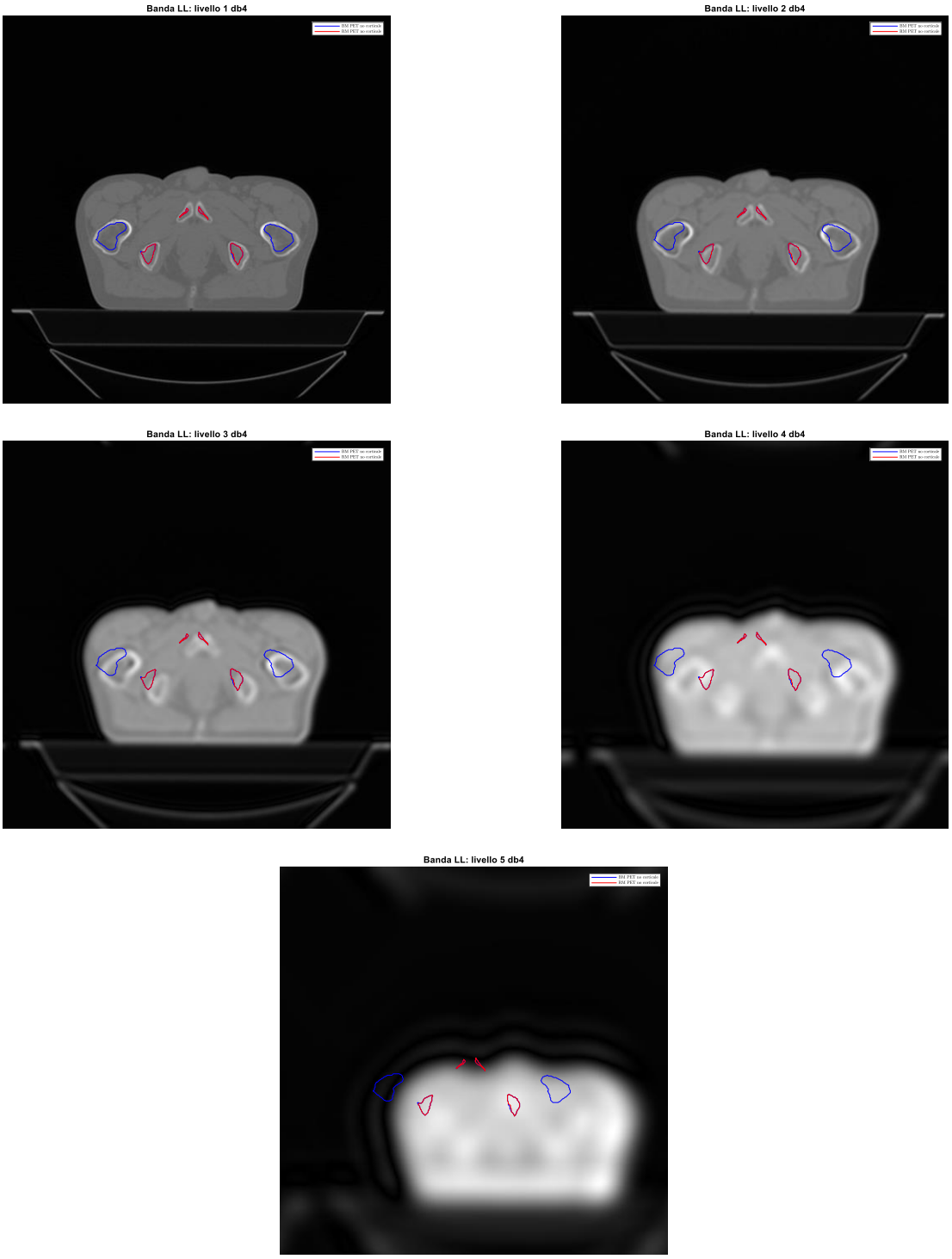


Figura 60 Bande LL dei livelli di decomposizione 1-5 con applicazione della wavelet madre Daubechies 4 sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

- Coiflet 1 banda LL (coefficienti di approssimazione, figura 61):

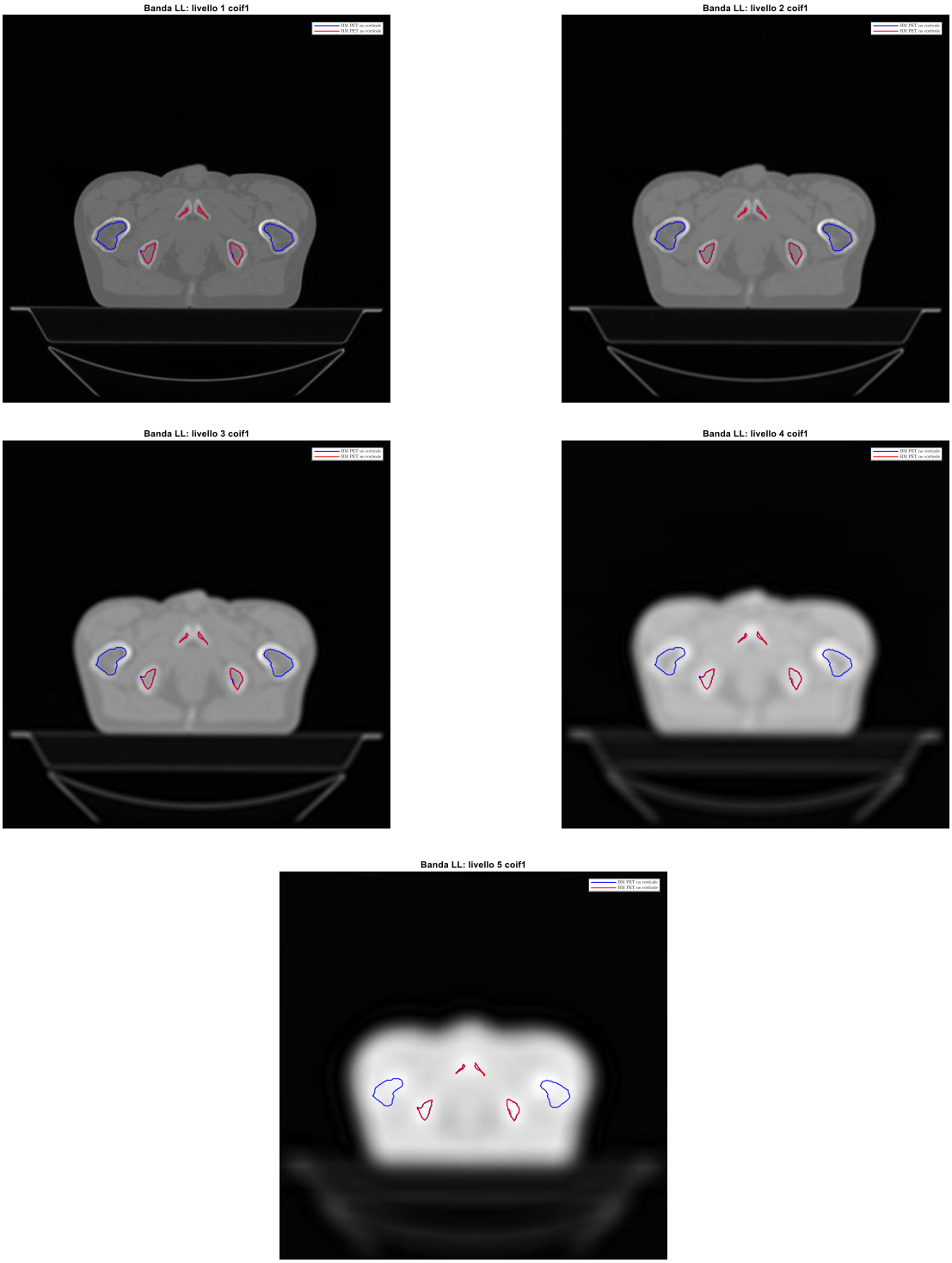


Figura 61 Bande LL dei livelli di decomposizione 1-5 con applicazione della wavelet madre Coiflet 1 sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

In generale si nota che per ogni banda si ha uno shift della struttura tanto maggiore quanto più è alto il livello di decomposizione. Tale shift è maggiore nel caso delle wavelet madri Haar e Daubechies 4 rispetto alla Coiflet 1.

