POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Biomedica

Tesi di Laurea Magistrale

PROSTATE CANCER DETECTION AND DIAGNOSIS

Relatore
Prof. Massimo Salvi

Candidato
Riccardo Russo

Co-Relatori
Dott. Marco Bologna
Dott.ssa Federica Amato

Luglio 2024

Indice

Elenco delle tabelle

Elenco delle figure

Elenco degli acronimi

- **ADC**: Coefficiente di diffusione apparente
- **AFMS**: Stroma Fibro-Muscolare Anteriore
- **AI**: Intelligenza Artificiale
- **ANN**: Artificial Neural Network
- **bpMRI**: Biparametric Magnetic Resonance Imaging
- **csPCa**: Clinically Significant Prostate Cancer
- **CZ**: Zona Centrale
- **DL**: Deep Learning
- **DWI**: Diffusion-Weighted Imaging
- **ESUR**: European Society of Urogenital Radiology
- **GS**: Punteggio di Gleason
- **HU**: Hounsfield Unit
- **IPB**: Iperplasia Prostatica Benigna
- **ISUP**: International Society of Urological Pathology
- **LR**: Learning Rate
- **ML**: Machine Learning
- **MONAI**: Medical Open Network for AI
- **M1**: Modello UNet per la segmentazione della prostata
- **M2**: Modello UNet per la segmentazione del csPCa
- **mpMRI**: Multiparametric Magnetic Resonance Imaging

- **MRI**: Risonanza Magnetica
- **MRBx**: Biopsie Guidate da RM
- **PET**: Imaging Nucleare
- **PI-CAI**: Prostate Imaging – Cancer Artificial Intelligence
- **PI-RADS**: Prostate Imaging Reporting and Data System
- **PReLU**: Parametric Rectified Linear Unit
- **PZ**: Zona Periferica
- **PSA**: Prostate-Specific Antigen
- **PSAD**: Prostate-Specific Antigen Density
- **ReLU**: Rectified Linear Unit
- **ROI**: Region Of Interest
- **RP**: Prostatectomia Radicale
- **RUMC**: Radboud University Medical Center
- **SysBx**: Biopsie Sistematiche
- **TC**: Tomografia Computerizzata
- **TRUS**: Ecografia Transrettale
- **T2W**: T2-Weighted Imaging
- **TZ**: Zona di Transizione
- **WD**: Weight Decay

Abstract

Il carcinoma prostatico è il tumore più diffuso tra gli uomini, con un'incidenza di 1 uomo su 6 e una mortalità che ne coinvolge 1 su 36 [1]. La diagnosi precoce e accurata è fondamentale per garantire un trattamento tempestivo ed efficace, migliorando le prospettive di guarigione e la qualità di vita dei pazienti. Negli ultimi anni, l'avvento dell'imaging ad alta risoluzione, in particolare delle scansioni biparametriche di risonanza magnetica (bpMRI), ha rivoluzionato l'approccio alla diagnosi e al monitoraggio del tumore alla prostata. Le immagini bpMRI forniscono informazioni dettagliate sulla morfologia e sulla composizione dei tessuti prostatici, consentendo una valutazione più accurata delle lesioni tumorali e una pianificazione più precisa dei trattamenti. Tuttavia, la valutazione delle immagini richiede un analisi accurata da parte di radiologi esperti. Infatti, le diverse caratteristiche morfologiche e la diversa intensità del segnale delle lesioni complicano il processo di identificazione e segmentazione rendendo il processo di diagnosi lungo, complesso e suscettibile a variazioni interpretative soggettive. In questo contesto, l'utilizzo dell'intelligenza artificiale e delle tecniche di apprendimento automatico hanno suscitato un crescente interesse come strumento ausiliario dimostrando di poter migliorare l'accuratezza e l'efficienza della diagnosi. Questa tesi ha come obiettivo quello di sviluppare un algoritmo completamente automatico per analizzare le scansioni addominali bpMRI di un paziente al fine di identificare e segmentare eventuali lesioni tumorali prostatiche, fornendo un supporto al radiologo nel processo di diagnosi. Per lo sviluppo dell'algoritmo, si è fatto uso del dataset pubblico della *PI-CAI challenge*, composto da dati multicentrici acquisiti tra il 2011 e il 2021. L'architettura implementata prevede l'impiego di due modelli UNet in successione: il primo è stato progettato per identificare e generare la maschera prostatica, mentre il secondo per identificare e generare la maschera delle eventuali lesioni tumorali presenti. Infine, le maschere ottenute vengono utilizzate per formulare la diagnosi. Durante l'analisi sperimentale sono state individuate diverse criticità e si sono sviluppate soluzioni mirate a garantire il corretto funzionamento dell'algoritmo in scenari d'uso reale. I risultati finali ottenuti *(Dice: 0.50 ± 0.27; Distanza di Hausdorff (95° percentile): 31.45 ± 32.27 pixel; RVD: 3.82 % ± 11.42; AUC=0.63; Corretti classificati*

positivi: 74%; Corretti classificati negativi: 33%) dimostrano la validità del metodo implementato rispetto alla soluzione base proposta dagli organizzatori della PI-CAI challenge *(Dice: 0.30 ± 0.39; Distanza di Hausdorff (95° percentile): 2.74 ± 4.24 pixel; RVD: -0.61 % ± 0.51; AUC=0.58; Corretti classificati positivi: 40%; Corretti classificati negativi: 83%).* I valori delle metriche confrontate sono stati calcolati sui casi appartenenti allo stesso Test set interno. Nonostante i progressi registrati grazie al metodo proposto, che ha portato a dei miglioramenti nelle performance di predizione e diagnosi, ci sono ancora ampi margini di miglioramento. Questi sono necessari per rendere l'algoritmo un'opzione affidabile e realmente utilizzabile nella pratica clinica.

Capitolo 1
Introduzione

Il seguente capitolo ha come obiettivo quello di presentare lo scopo della tesi e di fornire una panoramica utile per la comprensione delle strategie adottate per lo sviluppo dell'algoritmo.

1.1 Anatomia della Prostata

La ghiandola prostatica, parte importante dell'apparato uro-genitale maschile dei mammiferi, ha la funzione primaria di produrre il liquido prostatico essenziale per la sopravvivenza e la motilità degli spermatozoi [3]. Anatomicamente si trova nel compartimento sub-peritoneale tra il diaframma pelvico e la cavità peritoneale, posteriormente alla sinfisi pubica, anteriormente al retto e inferiormente alla vescica urinaria [2] (figura 1.1).

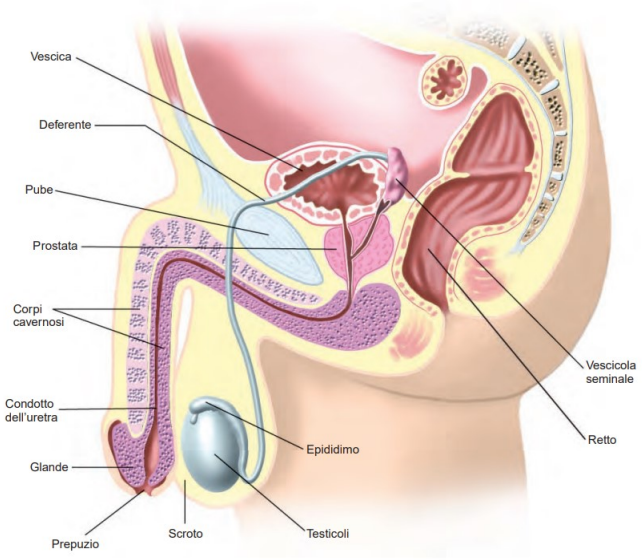


Figura 1.1: Posizione anatomica della prostata [2].

La prostata ha dimensioni approssimative di 3 cm in lunghezza, 4 cm in larghezza e 2,5 cm in spessore, con un peso stimato negli adulti tra 10 e 20

grammi [3]. Durante la pubertà si verifica un aumento iniziale delle dimensioni, mentre ulteriori incrementi si manifestano dopo i 50 anni, causando *iperplasia prostatica benigna (IPB)* [4]. In condizioni normali, la forma della prostata è piramidale, simile ad una "castagna", ma può assumere conformazioni diverse, come quella a “mezzaluna” o a “ciambella”, soprattutto in presenza di IPB [11]. È possibile classificare la struttura interna della prostata in diverse zone: due sezioni assiali destra e sinistra, identificate da una linea verticale tracciata attraverso il centro, e nelle sezioni anteriori e posteriori identificate da una linea orizzontale passante per il centro della ghiandola [2]. Le zone così identificate sono (figura 1.2):

I **La zona periferica (PZ)**: costituisce la parte esterna della prostata e contiene dal 70% all’80% del tessuto ghiandolare [1]. Essa si suddivide ulteriormente in sezioni anteriore (a), postero-mediale (mp) e postero-laterale (lp) [2] . Con l’avanzare dell’età, la PZ è spesso coinvolta in prostatiti croniche, atrofia post-infiammatoria e nel cancro alla prostata (PCa). Generalmente assume una forma a "ciambella" o "anello" intorno alla zona centrale [11].

II **La zona di transizione (TZ)**: costituisce il 5% del volume totale della prostata [1]. Si trova vicino all’apice della zona centrale e dei dotti eiaculatori. Con l’avanzare dell’età, la TZ è soggetta ad allargamenti atrofici, spesso causati da IPB, che tendono a comprimere la PZ [5]. La TZ è spesso descritta come una "mezzaluna" o "ferro di cavallo" [11].

III **La zona centrale (CZ)**: occupa circa il 25% del volume e contiene circa il 20% del tessuto ghiandolare [1]. Questa zona è soggetta a cambiamenti atrofici con l’avanzare dell’età. La CZ, descritta come un "cono", occupa la parte centrale della prostata e si estende verso l’uretra [9].

IV **Lo stroma fibro-muscolare anteriore (AFMS)**: indicato anche con la notazione **AS**, costituisce meno dell’1% del volume totale della prostata [1]. Questa zona è priva di tessuto ghiandolare. L’AFMS è più diffuso anteriormente e può assumere una forma più allungata o irregolare (a "cuneo"), a seconda della struttura specifica della prostata [9].

La CZ e la TZ costituiscono la cosiddetta **ghiandola centrale**.

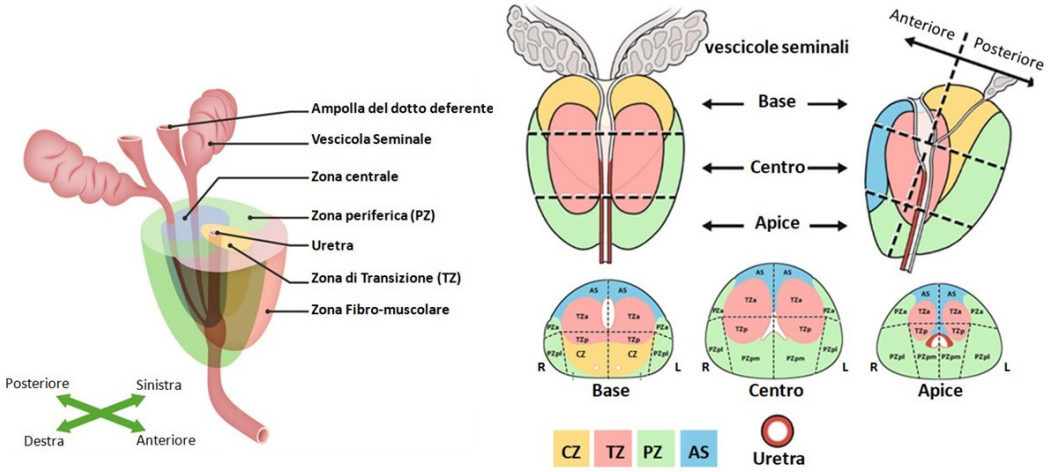


Figura 1.2: Zone anatomiche della prostata [2].

1.2 Classificazione del cancro

La prostata può essere affetta da diverse condizioni patologiche tra cui il tumore. Questa patologia si manifesta quando le cellule prostatiche iniziano a proliferare in modo incontrollato, formando un tumore che può rimanere localizzato o metastatizzare. La metastasi è il processo attraverso il quale le cellule tumorali si diffondono dal sito primario ad altre parti del corpo formando nuovi focolai di tumore. Questo può avere conseguenze potenzialmente letali se non diagnosticato e trattato tempestivamente. Il processo di diagnosi e valutazione del cancro alla prostata si avvale di diversi strumenti, tra cui il sistema di classificazione Pi-RADS (Prostate Imaging Reporting and Data System), la scala ISUP (International Society of Urological Pathology) e il Punteggio di Gleason (GS).

Il sistema Pi-RADS fornisce una guida standardizzata per interpretare le immagini della risonanza magnetica (MRI) della prostata, assegnando un punteggio da 1 a 5 alle lesioni rilevate. Questo punteggio riflette la probabilità di cancro alla prostata, con punteggi più alti che indicano un maggiore rischio di malignità [12] (tabella: 1.1; figura: 1.3).

Tabella 1.1: Tabella dei punteggi Pi-RADS.

Scala Pi-RADS	**Probabilità di cancro alla prostata**
1	Estremamente bassa
2	Bassa
3	Intermedia
4	Alta
5	Molto alta

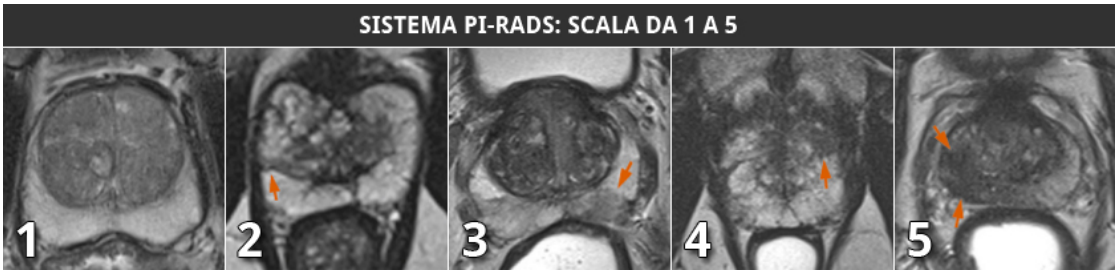


Figura 1.3: Scala Pi-RADS [10].

Il Punteggio di Gleason (GS) è utilizzato per valutare il grado di aggressività del tumore prostatico basandosi sul grado di differenziazione del tessuto ghiandolare. Questo punteggio viene determinato attraverso l'analisi microscopica dei campioni di tessuto prelevati durante la biopsia e viene assegnato in base alla somma di due valori, ciascuno compreso tra 1 e 5, che rappresentano i due pattern cellulari predominanti nel tumore. Nello specifico, il punteggio viene determinato osservando la struttura delle cellule tumorali e confrontandole con quelle normali. Il primo numero rappresenta il grado di differenziazione predominante (il tessuto tumorale di maggiore estensione), mentre il secondo numero rappresenta il secondo grado predominante (se presente). Questi due numeri vengono successivamente sommati per ottenere il Punteggio di Gleason totale che può variare da 2 a 10. Per aiutare a distinguere meglio il grado di aggressività del tumore sono stati sviluppati *i Grade Group*. Questi sono cinque gruppi distinti in cui i tumori possono essere classificati, fornendo una guida più chiara per la prognosi e il trattamento [13] (tabella: 1.2).

Tabella 1.2: Tabella riassuntiva dei Grade Group e dei corrispondenti Gleason Score.

Grade Group	**Gleason Score**
1	≤ 6
2	$3 + 4 = 7$
3	$4 + 3 = 7$
4	$4 + 4 = 8$; $3 + 5 = 8$; $5 + 3 = 8$
5	$9 - 10$

Infine, la scala ISUP valuta i risultati delle biopsie prostatiche assegnando un grado di aggressività da 1 a 5 al tumore (tabella: 1.3). Questa classificazione è una derivazione diretta del Gleason score, concepita per semplificarne l'interpretazione e incrementarne l'accuratezza prognostica [14].

Tabella 1.3: Tabella dei gradi ISUP.

Grado ISUP	**Grado di aggressività**
1	Basso grado con minima aggressività
2	Basso grado con un lieve aumento dell'aggressività
3	Grado intermedio con moderata aggressività
4	Grado alto con un'aggressività significativa
5	Tumori altamente aggressivi con alto rischio di diffusione e metastasi

1.3 Imaging nel tumore alla prostata

L'imaging medico svolge un ruolo cruciale nel processo di identificazione, diagnosi, pianificazione del trattamento e monitoraggio della risposta terapeutica nel tempo. Tra le principali tecniche comunemente utilizzate in questo ambito vi sono la risonanza magnetica (MRI), l'ecografia transrettale (TRUS), la tomografia computerizzata (TC) e l'imaging nucleare (PET). Ognuna di queste offre vantaggi specifici in termini di risoluzione, penetrazione e capacità di differenziazione tra tessuto sano e maligno [15]. In questo progetto viene trattata l'MRI, essendo la tecnica di acquisizione utilizzata per la formazione del dataset di cui si è fatto uso per lo sviluppo della tesi. L'MRI ha suscitato un crescente interesse grazie alla sua capacità di generare immagini dettagliate della prostata e dei tessuti circostanti, senza l'uso di radiazioni e con un migliore contrasto dei tessuti molli.

In questo studio si è fatto uso della MRI biparametrica (bpMRI), una tecnica di imaging che, sebbene fornisca meno informazioni diagnostiche rispetto alla multiparametrica (mpMRI) [7], è considerata più adatta per effettuare screening di massa [6]. La bpMRI si differenzia dal protocollo della mpMRI in quanto non prevede l'acquisizione di immagini dinamiche con contrasto (DCE-MRI). Questa scelta presenta diversi vantaggi, tra cui la riduzione dei tempi dell'esame [8], dei costi e l'eliminazione degli effetti collaterali per il paziente derivanti dall'uso del mezzo di contrasto. In particolare, le tecniche di Imaging di cui si è fatto uso sono:

- *Risonanza magnetica T2-pesata (MRI T2W)*: La tecnica evidenzia le differenze nei tempi di rilassamento T2 dei tessuti e consente di visualizzare in dettaglio l'anatomia della prostata e individuare eventuali lesioni tumorali in essa presenti. Tipicamente eseguita su tre piani, assiale coronale e sagittale, la T2 consente di distinguere le tre zone prostatiche: PZ, TZ e CZ.[1] Questo è importante in quanto le cellule del cancro, nelle diverse zone, appaiono diverse in termini di intensità e omogeneità. I tumori della prostata sono generalmente di forma rotonda o irregolare e presentano un segnale a bassa intensità circondato da uno ad alta intensità nella PZ. Tuttavia, è importante evidenziare che questo tipo di segnale può derivare anche da patologie benigne. Il limite principale di questa acquisizione è che non consente di rilevare le lesioni presenti al di fuori della PZ, poiché nella TZ le caratteristiche del segnale tumorale si confondono con quelle del tessuto sano. Infatti, i tumori localizzati nella TZ possono manifestare margini mal definiti, un segnale T2 medio-basso, comunemente noto come "segno di disegno a carboncino cancellato", e una forma non circoscritta, che può essere lenticolare o fusiforme.[17]

- *MRI diffusione-pesata (MRI DWI)*: utilizzando gradienti di campo magnetico, la DWI misura la diffusione delle molecole d'acqua nei tessuti molli, fornendo informazioni sulla densità cellulare e sull'integrità strutturale dei tessuti [1]. I tumori prostatici, con la loro maggiore densità cellulare, ostacolano la diffusione dell'acqua, producendo un segnale di alta intensità rispetto al tessuto sano. Questa differenza di intensità del segnale consente di identificare e localizzare le lesioni tumorali. Le DWI vengono generate controllando un parametro specifico detto *b-value*, che rappresenta il fattore di forza e la tempistica dei gradienti di campo. Valori di b più elevati permettono una migliore discriminazione

tra tessuto sano e tessuto tumorale. Secondo le linee guida standard, come il *PI-RADS v2.1*, è suggerito l'utilizzo di valori superiori a 1400 s/mm2 per ottimizzare l'identificazione delle lesioni tumorali. La qualità delle immagini DWI dipende dall'SNR e dalla presenza di artefatti, che possono essere minimizzati con l'ottimizzazione dei parametri di acquisizione ed una corretta esecuzione dell'esame.[17]

- *Mappe del coefficiente di diffusione apparente (MRI ADC):* generate combinando immagini DWI calcolate usando *b-value* differenti, forniscono ulteriori informazioni sulla diffusione delle molecole d'acqua nei tessuti [1]. I tumori maligni presentano tipicamente un ADC inferiore rispetto ai tessuti benigni, consentendo una diagnosi più accurata. La mappa ADC viene sempre calcolata utilizzando valori b<1000 s/mm2. Il primo valore utilizzato è tipicamente b50 al fine di evitare la trasparenza dei vasi ed escludere i segnali vascolari [17]. Diversi studi hanno inoltre evidenziato una relazione tra il valore del coefficiente di diffusione apparente (ADC) e il punteggio Gleason (GS). In particolare emerge come la diminuzione dei valori ADC, indicata da un segnale basso, è significativamente correlata all'aumento del GS [23], [24], [25].

Nella figura 1.4 sono riportate le scansioni T2W, ADC e DWI di due pazienti differenti.

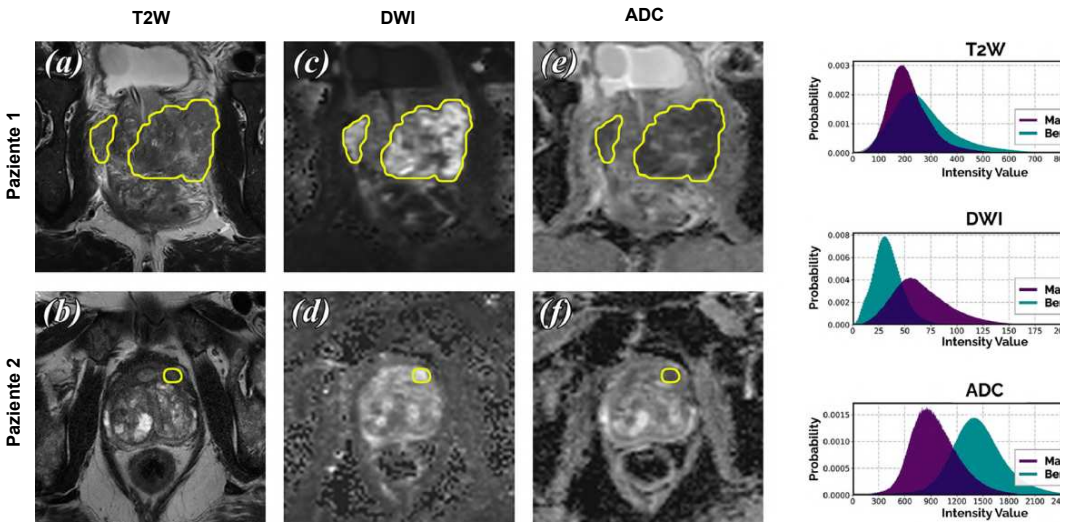


Figura 1.4: Esempi di scansioni bpMRI della prostata: (a-b) immagini T2W, (c-d) immagini DWI con valore b elevato, (e-f) mappe ADC. I contorni gialli indicano le lesioni tumorali [47].

Nella tabella 1.4 sono riassunte le caratteristiche del cancro nelle zone PZ e TZ della prostata.

Tabella 1.4: Caratteristiche del tumore nelle zone PZ e TZ della prostata.

Zona	Caratteristiche distintive	Tipologia di Tumore	Rilevazione tramite MRI	Caratteristiche MRI
Periferica (PZ)	-Forma rotonda o mal definita -Vicine alla superficie della prostata.	Tumori maligni (Adenocarcinoma Tra i più comuni)	-Facili da diagnosticare	-Elevata intensità del segnale su DWI a valori b elevati -Bassa intensità di segnale sulle ADC -Segnale T2 medio-basso
Transizione (TZ)	-Margini mal definiti -Forma non circoscritta : lenticolare o fusiforme -invasione delle strutture circostanti	Tumori meno aggressivi e meno comuni	-Difficili da diagnosticare	

Un'analisi comparativa ha dimostrato che l'uso combinato della T2w e

della DWI fornisce una sensibilità significativamente superiore (81%) rispetto all'uso della solo T2w (54%). Anche se l'approccio combinato ha mostrato una leggera riduzione della specificità (84% rispetto al 91%), la sua capacità di individuare il tumore nella PZ della prostata è risultata migliore[53]. Per migliorare la qualità delle acquisizioni e ridurre la variabilità tra i diversi centri è fondamentale adottare dei protocolli di acquisizione standardizzati. Le linee guida del PI-RADS, sviluppate dalla Società Europea di Radiologia Urogenitale (ESUR) nel 2012 e aggiornate nel 2016 con la versione 2.1, sanciscono i requisiti minimi e ottimali per l' mpMRI della prostata. Tuttavia queste linee guida si concentrano principalmente sulle specifiche tecniche e non forniscono istruzioni dettagliate sulla preparazione del paziente e sulla mitigazione degli artefatti che possono influenzare la qualità dell'immagine [16] [62].

1.4 Variabili cliniche e di acquisizione

Le variabili cliniche e di acquisizione svolgono un ruolo cruciale nel processo di diagnosi e nella pianificazione del trattamento del cancro alla prostata. Le variabili cliniche comunemente utilizzate sono [18]:

- *L'età del paziente*: può influenzare il rischio di sviluppo della malattia.
- Il *PSA* (antigene prostatico specifico): è una proteina prodotta dalle cellule sane e tumorali della ghiandola prostatica. Il test del PSA misura il livello di questa proteina nel sangue espresso in ng/mL. Sebbene un valore elevato di PSA sia spesso associato al cancro alla prostata, è importante considerare che diverse condizioni benigne, come la prostatite e l'IPB, possono anch'esse provocarne un aumento.
- Il *volume della prostata*: espresso in mL, è un fattore che può influenzare i livelli di PSA. Si tratta di un valore spesso approssimato.
- Il *PSAD (densità di PSA)*: espressa in ng/mL^2, rappresenta il rapporto tra il livello ematico di PSA e il volume della ghiandola prostatica. Questa fornisce una valutazione più accurata del rischio di cancro alla prostata rispetto al solo uso del PSA.

Le variabili di acquisizione determinano la qualità e l'affidabilità delle immagini diagnostiche [26]. In particolare:

- *Produttore e modello scanner*: possono influenzare la risoluzione e la qualità delle immagini.
- *Valore b di diffusione*: rappresenta la sensibilità del segnale alla diffusione molecolare dei tessuti. Questo può fornire informazioni cruciali sulla struttura e la composizione dei tessuti prostatici.
- *L'histopath_type:* indica la procedura utilizzata per campionare il tessuto della lesione per l'analisi microscopica o istopatologica. Tra le procedure tipicamente utilizzate vi sono le biopsie sistematiche (SysBx), biopsie guidate da RM (MRBx), o una combinazione di entrambe. Un'altra procedura significativa è la prostatectomia radicale (RP), un intervento chirurgico che comporta la rimozione completa della ghiandola prostatica e di eventuali tessuti circostanti.

1.5 Definizione degli obiettivi della tesi

La seguente tesi si propone di raggiungere diversi obiettivi fondamentali per lo sviluppo nel campo della diagnosi del tumore prostatico attraverso l'utilizzo di algoritmi di deep learning. Gli obiettivi delineati sono i seguenti:

1. Sviluppare un algoritmo completamente automatico capace di analizzare le scansioni addominali bpMRI di un paziente al fine di identificare e segmentare le eventuali lesioni tumorali presenti nella prostata e fornire una diagnosi della positività del paziente.
2. Ottimizzare il metodo implementato rendendolo robusto alle variazioni inter/intra paziente presenti nei casi d'uso reali, al fine di garantire il corretto funzionamento dell'algoritmo e migliorare la qualità delle immagini generate, facilitando l'elaborazione da parte dei modelli di deep learning.
3. Ottimizzare i processi di training dei modelli per garantire una maggiore efficienza e accuratezza nell'apprendimento e nella generazione delle predizioni. Questo include l'ottimizzazione della strategia, delle funzioni e dei parametri dell'algoritmo di apprendimento.

Capitolo 2
Analisi dello stato dell'arte

Il seguente capitolo ha come obiettivo quello di analizzare le principali tecniche utilizzate in letteratura per fornire un supporto alla diagnosi e alla segmentazione automatica del tumore prostatico.

2.1 AI nella segmentazione dei tumori

Nel contesto dell'analisi di immagini mediche, le tecniche tradizionali come snake [20], region growing e modelli geometrici deformabili presentano diverse limitazioni che possono influenzare la loro efficacia nell'ambito di dataset eterogenei e complessi come quelli formati da scansioni bpMRI multicentriche. In particolare queste tecniche richiedono l'estrazione manuale di feature e la definizione di regole e parametri specifici per ciascuna applicazione [19]. Questo approccio può essere computazionalmente oneroso e lungo e richiede una notevole expertise da parte degli operatori medici per ottenere risultati accurati. Inoltre, queste metodologie possono non essere sufficientemente flessibili per catturare la variabilità delle caratteristiche delle lesioni.

L'approccio del deep learning invece si basa sull'estrazione automatica delle feature direttamente dai dati grezzi e sull'adattamento del modello alle variazioni presenti nei dati di input. Questo lo rende particolarmente adatto per gestire la complessità e l'eterogeneità dei dataset. Inoltre, i modelli di deep learning possono essere addestrati su grandi quantità di dati senza la necessità di specificare manualmente le feature da utilizzare, consentendo una maggiore generalizzazione e adattabilità del modello. L'impiego del deep learning nell'analisi delle immagini mediche comporta diversi vantaggi, quali l'automatizzazione della segmentazione, una maggiore robustezza alle variazioni dei dati e la possibilità di incorporare informazioni complesse e multidimensionali per migliorare le performance dell'algoritmo. Sebbene questi modelli siano computazionalmente molto onerosi da addestrare, una volta ottenuti la loro applicazione risulta essere vantaggiosa in termini di tempi e costi rispetto alle tecniche di segmentazione tradizionali [27].

2.2 PI-CAI Challenge

La *PI-CAI (Prostate Imaging Cancer AI) Challenge* [45] segna un significativo avanzamento rispetto alla *ProstateX Challenge* [46], precedentemente considerata come punto di riferimento pubblico per la diagnosi di cancro alla prostata clinicamente significativo (csPCa). La ProstateX Challenge è limitata dalla dimensione ridotta del dataset, dalla scarsa diversità dei casi (provenienti da un unico centro e fornitore di risonanza magnetica) e dal formato di valutazione debole che rendeva il test set pubblicamente disponibile anziché "invisibile", compromettendo la possibilità di trarre conclusioni in modo affidabile. La PI-CAI challenge si propone di superare questi limiti raccogliendo oltre 10.000 esami bpMRI multicentrici della prostata e fornendo una solida piattaforma per la convalida degli algoritmi di intelligenza artificiale proposti. La challenge è costituita da due principali studi:

1. Studio sui lettori: l'obiettivo è quello di stimare le prestazioni del radiologo medio.

2. Studio AI (Grand Challenge): l'obiettivo è analizzare le scansioni bpMRI al fine di generare la maschera delle eventuali lesioni prostatiche e fornire una diagnosi.

L'obiettivo finale della PI-CAI Challenge è quello di confrontare gli output degli algoritmi implementati con quelli generati dai radiologi coinvolti nello studio sui lettori, al fine di valutare la fattibilità clinica delle moderne soluzioni di intelligenza artificiale per il rilevamento e la diagnosi di csPCa.

Dall'analisi dei principali studi presenti in letteratura [47] e delle strategie adottate dai gruppi che si sono classificati nei primi cinque posti alla challenge [29] [30] [31] [32] [33] emerge un'architettura di base composta da:

1. un modello per identificare e generare la maschera prostatica;

2. un modello per identificare e generare la maschera delle lesioni;

3. modelli ausiliari per ridurre il numero di falsi positivi nella maschera delle lesioni;

4. una "Decision Fusion Node": funzione che consente di combinare i diversi output e produrre un'unica mappa di previsione.

Inoltre, si evidenzia che:

1. Il confronto con gli studi condotti prima della PI-CAI Challenge è significativamente limitato poiché questi hanno fatto uso di dataset poco numerosi costituiti da acquisizioni provenienti da un singolo centro. Inoltre, i dataset e i modelli impiegati in questi studi non sono facilmente accessibili.

2. La maggior parte dei gruppi di ricerca ha fatto uso del framework nnU-Net [21] e del metodo di cross-validazione per il training dei modelli [22].

3. Nessun gruppo ha fatto uso delle sequenze di acquisizione "facoltative" (T2W sagittale e coronale).

4. Nessun gruppo ha implementato un metodo di co-registrazione automatica delle scansioni.

5. Diversi gruppi hanno fatto uso di set aggiuntivi (per esempio: Prostate158).

6. Diversi studi clinici evidenziano come gli indici PSA e PSAd presi singolarmente non siano sufficienti per fare una diagnosi attendibile. Tuttavia, alcuni gruppi di ricerca hanno fatto uso di questi indici, quando presenti, per migliorare la precisione della classificazione.

Una nota rilevante emersa dall'analisi dei documenti riguarda l'assenza dei valori associati alle metriche di valutazione. Sebbene queste metriche mostrino una notevole variazione con minimi cambiamenti nei contorni e/o nel volume delle lesioni [47], è importante fornire tali valori per favorire il confronto e una valutazione oggettiva e completa delle prestazioni del metodo.

Capitolo 3
Materiali e Metodi

3.1 PI-CAI Challenge dataset

Il dataset PI-CAI Challenge [28] comprende circa 10.000 scansioni multicentriche suddivise in due set principali: un set pubblicamente accessibile di 1500 casi, e un set privato di 8707 casi. Quest'ultimo è stato utilizzato dagli organizzatori della sfida, durante la fase di sviluppo chiuso, per classificare le soluzioni proposte e riaddestrare i modelli più performanti. Tutti i dati sono completamente anonimizzati e resi disponibili con una licenza non commerciale *CC BY-NC 4.0*. Il dataset include anche 328 casi provenienti dalla ProstateX Challenge. Tra il materiale fornito è incluso un file denominato "marksheet.csv", il quale contiene le informazioni cliniche e di acquisizione relative ad ogni singolo caso del dataset.

Gli esami includono [1]:

1. Variabili cliniche: età del paziente*, volume della prostata°, livello di PSA°, densità di PSA°.

2. Variabili di acquisizione: produttore dello scanner*, nome del modello dello scanner*, valore di diffusione b*.

3. Scansioni bpMRI assiali: T2W, ADC, DWI con valore b elevato ($b \geq 1000$ s/mm^2);

4. sequenze opzionali (una/nessuna/entrambe): T2W sagittale e coronale.

5. Annotazione manuale multi-classe delle lesioni csPCa effettuate da un team di radiologi supervisionato da radiologi più esperti. Dei 1500 casi del set pubblico, solo 425 sono positivi al csPCa e, di questi, solo 220 riportano l'annotazione manuale tracciata dal radiologo.

[1] la notazione "°" indica un valore disponibile solo se riportato durante la routine clinica; la notazione "*" indica un valore sempre disponibile

6. Annotazione binaria delle eventuali lesioni csPCa fornita per ciascun paziente del dataset. Questa è stata generata utilizzando un metodo di apprendimento semi-supervisionato basato sull'uso del report diagnostico [38]. I casi negativi presentano una maschera nulla.

7. Annotazione della prostata nelle aree PZ e TZ, ottenuta mediante modelli AI, fornita per ciascun paziente del dataset.

8. Annotazione dell'intera ghiandola prostatica fornita per ciascun paziente del dataset.

Nella tabella 3.1 sono riassunte le caratteristiche del dataset utilizzato.

Tabella 3.1: Caratteristiche del dataset pubblico

Numero di pazienti	1476
Numero di casi	1500
PCa benigno o indolente	1075
csPCa (ISUP >1)	425
Età media (anni)	66 (IQR: 61-70)
PSA mediano (ng/ml)	8,5 (IQR: 6-13)
Volume medio della prostata (ml)	57 (IQR: 40-80)
Numero di lesioni MRI positive	1087
PI-RADS 3	246 (2396)
PI-RADS 4	438 (40%)
PI-RADS 5	403 (37%)
Numero di lesioni basate su ISUP	776
ISUP 1	311 (40%)
ISUP 2	260 (34%)
ISUP 3	109 (14%)
ISUP 4	41 (5%)
ISUP 5	55 (7%)

Nella figura 3.1 è riportato il boxplot delle dimensioni delle lesioni tumorali annotate nelle maschere manuali del dataset.

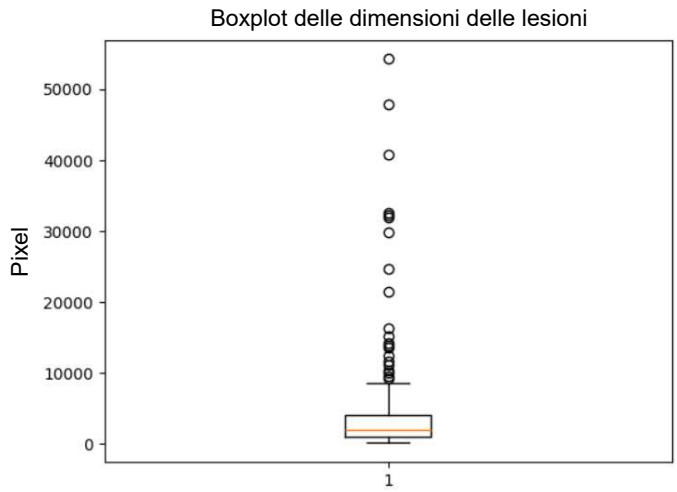


Figura 3.1: Boxplot delle dimensioni delle lesioni del dataset. Il plot tiene conto solo delle dimensioni delle lesioni calcolate facendo uso delle maschere manuali ridimensionate utilizzando un fattore di resampling di (0.5, 0.5, 3).

Nella tabella 3.2 è riportato, per ciascuna classe ISUP >1, il numero di casi per i quali è stata fornita la maschera manuale ("human_expert") o esclusivamente la maschera AI.

Classe	human_expert	AI
2	131	103
3	51	48
4	20	20
5	18	34

Tabella 3.2: Tabella del numero di casi positivi (ISUP > 1) del dataset con annotazioni manuali (human_expert) o esclusivamente annotazioni AI.

Nella figura 3.2 sono riportate le tabelle descrittive delle caratteristiche del dataset ricavate dal report dignostico fornito (file "marksheet.csv").

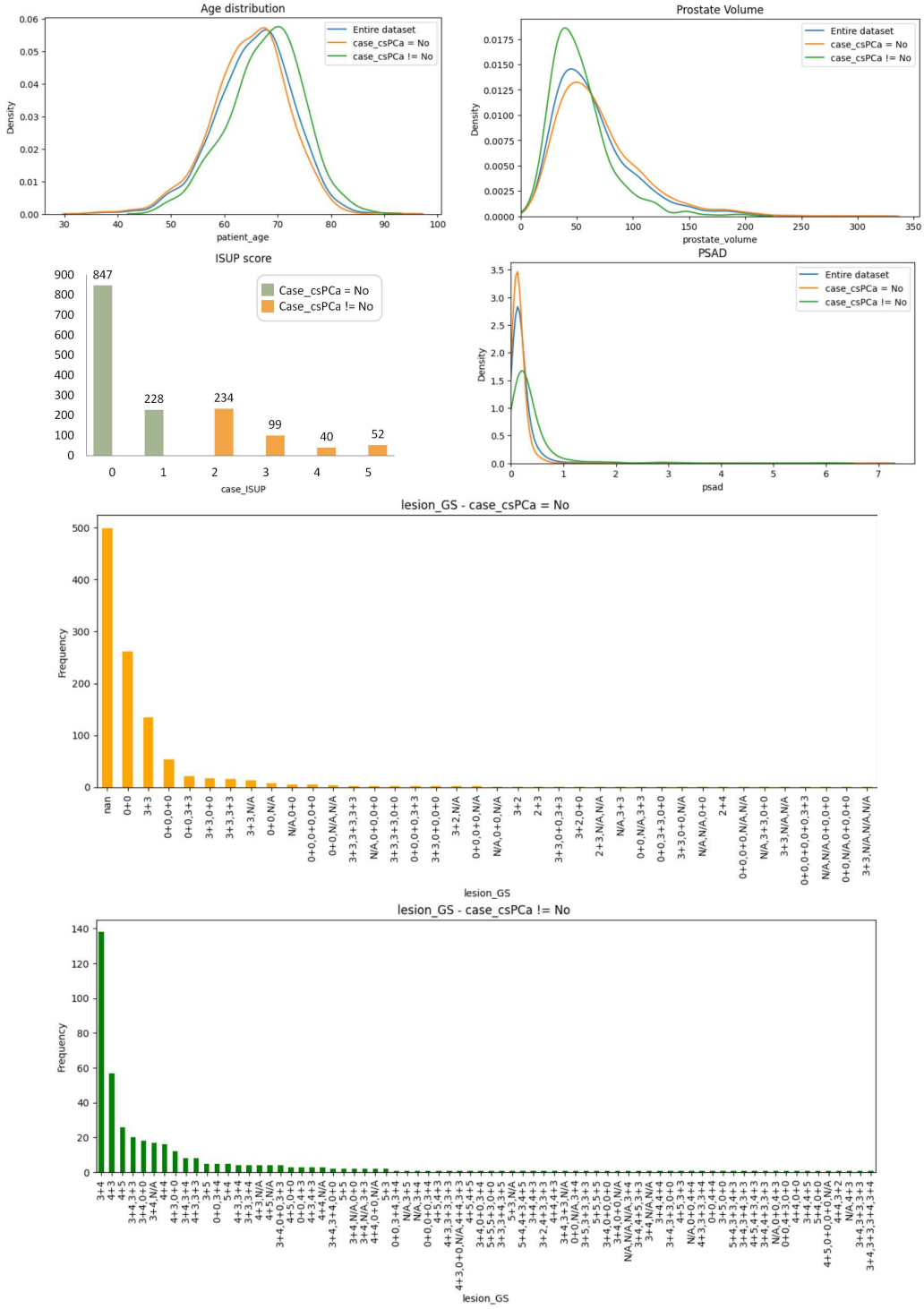


Figura 3.2: Analisi preliminare delle caratteristiche del dataset

Dall'analisi preliminare del dataset non emergono differenze significative legate all'età dei pazienti. I casi positivi mostrano tipicamente un valore di Psad superiore rispetto ai pazienti negativi. Il dataset risulta sbilanciato, con un numero maggiore di casi positivi con ISUP = 2 rispetto a quelli con ISUP 3, 4 e 5. Si osserva inoltre che la dimensione del volume della prostata dei casi classificati positivi è generalmente inferiore rispetto a quelli classificati come negativi. Per la maggior parte dei casi negativi, il punteggio di Gleason non è riportato (Nan). I casi positivi (con ISUP $>$ 1) mostrano punteggi di Gleason generalmente più elevati rispetto a quelli dei casi negativi.

Il dataset è costituito da casi acquisiti tra il 2011 e il 2021, di conseguenza potrebbero esserci delle differenze inter-paziente dovute ai progressi tecnologici e ai protocolli di acquisizione utilizzati in questo arco temporale.

Le immagini (T2W, DWI, ADC) di ciascun caso del set pubblico non sono registrate tra loro. Anche se nella maggior parte dei casi le scansioni sono allineate in modo ragionevole, in alcuni di essi si osservano deviazioni significative.

È importante evidenziare che i valori di intensità assoluta delle scansioni ADC fornite non sono clinicamente significativi a causa di protocolli di acquisizione non standardizzati e ridimensionamenti incoerenti delle immagini tra i diversi centri [35]. Inoltre, i valori ADC assoluti possono variare notevolmente a seconda di diversi fattori, come la forza del magnete, il valore b selezionato e la variabilità inter-paziente [34].

Le annotazioni delle lesioni del csPCa sono state effettuate utilizzando le sequenze bpMRI assiali (T2W, DWI, ADC) mediante *ITK-SNAP v3.80*. Tuttavia, a seconda dell'annotatore o del centro, alcune sono state create con la risoluzione spaziale e l'orientamento dell'immagine T2W, mentre altre con la risoluzione e l'orientamento delle immagini DWI/ADC. Tutte le annotazioni sono state convertite e fornite con le stesse dimensioni e risoluzione spaziale delle corrispondenti immagini T2W.

I pazienti con MRI negativa (lesioni benigne o portatrici di PI-RADS 1-2) generalmente non vengono sottoposti a biopsie o prostatectomia radicale e mancano di prove istologicamente confermate per l'assenza di csPCa. Tuttavia, è importante sottolineare che una piccola percentuale ($<$ 1% presso RUMC [36]) di csPCa potrebbe non essere rilevata. Inoltre le biopsie possono essere soggette a sottocampionamento del csPCa, soprattutto in presenza di lesioni di piccole dimensioni [37].

3.2 Descrizione del metodo implementato

Nella figura 3.3 è illustrato il flow chart descrittivo dell'architettura dell'algoritmo implementato.

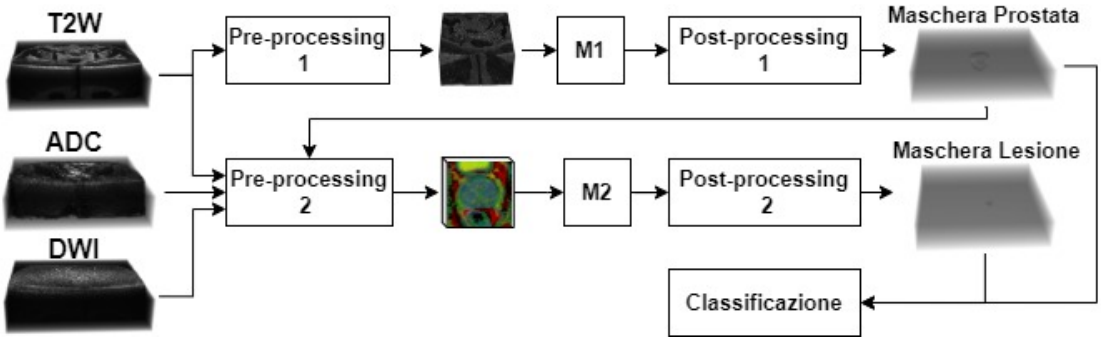


Figura 3.3: Flow chart dell'algoritmo.

Questa è costituita da 3 blocchi principali:

1. Blocco di segmentazione della prostata (figura: 3.4): la scansione T2w del paziente viene elaborata (*Pre-processing 1*) e fornita in input al modello *M1* che ha come obiettivo quello di generare la maschera prostatica. Questa viene sottoposta alle operazioni di *post-processing 1* e successivamente fornita in input al blocco di segmentazione delle lesioni tumorali.

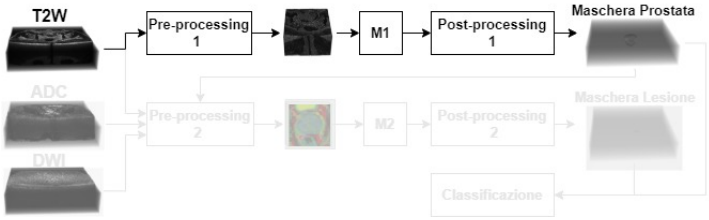


Figura 3.4: Flow chart dell'algoritmo: blocco di segmentazione della prostata.

2. Blocco di segmentazione delle lesioni tumorali (figura: 3.5): le immagini T2w, ADC, DWI e la maschera prostatica vengono utilizzate per generare, mediante la pipeline di *pre-processing 2*, un volume 4D che viene fornito in input al *modello M2* che ha come obiettivo quello di generare le maschere delle eventuali lesioni tumorali presenti nella prostata. Questa viene sottoposta alle operazioni di *post-processing 2* e successivamente fornita in input al blocco finale di classificazione.

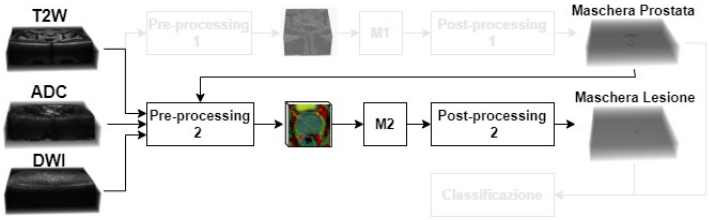


Figura 3.5: Flow chart dell'algoritmo: blocco di segmentazione delle lesioni.

3. Blocco di classificazione (figura: 3.6): fa uso delle maschere generate per effettuare la diagnosi di positività al csPCa del paziente.

Figura 3.6: Flow chart dell'algoritmo: blocco di classificazione.

3.3 Hardware/Software utilizzato

In questa sotto sezione sono descritti gli strumenti Software e Hardware utilizzati durante lo sviluppo della tesi.

Risorse Hardware

Il server fornito dall'azienda SynbrAIn per lo sviluppo dell'algoritmo è dotato di una Unità di Elaborazione Grafica (GPU) della serie NVIDIA GeForce RTX 3090, con una capacità di allocazione di 24 GB. Quest'ultima svolge un ruolo determinante nell'addestramento dei modelli. Tuttavia, questa configurazione si è rivelata appena sufficiente per la complessità del compito in questione, limitando lo sviluppo di modelli con architetture più complesse e/o profonde rispetto a quelle addestrate.

Docker

Docker è una piattaforma open-source che semplifica la creazione, la distribuzione e l'esecuzione di applicazioni tramite contenitori software. Un

contenitore Docker (o container) è un'unità standardizzata che contiene tutti gli elementi essenziali per eseguire un'applicazione, comprese librerie, file di configurazione e codice. Questo approccio agevola lo sviluppo e la distribuzione dei microservizi, assicurando la riproducibilità su qualsiasi ambiente, indipendentemente dall'hardware sottostante [51]. Per lo sviluppo di questa tesi si è fatto uso di Docker per creare un ambiente di sviluppo sul server aziendale. Nello specifico si sono definite in un Dockerfile le istruzioni per generare un ambiente virtuale, denominato immagine, insieme a tutte le dipendenze necessarie per l'esecuzione del servizio. Questo ha permesso di standardizzare e automatizzare il processo di configurazione dell'ambiente di sviluppo.

Jupyter Notebook

Per lo sviluppo del codice si è fatto uso di Jupyter Notebook, un ambiente di sviluppo interattivo basato sul web che consente agli utenti di creare e condividere documenti contenenti codice eseguibile, testo formattato e rappresentazioni grafiche.

Monai

MONAI [52], acronimo di Medical Open Network for AI, è una libreria oper-source progettata per affrontare le sfide nel campo dell'elaborazione delle immagini mediche. Essa fornisce una vasta gamma di strumenti e funzionalità ottimizzate per la manipolazione, l'analisi e l'apprendimento automatico di dati, semplificando notevolmente il processo di sviluppo e valutazione dei modelli. La sua architettura modulare e la compatibilità con i principali framework di deep learning come PyTorch l'hanno resa una scelta ideale per lo sviluppo di questa tesi.

3.4 Metriche di valutazione

Per valutare le performance si è fatto uso delle seguenti metriche: Dice, Distanza di Hausdorff al 95° percentile, Indice di Dice del Volume Relativo (RVD), Precisione, Sensibilità. Per il calcolo delle metriche si sono prima definiti (figura 3.7):

1. VP (Veri Positivi) : pixel positivi predetti come positivi;

2. VN (Veri Negativi) : pixel negativi predetti come negativi;
3. FP (Falsi Positivi) : pixel negativi predetti come positivi;
4. FN (Falsi Negativi) : pixel positivi predetti come negativi.

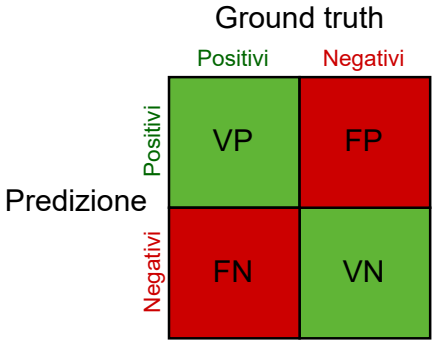


Figura 3.7: Tabella delle previsioni dell'algoritmo rispetto al Ground Truth. VP indica i Veri Positivi, VN i Veri Negativi, FP i Falsi Positivi e FN i Falsi Negativi.

Dice:

$$DICE = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.1}$$

Il Dice può assumere un valore compreso nell'intervallo [0, 1] e consente di valutare la precisione della segmentazione della maschera generata, rispetto a quella fornita, misurandone il grado di sovrapposizione. Un valore più alto indica una maggiore concordanza tra le segmentazioni, con valori vicini a 1 che indicano una sovrapposizione significativa. Tuttavia, non fornisce informazioni specifiche sulla sovrastima o sottostima delle segmentazioni.

Distanza di Hausdorff al 95° percentile:

$$\boldsymbol{HD_{95}} = d_h(X,Y) = max\{d_{XY}, d_{YX}\} = max\left\{P_{95_{x\in X}} min_{y\in Y} d(x,y), P_{95_{y\in Y}} min_{x\in X} d(x,y)\right\} \tag{3.2}$$

La distanza di Hausdorff è una stima peggiorativa. Essa rappresenta il massimo valore della minima distanza dei punti della curva X rispetto alla curva Y, e di Y rispetto a X; ossia il massimo errore di posizionamento che si ha dalla curva ottenuta rispetto alla curva di riferimento. Il calcolo al 95° percentile (HD_{95}) permette di ridurre il contributo di eventuali outliers.

Essa fornisce una stima robusta della massima discrepanza tra i punti dei due insiemi, utile per valutare la discrepanza spaziale tra le segmentazioni. Un valore più basso indica una maggiore concordanza tra le due segmentazioni.

RVD

$$\text{RVD} = \frac{V_a - V_m}{V_m} \tag{3.3}$$

Dove: V_m= Vol. Segmentazione Manuale, V_a = Vol. Seg. Automatica. RVD è una metrica utilizzata per valutare se una segmentazione sovra o sotto stima il volume di riferimento. Tuttavia, non fornisce informazioni sulla sovrapposizione delle segmentazioni. Un RVD pari a zero indica una perfetta coincidenza tra il volume della maschera generata e quella di riferimento, mentre valori positivi o negativi indicano rispettivamente sovra o sotto-stime del volume.