La causa di questo shift della struttura è da ricercare nelle proprietà delle wavelet madri utilizzate. Difatti, è possibile vedere come tali funzioni siano asimmetriche (figura 62). In particolare, la wavelet Haar è antisimmetrica, mentre la wavelet Coiflet 1 è quasi simmetrica, motivo per cui il suo shift sulla struttura è notevolmente molto meno accentuato.

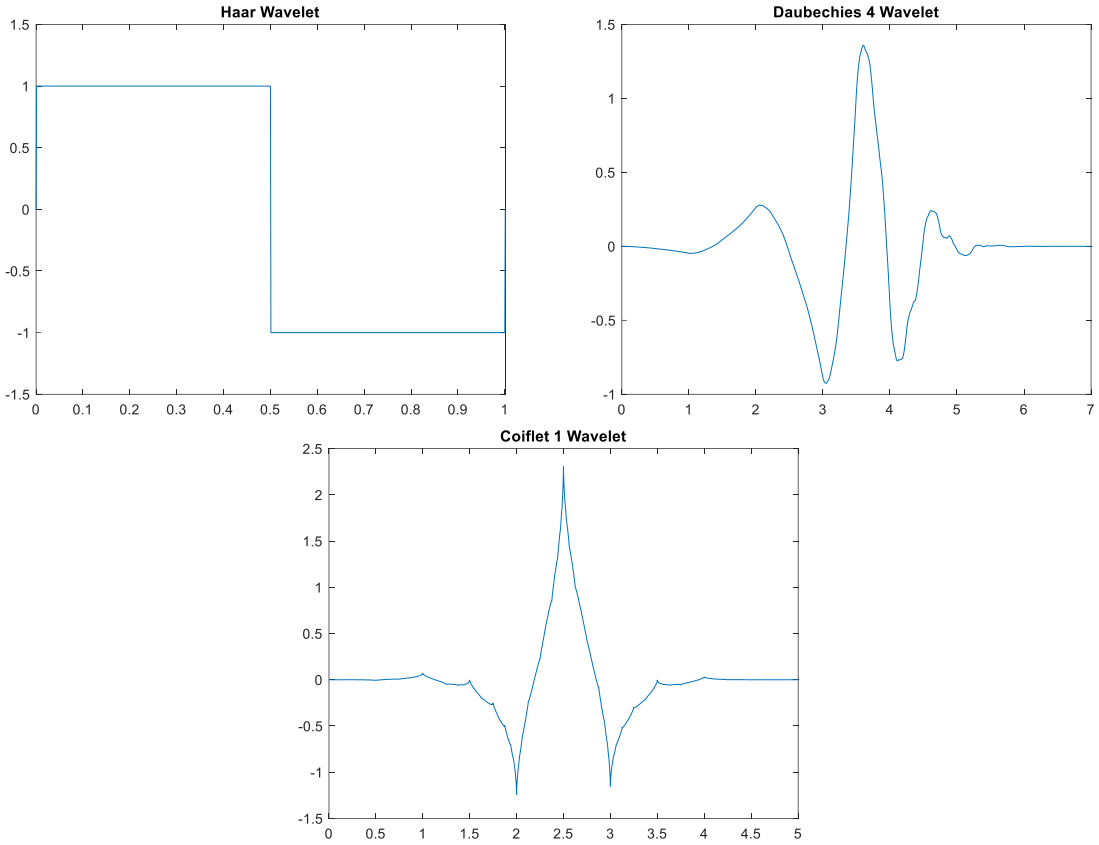


Figura 62 Wavelet madri Haar, Daubechies 4 e Coiflet 1: si nota l'asimmetria delle funzioni.

Tra le wavelet disponibili in MATLAB per il calcolo della DWT si è dunque deciso di optare per l'unica che presentasse la proprietà di simmetria, ossia quella di Meyer. In realtà la wavelet di Meyer non è a supporto compatto (cioè non è diversa da zero in un insieme chiuso e limitato di punti) e pertanto non è utilizzabile per il calcolo della DWT (e quindi anche della SWT) in MATLAB. Tuttavia, esiste una sua buona approssimazione nota come Discrete Meyer che può essere utilizzata (figura 63).

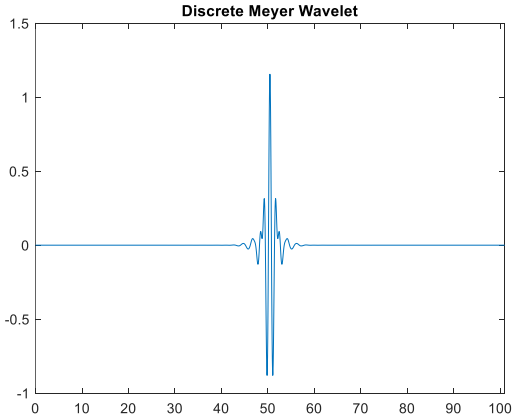


Figura 63 Wavelet madre Meyer discreta: si nota la simmetria della funzione.

È possibile, quindi, vedere di seguito come l'uso della wavelet Discrete Meyer non porti ad alcuno shift della struttura analizzata a nessun livello di decomposizione (le immagini sono relative ai coefficienti di approssimazione, figura 64):

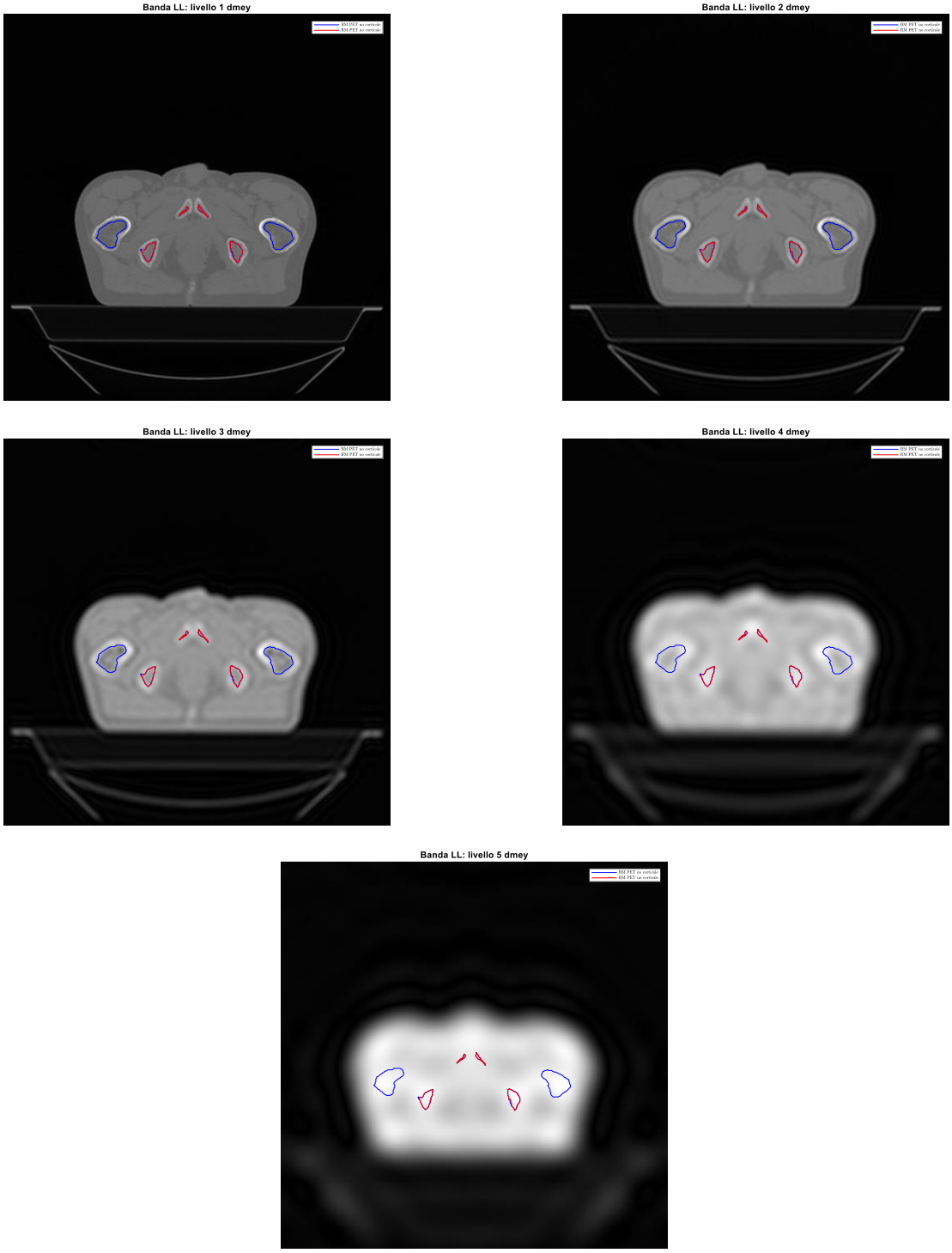


Figura 64 Bande LL dei livelli di decomposizione 1-5 con applicazione della wavelet madre di Meyer discreta sulla slice -201.5 del soggetto 1. In rosso i contorni del RM, in blu quelli del BM.