Precisione:

$$\text{Precisione} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.4}$$

La precisione consente di valutare, con un valore compreso nell'intervallo [0, 1], quanti dei pixel, individuati come positivi, siano corretti. Un valore di Precisione alto indica che il modello classifica correttamente la maggior parte dei pixel positivi, riducendo al minimo i falsi positivi.

Sensibilità:

$$\text{Sensibilità} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.5}$$

Anche conosciuta come True Positive Rate (TPR) o Recall, è una misura della capacità di un modello di individuare e classificare correttamente tutti i pixel positivi. Un valore più alto indica una capacità maggiore del modello di rilevare correttamente i pixel positivi, il che è particolarmente importante in applicazioni mediche e diagnostiche.

3.5 Deep Learning

Il Deep Learning (DL), insieme al Machine Learning (ML), è un sottoinsieme del campo dell'intelligenza artificiale (AI) (figura 3.8).

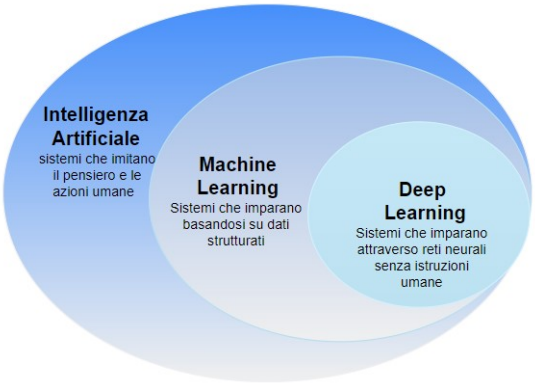


Figura 3.8: AI vs ML vs Deep-Learning [40].

Mentre il Machine Learning tradizionale richiede un'attenta estrazione manuale delle features, il DL, avvicinandosi al meccanismo di apprendimento del cervello umano, effettua questo processo in maniera automatica (figura 3.9) [40].

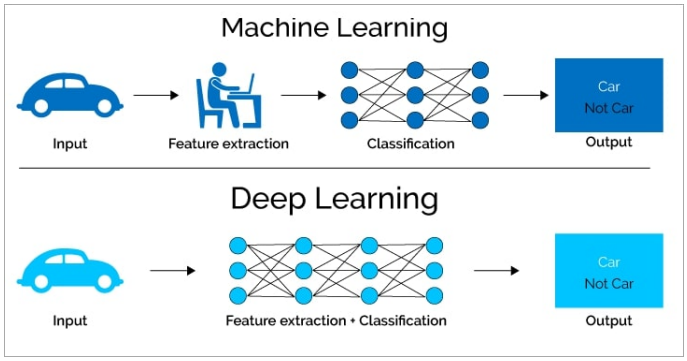


Figura 3.9: Deep Learning vs Machine Learning [40].

Le reti neurali artificiali sono strutture flessibili utilizzate per l'elaborazione e l'apprendimento da dati complessi. Queste sono composte da neuroni artificiali, ovvero da modelli matematici che elaborano un insieme di input, ciascuno ponderato da un coefficiente specifico e da un termine di bias, effettuano combinazioni lineari e applicano funzioni di attivazione al fine di generare un output [41] (figura 3.10).

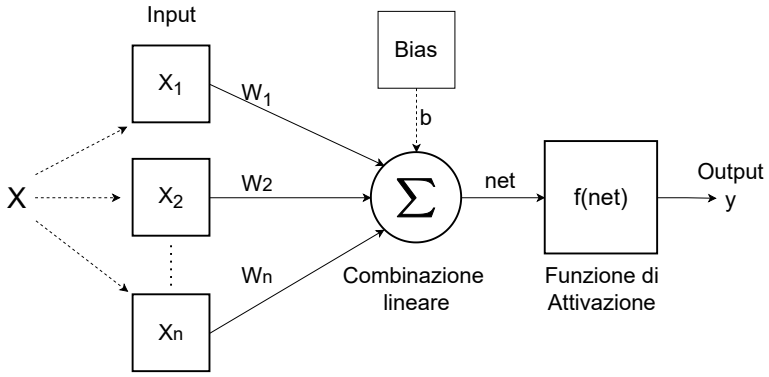


Figura 3.10: Modello matematico del neurone artificiale [41].

A differenza delle reti neurali artificiali, il cervello umano mostra una capacità di adattamento e apprendimento continuo. Mentre una rete neurale restituisce sempre lo stesso output per uno specifico input dopo l'addestramento, il cervello umano può generare output diversi per lo stesso input in contesti o momenti differenti. Questa plasticità e capacità di apprendimento continuo del cervello umano sono oggetto di ricerca nel campo dell'intelligenza artificiale [48].

Un'Artificial Neural Network (ANN) [42] (figura 3.11) è costituita da una serie di unità di elaborazione note come neuroni artificiali, organizzati in:

- Strato di input: riceve i dati in ingresso alla rete.

- Strati nascosti "Hidden layers": sono responsabili dell'elaborazione dei dati. All'aumentare del numero di strati aumenta la "profondità" della rete ed il numero di connessioni.

- Strato di output: fornisce i risultati finali della rete.

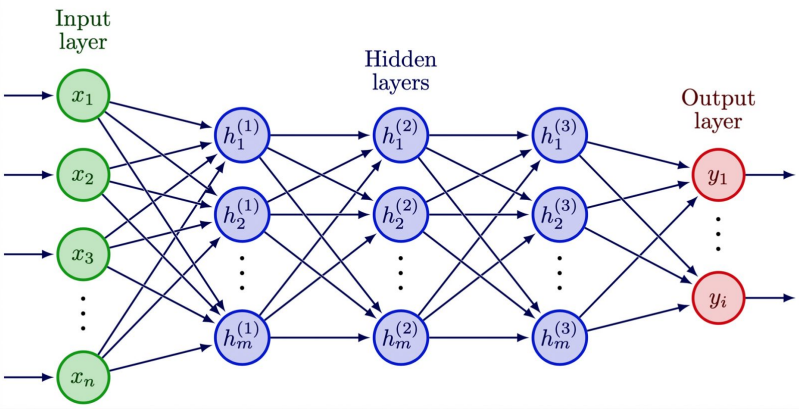


Figura 3.11: Schema strutturale di una Artificial Neural Network [42].

Le connessioni tra i neuroni nei diversi strati sono caratterizzate da pesi che determinano l'importanza delle informazioni trasmesse. Durante il processo di addestramento (figura 3.12), viene utilizzata una Loss function per valutare l'errore tra l'output predetto e quello atteso. Sulla base di questa valutazione viene effettuata una regolazione dei pesi al fine di minimizzare l'errore calcolato [41].

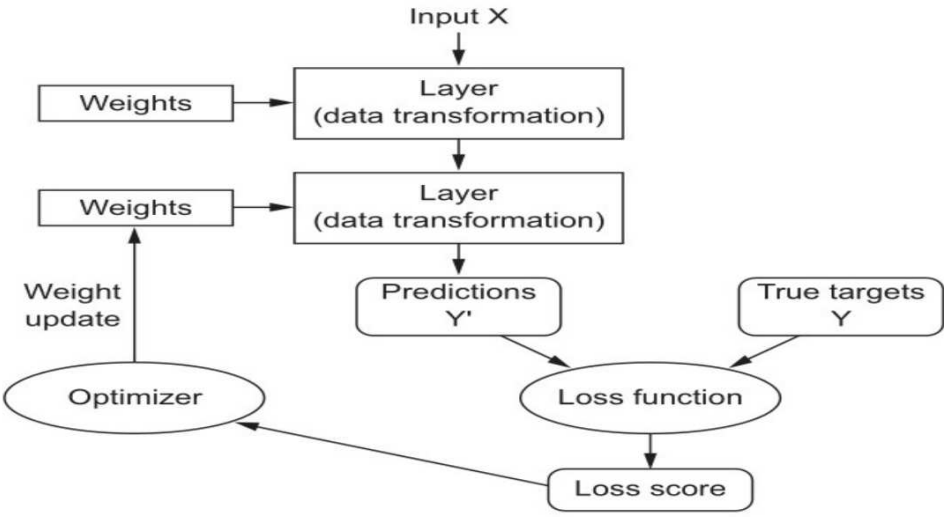


Figura 3.12: Flow-chart processo di addestramento [41].

Le Convolutional Neural Network (CNN) [54] rappresentano una delle architetture più diffuse per l'elaborazione delle immagini. Queste reti sono progettate per estrarre caratteristiche a più livelli dalle immagini tramite l'applicazione di filtri convoluzionali. Grazie alla loro struttura e alle funzioni implementate, le CNN sono in grado di eseguire una vasta gamma di

compiti, tra cui il rilevamento, la classificazione e la segmentazione. Ciò le rende estremamente versatili e adattabili. Nel contesto di questo progetto di tesi, è stata adottata l'architettura UNet, appartenente alla famiglia delle CNN, per via della sua semplicità e dei bassi costi computazionali.

Architettura UNet

L'architettura UNet si distingue per la sua caratteristica forma a "U", che integra l'encoder (attraverso convoluzioni) con il decoder (tramite upsampling) (figura 3.13).

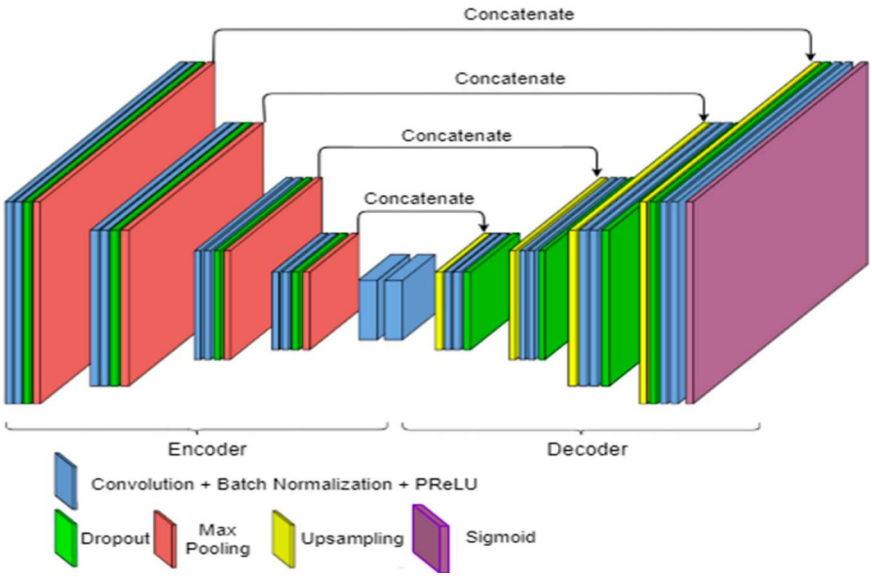


Figura 3.13: Esempio architettura UNet [44].

Encoder

Questo blocco comprende una serie di strati convoluzionali che riducono progressivamente la dimensione spaziale dell'input. L'utilizzo delle convoluzioni consente alla rete di catturare pattern e caratteristiche significative presenti nell'immagine nelle diverse scale, partendo dai dettagli a più basso livello fino a quelli a più alto livello. Vengono definiti tre parametri:

- la dimensione del kernel: rappresenta la dimensione della finestra usata per la convoluzione;

- stride: indica di quanto il kernel si sposta lungo l'immagine durante il processo di convoluzione;
- zero-padding: consiste nell'aggiungere o meno una cornice di zeri intorno ai bordi dell'immagine prima di effettuare la convoluzione al fine di ottenere un output della stessa dimensione dell'input.

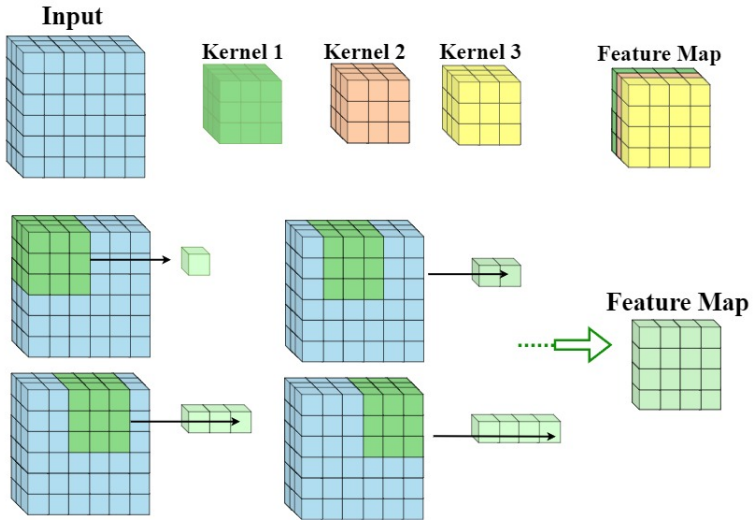


Figura 3.14: Esempio di operazione di Convoluzione 3D. L'esempio vede come input una matrice 6x6x3, kernel di dimensioni 3x3x3 e stride 1.

Di seguito è riportata una rappresentazione numerica dell'operazione di convoluzione. Per agevolarne la comprensione, si è scelto di presentare un esempio bidimensionale (il funzionamento in 3D è analogo).

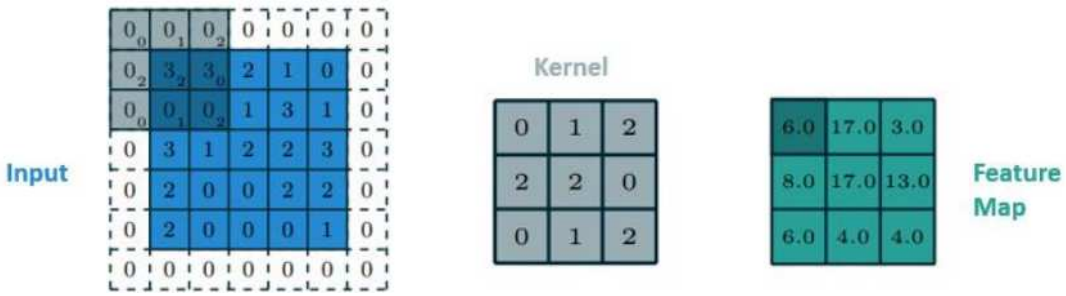


Figura 3.15: Rappresentazione numerica dell'operazione di convoluzione. Nell'esempio si è fatto uso di un immagine di input 5×5, kernel 3×3, zero padding con dimensione 1 e stride di 2. [43].

Le operazioni di pooling hanno come obiettivo quello di ridurre gradualmente le dimensioni spaziali delle features map. Questo processo funge da

feature selection e consente di mantenere solo le informazioni più rilevanti. Esistono diversi meccanismi di pooling, tra i più comuni vi sono (3.16):

- max pooling: seleziona il valore massimo in una regione specifica della features map, preservando le caratteristiche dominanti dell'immagine;
- average pooling: calcola la media dei valori in una regione specifica della features map;
- sum pooling: effettua la somma dei valori in una regione specifica della features map, utile per mantenere informazioni sulla magnitudo o la distribuzione dei valori presenti nell'immagine.

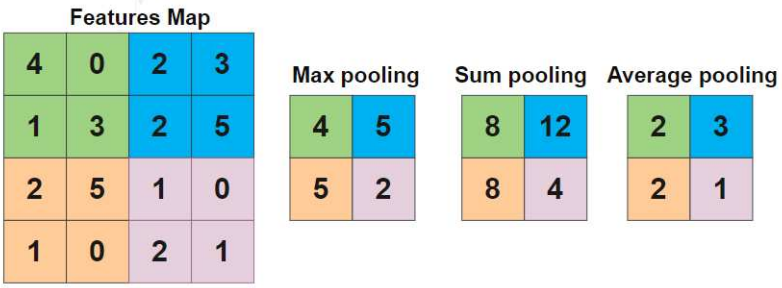


Figura 3.16: Rappresentazione numerica delle operazioni di pooling.

Decoder

Dopo l'encoding, la UNet passa alla fase di decodifica, che coinvolge l'upsampling delle feature map attraverso strati di upsampling e convoluzioni trasposte. Questa operazione consente alla rete di ricostruire l'immagine segmentata incorporando le informazioni apprese durante l'encoding. L'upsampling aiuta a ripristinare la risoluzione spaziale dell'immagine, consentendo alla rete di recuperare i dettagli persi durante la fase di riduzione dimensionale.

Funzione di attivazione

Le funzioni di attivazione svolgono un ruolo cruciale nel processo di apprendimento della rete, introducendo non linearità e complessità nel modello. Queste funzioni determinano l'attivazione o meno di un neurone in base

alla somma ponderata dei suoi input, contribuendo così alla capacità della rete di apprendere e rappresentare pattern complessi nei dati. La PReLU, o Parametric Rectified Linear Unit, è una funzione di attivazione utilizzata comunemente nelle reti neurali artificiali. A differenza della ReLU (Rectified Linear Unit), che ha un valore costante pari a zero per tutti i valori negativi dell'input, la PReLU ha una pendenza variabile per i valori negativi (figura 3.17). La formula matematica della PReLU è definita come:

$$f(x) = \begin{cases} y & \text{se } y > 0 \\ ay & \text{se } y \leq 0 \end{cases}$$

Dove a è un parametro che regola i valori negativi, consentendo un miglior adattamento ai dati e facilitando l'apprendimento di rappresentazioni più complesse. In questo progetto è stato adottato il valore predefinito impostato dalla libreria MONAI pari a 0,25.

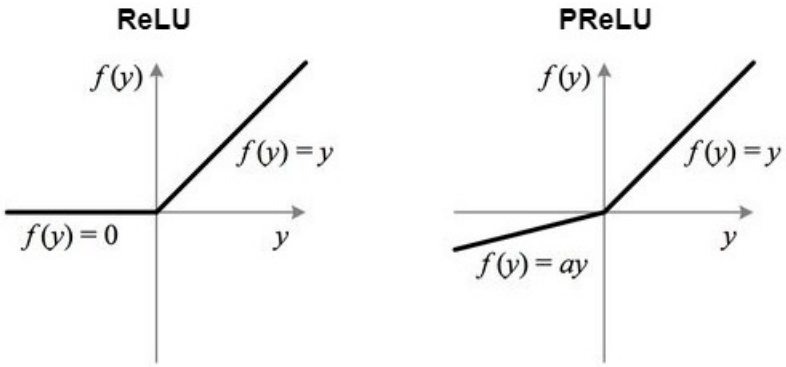


Figura 3.17: Funzione di attivazione *ReLU* vs *PReLU* [49].

L'uso della PReLU anziché la ReLU può contribuire a evitare il problema della "morte dei neuroni" (neurons dying) durante l'addestramento, in cui alcuni neuroni non aggiornano più i loro pesi a causa di valori negativi costanti nell'input, migliorando così la capacità della rete di adattarsi a una varietà più ampia di dati.

Loss function

La loss function, o "funzione di costo", è essenziale nell'addestramento delle reti neurali per compiti di apprendimento supervisionato. Durante il backpropagation la loss function calcola la discrepanza tra l'output predetto e l'etichetta di ground truth e, sulla base di essa, l'ottimizzatore aggiorna i

parametri del modello. L'obiettivo è minimizzare il valore di loss in modo che la segmentazione prodotta dalla rete si avvicini il più possibile alla segmentazione manuale, migliorando così l'accuratezza complessiva del modello. Tra le funzioni più utilizzate vi sono: DiceLoss, Binary Cross-Entropy Loss (BCELoss), DiceFocalLoss.

DiceLoss: calcola il coefficiente di similarità tra la maschera di segmentazione predetta e quella ground truth.

$$\text{DiceLoss} = 1 - \frac{2 \times \sum_i^N p_i \times g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{3.6}$$

- dove p_i rappresenta la probabilità predetta di appartenenza della classe i
- g_i rappresenta la probabilità effettiva di appartenenza alla classe i.
- N indica il numero totale di classi.
- Vantaggi: affronta meglio gli squilibri di classe e tiene conto della localizzazione dei pixel.
- Svantaggi: é sensibile a predizioni parziali o sfalsate. Pertanto, se il modello produce previsioni incomplete o imprecise, la Dice Loss potrebbe non essere in grado di penalizzarle in modo efficace.

Binary Cross-Entropy Loss (BCELoss): calcola la discrepanza tra le probabilità predette e le etichette binarie di ground truth.

$$\text{BCELoss} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \right) \tag{3.7}$$

Dove:

- N è il numero totale di pixel nell'immagine;
- y_i è l'etichetta binaria di ground truth per il pixel i;
- p_i è la probabilità predetta per il pixel i.
- Vantaggi: Semplice da implementare e calcolare.
- Svantaggi: Non tiene conto della struttura spaziale delle previsioni.

Rispetto alla DiceLoss, la BCELoss potrebbe soffrire di un problema di squilibrio di classe, specialmente se le classi positive sono rare rispetto alle classi negative.

FocalLoss: progettata per affrontare il problema dello squilibrio di classe.

$$\text{FocalLoss}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{3.8}$$

- dove p_t rappresenta la probabilità predetta per la classe vera, cioè:

$$p_t = \begin{cases} p & \text{se la classe è positiva} \\ 1-p & \text{se la classe è negativa} \end{cases}$$

- α_t: fattore di bilanciamento tra classi utilizzato per dare maggior peso alla classe meno rappresentata.
- γ: parametro che riduce il peso dei campioni ben classificati e aumenta quello dei campioni difficili (valori tipici sono $\gamma = 2$).
- Vantaggi: affronta efficacemente lo squilibrio di classe dando maggior peso ai campioni difficili da classificare. Questo porta il modello a concentrarsi sui campioni meno frequenti e difficili, migliorando la capacità di generalizzazione su dati sbilanciati.
- Svantaggi: l'implementazione richiede la scelta dei parametri α e γ, che possono necessitare di ottimizzazioni specifiche per ogni dataset. Inoltre, se non tarati correttamente, possono portare a una penalizzazione eccessiva dei campioni ben classificati, compromettendo la performance complessiva del modello.

DiceFocalLoss: combinazione tra la DiceLoss e la FocalLoss, è progettata per affrontare il problema degli squilibri di classe. La formula del DiceFocalLoss:

$$DiceFocalLoss = (1-\beta) \cdot DiceLoss + \beta \cdot FocalLoss \tag{3.9}$$

- $DiceLoss$: misura della sovrapposizione tra le maschere predette e le maschere di ground truth.
- $FocalLoss$: progettata per ridurre l'importanza degli esempi ben classificati e concentrarsi sui casi più difficili.

- β: parametro che regola il peso relativo della DiceLoss rispetto alla FocalLoss. Un valore più elevato enfatizza maggiormente la FocalLoss, mentre un valore più basso dà più peso alla DiceLoss.

- Vantaggi: affronta meglio gli squilibri di classe, concentrandosi sui casi più difficili.

- Svantaggi: computazionalmente più onerosa.

Ottimizzatore

Nel processo di addestramento dei modelli di segmentazione della prostata e delle lesioni tumorali, si è fatto uso dell'ottimizzatore Adam per regolare i pesi delle reti neurali durante la fase di addestramento. Esso è uno dei più utilizzati in quanto particolarmente adatto per l'addestramento di reti neurali profonde su grandi set di dati, come quelli utilizzati in questo progetto di tesi. L'ottimizzatore Adam aggiorna i pesi del modello in base al gradiente della funzione di loss. Durante la fase di backpropagation, il gradiente viene calcolato rispetto a ciascun parametro del modello e l'ottimizzatore utilizza questa informazione per regolare i pesi in modo da minimizzare la funzione di loss complessiva.

Due iperparametri chiave sono il tasso di apprendimento *learning rate* e il peso della penalità L2 *weight decay*. Il learning rate regola la dimensione dei passi durante l'ottimizzazione dei pesi del modello influenzando la velocità di convergenza e la stabilità dell'addestramento. Il weight decay agisce come una tecnica di regolarizzazione limitando la complessità del modello mediante la penalizzazione dei pesi grandi. Una scelta accurata di questi parametri è essenziale per evitare fenomeni come l'overfitting o l'underfitting.

Per lo sviluppo delle reti si è fatto uso del modello "UNet" presente nella libreria Monai; questo è stato modificato in base all'obiettivo e alle performance ottenute.

Capitolo 4

Sviluppo dell'algoritmo di segmentazione della prostata

Il processo di segmentazione della prostata, evidenziato in verde nel flow-chart 4.1, costituisce la componente iniziale dell'algoritmo implementato per l'analisi delle scansioni MRI della prostata.

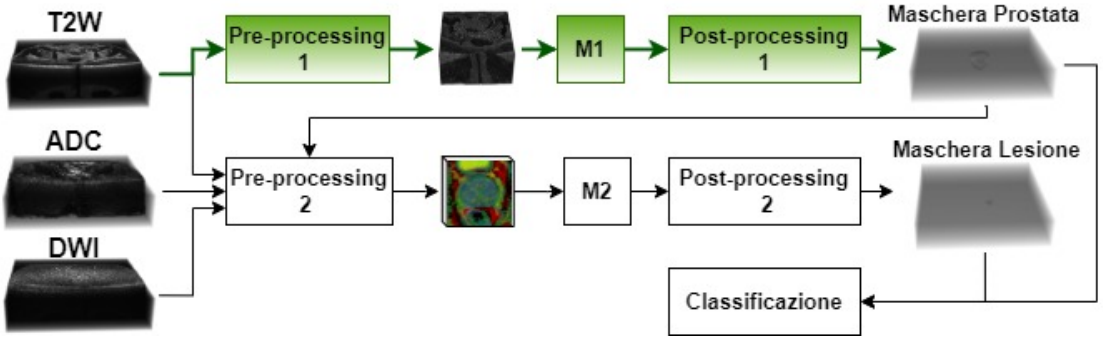


Figura 4.1: Flow chart dell'architettura dell'algoritmo. Le componenti descritte nel capitolo seguente sono evidenziate in verde.

Il suo scopo è quello di identificare e segmentare la prostata (figura 4.2).

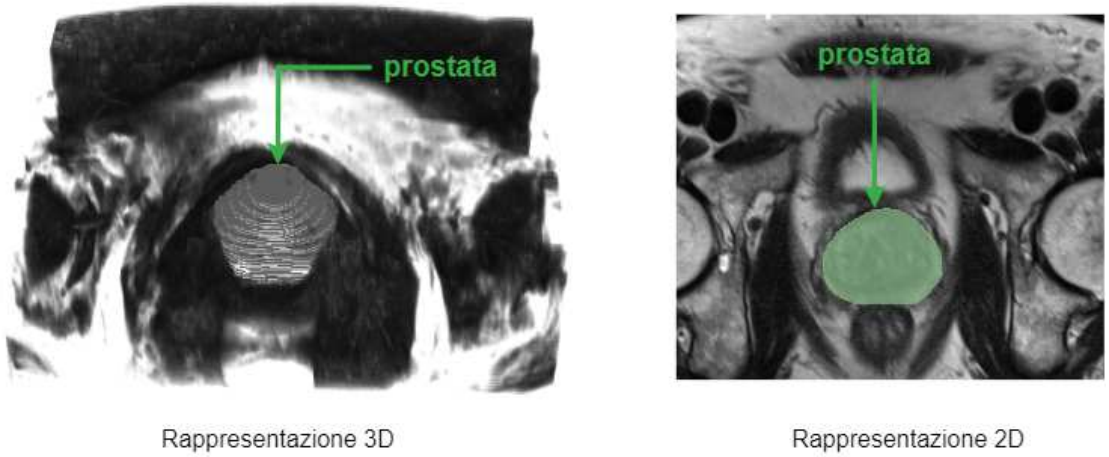


Figura 4.2: Rappresentazione 3D e 2D della segmentazione della prostata. Figure ottenute sovrapponendo la maschera di segmentazione della prostata sulla scansione T2w.

La maschera generata assume un ruolo essenziale nella fase di *pre-processing 2* dei volumi dati in input al modello *M2* per la segmentazione delle lesioni tumorali. L'implementazione del modello *M1* ha come *vantaggio* principale quello di rendere l'algoritmo robusto alle variazioni di posizionamento spaziale della prostata che possono verificarsi durante l'acquisizione in contesti reali (figura 4.3).

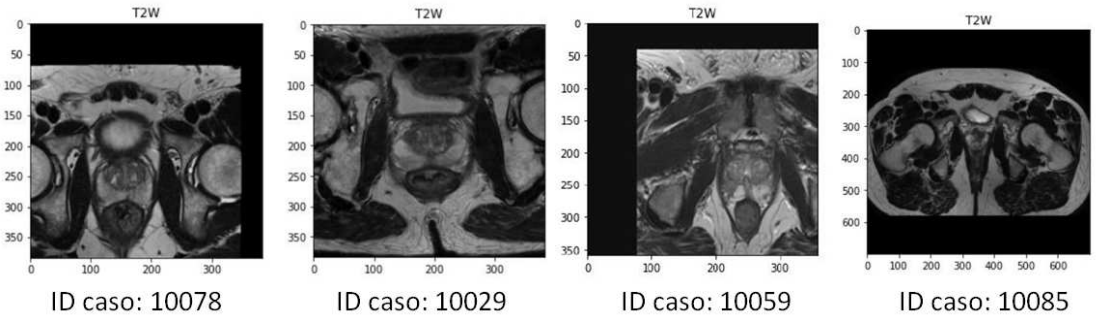


Figura 4.3: Confronto delle scansioni T2w di pazienti diversi. Le scansioni sono state sottoposte a resampling con un fattore di scala di (0.5, 0.5, 3).

Questo capitolo analizza il processo completo di sviluppo e ottimizzazione dell'algoritmo di segmentazione della prostata, valutando criticamente ciascuna fase e le performance ottenute (flow chart 4.4).

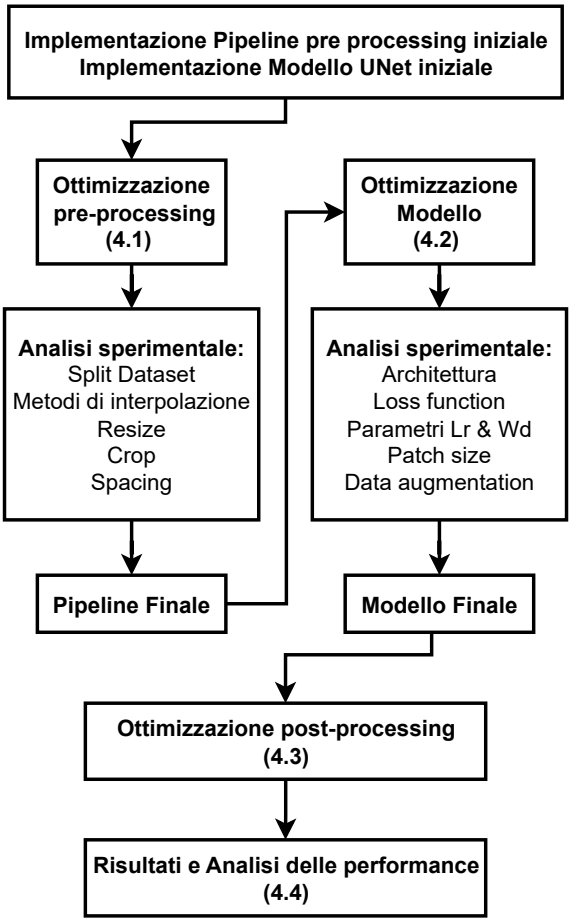


Figura 4.4: Flow chart descrittivo del capitolo 4.

4.0.1 Pipeline di *Pre-Processing 1* iniziale

Sulla base della letteratura esistente, si è delineata la pipeline di pre-processing iniziale illustrata nella figura 4.5. Per questa task si è scelto di lavorare con i volumi al fine di ottimizzare i tempi computazionali. Lavorare in 2D, infatti, richiederebbe i processi di estrazione delle slice e ricostruzione del volume, determinando un aumento significativo del tempo complessivo di inferenza. Inoltre, l'utilizzo del volume, sebbene più oneroso a livello computazionale, consente l'estrazione di features tridimensionali migliorando, almeno sulla base di considerazioni teoriche [57], la qualità della segmentazione.

Le maschere sono state sottoposte alle stesse operazioni di pre-processing implementate per la scansione T2w fatta eccezione per l'operazione di normalizzazione che è sostituita dall'operazione di *Thresholding con soglia 0,5*.

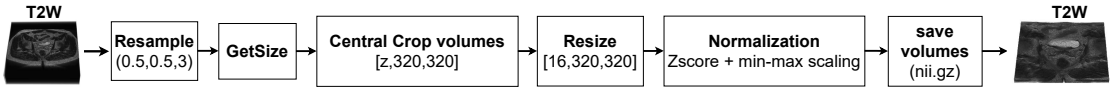


Figura 4.5: Pipeline di pre-processing iniziale.

4.0.2 Modello UNet *M1* iniziale

Per una valutazione quantitativa dell'impatto delle operazioni di pre-processing sulle performance, è stato impiegato un modello UNet i cui iperparametri interni, riportati nella tabella 4.1, sono stati ottimizzati nella fase successiva del lavoro (sezione 4.2).

Tabella 4.1: Tabella riassuntiva dei parametri e delle funzioni implementate per ottenere il modello UNet M1 di partenza impiegato per l'analisi delle operazioni di pre-processing.

Architettura	Unet3D: (spatial_dims=3, in_channels=1, out_channels=1, channels=(16,32,64,128,256)) strides=((2,2,2),(2, 2, 1),(2, 2, 1),(2, 2, 2)), num_res_units=2, norm=Norm.BATCH)
Patch	spatial_size:(192,192,16) pos=2, neg=0, num_samples=2
Loss function	DiceLoss (to_onehot_y=True, softmax=True)
Optimizer	torch.optim.Adam(model.parameters(), lr=1e-03, weight_decay=5e-06)
Batch size	8
Patience	10

4.1 *Pre-processing 1*: ottimizzazione

In questa sezione viene trattato il processo di ottimizzazione della pipeline di *pre-processing 1*. L'obiettivo è definire una pipeline robusta alle variazioni presenti in un dataset reale, garantendo al contempo la generazione di volumi di buona qualità da utilizzare come input per il modello.

4.1.1 Tipologia della scansione usata

Per questa task si è fatto uso soltanto delle scansioni T2w. Questa modalità di scansione infatti, come già discusso nella sezione introduttiva (sezione 1.3), permette di delineare in maniera efficace ed efficiente la ghiandola prostatica. Inoltre, l'utilizzo di una sola immagine comporta diversi vantaggi:

1. implementazione di un modello 3D, più leggero rispetto alla controparte 4D (minore utilizzo di memoria CPU e GPU);
2. semplificazione della pipeline di pre-processing;
3. riduzione del tempo di inferenza.

4.1.2 Tipologia della maschera usata

Per allenare il modello UNet *M1* si è fatto uso delle maschere di "Guerbet23" fornite dagli organizzatori della challenge. Queste sono state ottenute mediante un modello di intelligenza artificiale (AI) addestrato su maschere manuali ed orientato a ottenere una segmentazione accurata della prostata, privilegiando la segmentazione anatomica rispetto alla precisione zonale fine *(Dice 0,8968 ± 0,0547)* [45]. Pertanto, il modello implementato risentirà di un *bias iniziale* nelle performance da attribuire alla qualità delle maschere utilizzate per il suo addestramento. Le maschere in questione delineano l'intera ghiandola prostatica (figura 4.2).

4.1.3 Divisione del dataset

La divisione del dataset in Train, Validation e Test set riveste un ruolo importante nella fase di costruzione del modello. Durante l'analisi sperimentale sono state eseguite diverse prove per ottimizzare le dimensioni dei

set e garantire che ognuno avesse una numerosità adeguata. I set di Train e Validation sono stati formati includendo solo i casi *negativi* (ISUP < 2), mentre i casi *positivi* (ISUP > 1) sono stati inseriti soltanto nel Test set (figura 4.6). Questo per far si che il Test set finale utilizzato nella fase di inferece, che vede l'applicazione dei due modelli in cascata, sia diverso da quelli di Train e Validation utilizzati durante le fasi di costruzione dei due modelli, garantendo una valutazione robusta delle performance dell'intero algoritmo. L'inclusione dei *positivi* nel Test set mira a valutare le performance del modello con input che presentano piccole variazioni nella regione di interesse rispetto ai casi usati per il suo addestramento. In questi, infatti, la prostata mostra delle variazioni di intensità dei pixel nelle aree in cui vi è una lesione tumorale. È importante sottolineare che la divisione dei casi è stata effettuata in modo casuale e ripetibile, utilizzando un *seed* fisso pari ad 1. Questo è fondamentale per il corretto confronto delle diverse prove effettuate durante la fase sperimentale. Infatti, come evidenziato da studi scientifici, le performance potrebbero essere influenzate dalla numerosità ma anche dalla diversa distribuzione dei dati [55].

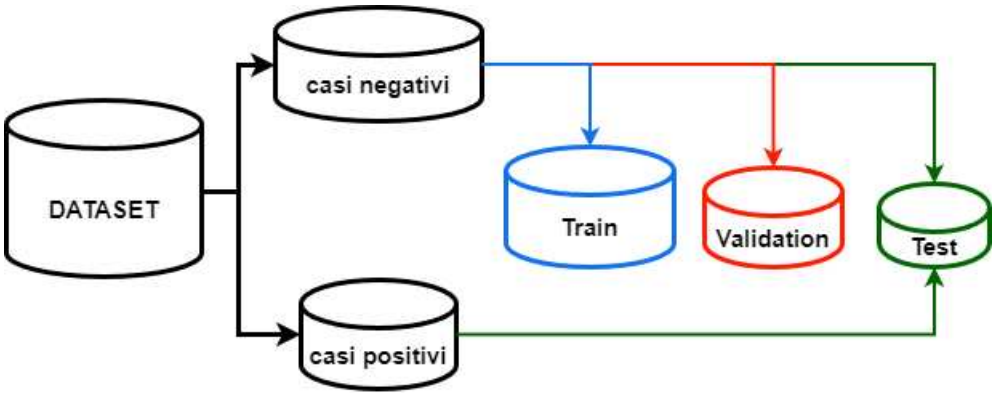


Figura 4.6: rappresentazione grafica della divisione del dataset.

4.1.4 Resample

I volumi del dataset presentano spacing diversi. Pertanto, all'inizio della pipeline di pre-processing, è essenziale implementare l'operazione di Resample la quale permette di uniformare le scansioni ad un unica risoluzione spaziale. Durante l'analisi sperimentale, sono stati esaminati i parametri impiegati per eseguire questa operazione, ovvero il *metodo di interpolazione* e lo *spacing*.

Metodi di interpolazione

Nel contesto dell'elaborazione delle immagini, l'interpolazione è un operazione matematica che permette di stimare i valori dei pixel in posizioni non discrete dell'immagine originale, generando immagini di diversa risoluzione.

Durante la fase sperimentale sono stati esaminati e confrontati due metodi comunemente utilizzati: *l'interpolazione lineare e l'interpolazione bicubica.*

L'interpolazione lineare stima il valore di un pixel intermedio utilizzando una combinazione lineare dei valori dei pixel adiacenti mediante la seguente formula:

$$I(x,y) = (1-\alpha)(1-\beta)\cdot I_{00} + \alpha(1-\beta)\cdot I_{10} + (1-\alpha)\beta\cdot I_{01} + \alpha\beta\cdot I_{11} \quad (4.1)$$

dove:

- $I(x,y)$ rappresenta il valore interpolato del pixel;
- I_{ij} sono i valori dei pixel noti più vicini al punto di interpolazione;
- α e β sono i coefficienti di interpolazione lineare.

L'interpolazione bicubica utilizza una funzione cubica per stimare i valori dei pixel intermedi mediante la formula:

$$I(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \quad (4.2)$$

dove:

- a_{ij} rappresenta i coefficienti dell'interpolazione bicubica.

Per valutare l'efficacia di questi metodi di interpolazione, si è fatto uso della metrica *Peak Signal-to-Noise Ratio (PSNR).*

Il PSNR (4.3) misura il rapporto tra l'energia del segnale e l'energia del rumore presenti nell'immagine interpolata. In questo contesto, l'energia del segnale è il valore massimo che il pixel può assumere, mentre l'energia del rumore è l'errore quadratico medio tra i pixel originali e i pixel interpolati. Solitamente rappresentato in scala logaritmica, il PSNR, è comunemente impiegato come indicatore della qualità di un'immagine. Nello specifico, un valore più elevato suggerisce una maggiore fedeltà all'immagine originale [56]. La formula per calcolare il PSNR è la seguente:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{4.3}$$

dove:

- MAX: rappresenta il valore massimo che un pixel dell'immagine può assumere;
- MSE(Mean Squared Error): è l'errore quadratico medio tra i pixel dell'immagine originale e i pixel dell'immagine interpolata.

Nell'equazione 4.4 è riportata la formula per il calcolo dell'errore quadratico medio (MSE), dove $I(i, j)$ rappresenta l'intensità del pixel nell'immagine originale e $K(i, j)$ rappresenta l'intensità del pixel nell'immagine interpolata.

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i, j) - K(i, j))^2 \tag{4.4}$$

Per le maschere è stata adottata l'interpolazione *NearestNeighbor* per preservare la loro natura binaria durante il processo di resampling. Questo metodo assegna a ciascun pixel della nuova maschera il valore del pixel più vicino nella maschera originale, evitando così la creazione di valori intermedi che potrebbero compromettere l'integrità della segmentazione.

Spacing

Lo spacing, ovvero la distanza tra i pixel lungo ciascuna dimensione dello spazio, influenza la risoluzione spaziale e la qualità complessiva delle immagini. La riduzione dello spacing determina un aumento della risoluzione spaziale, consentendo una migliore visualizzazione e interpretazione delle strutture anatomiche specialmente in presenza di quelle più piccole e complesse. Tuttavia, questo comporta anche un aumento dei costi computazionali. Al contrario, uno spacing più grande consente di ridurre i requisiti computazionali riducendo tuttavia la qualità dell'immagine. La scelta dello spacing ottimale pertanto dipende da una serie di fattori, tra cui le dimensioni delle immagini, la complessità delle strutture anatomiche di interesse e le risorse computazionali disponibili ed è dunque fondamentale valutarne l'impatto sulle performance del modello.

4.1.5 Resize

Questa operazione ha come obiettivo quello di uniformare le dimensioni dei volumi da fornire in input al modello *M1* garantendo il corretto funzionamento del codice. Per determinare le dimensioni ottimali, è stato adottato un approccio basato sulla moda della distribuzione delle dimensioni dei volumi presenti nel dataset. Durante l'analisi sperimentale, sono state considerate due diverse dimensioni per l'asse z: 16 e 32. Queste sono state selezionate per la loro caratteristica di essere divisibili per due, requisito fondamentale per il corretto funzionamento del modello. L'obiettivo è quello di valutare le dimensioni ottimali in termini di costi-benefici.

4.1.6 Crop intorno al centro della scansione

Nel contesto della segmentazione della ghiandola prostatica, l'operazione di crop ha come obiettivo quello di concentrare l'attenzione del modello su una regione di interesse più piccola al fine di migliorarne le prestazioni di predizione. Questa operazione si basa sulla costatazione che, generalmente, la prostata si trova al centro della scansione. Tuttavia, presenta un limite importante: essendo un'operazione a priori, potrebbe non essere adeguata per tutti i casi.

Durante l'analisi sperimentale si è valutato l'effetto dell'operazione di crop intorno al centro della scansione, utilizzando dimensioni fisse di 320x320 nel piano trasversale XY e una profondità z, pari al numero di slice del volume, lungo l'asse Z. Nella figura 4.7 è riportato il boxplot delle dimensioni spaziali delle scansioni T2w dell'intero dataset.

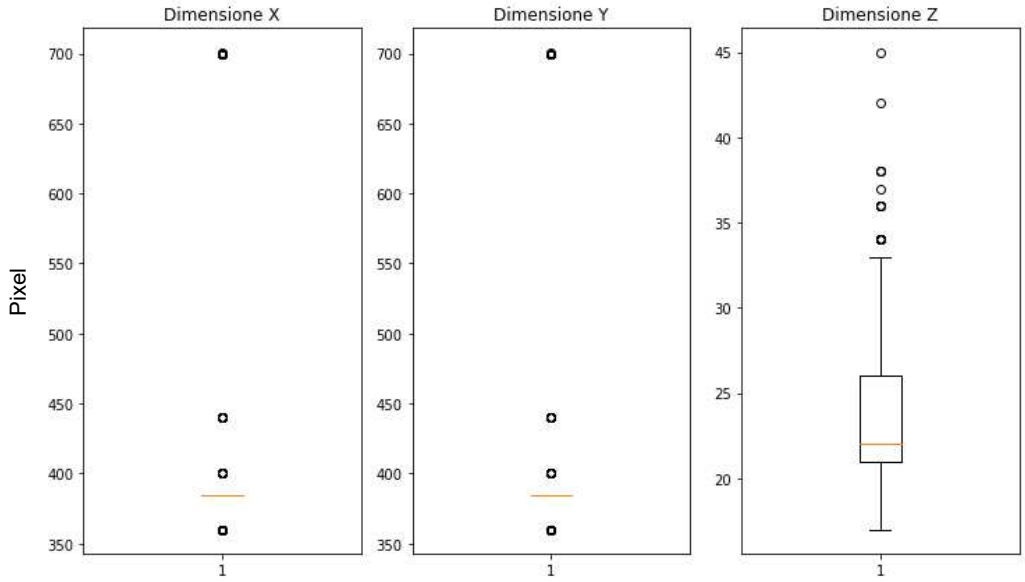


Figura 4.7: Boxplot delle dimensioni spaziali delle scansioni T2w dell'intero dataset. Tutte le scansioni sono state ridimensionate tramite resampling con fattore di scala (0.5, 0.5, 3).

Le dimensioni adottate sono ampiamente adeguate per l'inclusione dell'intera ghiandola prostatica. Infatti, essa occupa tipicamente un volume di circa 45–50 cm^3 e, in letteratura, sono comunemente utilizzate finestre ancora più piccole, come ad esempio [18,144,144] [47]. L'utilizzo di questa finestra inoltre permette di evidenziare dei limiti che, per questo specifico dataset, non emergerebbero utilizzando finestre più grandi, come quella di [z,400,400] implementata nella fase di inference.

Nelle figure 4.8 e 4.9 sono ripotati i flow chart delle due pipeline di pre-processing confrontate. Come è possibile osservare dai diagrammi di flusso, l'operazione di crop non esclude l'operazione di Resize. Questa infatti continua ad essere necessaria per uniformare la dimensione z dei volumi prima che questi siano forniti in input al modello. Inoltre essa garantisce il corretto funzionamento dell'algoritmo anche per i casi con dimensioni più piccole a quelle della finestra utilizzata.

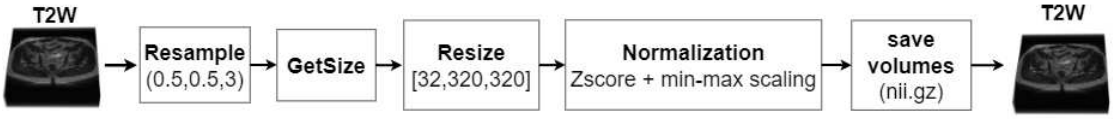


Figura 4.8: Flow chart della pipeline di *pre-processing prova 1*. Questa non include l'operazione di crop.

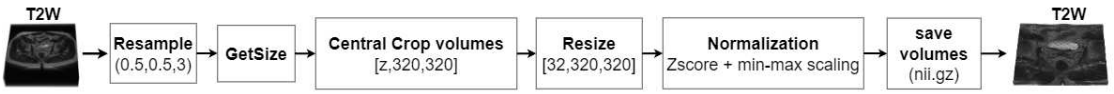


Figura 4.9: Flow chart della pipeline di *pre-processing prova 2*. Questa include al suo interno l'operazione di crop.

4.1.7 Normalizzazione

Nell'ambito del trattamento delle scansioni della prostata è stata implementata un'operazione di normalizzazione al fine di uniformare l'intensità dei pixel tra i diversi casi del dataset. Questa prevede due operazioni matematiche effettuate in cascata: z score e min-max scaling.
La *normalizzazione z score* (4.5) consiste nel trasformare i dati in modo che abbiano una distribuzione con media nulla e deviazione standard unitaria. L'operazione è stata implementata mediante la seguente formula:

$$z = \frac{x - \mu}{\sigma} \tag{4.5}$$

- x rappresenta il valore originale del pixel;
- μ è la media dei valori dei pixel non nulli nel volume;
- σ è la deviazione standard dei valori dei pixel non nulli nel volume.

Escludere i pixel nulli, ovvero quelli con valore pari a 0, dal calcolo della media e della deviazione standard è essenziale per garantire che la normalizzazione rifletta accuratamente le caratteristiche dell'area di interesse delle immagini. Questo approccio evita l'introduzione di bias nei risultati della normalizzazione che potrebbero essere generati dalla presenza degli zeri nel volume.
Il *min-max scaling* (4.6) è un secondo metodo di normalizzazione in cui i

valori dei pixel vengono scalati nell'intervallo specificato [0-1]. L'operazione è stata implementata mediante la seguente formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4.6}$$

- x è il valore originale del pixel;
- x' è il valore normalizzato.

Insieme, queste normalizzazioni contribuiscono a migliorare l'efficacia del processo di segmentazione riducendo le variabilità introdotte dalle diverse modalità di acquisizione adottate nei diversi centri.

4.2 Sviluppo del modello *M1*

In questa sezione, viene esaminato il processo di sviluppo del modello di segmentazione della prostata *M1*. Questo include la scelta dell'architettura e l'ottimizzazione dei parametri.

4.2.1 Strategia di addestramento: Metodo Patch

Il *metodo delle patch* è una tecnica utilizzata nell'ambito del deep learning per gestire in modo efficiente le immagini di grandi dimensioni. In questo contesto, le *patch* sono delle piccole porzioni di immagini estratte da una più grande ed utilizzate come input per addestrare il modello.

Questa tecnica ha come obiettivo quello di concentrare l'attenzione della rete su regioni più piccole dell'immagine facilitando così un apprendimento efficiente e focalizzato.

Nella figura sottostante è riportato un esempio relativo all'estrazione random delle patch.

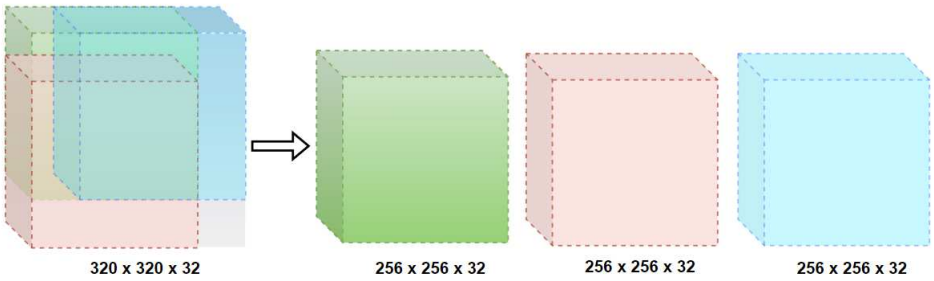


Figura 4.10: Schema estrazione random di 3 patch di dimensioni [256x256x32], raffigurate in verde rosa e azzurro, da un volume di dimensioni [320x320x32].

La procedura di estrazione delle patch è guidata dall'uso delle label positive e negative, dove le label positive indicano la presenza della ghiandola nella sezione dell'immagine. Queste sono state utilizzate solamente nella fase di training per consentire al modello di modificare i propri parametri interni sulla base delle features estratte da porzioni più piccole e significative dell'immagine. Pertanto mentre nella fase di Training il modello prende in input n sottovolumi (patch) estratti randomicamente dall'immagine, nella fase di Validation e Testing il modello prende in input l'intero volume. Nel corso della fase sperimentale, si è dedicata particolare attenzione all'ottimizzazione delle dimensioni delle patch impiegate nel processo di addestramento del modello.

4.2.2 Architettura UNet

La struttura del modello UNet utilizzata per questa task prevede l'uso di convoluzioni tridimensionali che consentono alla rete di catturare le correlazioni spaziali tra i voxel del volume T2w. La rete UNet implementata restituisce come output una mappa di probabilità dove i valori più alti indicano la presenza della prostata.

4.2.3 Loss function

Durante l'analisi sperimentale sono state testate diverse *Loss function* (3.5) messe a disposizione dalla libreria MONAI. Per determinare la funzione più adatta per la task si sono confrontate le performance ottenute sul Validation set utilizzando il modello con l'architettura ottimale identificata nell'analisi precedente i cui parametri sono riportati nella tabella 4.2.

Tabella 4.2: Tabella riassuntiva dei parametri del modello utilizzato per la valutazione della loss function ottimale.

Architettura	Unet3D: spatial_dims=3, in_channels=1, out_channels=1, channels=(16,32,64,128,256)) strides=((2,2,2),(2, 2, 2),(2, 2, 2),(2, 2, 2)), num_res_units=2, norm=Norm.BATCH
Dimensione & Numero Patch	Patch size=(192,192,16) Numero= 3
Ottimizzatore	Adam (lr=0.001, wd=5e-06)
Batch size	8

4.2.4 Learning rate e Weight decay

Nell'ambito dell'ottimizzazione dei parametri, un aspetto importante riguarda la scelta dei valori di learning rate (LR) e il weight decay (WD). Questi parametri influenzano significativamente le prestazioni complessive del modello durante il processo di addestramento. Per determinare la combinazione ottimale di LR e WD, è stato adottato un approccio iterativo. Pertanto sono state valutate diverse combinazioni al fine di identificare quella che consentiva di massimizzare l'accuratezza e la stabilità del modello.

4.2.5 Operazioni di Data Augmentation

Le operazioni di data augmentation hanno come obiettivo quello di aumentare l'eterogeneità dei dati di addestramento generando variazioni realistiche nelle immagini e nelle etichette associate al fine di aumentare la robustezza del modello e di ridurre il rischio di overfitting.