Riassumendo si è scelto di impiegare la wavelet madre Discrete Meyer con 5 livelli di decomposizione e di estrarre le feature a partire dalle bande LL risultanti da ciascun livello.

Restano da stabilire con esattezza quali feature estrarre per ciascuna ROI individuata. Nel fare ciò, si è preso spunto da alcuni lavori già presenti in letteratura [23] [25] [26] riguardanti problemi di radiomica su immagini TC. Alla fine, si è deciso di calcolare unicamente 14 feature statistiche del primo ordine (media, deviazione standard, varianza, skewness, kurtosis, energia, entropia, mediana, massimo, minimo, range, deviazione media assoluta, valor quadratico medio e uniformità) in modo da avere nel complesso 70 feature per ROI. Ciò che si ottiene è una matrice di dimensione n x 75 (con n numero di ROI) dove le prime due colonne rappresentano, rispettivamente, l'ID del paziente e il numero della slice da cui proviene la ROI per la quale sono state calcolate le feature, la terza e la quarta colonna contengono le coordinate x e y del pixel centrale della ROI, si trovano poi i valori delle 70 feature e infine, nell'ultima colonna, è presente un valore che indica l'appartenenza di quella ROI a RM, YM o al bordo tra i due (0 = YM, 1 = RM, 2 = bordo).
Di seguito è possibile vedere un esempio della matrice ricavata:

				Livello 1			**Livello 2**			**…**	**Livello 5**			
ID paziente	Slice	X	Y	Feature 1	…	Feature 14	Feature 15	…	Feature 28	…	Feature 57	…	Feature 70	Classe
1	-141.5	166	299	3.52e+03	…	0.0522	7.10e+03	…	0.0476	…	5.95e+04	…	0.0476	0
1	-147.5	189	259	3.71e+03	…	0.0567	7.50e+03	…	0.0476	…	5.94e+04	…	0.0476	1
1	-150.5	185	260	3.74e+03	…	0.0476	7.62e+03	…	0.0476	…	5.99e+04	…	0.0476	2
2	-162	171	268	3.53e+03	…	0.0476	7.10e+03	…	0.0476	…	5.80e+04	…	0.0476	0
3	-199	166	210	4.08e+03	…	0.0476	8.16e+03	…	0.0476	…	6.50e+04	…	0.0522	1

5.5 Construction set e analisi separabilità

A partire dai nuovi dati ottenuti è stato estratto un construction set per la struttura LPBM andando a selezionare 5 slice per ciascuno dei primi 40 pazienti in maniera stratificata, così come è stato fatto nel paragrafo 4.1.
Anche in questo caso il metodo di normalizzazione adottato è stato quello del min-max scaling, tuttavia l'elevata presenza di outlier portava a dei boxplot molto ristretti per alcune variabili (figura 65).

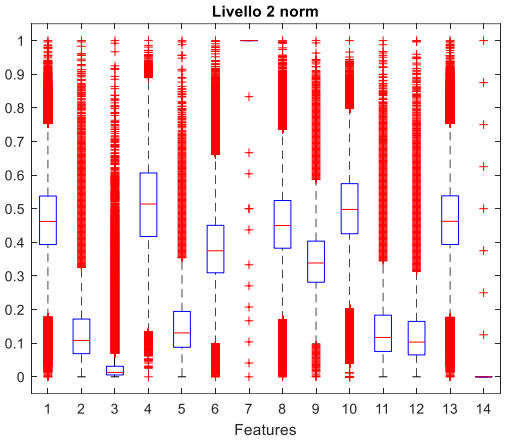


Figura 65 Boxplot delle feature statistiche del 1° ordine calcolate per il 2° livello di decomposizione dopo normalizzazione con min-max scaling: si nota la presenza di boxplot molto ristretti come per il caso della terza feature.

Al fine di evitare di dare un peso minore a variabili che presentassero boxplot meno ampi e possibili problemi di instabilità numerica dovuta a valori troppo piccoli, si è scelto di utilizzare come minimo

il valore corrispondente al 1° percentile e come massimo quello corrispondente al 99° percentile del boxplot di ciascuna variabile non normalizzata ed effettuare così il min-max scaling. Il risultato è quello di ridurre l'effetto degli outlier ed avere boxplot maggiormente ampi, mantenendo dei range non troppo diversi tra le variabili (figura 66).

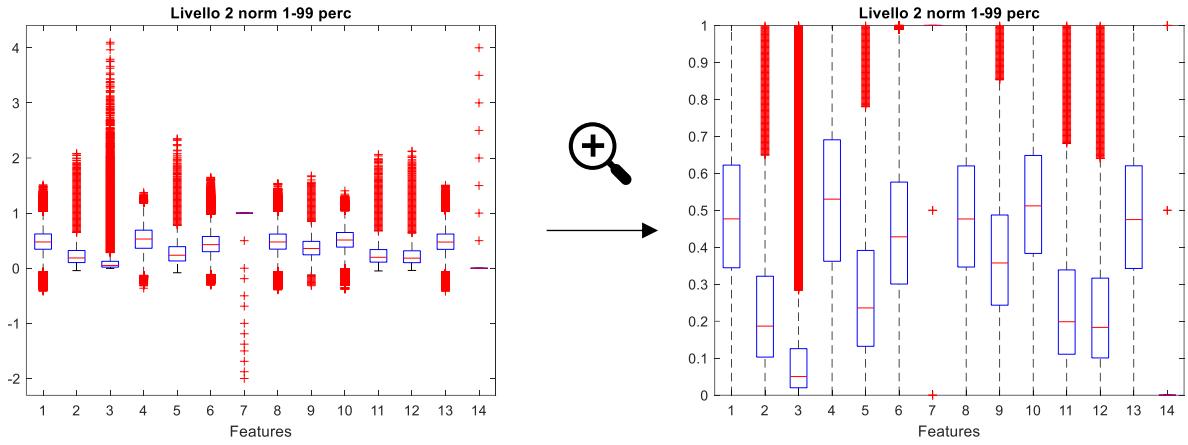


Figura 66 Boxplot delle feature statistiche del 1° ordine calcolate per il 2° livello di decomposizione dopo normalizzazione con min-max scaling e uso del 1° e del 99° percentile come valori di minimo e massimo rispettivamente. A sinistra i boxplot per intero che mostrano range simili tra le variabili, a destra uno zoom degli stessi nell'intervallo 0-1 che mostra l'effetto di distensione dei boxplot rispetto alla normalizzazione precedente.

La prima analisi condotta è stata quella relativa alla separabilità dei dati contenuti nel construction set.