Le operazioni di data augmentation implementate sono:

- Flip random lungo l'asse Y: Questa operazione esegue il flip delle immagini e delle etichette associate lungo l'asse Y, con una probabilità del 50%, con l'obiettivo di aumentare la diversità del dataset.
- Rotazione random ± 15 gradi: applicata con una probabilità del 50% consente di simulare le variazioni di orientamento, migliorando così la capacità di generalizzazione del modello rendendolo robusto ai diversi orientamenti della prostata.
- Estrazione randomizzata delle patch: viene eseguito il cropping casuale delle immagini e delle etichette basato sulla presenza di label positive

e negative. Questa procedura randomizzata fa si che le patch generate ad ogni iterazione siano differenti tra loro. Ciò consente al modello di concentrarsi su regioni più piccole e di volta in volta diverse, migliorando l'efficienza dell'addestramento e riducendo il rischio di overfitting.

Di seguito è riportata una tabella riassuntiva delle operazioni di data augmentation implementate e dei loro parametri interni.

Tabella 4.3: Tabella riassuntiva delle operazioni di Data Augmentation implementate.

Operazione	**Parametri**
Flip lungo l'asse Y	Probabilità $= 0.5$
Rotazione random	Probabilità $= 0.5$, Rotazione $= \pm 15$ gradi
Estrazione delle patch	Dimensione spaziale $= (256, 256, 32)$, Campioni positivi $= 3$, Campioni negativi $= 0$

4.3 *Post-processing 1*: ottimizzazione

In questa sezione, viene approfondito il processo di ottimizzazione della pipeline di post-processing delle maschere generate dal modello *M1* con l'obiettivo di migliorarne la precisione e l'accuratezza.

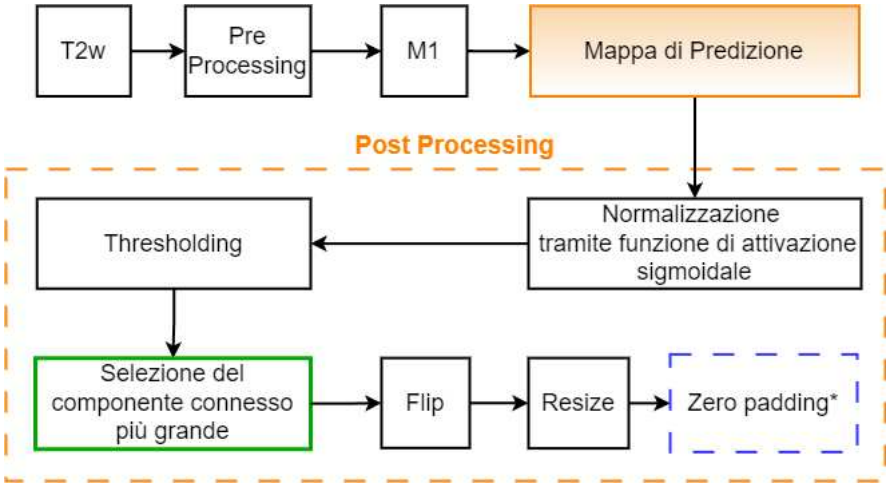


Figura 4.11: flow chart post processing.

Il processo di post processing, descritto nella figura 4.11, inizia con la normalizzazione della mappa di predizione ottenuta come output dal modello, tramite l'applicazione della *funzione di attivazione sigmoidale*, per garantire che i valori siano compresi nell'intervallo [0,1].

Successivamente, viene eseguita un'*operazione di thresholding* per convertire la mappa di predizione in una maschera binaria. Questo avviene impostando a "1" tutti i valori della maschera che superano un determinato valore soglia e a "0" quelli al di sotto della soglia stabilita. Durante l'analisi sperimentale è emerso che un valore soglia più alto rispetto a "0.5" consente di eliminare eventuali predizioni ambigue. In particolare, per questa specifica task e per questo modello, si è identificato mediante approccio iterativo un valore soglia ottimale pari a "0.7".

Il processo prosegue effettuando la "selezione del componente connesso più grande" il cui flow chart descrittivo della funzione implementata è riportato nella figura 4.12. Questa viene eseguita iterando attraverso ogni slice della maschera, identificando ed estraendo la componente bidimensionale connessa più grande presente. Completato il ciclo su tutte le slice del volume, il processo passa alla fase successiva in cui viene mantenuta solo

la componente volumetrica più grande presente nella maschera, eliminando così eventuali piccoli artefatti isolati.

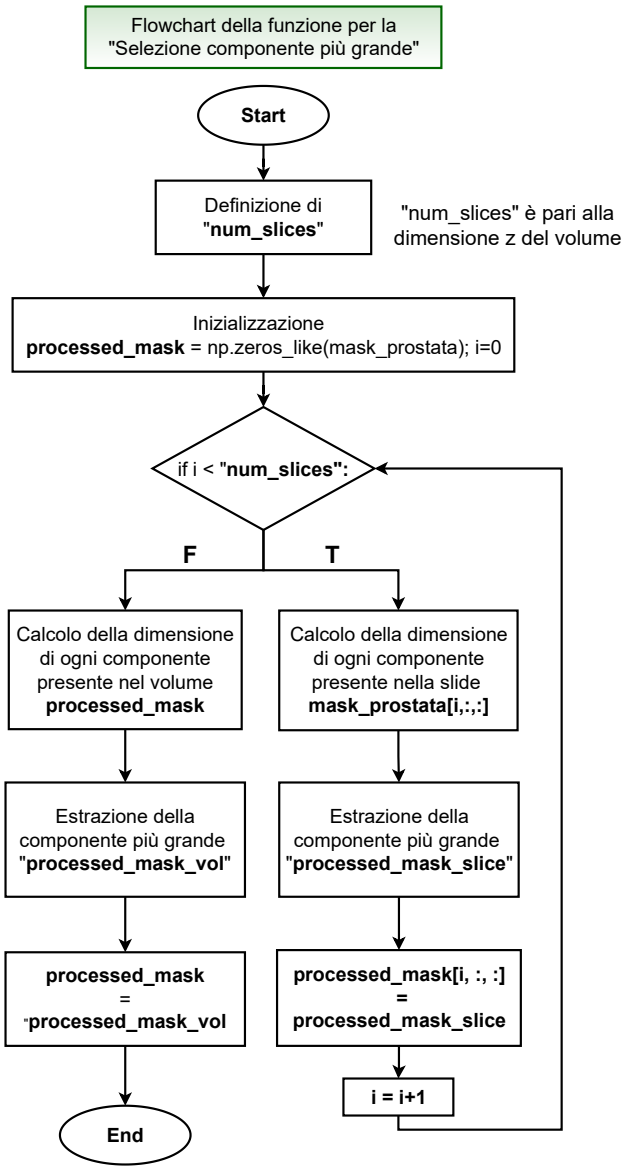


Figura 4.12: Flowchart della funzione di post-processing per la selezione del componente connesso più grande.

Nella figura 4.13 è riportato un esempio visivo dell'effetto dell'applicazione

della funzione descritta.

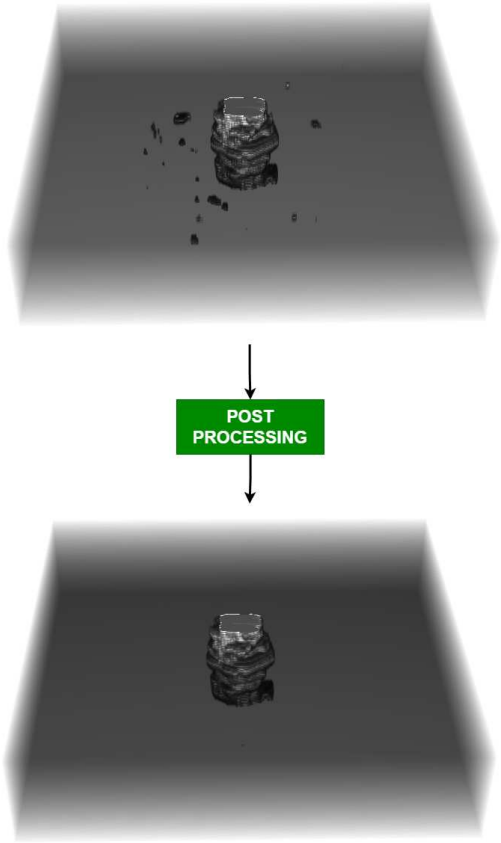


Figura 4.13: Rappresentazione grafica dell'effetto del post-processing sul caso ID 11465_1001489.

Infine, per garantire che la maschera risultante sia allineata correttamente con la scansione T2w, vengono effettuate le operazioni "opposte" a quelle eseguite in fase di pre-processing sulla scansione. Le operazioni sono le seguenti:

- Flip verticale: compensa la rotazione interna effettuata durante l'applicazione del modello;
- Resize: necessaria per ripristinare le dimensioni delle maschere al formato dell'immagine T2w post resample;
- Zero padding: operazione effettuata solo *nella fase di inference* per compensare l'operazione di crop effettuata sui volumi con dimensione,

nel piano trasversale XY, superiore a 400x400. Questa operazione non ha effetto sui volumi con dimensioni della sezione trasversale inferiori.

4.4 Risultati e analisi delle performance

In questa sezione sono riportati e commentati i risultati relativi alle analisi sperimentali effettuate durante le fasi di ottimizzazione della pipeline di pre-processing, del modello e del post-processing. Infine vengono analizzate le performance finali dell'algoritmo implementato. Per semplificarne la lettura ed il confronto, per le diverse analisi effettuate si è scelto di riportare i valori numerici ed i relativi plot delle prove ritenute più significative.

4.4.1 *Pre-processing 1*

Divisione del dataset

Per la suddivisione del dataset nei set di Train Validation e Test, è stato adottato il criterio standard comunemente utilizzato nella letteratura [39]. In particolare, la divisione dei casi negativi è stata effettuata seguendo le proporzioni del 70%, 20% e 10%. Inoltre sono stati inclusi 50 casi positivi nel set di Test. Pertanto, si è fatto uso di un dataset costituito complessivamente da 1050 casi, suddivisi come segue: *Train (67%)*, *Validation (19%)*, e *Test (14%)* (figura 4.14).

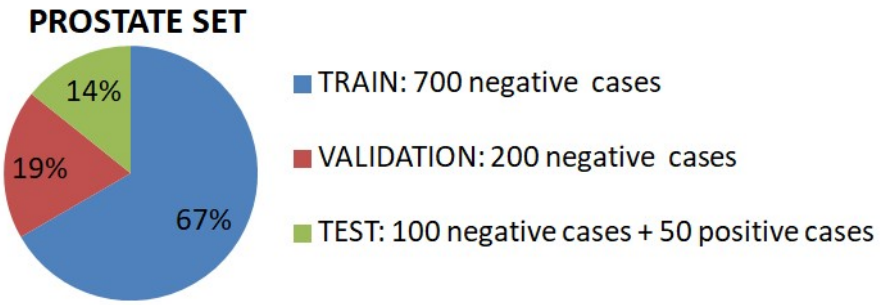


Figura 4.14: Divisione del dataset: Train (67%), Validation (19%), Test (14%).

Per questa specifica task, la dimensione complessiva dei set si è rivelata essere ottimale.

Metodi di interpolazione

I risultati dell'analisi sperimentale effettuata per la valutazione ed il confronto dei due metodi di interpolazione *Lineare e Bicubica*, discussi nella sezione 4.1.4, sono riportati nelle figure 4.15 & 4.16.

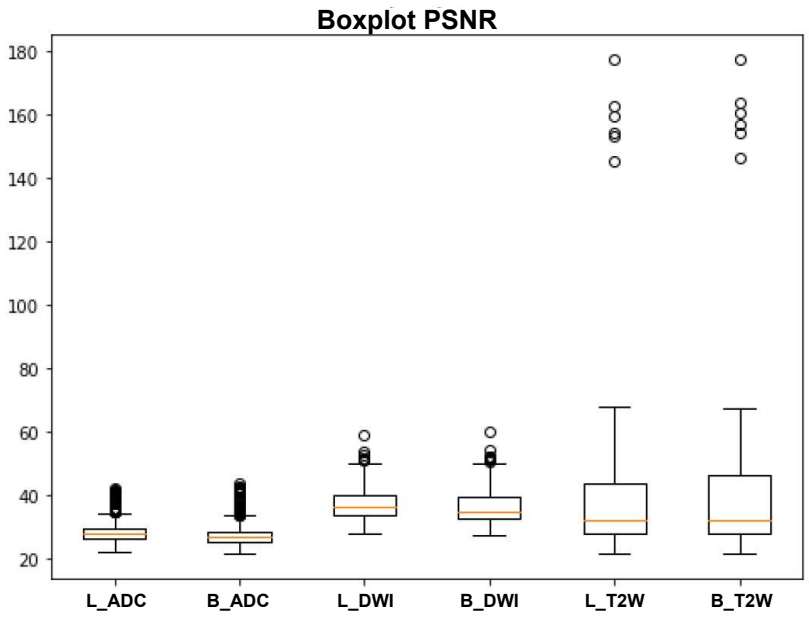


Figura 4.15: Boxplot dei valori di PSNR ottenuti per ciascun metodo di interpolazione (L *Linerare* & B *Bicubica*) sulle scansioni T2W, ADC, DWI.

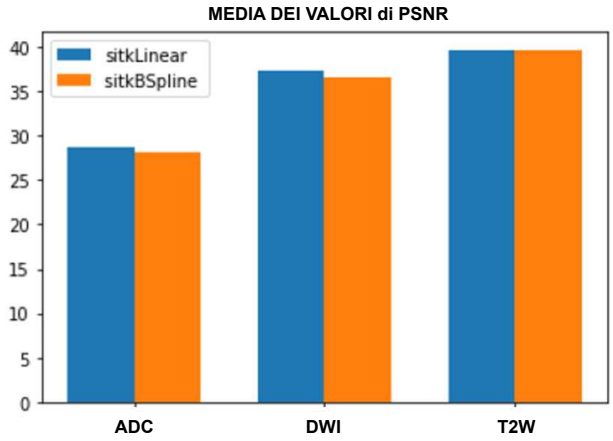


Figura 4.16: Diagramma a barre dei valori medi di PSNR ottenuti per ciascun metodo di interpolazione (*Linerare* & *Bicubica*) sulle scansioni ADC, DWI, T2W.

I risultati ottenuti indicano che entrambi i metodi consentono di ottenere prestazioni simili in termini di qualità dell'immagine interpolata. Tuttavia, l'interpolazione lineare si è dimostrata leggermente superiore alla bicubica in termini di *PSNR medio* e di tempi di esecuzione. Di conseguenza, si è fatto uso dell'interpolazione lineare per eseguire il resample dei volumi contribuendo così all'ottimizzazione complessiva del processo.

Spacing

Per quanto riguarda lo *spacing*, i Test condotti per questa specifica applicazione suggeriscono che questo non influenzi significativamente le performance del modello. Pertanto, nell'effettuare l'operazione di resample si è optato per uno spacing pari a [0.5,0.5,3], come suggerito dagli organizzatori della challenge, essendo questa la configurazione media comunemente utilizzata dai diversi centri di acquisizione. Questa, nonostante comporti risultati analoghi a quelli ottenuti con spacing [1,1,1], è stata preferita per semplificare lo sviluppo del codice negli step successivi dove l'utilizzo di uno spacing più piccolo riveste un ruolo di maggiore importanza nella segmentazione delle lesioni.

Resize

I risultati dell'analisi sperimentale condotta per la valutazione dell'operazione di *resize*, discussa nella sezione 4.1.5, sono riportati nella tabella 4.4 e

nella figura 4.17.

Tabella 4.4: Tabella di valutazione dell'impatto della dimensione z nell'operazione di resize sulle prestazioni del modello. I valori delle metriche sono stati calcolati sui casi del Test set interno. In verde è evidenziata la prova considerata migliore.

TEST SET interno	Z=16		Z=32	
Metrica	Mean	± Std	Mean	± Std
DICE	0.92	±0.05	0.93	± 0.0321
HD_{95}	8.72	± 17.52	6.43	± 15.0203
RVD	0.02	± 0.16	-0.0042	± 0.0682
PRECISIONE	0.91	± 0.08	0.94	± 0.0449

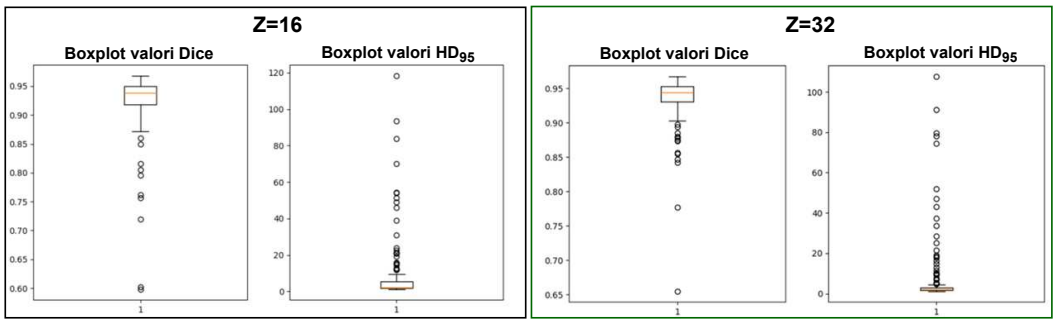


Figura 4.17: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sui relativi Test set interni. A sinistra, nel blocco nero, sono riportati i plot relativi all'analisi effettuata con resize di dimensioni [320,320,16]. A destra, nel blocco verde, sono riportati i plot relativi all'analisi effettuata con resize di dimensioni [320,320,32].

I risultati evidenziano come l'utilizzo della dimensione pari a 32 sia preferibile nonostante questa comporti un costo computazionale leggermente superiore. Questi risultati sono motivati dal fatto che dimensioni più grandi consentono di preservare o, in alcuni casi, ampliare i dettagli anatomici presenti nell'immagine. Questo consente al modello di estrarre maggiori informazioni e, di conseguenza, classificare i pixel con maggiore precisione. Pertanto, dall'analisi sperimentale condotta, è emerso che la dimensione [320,320,32] rappresenta il compromesso ottimale tra costi computazionali e performance di predizione. Infatti, dimensioni più grandi comportano

un aumento dei costi computazionali e della memoria occupata, mentre dimensioni più piccole riducono le performance in termini di precisione della segmentazione.

Crop

I risultati dell'analisi sperimentale condotta per la valutazione dell'operazione di *crop*, discussa nella sezione 4.1.6, sono riportati nella tabella 4.5 e nella figura 4.18.

Tabella 4.5: Tabella di valutazione dell'impatto dell' operazione di *crop centrale* sulle prestazioni del modello. I valori delle metriche sono stati calcolati sui relativi Test set interni. In verde è evidenziata la prova considerata migliore.

Test: CROP	Senza Crop		Con Crop	
Metrica	Mean	± Std	Mean	± Std
DICE	0.9088	±0.1084	0.9361	± 0.0321
HD_{95}	3.4393	± 4.7472	6.4325	± 15.0203
RVD	-0.0432	± 0.1941	-0.0042	± 0.0682
PRECISIONE	0.9427	± 0.05769	0.9402	± 0.0449

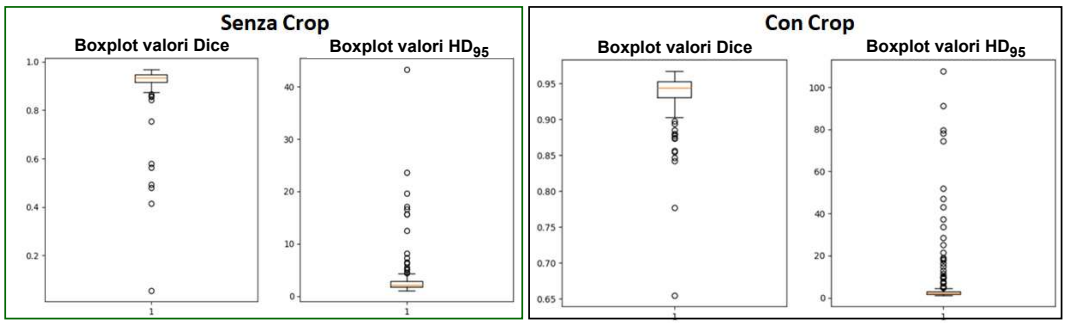


Figura 4.18: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sui Test set interni. A sinistra, nel riquadro verde, sono riportati i plot relativi all'analisi effettuata senza l'applicazione dell'operazione di crop (*Pipeline di pre-processing 1* 4.8). A destra, nel riquadro nero, sono riportati i plot relativi all'analisi effettuata applicando il crop (*Pipeline di pre-processing 2* 4.9).

Come discusso precedentemente, la prostata non è sempre al centro della scansione (figura 4.3). Di conseguenza il crop effettuato potrebbe non inglobare correttamente l'intero volume della ghiandola compromettendo, come evidenziato dai risultati ottenuti (tabella 4.5) ed in particolare dai valori delle metriche dei casi identificati come *"outlier"* nel boxplot (figura 4.18), la precisione e l'affidabilità del processo di segmentazione.

Le scansioni che riportano l'intero addome sono quelle in cui il modello riscontra maggiori difficoltà e rappresentano gli *outlier* precedentemente menzionati. Queste presentano tipicamente una dimensione di Zx700x700 pixel dopo il resample (figura 4.3). Pertanto, al fine di migliorarne le prestazioni di segmentazione, in fase di inference, è stata implementata l'operazione di crop centrale di [z',400,400], riducendo significativamente la dimensione delle immagini e allo stesso tempo garantendo, per questo specifico dataset, l'inclusione dell'intera regione di interesse (figura 4.19). Questo approccio è stato adottato per mitigare le problematiche legate alla complessità delle immagini di grandi dimensioni migliorando la qualità della segmentazione e, di conseguenza, le performance generali dell'algoritmo.

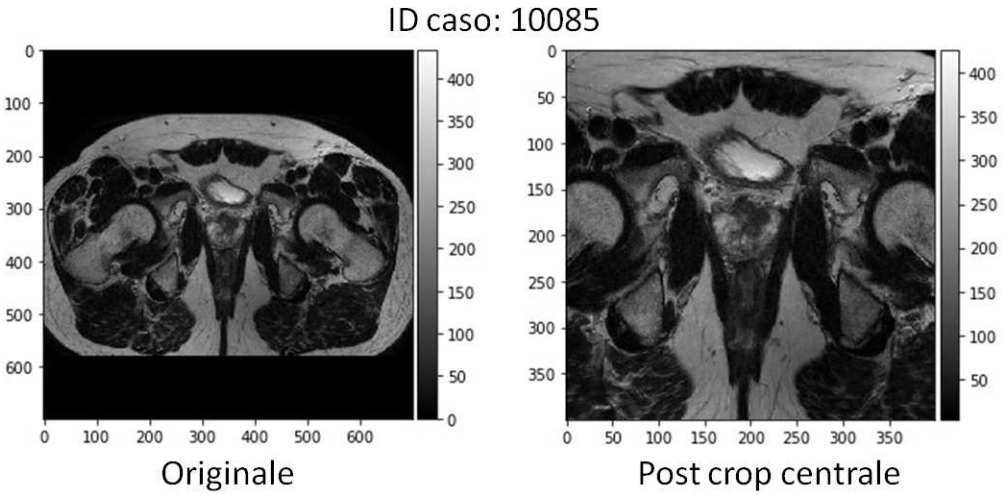


Figura 4.19: Confronto della scansione T2w dell'intero addome prima e dopo l'applicazione dell'operazione di crop centrale [z',400,400]. La scansione è stata sottoposta ad un resampling iniziale con un fattore di scala di (0.5, 0.5, 3).

Sebbene il crop di dimensioni fisse [z,400,400] possa essere una soluzione valida per il dataset specifico utilizzato, la sua applicazione potrebbe causare problemi con set "esterni" in cui le immagini sono acquisite in modo differente.

Al fine di garantire la robustezza del sistema in scenari d'uso reale, è preferibile evitare l'adozione di un crop con dimensioni predefinite e migliorare la capacità del modello di adattarsi a diverse condizioni di input mediante opportune modifiche del dataset di partenza e/o del modello utilizzato facendo sempre attenzione al rapporto costi/benefici. Nell'ambito di questo lavoro di tesi, si è scelto pertanto di escludere l'utilizzo del crop durante le fasi di addestramento validazione e test del modello al fine di sviluppare una pipeline robusta e affidabile in applicazioni reali.

Pipeline di pre-processig finale

Di seguito sono elencate le operazioni implementate nelle pipeline di preprocessing definite in seguito al processo di ottimizzazione.

Pipeline di pre-processing finale delle scansioni T2w:

1. Resample (0.5,0.5,3) [1] con interpolazione lineare;
2. Resize [32,320,320]; [2]
3. Normalizzazione: Zscore + min-max scaling.

Pipeline di pre-processing finale delle maschere utilizzate per l'addestramento e la valutazione delle performance del modello:

1. Resample (0.5,0.5,3) con interpolazione NearestNeighbor;
2. Resize [32,320,320];
3. Thresholding con soglia 0,5.

Pipeline di pre-processing delle scansioni T2w implementata nella fase di *inference:*

1. Resample (0.5,0.5,3) con interpolazione lineare;
2. Crop [z',400,400] [3], z'= numero slice del volume;
3. Resize [32,320,320];
4. Normalizzazione: Zscore + min-max scaling.

[1] (x,y,z)
[2] (z,x,y)
[3] (z,x,y)

4.4.2 Modello *M1*

In questa sottosezione sono riportati e commentati i risultati relativi alle analisi sperimentali effettuate nella fase di sviluppo e ottimizzazione del modello *M1*.

Architettura del modello

Per quanto concerne l'ottimizzazione dell'architettura, tra le varie prove effettuate si è riportato il confronto tra l'architettura del modello di partenza (utilizzato nella fase di ottimizzazione della pipeline di pre-processing) e l'architettura del modello finale ottenuto in seguito al processo di ottimizzazione. Nella tabella 4.6 sono riportati i parametri riassuntivi delle due architetture confrontate. Per la valutazione quantitativa delle performance sono riportate: nella tabella 4.7 le metriche calcolate sul Test set mentre nella figura 4.20 i boxplot dei valori di Dice e Distanza di Haussdorff(95° percentile).

Tabella 4.6: Tabella riassuntiva dei parametri delle architetture dei modelli confrontati.

Architettura Iniziale	Unet3D: spatial_dims=3, in_channels=1, out_channels=1, channels=(16,32,64,128,256)) strides=((2,2,2),(2, 2, 1),(2, 2, 1),(2, 2, 2)), num_res_units=2, norm=Norm.BATCH
Architettura Finale	Unet3D: spatial_dims=3, in_channels=1, out_channels=1, channels=(16,32,64,128,256)) strides=((2,2,2),(2, 2, 2),(2, 2, 2),(2, 2, 2)), num_res_units=2, norm=Norm.BATCH
Parametri in comune	
Dimensione & Numero Patch	Patch size=(192,192,16) Numero= 3
Loss function	DiceLoss
Ottimizzatore	Adam (lr=0.001, wd=5e-06)
Batch size	8

Tabella 4.7: Tabella di valutazione dell'impatto dell'architettura sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni del modello iniziale con quelle del modello finale implementato. I valori delle metriche sono stati calcolati sui Test set interni. In verde è evidenziata la prova considerata migliore.

Test: Architettura	**Architettura iniziale**		**Architettura finale**	
Metrica	Mean	± Std	Mean	± Std
DICE	0.9088	±0.1084	0.9177	± 0.084
HD_{95}	3.4393	± 4.7472	3.12	± 2.7052
RVD	-0.0432	± 0.1941	0.022	± 0.083
PRECISIONE	0.9427	± 0.05769	0.9245	± 0.05579

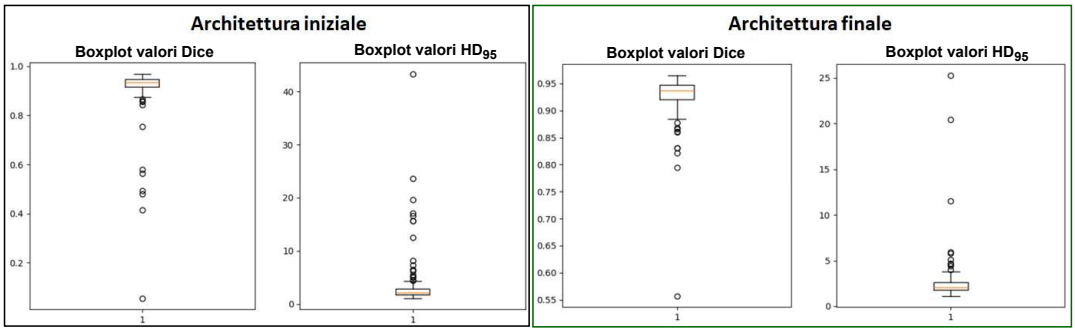


Figura 4.20: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sui Test set interni. A sinistra, nel riquadro nero, sono riportati i plot relativi all'analisi effettuata mediante l'utilizzo *"dell'architettura iniziale"* (tabella 4.6). A destra, nel riquadro verde, sono riportati i plot relativi all'analisi effettuata mediante l'utilizzo *dell'architettura finale* (tabella 4.6).

Sebbene la segmentazione precisa in termini di pixel non sia essenziale per questa specifica task, è molto importante identificare correttamente la ghiandola prostatica e la sua estensione lungo le tre direzioni spaziali X Y Z. Nel valutare le performance del modello e nel processo di identificazione dell'architettura ottimale, assume molta importanza l'utilizzo dei boxplot. Questi consentono di visualizzare e valutare le performance generali dei modelli anche nei casi più complessi identificati come "outlier". L'analisi dei risultati evidenzia come l'implementazione dell'architettura "finale" consenta

di ottenere performance simili in termini di prestazioni medie, ma presenta un notevole miglioramento nei casi "outlier" identificati .

Loss function

Per lo sviluppo del modello è stata implementata la funzione di perdita DiceLoss dalla libreria MONAI. Questa è stata preferita rispetto ad altre opzioni disponibili, come ad esempio la BCELoss o la DiceFocalLoss (descritti nella sotto sezione 3.5), in quanto ha permesso di ottenere risultati migliori. Per un confronto numerico, è possibile consultare la tabella 4.8, dove, per ciascuna tipologia di funzione, è riportato il valore del Dice medio calcolato sul Validation set.

Tabella 4.8: Tabella di valutazione dell'impatto della funzione di Loss sulle prestazioni del modello. I valori delle metriche sono stati calcolati sul Validation set interno. In verde è evidenziata la prova considerata migliore. Questi risultati sono stati ottenuti utilizzando il modello con *architettura finale* descritta nella tabella 4.6.

Loss function	Dice	
DiceFocalLoss	0.9012	±0.06
BCELoss	0.9166	±0.0808
DiceLoss	0.9343	±0.0321

Learning rate (LR) e Weight decay (WD)

La tabella 4.9 riporta i valori medi di Dice calcolati sul Validation set mediante alcune tra le diverse configurazioni di learning rate (LR) e weight decay (WD) testate durante la fase di ottimizzazione. In particolare, si nota che la riga 9 corrisponde a una configurazione che ha raggiunto il valore migliore di metrica *DICE (0.9381)*, impiegando un *learning rate di 1e-03 e un weight decay di 1e-04.*

Tabella 4.9: Tabella riassuntiva dei valori medi di Dice calcolati sul Validation set, utilizzando diversi valori di learning rate (lr) e weight decay (wd). Questi risultati sono stati ottenuti utilizzando il modello con "architettura finale" descritta nella tabella 4.6.

Prova	DICE	Configurazione
1	0.7459	Lr: 5e-03, wd: 1e-03
2	0.8334	Lr: 5e-03, wd: 5e-04
3	0.8646	Lr: 5e-03, wd: 1e-04
4	0.8985	Lr: 5e-03, wd: 5e-05
5	0.9043	Lr: 5e-03, wd: 1e-05
6	0.9241	Lr: 5e-03, wd: 5e-06
7	0.8957	Lr: 1e-03, wd: 1e-03
8	0.9058	Lr: 1e-03, wd: 5e-04
9	0.9381	Lr: 1e-03, wd: 1e-04
10	0.9225	Lr: 1e-03, wd: 5e-05
11	0.9316	Lr: 1e-03, wd: 1e-05
12	0.9343	Lr: 1e-03, wd: 5e-06
13	0.9293	Lr: 5e-04, wd: 1e-03
14	0.9224	Lr: 5e-04, wd: 5e-04
15	0.9379	Lr: 5e-04, wd: 1e-04
16	0.9373	Lr: 5e-04, wd: 5e-05
17	0. 9351	Lr: 5e-04, wd: 1e-05
18	0.9322	Lr: 5e-04, wd: 5e-06

Nella tabella 4.10 sono riportati i valori delle metriche di valutazione calcolate sul Test set interno utili per una valutazione quantitativa dell'impatto dei valori di lr e wd sulle prestazioni del modello. Nella tabella sono riportati i valori ottenuti facendo uso della combinazione iniziale (lr= 1e-03 e wd= 5e-06) e quella ottimale (lr= 1e-03 e wd= 1e-04).

Tabella 4.10: Tabella di valutazione dell'impatto dei valori di lr & wd sulle prestazioni del modello. I valori delle metriche sono stati calcolati sul Test set interno. In verde è evidenziata la prova migliore. Questi risultati sono stati ottenuti utilizzando il modello con *architettura finale* descritta nella tabella 4.6.

Test: lr & wd	**Lr=1e-03; wd=5e-06**		**Lr=1e-03; wd=1e-04**	
Metrica	Mean	± Std	Mean	± Std
DICE	0.9177	± 0.084	0.9185	± 0.075
HD_{95}	3.12	± 2.7052	3.08	± 2.60
RVD	0.022	± 0.083	0.017	± 0.080
PRECISIONE	0.9245	± 0.05579	0.9270	± 0.0527

Patch size

Dalle analisi sperimentali effettuate è emerso che l'adozione delle patch risulta essere particolarmente vantaggioso in quanto, oltre a ridurne i costi computazionali, consente di velocizzare il tempo complessivo di addestramento ottenendo performance paragonabili a quelle ottenute mediante trainig effettuato fornendo in input l'intero volume. Inoltre, l'uso delle patch nella sola fase di training consente di semplificare notevolmente lo sviluppo del codice e di ridurre i tempi ed i costi computazionali in fase di inference. Questo approccio infatti non richiede: l'implementazione di funzioni di pre-processing necessarie per l'estrazione delle patch, l'applicazione ripetuta N volte del modello su ciascun caso e l'implementazione di funzioni di post-processing necessarie per la combinazione degli output del modello in un unica maschera prostatica.

Per quanto riguarda il numero di patch estratte da ciascun volume si è determinato iterativamente che il valore ottimale per entrambe le configurazioni risulta essere pari a 3. Nella tabella 4.11 sono riportati i valori delle metriche calcolate utilizzando patch di dimensioni (192x192x16), e quelli ottenuti utilizzando le dimensioni ottimali pari a (256x256x32), ovvero quelle che, tra tutte le prove effettuate utilizzando proporzioni differenti, hanno permesso di ottenere i risultati migliori.

Tabella 4.11: Tabella di valutazione dell'impatto del *patch size* sulle prestazioni del modello. I valori delle metriche sono stati calcolati sul Test set interno. In verde è evidenziata la prova considerata migliore.

Test: Patch Size	**(192,192,16)**		**(256,256,32)**	
Metrica	Mean	± Std	Mean	± Std
DICE	0.9185	± 0.075	0.9225	± 0.0402
HD_{95}	3.08	± 2.60	2.71	± 1.83
RVD	0.017	± 0.080	0.015	± 0.060
PRECISIONE	0.9270	± 0.0527	0.918	± 0.043

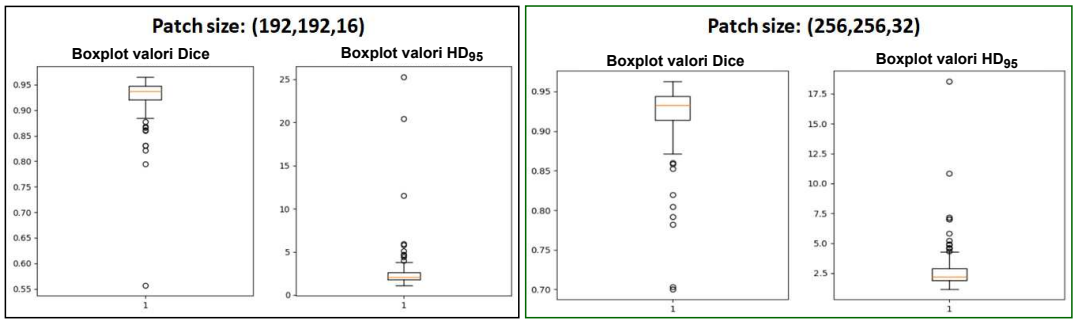


Figura 4.21: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sul Test set interno. A sinistra, nel riquadro nero, sono riportati i plot relativi all'analisi effettuata mediante *patch size (192,192,16)*. A destra, nel riquadro verde, sono riportati i plot relativi all'analisi effettuata mediante *patch size (256,256,32)*.

I risultati indicano che fornendo maggiori informazioni spaziali al modello, mediante l'utilizzo di patch di dimensioni leggermente più grandi, si ottiene un miglioramento delle prestazioni. Questo si evidenzia soprattutto nei casi "critici" precedentemente individuati e analizzati, ovvero quelli in cui le scansioni presentano l'intero addome.

Modello Finale

Per fornire una visione complessiva delle caratteristiche del modello UNet *M1* implementato, la cui struttura finale è descritta nella figura 4.22, si sono riportati i parametri principali nella tabella 4.12. Nella tabella 4.13 sono invece riportati i parametri settati per il training finale.

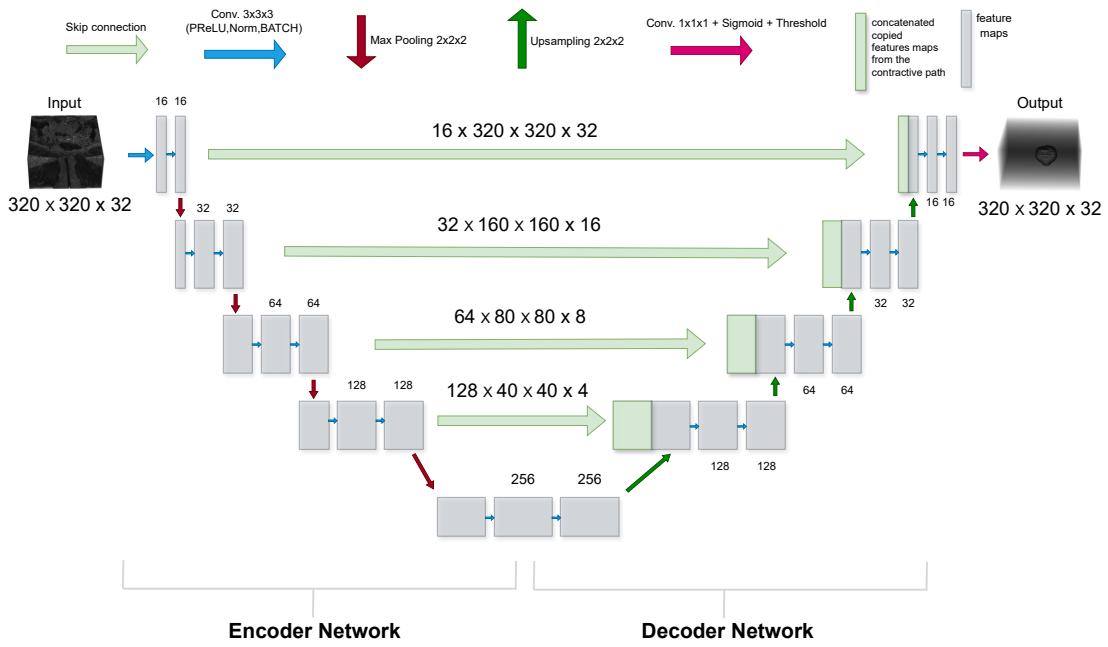


Figura 4.22: Rappresentazione grafica dell'architettura del modello UNet *M1*.

Tabella 4.12: Parametri principali dell'architettura UNet *M1*.

Parametri del modello	
spatial_dims	3
in_channels	1
out_channels	1
channels	(16, 32, 64, 128, 256)
strides	((2, 2, 2), (2, 2, 2), (2, 2, 2), (2, 2, 2))
kernel_size	(3, 3, 3)
num_res_units	2
norm	Norm.BATCH
padding	(1, 1, 1)
activation function	PReLU
dim input volume	320x320x32
dim output volume	320x320x32

Tabella 4.13: Parametri principali del training del modello UNet *M1*

Parametri di training	
wks	0
BATCH_SIZE	8
num_iterations (max)	200
patience	20
loss_function	DiceLoss(to_onehot_y=True, softmax=True)
optimizer	torch.optim.Adam(lr=1e-03, weight_decay= 1e-04)
data_augmentation	random flip verticale, rotazione ±15° random, estrazione random 3 patch [256,256,32]

4.4.3 *Post-processing 1*

Per una valutazione quantitativa della funzione di post-processing *Selezione del componente connesso più grande* implementata (sezione 4.3), di seguito è riportata la tabella riassuntiva delle metriche calcolate sul Test set ed i boxplot dei valori di Dice e Distanza di Hausdorff (95° percentile).

Tabella 4.14: Tabella di valutazione dell'effetto dell'operazione di post-processing *"Selezione del componente connesso più grande"* sulle maschere. I valori delle metriche di valutazione sono stati calcolati sul Test set interno. In verde è evidenziata la prova considerata migliore.

TEST SET interno	**Senza post processing**		**Con post processing**	
Metrica	Mean	± Std	Mean	± Std
DICE	0.9225	±0.0402	0.9239	± 0.0387
HD_{95}	2.71	± 1.83	2.53	± 1.18
RVD	0.015	± 0.086	-0.0004	± 0.088
PRECISIONE	0.918	± 0.043	0.927	± 0.035
SENSIBILITÀ	0.93	± 0.06	0.924	± 0.06

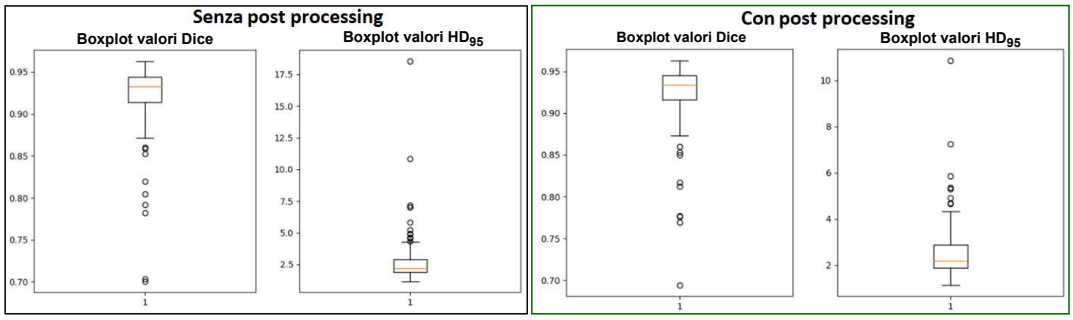


Figura 4.23: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sul Test set interno. A sinistra, nel riquadro nero, sono riportati i plot relativi all'analisi effettuata *"senza post processing"*, ovvero senza l'applicazione della funzione per la *"Selezione del componente connesso più grande"*. A destra, nel riquadro verde, sono riportati i plot relativi all'analisi effettuata implementando tale operazione di *post processing*.

Da un analisi visiva delle maschere è emerso che la divisione dell' operazione di post-processing in due step, prima in 2D sulle singole slice e poi in 3D sull'intero volume, consente di rimuovere alcuni *falsi positivi* che non sarebbero rimossi effettuando direttamente l'operazione sul volume (un esempio è riportato nella figura 4.24). Sebbene siano necessari due step, il rapporto costo/benefici è favorevole all'implementazione del doppio filtro, essendo il volume costituisco tipicamente da poche slice.

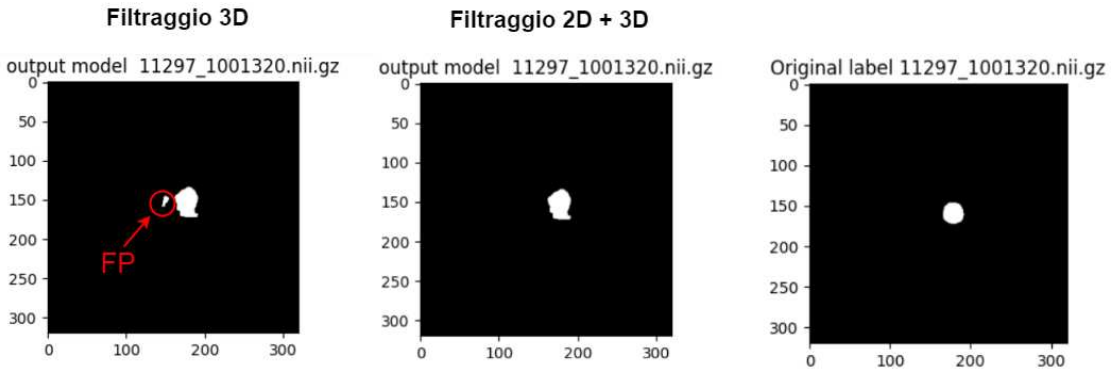


Figura 4.24: Confronto visivo dell'effetto del solo filtraggio 3D vs combinazione 2D+3D.

A differenza delle classiche operazioni di post processing, la funzione implementata garantisce la presenza di un unico "oggetto" nella maschera generata. Inoltre, la funzione ingloba al suo interno un criterio per identificare e rimuovere, in modo automatico, tutti gli artefatti senza l'ausilio di parametri esterni forniti manualmente.

4.4.4 Algoritmo finale di segmentazione della prostata

In questa sottosezione sono analizzate le performance finali dell'algoritmo implementato per la segmentazione della prostata.

Come ben visibile dagli esempi riportati nelle figure (4.25 e 4.26), le maschere presentano dei Falsi negativi che sarebbe possibile rimuovere mediante l'applicazione di operazioni di post-processing aggiuntive, come ad esempio *remove_small_holes* dalla libreria Skimage, migliorando così la qualità della segmentazione prodotta. Tuttavia, l'introduzione di queste operazioni comporterebbe un aumento dei costi e dei tempi computazionali che non sarebbe giustificato rispetto all'obiettivo della task. L'obiettivo infatti non è ottenere una segmentazione precisa in termini di pixel, ma una maschera di buona qualità che permetta di identificare orientativamente il centro della prostata nel piano trasversale XY. Questo è essenziale per eseguire l'operazione di crop nella fase successiva del processo.

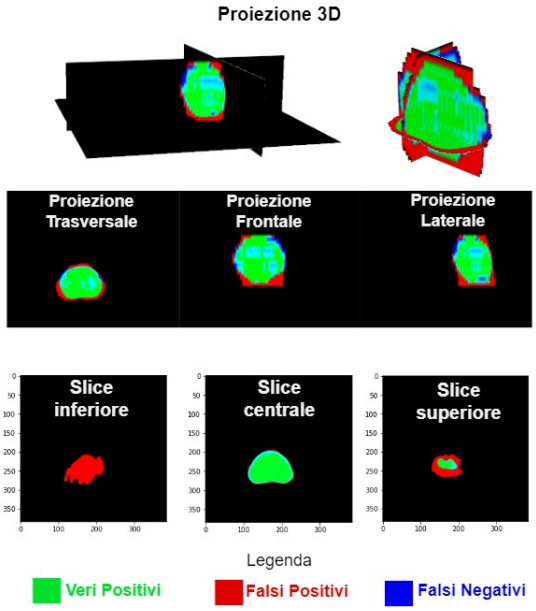


Figura 4.25: Rappresentazione grafica dei risultati. Il paziente preso in esame ha il seguente ID 11241_1000245. I pixel rappresentati in verde sono i *"Veri Positivi"*, quelli in rosso i *"Falsi Positivi"* e in blu i *"Falsi Negativi"*.

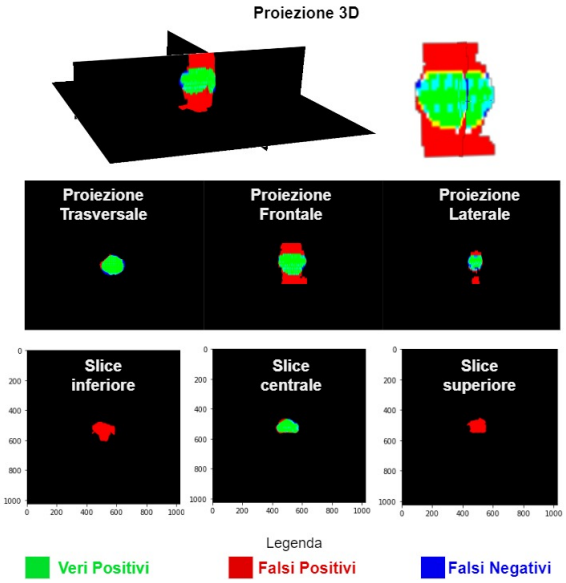


Figura 4.26: Rappresentazione grafica dell'effetto del post-processing sul paziente ID 11465_1001489. I pixel rappresentati in verde sono i *"Veri Positivi"*, quelli in rosso i *"Falsi Positivi"* e in blu i *"Falsi Negativi"*.

La maschera ottenuta, nonostante la presenza di qualche "falso positivo" consente di effettuare questa operazione con successo. Dall'analisi dei risultati, sebbene i processi di ottimizzazione hanno permesso di migliorarne notevolmente le prestazioni, emerge la difficoltà del modello nel segmentare correttamente la ghiandola lungo l'asse Z. Tale limite è ben visibile osservando le proiezioni laterali delle maschere e si manifesta in maniera più accentuata nelle scansioni in cui è presente l'intero addome (figura 4.26). In particolare, il modello tende a sovrastimare la dimensione della prostata lungo l'asse Z, mentre nel piano trasversale XY la segmentazione risulta piuttosto precisa. Quest'ultimo risulta essere un fattore importante per la determinazione della dimensione della finestra utilizzata per eseguire il crop delle scansioni nella pipeline di *pre-processing 2*. Poiché la dimensione z della segmentazione è utilizzata per identificare il numero di slice da estrarre, in caso di sovrastima, viene estratta e fornita ad *M2* una fetta di volume contenente l'intera ghiandola prostatica e parte del tessuto circostante. Pertanto, per gli obiettivi della task, la sovrastima del valore z può essere considerato un *limite accettabile* e preferibile rispetto ad una sua sottostima.

Di seguito, per la valutazione quantitativa delle prestazioni dell'intero algoritmo implementato, è riportata la tabella dei valori delle metriche calcolate sui set di Train, Validation e Test ed i relativi boxplot dei valori di

Dice e Distanza di Hausdorff (95° percentile).

Metrica	Train	Validation	Test
Dice	0.923 ± 0.03	0.928 ± 0.035	0.937 ± 0.025
HD_{95}	2.539 ± 1.18	2.56 ± 1.91	2.21 ± 1.15
RVD	-0.0004 ± 0.088	0.002 ± 0.077	-0.0028 ± 0.05
Precisione	0.927 ± 0.035	0.929 ± 0.038	0.94 ± 0.029
Sensibilità	0.924 ± 0.066	0.925 ± 0.067	0.924 ± 0.066

Figura 4.27: Tabella di valutazione delle performance finali dell'intero processo di segmentazione della prostata. Per ciascuna metrica, calcolata su Train Validation e Test set, è riportato il valore medio ± la deviazione standard (es: 0.923 ± 0.03).

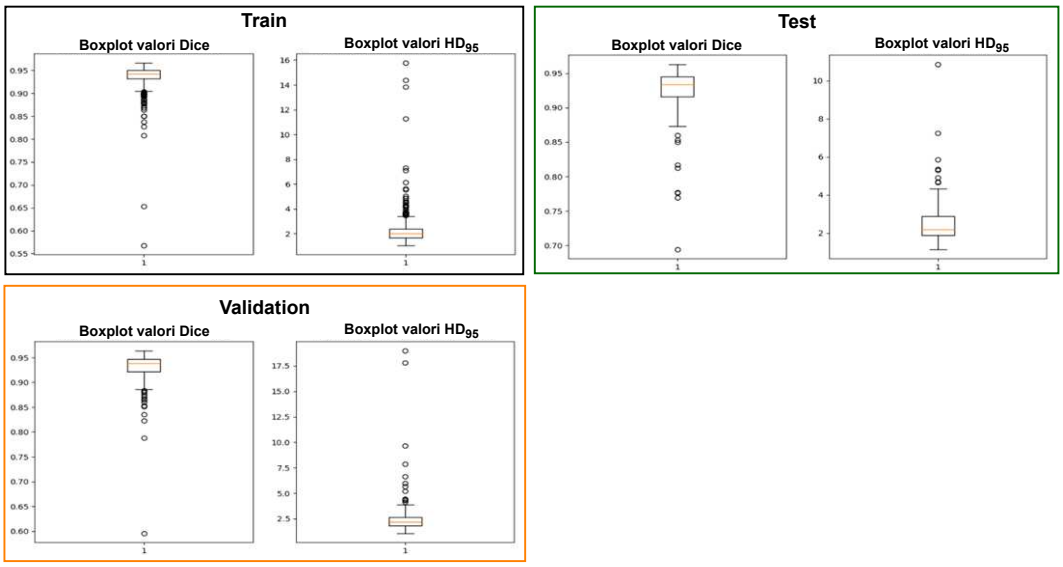


Figura 4.28: Boxplot dei valori di Dice e distanza di Hausdorff (95° percentile) calcolati sui set di Train, Validation e Test. Nel riquadro nero sono riportati i plot relativi all'analisi effettuata sul *Train set*. Nel riquadro arancione sono riportati i plot relativi all'analisi effettuata sul *Validation set*. Nel riquadro verde sono riportati i plot relativi all'analisi effettuata sul *Test set*.

L'analisi delle prestazioni del modello ottenuto conferma che il solo utilizzo della scansione T2w è sufficiente per il compito in questione. Infine, i risultati del Test set non presentano bias legati all'intensità dei pixel delle immagini evidenziando pertanto come la presenza delle lesioni tumorali nella prostata non influenzi negativamente le prestazioni del modello.