Si è proceduto in due modi:

- PCA (Principal Component Analysis): utilizzando le prime due componenti principali, ossia quelle due combinazioni lineari delle variabili originarie, che permettono di conservare la maggior parte della varianza dei dati in esame (figura 67);

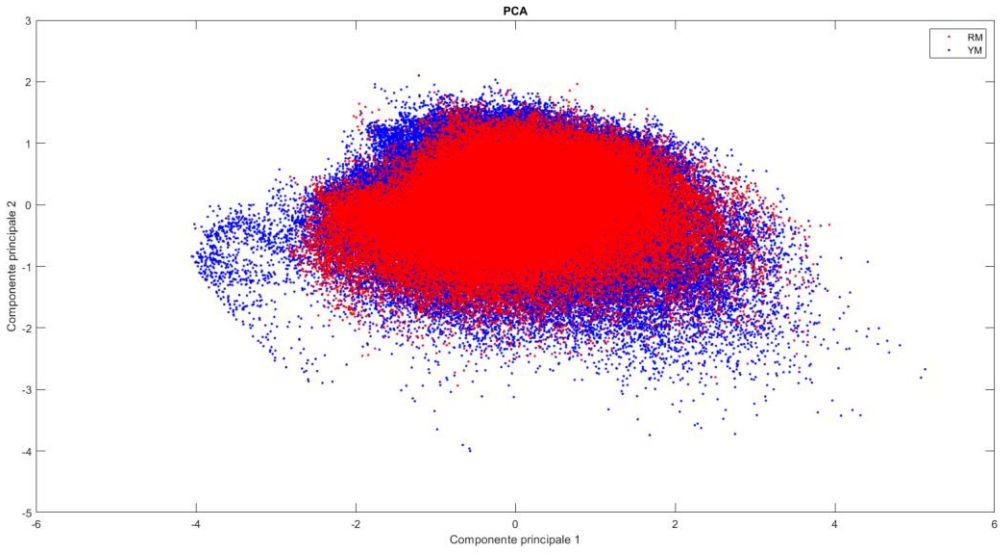


Figura 67 Applicazione della PCA ai dati del construction set. In rosso le osservazioni corrispondenti a RM, in blu quelle corrispondenti all'YM.

- t-SNE (t-distributed Stochastic Neighbor Embedding): riorganizzando i punti in modo che oggetti vicini nello spazio originale risultino vicini anche in uno spazio a 2 dimensioni, mentre oggetti lontani risultino lontani, cercando di preservare la struttura locale dei dati (figura 68).

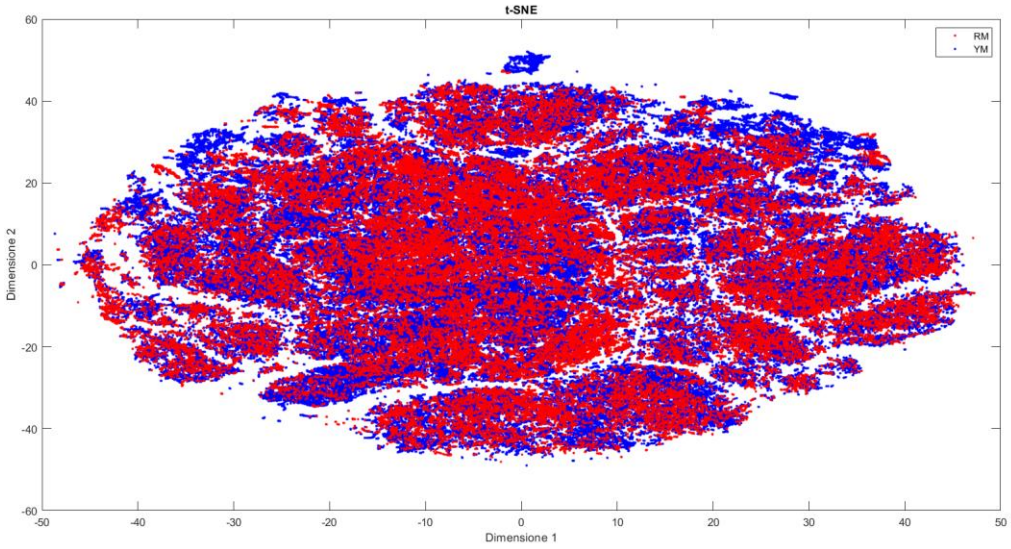


Figura 68 Applicazione del t-SNE ai dati del construction set. In rosso le osservazioni corrispondenti a RM, in blu quelle corrispondenti all'YM.

In entrambi i casi appare chiaro come la separabilità dei dati nelle due classi risulti difficile nello spazio a 2 dimensioni. Tuttavia, il fatto che non ci sia separabilità in questo spazio a dimensionalità ridotta non comporta che non possa sussistere separabilità in uno spazio a dimensione maggiore. Ciò che al più sarebbe vero è infatti il contrario.

5.6 Training set, test set e classificazione

Si è quindi proceduto all'ottenimento di un nuovo training set per la struttura LPBM a partire dal construction set stratificato.
Più nel dettaglio, per ciascun paziente sono state estratte 3 ROI da ogni cluster ottenuto tagliando i dendrogrammi di ognuna delle sue 5 slice presenti nel construction set con il taglio in base alla variabilità e laddove un cluster contasse meno di 3 ROI sono state selezionate tutte quelle presenti al suo interno. Questa operazione è stata svolta separatamente sia per la classe RM che per la classe YM. A questo punto, con le ROI così individuate sono stati realizzati due ulteriori dendrogrammi, uno per ogni classe, che sono stati tagliati ancora una volta con il metodo di taglio in base alla variabilità (figura 69).

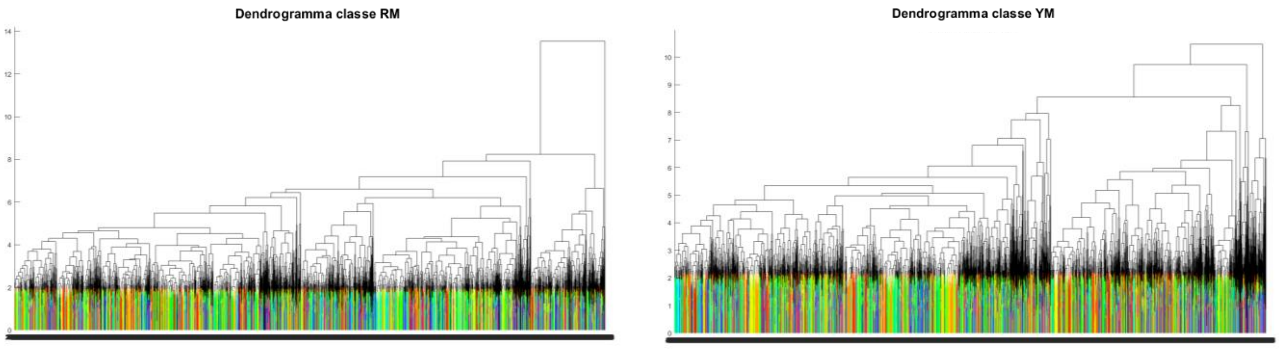


Figura 69 A sinistra dendrogramma delle ROI di RM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster. A destra dendrogramma delle ROI di YM selezionate con indicazione dei cluster formatisi con il taglio sulla base della variabilità intra-cluster.

Da ciascun cluster ottenuto, sono state estratte in maniera proporzionale alla sua numerosità un certo numero di ROI, in modo da costruire un training set che presentasse circa 5000 ROI di ognuna delle due classi. Gli elementi non inseriti nel training set sono andati a costituire il test set. Il metodo impiegato è dunque lo stesso applicato in precedenza con i dati delle feature di primo e secondo ordine nel paragrafo 4.3.

Con questo nuovo training set è stato allenato un DT andando ad utilizzare tutte le feature presenti. Il classificatore è stato poi testato separatamente sul training set, sul test set (ossia le ROI non contenute nel training set ma selezionate nella fase di estrazione di 3 ROI per cluster) e sulle ROI restanti del construction set di partenza.

DT feature di 1° e 2° ordine

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	91.65 %	8.22 %
	YM	8.35 %	91.78 %

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	52.70 %	48.39 %
	YM	47.30 %	51.61 %

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	53.92 %	49.69 %
	YM	46.08 %	50.31 %

DT feature wavelet

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	94.37 %	4.69 %
	YM	5.63 %	95.31 %

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	64.42 %	35.23 %
	YM	35.58 %	64.72 %

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	59.63 %	38.29 %
	YM	40.37 %	61.71 %

Se si confrontano i risultati ottenuti a partire dal dataset precedente con quelli derivanti dalle feature estratte con decomposizione wavelet, si può notare un miglioramento di sensibilità e specificità sui dati non di allenamento con un minimo del 5.7 % per la sensibilità e del 11.4 % per la specificità. Sembra quindi che la decomposizione wavelet porti a un miglioramento delle performance di classificazione nel problema di individuazione del midollo attivo.

Da questi risultati appare esserci un certo overfitting del DT, quindi si è deciso di costruire una random forest in modo da cercare di alleviare questo comportamento.

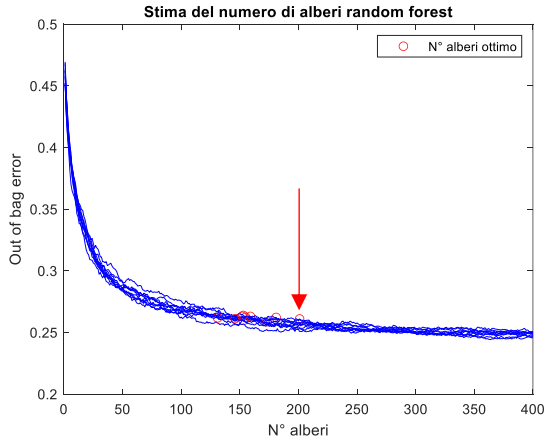


Figura 70 Out of bag error per la stima del numero di alberi ottimale della random forest.

Anzitutto è necessario ottimizzare il numero di alberi da utilizzare, quindi è stato ripetuto per 10 volte l'allenamento di una random forest con un numero molto elevato di alberi (400 alberi) e si è

calcolato l'errore degli out of bag, cioè l'errore su quelle osservazioni che vengono escluse dall'allenamento di ognuno degli alberi della random forest separatamente dagli altri (figura 70).
Per ciascuna ripetizione è stato selezionato un numero di alberi ottimo andando a individuare il primo numero di alberi il cui out of bag error non fosse superiore al 5% di quello della random forest con 400 alberi.
Alla fine, si è scelto il valore maggiore (201 alberi) tra quelli risultati da ogni ripetizione, perché rappresenta il numero di alberi che assicura più probabilmente il compromesso tra errore e numero di alberi a prescindere dalla ripetizione.
Dunque, si è allenata la RF con 201 alberi ed è stata verificata sul training set, sul test set e su quelle ROI del construction set non incluse né nel training set né nel test set.

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	100 %	0 %
	YM	0 %	100 %

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	76.33 %	23.02 %
	YM	23.67 %	76.98 %

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	65.09 %	26.47 %
	YM	34.91 %	73.53 %

I risultati sono effettivamente promettenti, infatti risulta evidente un miglioramento di sensibilità e specificità sui dati non di allenamento di un minimo dell'11.17 % per la sensibilità e del 23.22 % per la specificità.
A questo punto si è proceduto con una feature selection, in modo da vedere se la rimozione di alcune variabili possibilmente rumorose potesse portare ad un ulteriore miglioramento delle performance. Per fare ciò si è deciso di utilizzare la RF stessa andando ad osservare l'importanza attribuita a ciascuna feature.
Per imporre una soglia sull'importanza minima da conservare sono state ordinate le feature in ordine decrescente di importanza e ne è stata calcolata la cumulata normalizzata tra 0 e 1.
La soglia di importanza è stata scelta come la media tra l'importanza della feature che permetteva di conservare il 95 % dell'importanza complessiva e quella ad importanza immediatamente inferiore (figura 71).

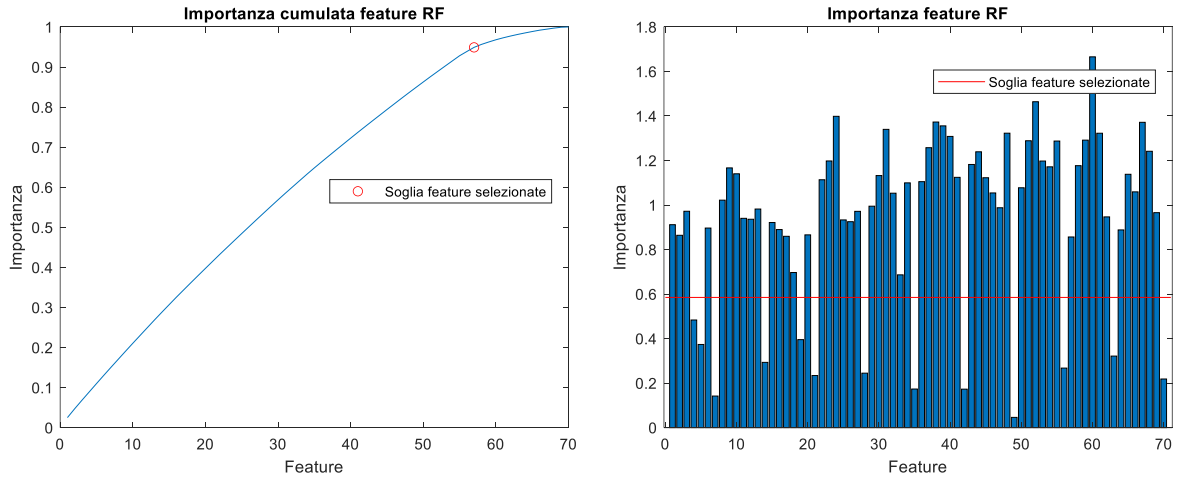


Figura 71 A sinistra importanza cumulata delle features normalizzata tra 0 e 1, a destra importanza delle feature e soglia di importanza.

In questo modo delle 70 feature totali ne vengono escluse 13, le quali è stato notato essere quelle che mostrano sempre importanza inferiore a prescindere dalla ripetizione della random forest.

Con le feature così selezionate è stata allenata nuovamente una random forest con 201 alberi, la quale è stata verificata sui tre set di dati presenti, come avvenuto in precedenza.

Training set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	100 %	0 %
	YM	0 %	100 %

Test set LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	76.66 %	22.37 %
	YM	23.34 %	77.63 %

ROI restanti LPBM strat		Classe reale	
		RM	YM
Classe predetta	RM	65.27 %	26.01 %
	YM	34.73 %	73.99 %

Il risultato della feature selection è stato un leggero miglioramento di qualche decimo di punto percentuale sia sulla sensibilità che sulla specificità, che, sebbene non sia particolarmente notevole, ha portato a una riduzione della complessità dell'algoritmo senza peggiorarne affatto le performance.

5.7 Valutazione performance maschere

A partire dalla RF realizzata e dalla feature selection effettuata, si è successivamente proceduto con l'ottenimento delle maschere di segmentazione dell'RM seguendo esattamente le stesse operazioni di post processing impiegate con le feature di 1° e 2° ordine.
Di seguito sono mostrati i risultati in termini di dice calcolati su queste nuove maschere e sono messi a confronto con quelli derivanti dalle feature precedenti (ottenuti mediante un classificatore DT e feature selection tramite GA, figura 72).