4.5 Confronto con la letteratura

Il confronto con la letteratura ha evidenziato una varietà di approcci e modelli utilizzati per la segmentazione della prostata.

- A differenza degli studi presenti in letteratura [1], nel presente lavoro, per l'addestramento del modello, si è fatto uso di maschere ottenute mediante modelli di intelligenza artificiale, fornite dagli organizzatori della challenge, le quali presentano minore precisione rispetto alle maschere manuali tracciate da radiologi esperti, introducendo un bias iniziale nelle performance.

- Il modello implementato fa uso della sola scansione T2w mentre in alcuni studi presenti in letteratura viene fatto uso anche delle sequenze ADC e DWI [45],[59].

- Nell'effettuare il resample delle scansioni si è fatto uso dell'*interpolazione Lineare* invece della Bicubica, ottenendo prestazioni leggermente migliori e riducendo il tempo computazionale di elaborazione (sottosezione 4.4.1: Metodi di interpolazione).

- *Rispetto alla UNet implementata, in contesti come le sfide scientifiche sono stati impiegati anche modelli più complessi e onerosi*, come ResNet 3D + Unet 3D [60], Dense-Unet [58] ed NNUnet [45] [59], framework noto per l'elevata capacità predittiva, flessibilità architetturale ed elevati costi computazionali.

- Durante la fase di training non è stato impiegato il metodo della cross-validazione. Infatti, sebbene consenta una migliore ottimizzazione della divisione del dataset, e di conseguenza delle prestazioni del modello, questo approccio richiede una considerevole quantità di risorse computazionali il che, considerando i risultati ottenuti, per questa specifica task presenta un *rapporto costi/benefici* sfavorevole.

Nonostante la qualità delle maschere usate per l'addestramento ed il minor uso di risorse computazionali, le prestazioni del modello sono paragonabili a quelle dei modelli presenti in letteratura.

Capitolo 5

Sviluppo dell'algoritmo di segmentazione delle csPCa

Il processo di segmentazione delle lesioni tumorali, rappresentato in verde nel flow chart 5.1, costituisce la componente principale dell'algoritmo implementato per l'analisi delle scansioni MRI della prostata.

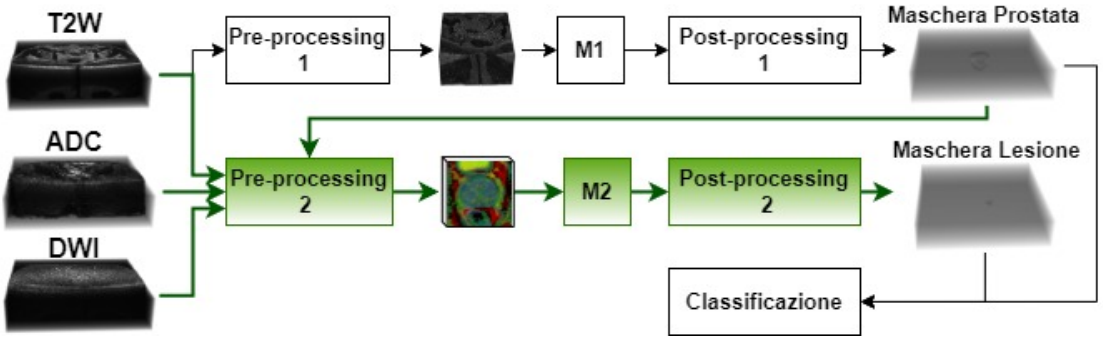


Figura 5.1: Flow chart dell'algoritmo. I processi descritti nel capitolo seguente sono evidenziati in verde.

Il suo *scopo* è quello di *identificare e segmentare* le eventuali lesioni tumorali presenti nella prostata (figura 5.2). La maschera delle lesioni tumorali generata viene impiegata nella fase finale del processo per effettuare la *classificazione* del paziente (sezione 6).

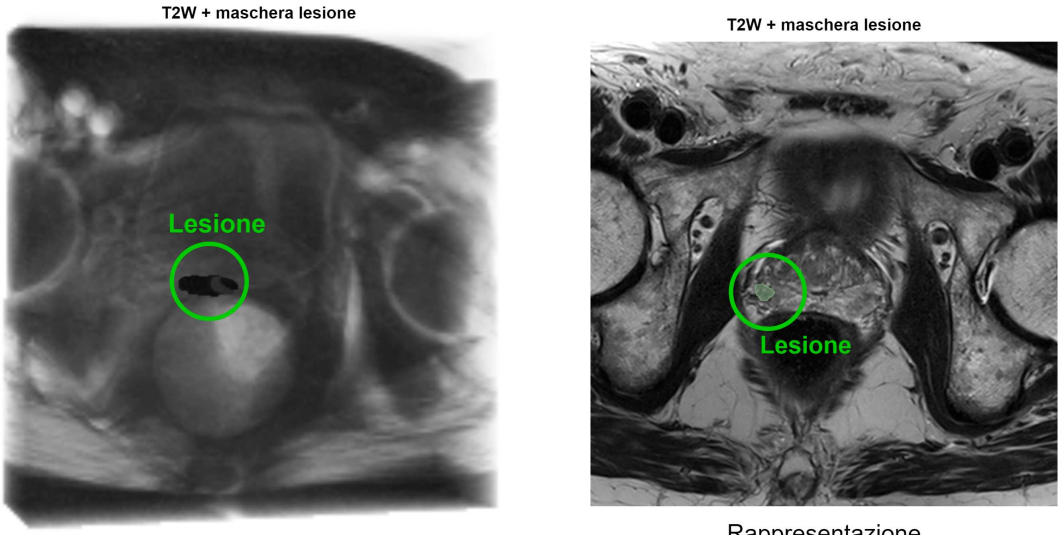


Figura 5.2: Rappresentazione 3D e 2D della segmentazione delle lesioni. Queste figure sono state ottenute sovrapponendo la maschera di segmentazione della lesione sulla scansione T2w.

Questo capitolo descrive il processo completo di sviluppo e ottimizzazione dell'algoritmo di segmentazione delle lesioni tumorali, valutando e analizzando criticamente ciascuna fase e le relative performance ottenute (flow chart 5.3).

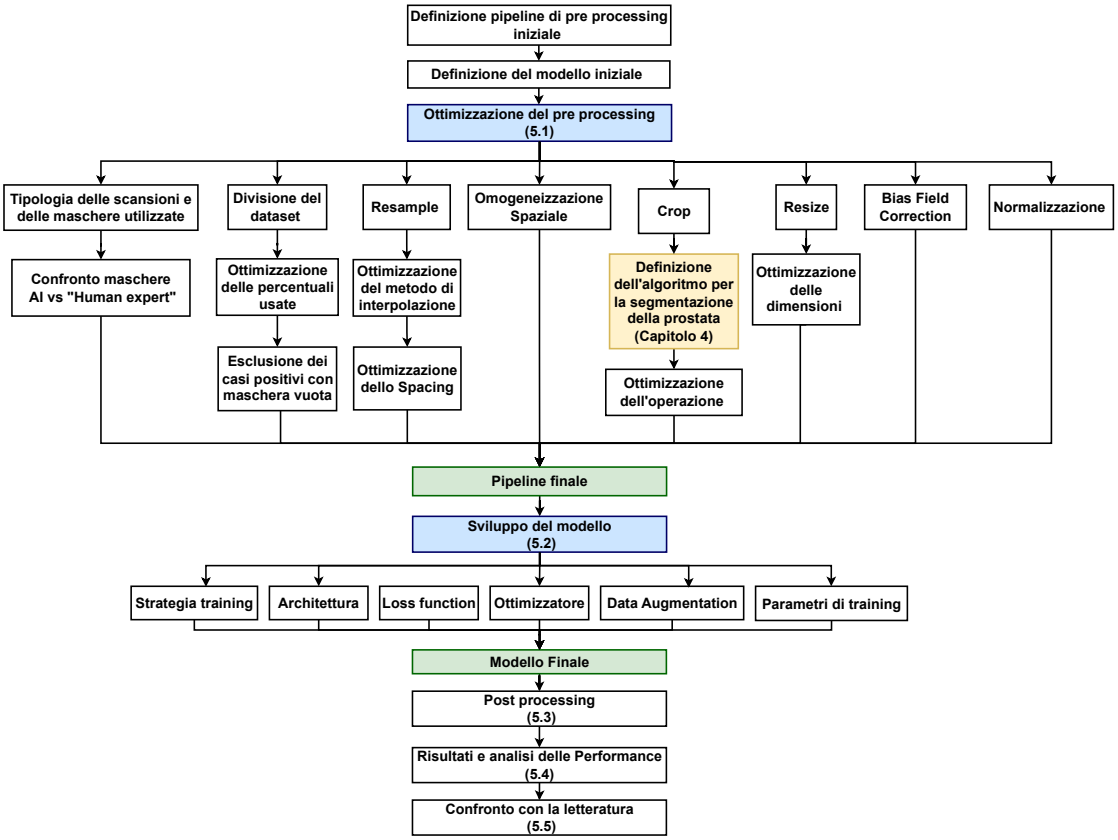


Figura 5.3: Flow chart del capitolo 5.

5.0.1 Pipeline di *pre-processing 2* iniziale

Di seguito sono elencate le operazioni implementate nelle pipeline di pre-processing iniziale definita sulla base della letteratura esistente.

Pipeline di pre-processing delle scansioni T2w ADC DWI:

1. *Resample (0.5,0.5,3)* [1] con interpolazione lineare.
2. *Crop* centrale [z,320,320] [2], z= numero slice del volume.
3. *Resize [16,256,256].* [3]

[1] (x,y,z)
[2] (z,x,y)
[3] (z,x,y)

4. *Normalizzazione*: Zscore + min-max scaling.

5. *Formazione del volume 4D ("rgb").*

Pipeline di pre-processing delle maschere delle lesioni usate nella fase di training:

1. *Resample (0.5,0.5,3)* [4] con interpolazione NearestNeighbor.

2. *Crop* centrale [z,320,320][5], z= numero slice del volume.

3. *Resize [16,256,256].* [6]

4. *Thresholding* con soglia 0,5.

5.0.2 Modello UNet *M2* iniziale

Per una valutazione quantitativa dell'impatto delle operazioni di pre-processing, è stato impiegato un modello UNet i cui parametri, riportati nella tabella 5.1, sono stati ottimizzati nella fase successiva del lavoro (sezione 5.2).

[4] (x,y,z)
[5] (z,x,y)
[6] (z,x,y)

Tabella 5.1: Tabella riassuntiva dei parametri e delle funzioni implementate per ottenere il modello UNet *M2* di partenza impiegato nella fase di ottimizzazione del pre-processing.

Architettura	model = UNet(spatial_dims=3, in_channels=3, out_channels=1, channels=(16,32,64,128), strides= ((2,2,2),(2,2,2),(2,2,2), num_res_units=2, norm=Norm.BATCH)
Loss function	DiceFocalLoss (include_background=False, to_onehot_y=False, sigmoid=False, softmax=False, other_act=None, squared_pred=True, jaccard=False, reduction='mean', smooth_nr=1e-05, smooth_dr=1e05, batch=True, gamma=2, focal_weight=None, weight=None, lambda_dice=1, lambda_focal=1)
Optimizer	Adam(lr=5e-4, weight_decay=5e-06)
Batch size	8
Data Augmentation	None
Patience	10

5.1 *Pre-processing 2*: ottimizzazione

In questa sezione viene trattato il processo di ottimizzazione della pipeline di pre-processing. L'obiettivo è definire una pipeline robusta alle variazioni presenti in un dataset reale, garantendo al contempo la generazione di volumi di buona qualità da utilizzare come input per il modello *M2*.

5.1.1 Tipologia delle scansioni usate

Per la task di segmentazione delle lesioni sono state utilizzate le scansioni T2w, ADC e DWI, combinate in un volume 4D, assimilabile a "RGB", in cui la quarta dimensione rappresenta la tipologia di scansione. Nella figura 5.4 è presentato un esempio di una slice estratta da ciascuna scansione e dal volume "rgb" risultante. Questo approccio permette di catturare le diverse caratteristiche delle lesioni evidenziate dalle diverse modalità di acquisizione, come discusso nella sezione 1.3.

Come effettuato in altri studi presenti in letteratura (sottosezione: 2.2), si è deciso di non utilizzare le scansioni *T2W sagittali e coronali*, in quanto

sono acquisizioni opzionali e non sempre disponibili. Inoltre, l'inclusione di queste ulteriori modalità comporterebbe un aumento significativo dei costi e dei tempi computazionali, senza garantire necessariamente un miglioramento delle prestazioni.

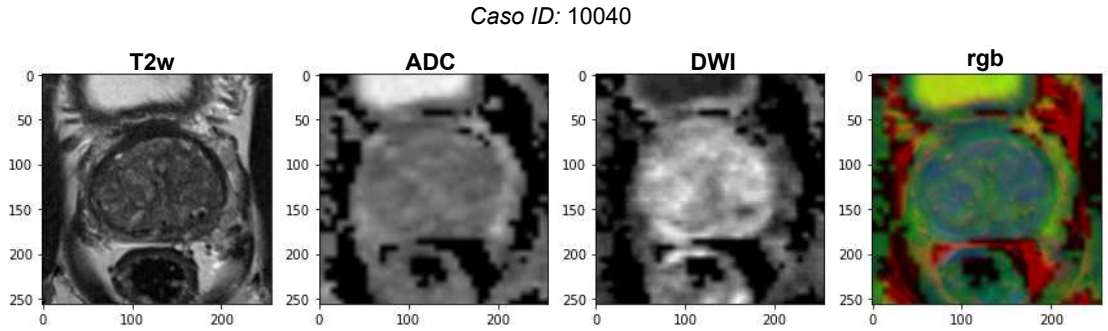


Figura 5.4: Esempio di slice estratte dalle scansioni T2w, ADC, DWI e dal volume 4D generato (rgb). Le immagini utilizzate appartengono al caso con ID 10040.

5.1.2 Tipologia delle maschere usate

Per questa task si è fatto uso delle seguenti maschere:

- *Maschere della prostata* di *"Guerbet23"*: impiegate per effettuare il pre-processing dei volumi utilizzati per lo sviluppo del modello *M2*. Durante questa fase si è preferito fare uso delle maschere fornite al fine di evitare eventuali bias legati alle performance del *modello M1*. Nella fase di *inference* invece viene fatto uso della maschera prostatica generata da *M1* al fine di valutare le performance complessive dell'algoritmo implementato.

- *Maschere AI* delle lesioni tumorali ottenute mediante metodo semi-supervisionato [38]. Queste sono state impiegate per il pre-addestramento del modello e per la valutazione delle performance nelle prime fasi di ottimizzazione.

- *Maschere Manuali "human expert"* delle lesioni tumorali tracciate da esperti radiologi. Queste sono state impiegate nel secondo training del modello e per la valutazione delle performance finali dell'algoritmo. Nelle maschere manuali, i pixel corrispondenti alle lesioni presentano intensità diverse a seconda della classe ISUP di appartenenza. Tuttavia,

il modello *M2* è progettato per identificare e segmentare le lesioni indipendentemente dalla loro classe. Pertanto, nella fase di pre-processing, le maschere sono state binarizzate.

5.1.3 Maschere AI vs Maschere "human expert"

Un passaggio importante per la definizione della strategia adottata è stato la valutazione della qualità delle maschere AI presenti nel dataset. Di queste sono state analizzate solo 225 maschere, corrispondenti ai casi per i quali è disponibile anche l'annotazione manuale ("human expert"). Questa valutazione è fondamentale per assicurare che il modello *M2* sia addestrato correttamente usando dati di buona qualità.

5.1.4 Divisione del dataset

La divisione del dataset in Train, Validation e Test set riveste un ruolo importante nella fase di costruzione del modello. Nella figura 5.5 è illustrata una rappresentazione grafica del processo di divisione del dataset.

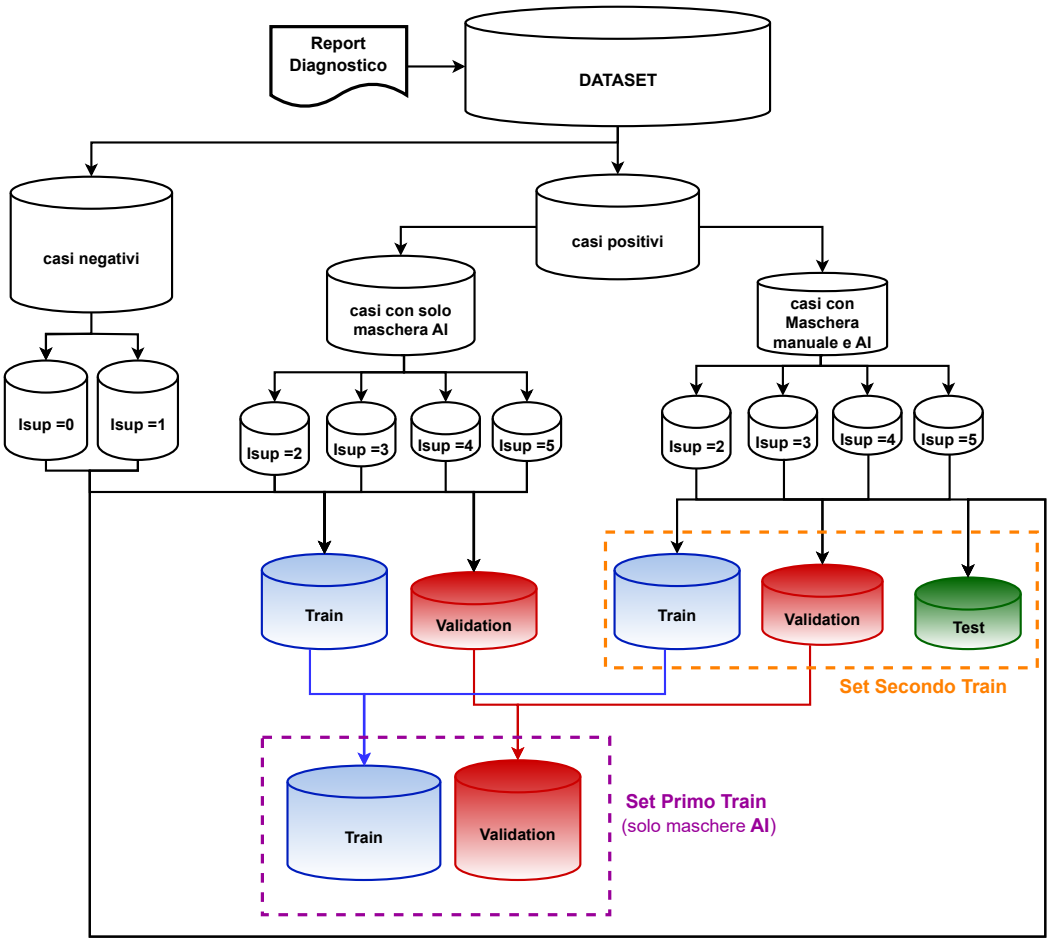


Figura 5.5: Rappresentazione grafica della divisione del dataset. "Set Primo Train" evidenziato nel riquadro viola include solo le maschere AI. I casi positivi nel "Set Secondo Train", contrassegnati dal riquadro arancione, includono solo le maschere manuali.

Come descritto nella sezione 3.1, mentre le maschere AI sono disponibili per tutti i casi del dataset, le maschere manuali sono disponibili solo per 225 dei 425 casi positivi.

Per sfruttare al meglio tutto il dataset a disposizione si è deciso di adottare la strategia del *Fine tuning*. In questa strategia il modello viene addestrato su due set di dati:

1. "Set Primo Train": evidenziato nel riquadro viola nella figura 5.5, è formato da n casi negativi e 425 positivi. Il set contiene solo le maschere AI.

2. "Set Secondo Train": evidenziato nel riquadro arancione, è formato da 225 casi positivi e da *n'* casi negativi. Il set contiene solo le maschere manuali.

Facendo uso del *Report diagnostico* si è effettuata una *divisione stratificata* del dataset. I set pertanto sono stati formati in maniera randomizzata ma garantendo, per ognuno di essi, la corretta rappresentazione di ciascuna classe ISUP. Durante l'analisi sperimentale sono state effettuate diverse prove per ottenere un corretto bilanciamento delle classi in ciascun set. Dall'analisi condotta e dai controlli eseguiti, è emerso che in alcuni casi il report diagnostico indicava la presenza di lesioni tumorali in maschere che in realtà risultavano essere vuote. Questi casi sono stati individuati e rimossi dal dataset.

È importante sottolineare che, come per la task di segmentazione della prostata, la divisione dei casi è stata effettuata in modo randomico e ripetibile, utilizzando *seed* fisso pari ad *1*, fondamentale per il corretto confronto delle diverse prove effettuate.

5.1.5 Identificazione di due nuove classi: "*piccole*" e "*grandi*" lesioni

Al fine di ottenere una valutazione più accurata delle performance del modello si è effettuata una *valutazione stratificata*. Nello specifico le metriche sono state calcolate e valutate sui casi positivi dell'intero set di riferimento e sui quattro subset, uno per ciascuna classe ISUP (>1), in cui sono stati suddivisi i casi in base alla loro classe di appartenenza. A questo scopo, sono state individuate ulteriori due classi basate sulle dimensioni, *"Piccole" e "Grandi"*, delle lesioni. La suddivisione è stata effettuata utilizzando come valore soglia il 25° percentile della distribuzione dei volumi delle lesioni presenti nel relativo set.

5.1.6 Resample & Normalizzazione

Nel contesto dell'ottimizzazione della pipeline di pre-processing, le considerazioni relative all'operazione di *Resample* e alla *Normalizzazione* sono analoghe a quelle effettuate per il processo di segmentazione della prostata. Per una trattazione più approfondita si rimanda alle sottosezioni 4.1.4

e 4.1.7, dove sono forniti dettagli specifici sulle metodologie e le scelte effettuate. Durante l'analisi sperimentale si è valutato l'impatto dello spacing sulla qualità complessiva delle immagini e, di conseguenza, sulle prestazioni del modello di segmentazione *M2*.

5.1.7 Omogeneizzazione spaziale

L'omogeneizzazione spaziale delle scansioni è un operazione fondamentale per evitare disallineamenti critici durante la transizione dalle immagini all'array matriciali.

Dall'analisi del dataset è emerso che le scansioni sono generalmente allineate nello spazio per garantire la sovrapposizione spaziale della prostata. Per ottenere questa sovrapposizione, le immagini spesso non condividono la stessa origine spaziale e/o lo stesso centro. Tipicamente le acquisizioni ADC e DWI mostrano coerenza nelle coordinate spaziali, nello spacing e nelle dimensioni differendo dalle T2W e dalle maschere prostatiche. Nella figura 5.6 è presentato un esempio di questa discrepanza dove, per ciascuna tipologia di scansione, viene mostrata una slice e le relative informazioni estratte dall'immagine.

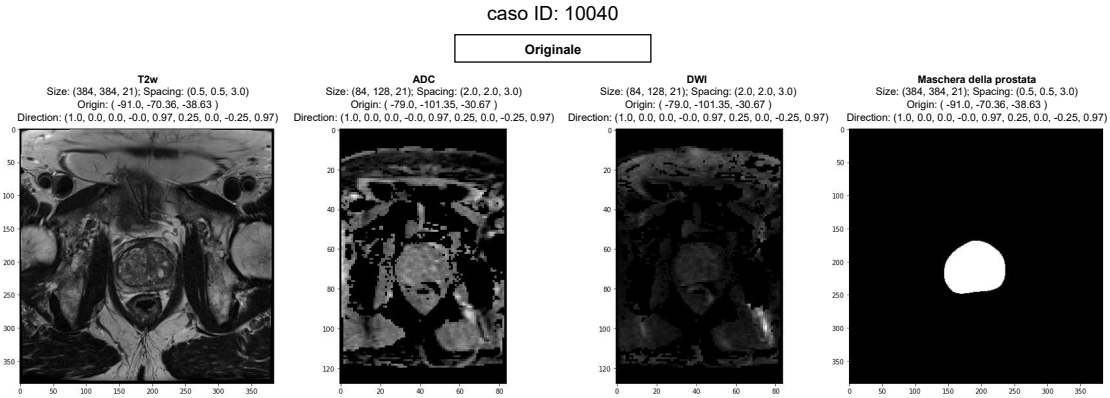


Figura 5.6: Illustrazione delle variazioni nelle caratteristiche delle scansioni: per ciascuna tipologia è presentata una slice con le informazioni estratte dall'immagine corrispondente.

Inoltre, per agevolarne la comprensione sono presentati nella figura 5.7 due esempi in cui sono mostrati i plot tridimensionali delle immagini T2w e ADC di due casi del dataset. La scansione DWI si sovrappone perfettamente all' ADC, così come la maschera prostatica alla T2w, pertanto queste non sono state riportate negli esempi visivi.

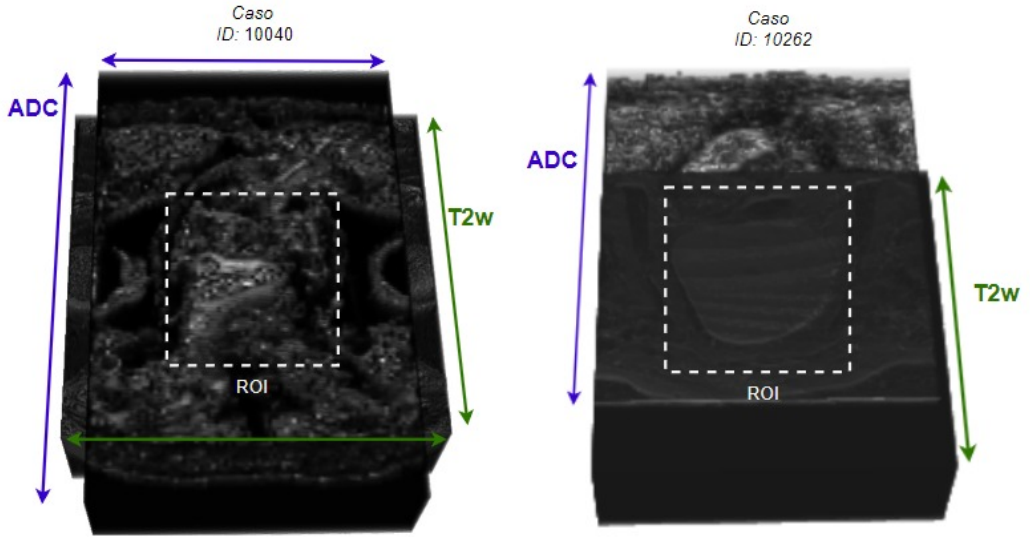


Figura 5.7: Plot tridimensionali delle immagini T2w e ADC relative ai casi con ID 10040 e 10262. Nelle figure, la *Roi* è evidenziata da un riquadro tratteggiato di colore bianco.

Quando viene eseguita la transizione dall'immagine all'array matriciale, necessaria per eseguire le successive operazioni di pre-processing, l'informazione spaziale non viene mantenuta causando un disallineamento critico tra le scansioni. Questo è evidente soprattutto nei casi in cui la prostata non occupa il centro dell'immagine.