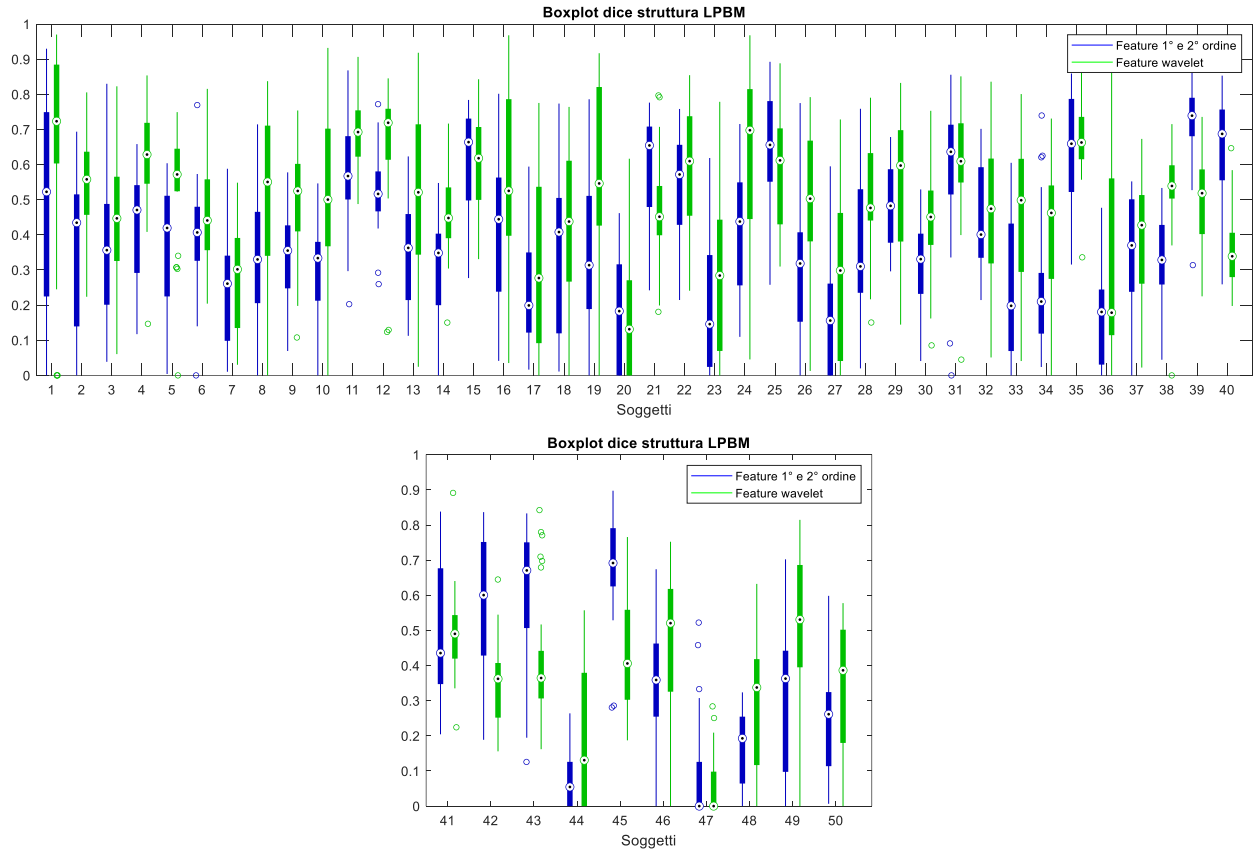


Figura 72 Boxplot della dice calcolata sulle maschere di segmentazione della struttura LPBM, in blu i boxplot relativi all'utilizzo delle feature di 1° e 2° ordine, in verde i boxplot relativi all'uso delle feature derivanti dalla decomposizione wavelet. In alto sono mostrati i risultati relativi ai pazienti inseriti nel construction set, in basso quelli relativi ai pazienti inseriti nel validation set.

In generale si apprezza un aumento delle performance di classificazione andando ad utilizzare le feature derivanti dalla decomposizione wavelet. Ciò è vero sia per i pazienti del construction set che per quelli del validation set.
Tuttavia, per alcuni soggetti, le performance peggiorano, dunque si è deciso di analizzare più nel dettaglio le segmentazioni PET di tutti i pazienti per comprendere le principali differenze tra i pazienti con miglioramenti e quelli con peggioramenti.
Nel fare ciò è emerso che in circa la metà dei pazienti la segmentazione del BM da PET presenta errori di classificazione sulla struttura LPBM in una o più slice per cui esse dovrebbero essere soggette ad una revisione ed escluse da quelle utili per la valutazione delle performance (figura 73).

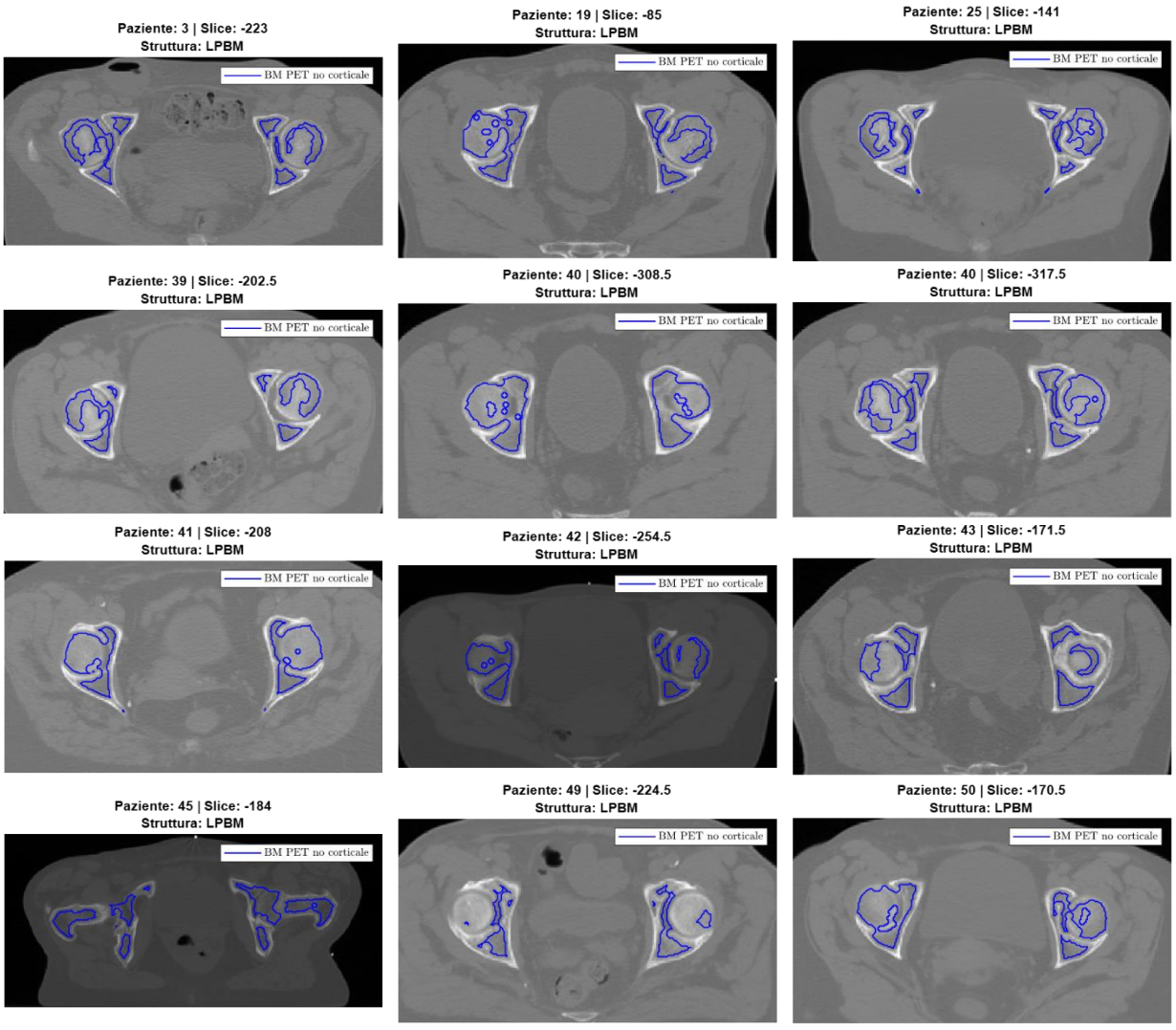


Figura 73 Esempi di segmentazioni errate del BM da PET per la struttura LPBM nel caso di diversi soggetti del construction set e del validation set.

Analizzando la segmentazione del RM ricavata dalla PET per i pazienti (15, 20, 21, 25, 31, 39, 40, 42, 43 e 45) che mostrano un calo di performance con feature derivanti dalla decomposizione wavelet, si è notato come questi siano caratterizzati dal possedere una elevata quantità di RM nella zona dell'acetabolo e del femore prossimale rispetto a quella presente nel pube e nell'ischio (figura 74).

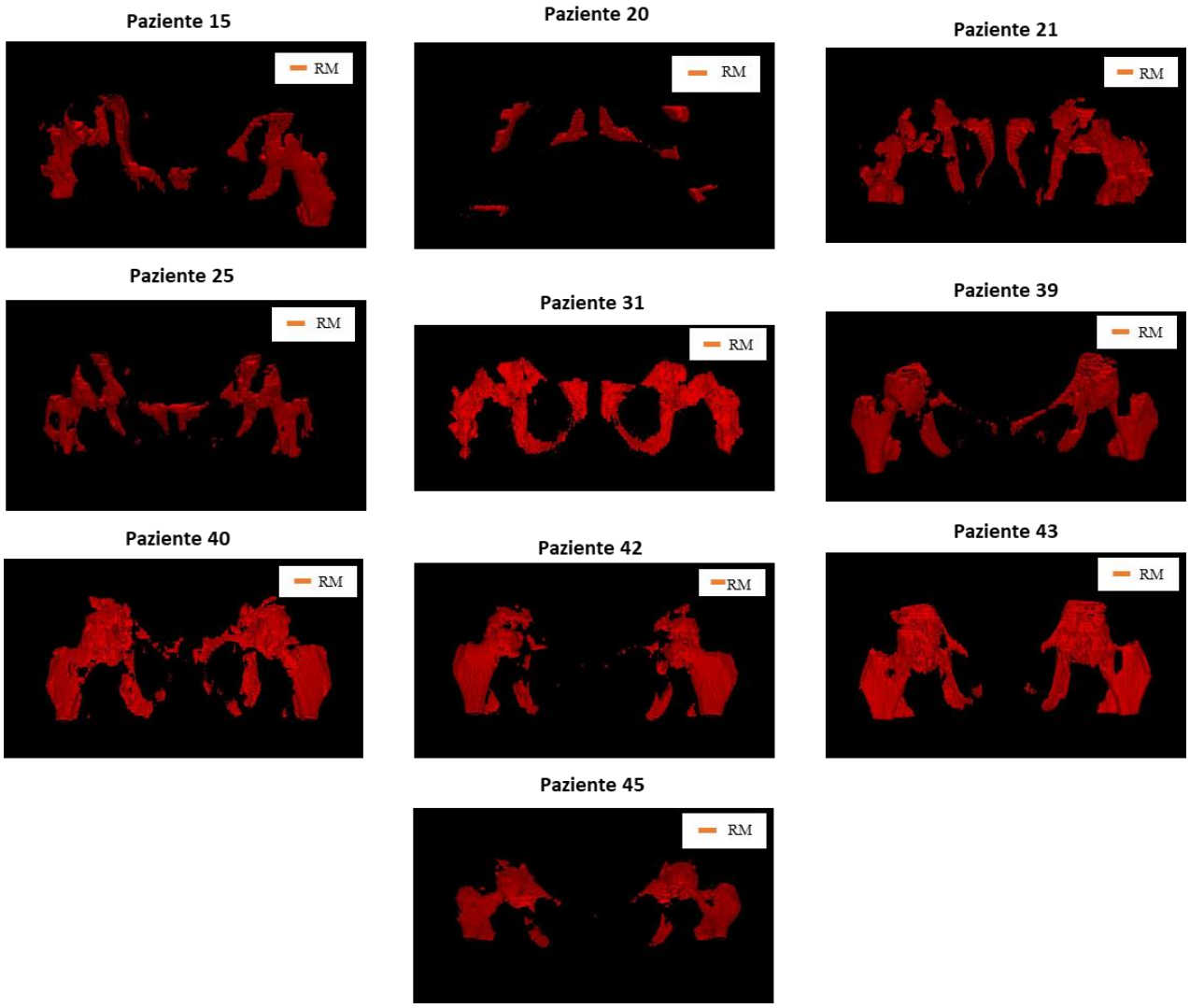


Figura 74 Esempi di segmentazione dell'RM da PET in 3D per la struttura LPBM nel caso di diversi soggetti del construction set e del validation set.

Il classificatore basato sulla decomposizione wavelet presenta la maggior parte delle difficoltà di segmentazione proprio in questa regione che tende generalmente a classificare come YM. Ecco, quindi, che si ha nel complesso un peggioramento della dice per la maggior parte delle slice presenti.

Un'eccezione è rappresentata dal paziente 20 che non presenta molto RM nell'area femorale, ma il quale è caratterizzato da una quantità insolitamente bassa di RM nella zona del pube e dell'ischio. Inoltre, anche i soggetti 36 e 47 che mostrano un leggero calo di performance mostrano una quantità molto bassa di RM nella zona del pube e dell'ischio (figura 75).

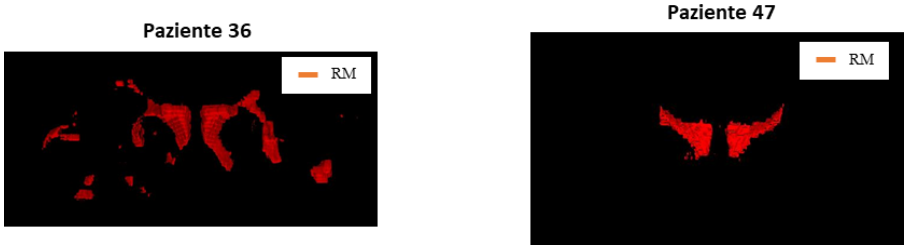


Figura 75 Segmentazione dell'RM da PET in 3D per la struttura LPBM nel caso dei soggetti 36 e 47.

In ultimo, merita una menzione a parte il paziente 41, il quale, sebbene mostri un miglioramento delle performance attraverso l'uso delle wavelet, è ricco di RM sia in pube ed ischio che in acetabolo e femore prossimale. Questo soggetto ha la particolare caratteristica di avere la quasi totalità (oltre il 99 %) del BM della struttura LPBM come midollo attivo e quindi bisognerebbe rivedere la sua segmentazione da PET al fine di escludere possibili errori verificatisi in precedenza (figura 76).

Figura 76 Segmentazione dell'RM da PET in 3D per la struttura LPBM nel caso del soggetto 41.

6 Conclusioni

Nel corso di questo lavoro di tesi sono state sfruttate le immagini di tomografia computerizzata per l'individuazione del midollo osseo attivo. In tal modo vengono risolti i problemi di costo e di difficoltà di accesso che caratterizzano le metodiche attualmente impiegate in questo campo. Dal punto di vista operativo, si è scelto di seguire la strada della radiomica andando a ricavare informazioni dai pixel delle immagini e implementando tecniche di machine learning.
A partire da una popolazione di 50 pazienti in cura presso l'oncologia dell'Ospedale Molinette di Torino, i quali sono stati sottoposti sia a PET che a TC, sono state distinte tre regioni anatomiche: midollo osseo lombosacrale (LSBM), midollo osseo iliaco (IBM) e porzione bassa del midollo osseo pelvico (LPBM). Lo studio è cominciato dapprima effettuando il training, il test e l'ottimizzazione dei parametri di classificatori diversi su tutte e tre le regioni anatomiche andando ad impiegare per il training un set di dati ottenuto mediante clustering con reti SOM e comprendente feature statistiche del 1° e del 2° ordine. I risultati ottenuti sono stati messi a confronto con quelli provenienti da un training set ottenuto mediante estrazione random. Nel complesso i metodi di classificazione hanno ottenuto performance abbastanza simili, con una leggera diminuzione della variabilità inter-paziente nel caso di training set ottenuto mediante clustering.
In seguito, si è rivolta l'attenzione alla struttura LPBM, in quanto è risultata affetta dalle maggiori difficoltà di classificazione. È stata effettuata un'analisi delle sue slice in termini di RM presente, sono state rilevate le sue criticità rispetto alle altre due regioni anatomiche e si è sviluppato un sistema che cercasse di aumentare la rappresentatività dei dati di training dell'intera struttura. Più nel dettaglio, è stato rivalutato il metodo di selezione delle slice contenute all'interno del construction set passando da una selezione random ad una stratificata e, in seguito, è stata modificata anche la tecnica di clustering adottata per l'estrazione del training set utilizzando dendrogrammi con un taglio sulla base della variabilità intra-cluster. Tuttavia, i risultati ricavati non hanno mostrato particolari differenze in termini di performance sulle maschere di segmentazione.
Successivamente, sono stati analizzati i misclassificati comuni tra diverse tipologie di classificatori impiegati, è stato implementato un metodo di classificazione basato sul majority voting di quest'ultimi e sono state valutate le sue prestazioni. In questo caso, lo sbilanciamento delle performance verso una sensibilità maggiore rispetto ai casi precedenti a discapito però della specificità ha condotto a non ritenere questa strada la più conveniente da seguire.
In ultimo, sono stati rivalutati i descrittori impiegati all'interno di questo studio e si è esplorato il campo delle feature di ordine superiore, prendendo in considerazione, in particolare, quelle derivanti dalla decomposizione wavelet. Sono state allenate alcune tipologie di classificatori e i risultati sono stati messi a confronto con quelli ottenuti a partire dai descrittori del 1° e del 2° ordine.
Il lavoro svolto ha evidenziato una generale migliore capacità delle feature derivanti dalla decomposizione wavelet di risolvere il problema di individuazione del midollo attivo sulle immagini di TC della struttura LPBM rispetto alle tradizionali feature statistiche del 1° e del 2° ordine.
In taluni pazienti si è riscontrata una certa difficoltà nell'identificazione del midollo osseo attivo. Questi ultimi presentano la caratteristica di possedere una segmentazione dell'RM da PET che mostra una elevata quantità di RM nella zona dell'acetabolo e del femore prossimale rispetto a

quella presente nel pube e nell'ischio. In questi soggetti la regione più critica da identificare come RM appare essere proprio quella del femore, per cui in studi successivi è possibile pensare di suddividere la regione LPBM in due parti distinte: la prima contenente il midollo di pube ed ischio, la seconda contenente quello dell'area femorale. Quindi impiegare un classificatore multi-classe che riconosca RM e YM distintamente per le due parti oppure adottare due distinti classificatori binari. In ultimo, dati i risultati promettenti ottenuti con le feature statistiche del 1° ordine in seguito a decomposizione wavelet, rappresenta sicuramente un ulteriore interessante punto di indagine quello di estrarre feature statistiche del 2° ordine dopo decomposizione wavelet e analizzare l'efficacia di queste ultime per la risoluzione del problema di identificazione di midollo osseo attivo da immagini di TC.

Bibliografia

[1] J. S. M. Blebea, M. M. Houseni, D. A. M. M. Torigian, C. M. Fan, A. M. Mavi, Y. P. Zhuge, T. B. Iwanaga, S. B. Mishra, J. P. Udupa, J. Zhuang, R. Gopal e A. M. Alavi, «Structural and Functional Imaging of Normal Bone Marrow and Evaluation of Its Age-Related Changes,» *Seminars in Nuclear Medicine,* vol. 37, n. 3, pp. 185-194, 2007.

[2] M. Chiarilli, A. Delli Pizzi, D. Mastrodicasa, M. Febo, B. Cardinali, B. Consorte, A. Cifaratti, V. Panara, M. Caulo e G. Cannataro, «Bone marrow magnetic resonance imaging: physiologic and pathologic findings that radiologist should know,» *La Radiologia medica,* vol. 126, n. 2, pp. 264-276, 2021.

[3] J. M. M. David, Y. P. Yue, K. M. Blas, A. M. Hendifar, P. M. Kabolizadeh e R. M. P. Tuli, «18F-FDG PET Predicts Hematologic Toxicity in Patients with Locally Advanced Anal Cancer Treated With Chemoradiation,» *Advances in Radiation Oncology,* vol. 4, n. 4, pp. 613-622, 2019.

[4] R. Carr, S. F. Barrington, B. Madan, M. J. O'Doherty, C. A. Saunders, J. van der Walt e A. R. Timothy, «Detection of Lymphoma in Bone Marrow by Whole-Body Positron,» *Blood,* vol. 91, n. 9, pp. 3340-3346, 1998.

[5] A. Andreychenko, P. S. Kroon, M. Maspero, I. Jürgenliemk-Schulz, A. A. De Leeuw, M. G. Lam, J. J. Lagendijk e C. A. van den Berg, «The feasibility of semi-automatically generated red bone marrow segmentations based on MR-only for patients with gynecologic cancer,» *Radiotherapy and oncology,* vol. 123, n. 1, pp. 164-168, 2017.

[6] H. Zaidi, A. Alavi e I. E. Naqa, «Novel Quantitative PET Techniques for Clinical Decision Support in Oncology,» *Seminars in nuclear medicine,* vol. 48, n. 6, pp. 548-564, 2018.

[7] P. Franco, C. Fiandra, F. Arcadipane, E. Trino, F. R. Giglioli, R. Ragona e U. Ricardi, «Incorporating 18FDG-PET-defined pelvic active bone marrow in the automatic treatment planning process of anal cancer patients undergoing chemo-radiation,» *Incorporating 18FDG-PET-defined pelvic active bone marrow in the automatic treatment planning process of anal cancer patients undergoing chemo-radiation,* vol. 17, n. 1, p. 710, 2017.

[8] C. Kim, N. Gupta, B. Chandramouli e A. Alavi, «Standardized uptake values of FDG: body surface area correction is preferable to body weight correction,» *Journal of nuclear medicine,* vol. 35, n. 1, pp. 164-167, 1994.

[9] S. Rosati, G. Balestra, P. Franco, C. Fiandra, F. Arcadipane, P. Silvetti, U. Ricardi e E. Gallio, «Radiomics for identification of active bone marrow from ct: An exploratory study,» 2018.

[10] T. G. Perk, N. A. Weisse, S. S. F. Yip e R. Jeraj, «A method for quantitative total marrow imaging (QTMI) with PET/CT,» *Biomedical physics & engineering express,* vol. 2, n. 5, p. 55006, 2016.

[11] N. Makris, R. Boellaard, C. Menke, A. Lammertsma e M. Huisman, «An automatic delineation method for bone marrow absorbed dose estimation in 89Zr PET/CT,» *EJNMMI physics,* vol. 3, n. 1, pp. 1-9, 2016.

[12] G. Sambuceti, M. Brignone, C. Marini, M. Massollo, F. Fiz, S. Morbelli, A. Buschiazzo, C. Campi, R. Piva, A. M. Massone, M. Piana e F. Frassoni, «Estimating the whole bone-marrow asset in humans by

a computational approach to integrated PET/CT imaging,» *European journal of nuclear medicine and molecular imaging,* vol. 39, n. 8, pp. 1326-1338, 2012.

[13] F. Kogan, S. M. Broski, D. Yoon e G. E. Gold, «Applications of PET-MRI in musculoskeletal disease,» *Journal of magnetic resonance imaging,* vol. 48, n. 1, pp. 27-47, 2018.

[14] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker e H. J. Aerts, «Radiomics: Extracting more information from medical images using advanced feature analysis,» *European journal of cancer ,* vol. 48, n. 4, pp. 441-446, 2011.

[15] S. T. Sepe, «Confronto di algoritmi di Deep Learning per l'individuazione del midollo osseo attivo in immagini CT».

[16] A. D. Chiara, «Radiomica per Individuazione di Midollo Osseo Emopoietico da Tomografia Computerizzato».

[17] A. SIlik, M. Noori, W. Altabey, R. Ghiasi e Z. Wu, «Comparative Analysis of Wavelet Transform for Time-Frequency Analysis and Transient Localization in Structural Health Monitoring,» *Structural Durability & Health Monitoring,* vol. 15, n. 1, pp. 1-22, 2021.

[18] V. Giurgiutiu, «Chapter 14 - Signal Processing and Pattern Recognition for Structural Health Monitoring with PWAS Transducers,» in *Structural Health Monitoring with Piezoelectric Wafer Active Sensors*, Elsevier Inc, 2014, pp. 807-862.

[19] S. G. Mallat, «A theory for multiresolution signal decomposition: the wavelet representation,» *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, n. 7, pp. 674-693, 1989.

[20] S. Mallat, «A Theory for Multiresolution Signal Decomposition: the wavelet representation,» in *Fundamental Papers in Wavelet Theory*, Princeton University Press, 2009, pp. 494-513.

[21] M. Vigni, J. Prats-Montalban, A. Ferrer e M. Cocchi, «Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA),» *Journal of Chemometrics,* vol. 32, n. 1, 2017.

[22] C. B. R. Ferreira e D. Borges, «Analysis of mammogram classification using a wavelet transform decomposition,» *Pattern Recognition Letters,* vol. 24, n. 7, pp. 973-982, 2003.

[23] H. J. Aerts, E. R. Velazquez, R. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies e P. Lambin, «Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,» *Nature communications,* vol. 5, n. 1, pp. 1-9, 2014.

[24] V. Srivastava e R. K. Purwar, «A Five-Level Wavelet Decomposition and Dimensional Reduction Approach for Feature Extraction and Classification of MR and CT Scan Images,» *Applied Computational Intelligence and Soft Computing,* 2017.

[25] Z. Hou, Y. Yang, S. Li, J. Yan, W. Ren, J. Liu, K. Wang, B. Liu e S. Wan, «Radiomic analysis using contrast-enhanced CT: predict treatment response to pulsed low dose rate radiotherapy in gastric carcinoma with abdominal cavity metastasis,» *Quantitative imaging in medicine and surgery,* vol. 8, n. 4, p. 410, 2018.

[26] P. Grossmann, O. Stringfield, N. El-Hachem, M. Bui, E. R. Velazquez, C. Parmar, R. T. Leijenaar, B. Haibe-Kains, P. Lambin, R. J. Gillies e H. J. Aerts, «Defining the biological basis of radiomic phenotypes in lung cancer,» *Elife,* vol. 6, 2017.