La funzione implementata ha come obiettivo quello di risolve in modo ottimale questa problematica. Per eseguire l'omogeneizzazione spaziale, le immagini T2W, ADC e DWI sono state ridimensionate, con padding di zeri, alle dimensioni massime per ciascuna direzione sfruttando la funzione 'Resample' del toolkit SimpleITK. È importante sottolineare che l'operazione di padding tiene conto della direzione spaziale lungo la quale aggiungere gli zeri.

Nella figura 5.8 è illustrato il flow chart descrittivo della funzione implementata.

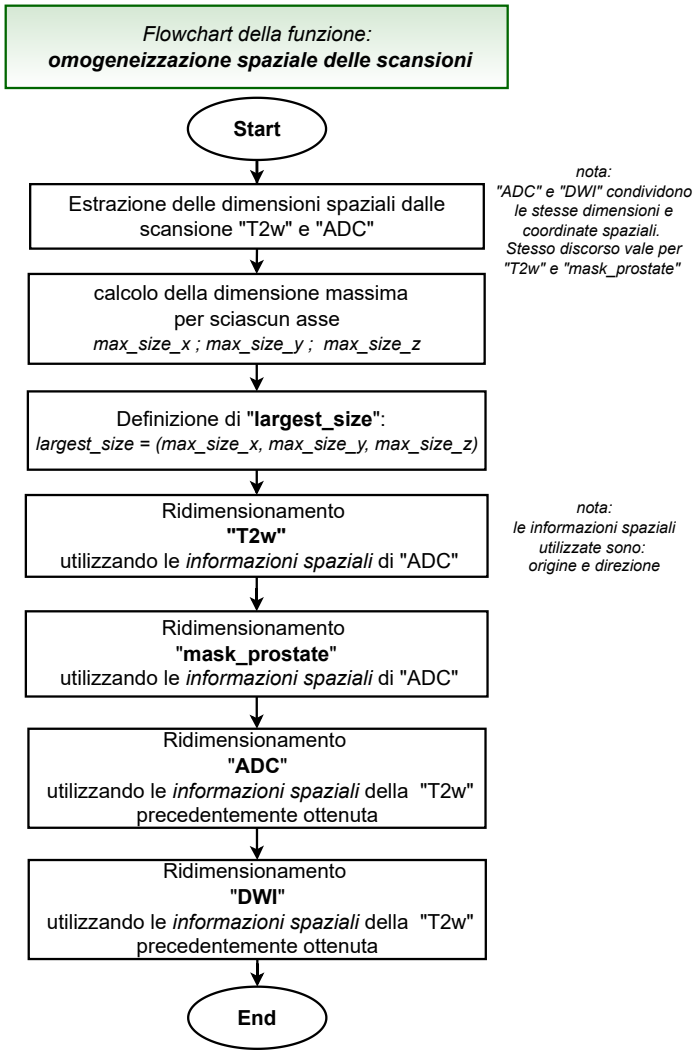


Figura 5.8: Flow chart della funzione utilizzata per effettuare l'omogeneizzazione spaziale delle immagini.

Nella figura 5.9 è illustrato un esempio dell'effetto dell'operazione di omogeneizzazione spaziale. Si specifica che tutte le operazioni sono state eseguite sulle immagini 3D. Per una migliore comprensione visiva, negli esempi viene mostrata una slice estratta da ciascun volume.

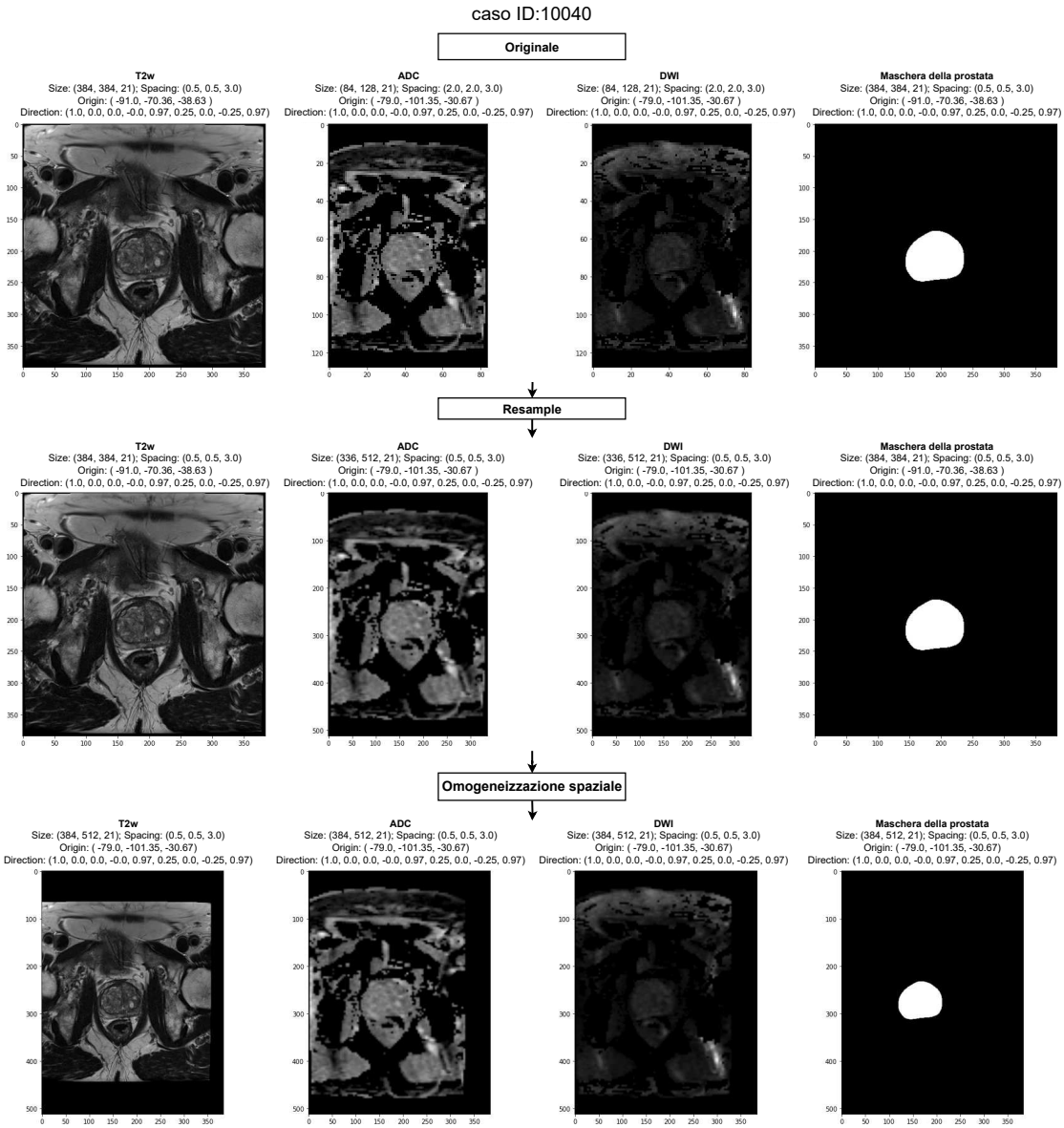


Figura 5.9: Esempio dell'effetto dell'operazione di omogeneizzazione spaziale. Nella figura sono riportate le slice estratte da ciascuna scansione e dalla maschera prostatica. A ciascuna slice sono associate le informazioni estratte dalle relative immagini originali, post resample e post omogeneizzazione spaziale.

5.1.8 Crop

L'operazione di crop ha come obiettivo quello di concentrare l'attenzione del modello sulla ROI al fine di migliorarne le prestazioni di predizione.

Nella figura 5.10 è presentato il flow chart descrittivo degli step logici seguiti durante il processo di ottimizzazione dell'operazione. Durante l'analisi sperimentale, sono state esaminate diverse problematiche e, per ciascuna di essa, è stata proposta ed implementata una soluzione.

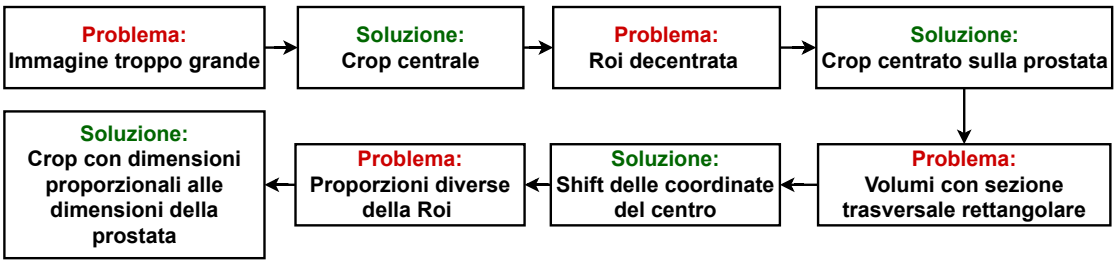


Figura 5.10: Flow chart degli step logici seguiti durante il processo di ottimizzazione dell'operazione di Crop.

La pipeline iniziale di pre-processing prevede, seguendo le pratiche adottate da diversi gruppi di ricerca [47] [31] [32], l'implementazione dell'operazione di crop centrato sulla scansione. Nello specifico si sono testate finestre di dimensioni [z,320,320] e [z,150,150] dove z rappresenta il numero di slice delle scansioni. L'obiettivo è quello di risolvere il problema legato alle grandi dimensioni dei volumi. Tuttavia, questa procedura presenta i limiti, discussi nel dettaglio nel capitolo 4 dedicato alla segmentazione della prostata, relativi al posizionamento spaziale della prostata e alle diverse dimensioni delle finestre di acquisizione. Di conseguenza si è fatto uso della maschera prostatica per eseguire il crop centrato sulla *ROI* con dimensioni [z',150,150], dove z' rappresenta il numero di slice in cui è presente la ghiandola. Questa operazione è stata a sua volta sottoposta ad un processo di analisi e ottimizzazione. Nei casi in cui la prostata è decentrata e si trova vicina ai bordi dell'immagine, questo approccio potrebbe generare volumi con sezione trasversale rettangolare. Questo rappresenterebbe un problema durante la fase successiva di resize, poiché potrebbero verificarsi distorsioni nella scansione. Per affrontare questa eventualità, è stata implementata all'interno della funzione di crop un controllo che, se necessario, modifica automaticamente le coordinate del centro, garantendo così la generazione di volumi con sezione trasversale quadrata.

Un'ulteriore modifica, testata durante l'analisi sperimentale, riguarda l'utilizzo di una finestra con dimensioni non fisse ma proporzionali a quelle della prostata. Nello specifico l'operazione prevede, come fase preliminare, il calcolo della dimensione ottimale da utilizzare per effettuare il crop. Questa viene determinata selezionando, facendo uso della maschera prostatica, la

massima tra le estensioni lungo le direzioni X (larghezza) e Y (lunghezza) della prostata. L'idea è quella di adattare la dimensione della finestra sul singolo paziente al fine di risolvere il problema delle proporzioni disomogenee della ROI tra i vari casi del dataset.La dimensione d della finestra è incrementata di un fattore pari a 10 pixel (5 per ogni direzione), per garantire un margine intorno alla ghiandola. Inoltre si sono stabiliti valori minimi e massimi di 90 e 150 al fine di garantire la robustezza dell'operazione ed evitare l'implementazione di finestre troppo piccole. La dimensione massima di 150 si è rivelata essere adatta per l'inclusione dell'intera prostata (come discusso nella sottosezione 4.1.6). La dimensione minima di 90 impedisce un'eccessiva deformazione delle immagini durante la fase di resize, evitando così una significativa perdita di risoluzione spaziale. Nella figura 5.11 è riportato il flow chart descrittivo degli step eseguiti per effettuare l'operazione di crop ottimizzata.

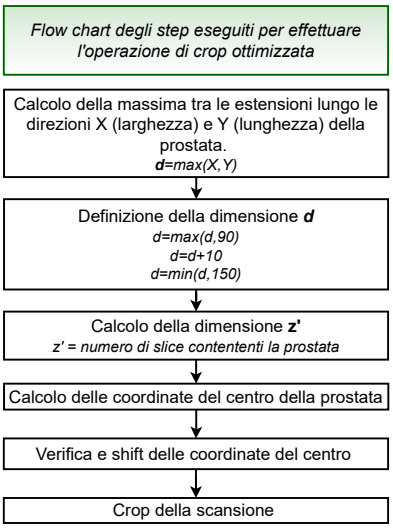


Figura 5.11: Flow chart degli step eseguiti per effettuare l'operazione di crop ottimizzata.

5.1.9 Resize

L'operazione di resize è stata implementata con l'obiettivo principale di standardizzare le dimensioni delle scansioni e renderle adatte per il corretto funzionamento del modello. Questa consente inoltre di ampliare la ROI aumentando di conseguenza le dimensioni delle lesioni. L'idea alla base del resize infatti è anche quella di agevolare il modello nell'analisi delle lesioni di *piccole dimensioni* presenti nei volumi. Durante l'analisi sperimentale si è

evidenziata la difficoltà del modello nel segmentare correttamente la lesione in particolare lungo l'asse Z. Di conseguenza, sono state condotte prove mirate ad identificare la dimensione z di ridimensionamento più appropriata.

5.1.10 Bias Field Correction

La Bias Field Correction [50] è un'operazione utilizzata per rimuovere le variazioni non uniformi nell'intensità del segnale nelle immagini mediche, causate da fattori tecnici e fisici quali, per esempio:

- non uniformità dei campi magnetici;
- diversa suscettibilità magnetica dei tessuti che causa alterazioni del campo magnetico locale;
- artefatti da movimento del paziente durante l'acquisizione.

Questo fenomeno si manifesta tipicamente come un gradiente di intensità che varia gradualmente all'interno dell'immagine, introducendo distorsioni dell'intensità dei pixel che potrebbero compromettere l'analisi accurata dei tessuti da parte del modello. L'effetto del *bias field* si manifesta negli istogrammi delle luminosità dei volumi attraverso dei picchi. Durante l'analisi sperimentale si è analizzata la presenza di tale artefatto nelle scansioni e si è valutato l'impatto, sulle prestazioni del modello, dell' applicazione dell'algoritmo *N4BiasFieldCorrection* della libreria ITK impiegato per la correzione.

5.2 Sviluppo del modello *M2*

In questa sezione, viene esaminato il processo di sviluppo del modello *M2* utilizzato per effettuare in modo automatico la segmentazione delle lesioni tumorali nella prostata. Questo include la scelta e l'ottimizzazione dell'architettura, delle funzioni e dei parametri utilizzati.

5.2.1 Strategia di addestramento: *fine-tuning*

Nella ricerca della precisione e dell'affidabilità nella segmentazione delle lesioni prostatiche, è stata adottata la strategia di addestramento del modello nota in letteratura come *fine-tuning*. Questa è stata implementata

con l'obiettivo di ottimizzare l'uso delle risorse disponibili e massimizzare le performance.

Secondo questa strategia, inizialmente il modello viene pre-addestrato utilizzando un dataset più ampio e contenente le maschere automatiche generate da un algoritmo di intelligenza artificiale (maschere AI). L'obiettivo di questa prima fase è consentire al modello di apprendere i pattern comuni e le caratteristiche principali delle lesioni prostatiche. In seguito, il modello è sottoposto al processo di fine-tuning utilizzando un secondo dataset più piccolo, composto da maschere manuali di maggiore qualità (maschere "human expert"). Questo consiste in un secondo training eseguito per consentire al modello di adattarsi meglio alle caratteristiche delle lesioni prostatiche, migliorando la sua capacità di segmentazione sui casi più complessi e dettagliati. L'adozione di questa strategia consente l'utilizzo di tutte le maschere fornite senza dover scartare le maschere AI dei 220 casi in cui è disponibile anche quella manuale.

5.2.2 Ricerca dell'architettura ottimale

La struttura del modello UNet utilizzata per questa task prevede l'uso di convoluzioni tridimensionali che consentono alla rete di catturare le correlazioni spaziali tra i voxel del volume. La rete UNet prende in input il volume 4D generato nella fase di pre-processing 2 e restituisce come output uno score dove i valori più alti indicano la presenza della lesione nella prostata. Durante il processo di ricerca dell'architettura ottimale si sono effettuate numerose prove volte a valutare l'impatto dell'architettura sulle performance di predizione. Le diverse configurazioni sono state sottoposte ad un processo di ottimizzazione dei parametri dell'ottimizzatore, poiché la combinazione ottimale di learning rate e weight decay potrebbe variare a seconda dell'architettura del modello.

5.2.3 Valutazione delle funzioni di Loss

Durante l'analisi sperimentale si sono valutate diverse Loss function messe a disposizione dalla libreria MONAI. Per determinare la funzione più adatta per la task si sono confrontate le performance ottenute utilizzando il modello con l'architettura ottimale identificata nell'analisi precedente.

5.2.4 Analisi degli Ottimizzatori

Nel contesto dell'ottimizzazione del modello, un aspetto cruciale riguarda la selezione dell'ottimizzatore e la calibrazione dei suoi parametri interni, quali il learning rate (LR) e il weight decay (WD). Durante l'analisi sperimentale, si sono valutati diversi tipi di ottimizzatori attraverso un approccio iterativo mirato a individuare la combinazione ottimale dei parametri interni.

5.2.5 Operazioni di Data Augmentation

Durante la fase di sviluppo del modello sono state provate diverse operazioni di data augmentation.

- Rotazione random: Questa operazione esegue una rotazione delle immagini e delle etichette associate nel piano XY con una probabilità del 50%. L'obiettivo è quello di rendere il modello robusto elle variazioni di orientamento.
- Flip random lungo l'asse Y: questa operazione esegue il flip delle immagini e delle etichette associate lungo l'asse Y con una probabilità del 50%. L'obiettivo di questa operazione è quello di aumentare la diversità del dataset.
- Simulazione del disallineamento delle scansioni: questa operazione esegue uno shift di un numero random di pixel compreso tra 0 e 5. L'operazione viene eseguita solo sul 2° e 3° canale del volume, ovvero sulle scansioni ADC e DWI, con l'obiettivo di rendere il modello robusto alla non co-registrazione delle scansioni.
- Zoom in-out random: Questa operazione esegue lo Zoom in-out con una probabilità del 50%. L'obiettivo è migliorare la qualità della segmentazione delle lesioni di piccole dimensioni.

5.3 *Post-processing 2*: ottimizzazione

In questa sezione sono analizzate le operazioni di post-processing effettuate sulle maschere generate dal modello *M2*.

Il processo di post-processing inizia con la normalizzazione nell'intervallo [0,1] della mappa di predizione, ottenuta come output dal modello, tramite

l'applicazione della funzione di attivazione sigmoidale. Per garantire che la maschera risultante sia perfettamente sovrapponibile con la scansione T2w, come richiesto dalla challenge, vengono eseguite le seguenti operazioni:

1. *Flip verticale*: utile per compensare la rotazione interna effettuata durante l'applicazione del modello.

2. *Resize*: la maschera delle lesioni viene ridimensionata alle dimensioni delle scansioni assunte dopo l'operazione di crop [d,d,z'].

3. *Zero padding*: operazione inversa al crop centrato sulla prostata. La funzione implementata restituisce come output la maschera delle lesioni avente le stesse dimensioni della scansione T2w post resample.

4. *Thresholding*: Tutti i pixel con un valore inferiore ad una determinata soglia vengono posti a 0. Questa operazione ha come obiettivo quello di eliminare i falsi positivi e migliorare la precisione complessiva delle predizioni. Durante l'analisi sperimentale si sono valutati gli effetti dell'utilizzo di diversi valori. Per il calcolo delle metriche la maschera viene binarizzata ponendo pari ad 1 tutti i pixel con valore superiore alla soglia identificata.

5. *Resample*: l'immagine della maschera delle lesioni viene ridimensionata alle dimensioni e allo spacing dell'immagine T2w originale utilizzando l'interpolazione *nearest neighbor*. Questo processo garantisce la coerenza spaziale della predizione rispetto all'immagine di riferimento.

6. *Sovrascrizione dei metadati* dell'immagine generata con quelli della scansione T2w.

5.4 Risultati e analisi delle performance

In questa sezione sono riportati e commentati i risultati relativi alle analisi sperimentali effettuate durante le fasi di ottimizzazione. Infine vengono analizzate le performance finali dell'algoritmo implementato. Per semplificarne la lettura ed il confronto, per le diverse analisi effettuate si è scelto di riportare solo le prove ritenute più significative.

5.4.1 *Pre-processing 2*

Nella seguente sotto sezione sono riportati i risultati delle analisi sperimentali effettuate durante la fase di ottimizzazione della pipeline di *pre-processing 2*. In questa fase i modelli confrontati non sono stati sottoposti al processo di *fine tuning* al fine di ottimizzare l'uso delle risorse e del tempo computazionale. L'obiettivo infatti è quello di confrontare e valutare l'impatto delle diverse operazioni sulle prestazioni del modello *M2*.

Confronto maschere AI vs manuali "human expert"

Di seguito, nella tabella 5.2 e nella figura 5.12, sono riportati i valori delle metriche di valutazione ed i boxplot dei valori di Dice e distanza di Hausdorff (95° percentile) utilizzati per il confronto delle maschere.

Tabella 5.2: Tabella di valutazione della qualità delle maschere AI fornite mediante confronto con le maschere manuali.

Metrica	Mean	± Std
DICE	0.6526	± 0.1973
HD_{95}	9.5079	± 10.751
RVD	-0.1639	± 0.3792
PRECISIONE	0.7572	± 0.1195

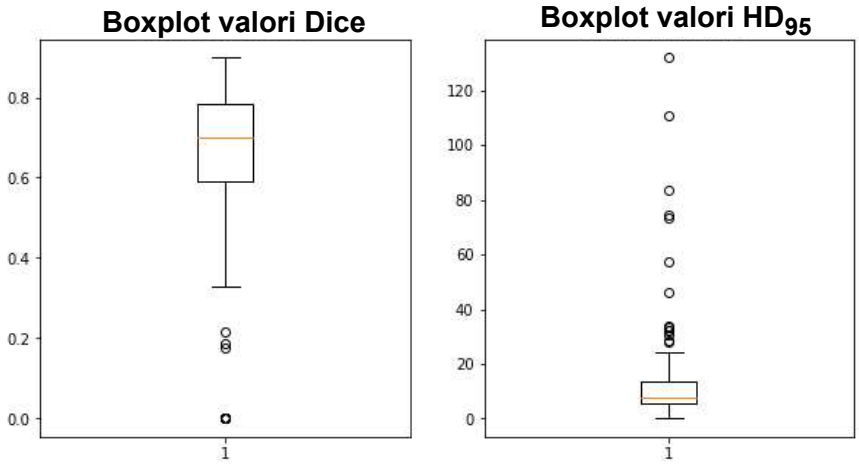


Figura 5.12: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) dell'analisi effettuata per il confronto delle maschere AI con quelle manuali.

Dall'analisi effettuata emerge come la qualità delle maschere AI, pur essendo state generate tramite un metodo semi-supervisionato, non sia particolarmente elevata. Per un confronto visivo, nella figura 5.13 è riportato un esempio di una maschera AI e di una manuale sovrapposte a una scansione T2w.

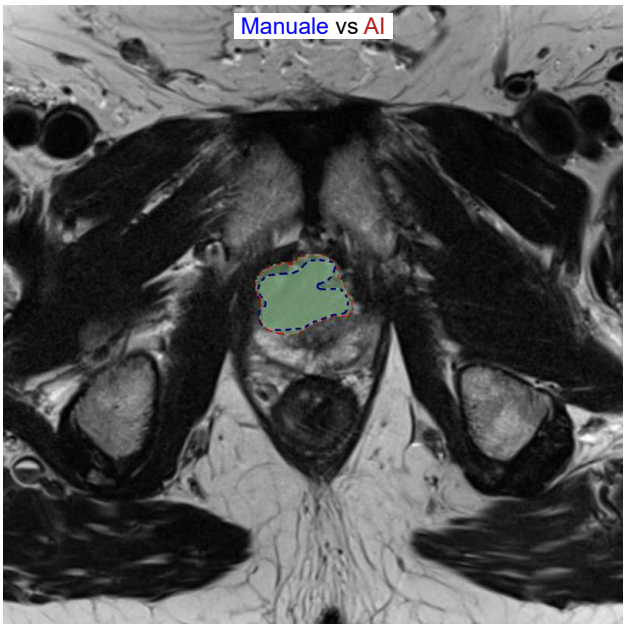


Figura 5.13: Confronto visivo di una maschera AI e di una manuale sovrapposte a una scansione T2w. Le differenze tra le due maschere sono evidenziate dai contorni tratteggiati, in rosso per la AI e in blu per la Manuale, evidenziando le discrepanze nella segmentazione della lesione tumorale.

Utilizzando solo le maschere AI per la costruzione del modello, si verificherebbe un significativo bias iniziale nelle prestazioni legato alla qualità delle maschere usate. D'altra parte, l'uso delle sole maschere manuali renderebbe il set di dati troppo piccolo e il modello potrebbe non essere in grado di generalizzare correttamente. Questo inoltre comporterebbe un utilizzo non ottimale del dataset a disposizione. Di conseguenza, si è adottata una strategia che prevede l'utilizzo di entrambe le maschere per l'addestramento del modello. In particolare i casi con le maschere AI sono stati utilizzati per il primo allenamento del modello. Successivamente, mediante fine-tuning, eseguito solo sui casi con maschere manuali, si è migliorata la sensibilità e la precisione della segmentazione delle lesioni tumorali.

Divisione del dataset

Di seguito sono riportati i risultati delle prove più significative effettuate durante il processo di ottimizzazione della divisione del dataset. I modelli implementati, i cui parametri sono riportati nella tabella 5.1, sono stati ottenuti utilizzando i rispettivi set preparati mediante la pipeline di *preprocessing iniziale* (sotto sezione 5.0.1). Le percentuali indicate sono state

utilizzate per stimare, per ciascuna classe ISUP, il numero di casi da inserire nei relativi set. I set delle prove confrontate sono così costituiti:

- "Solo casi positivi": set formati da soli casi positivi diviso in:
  - *Train set primo training*: 70% dei 205 casi positivi con solo maschera AI più i casi del *Train set Human*.
  - *Validation set primo training*: 30% dei 205 casi positivi con solo maschera AI più i casi del *Validation set Human*.
  - *Train set Human*: 70% dei 220 casi positivi con maschera manuale.
  - *Validation set Human*: 20% dei 220 casi positivi con maschera manuale.
  - *Test set*: 10% dei 220 casi positivi con maschera manuale.
- "Inclusione negativi": set formati da casi positivi e negativi. I negativi sono circa pari al 30% dei positivi.
  - *Train set primo training*: 70% dei 205 casi positivi con solo maschera AI più l' 70% dei 60 casi negativi più i casi del *Train set Human*.
  - *Validation set primo training*: 30% dei 205 casi positivi con solo maschera AI più il 20% degli 60 casi negativi più i casi del *Validation set Human*.
  - *Train set Human*: 70% dei 220 casi positivi con maschera manuale più il 70% dei 66 casi negativi.
  - *Validation set Human*: 20% dei 220 casi positivi con maschera manuale più il 20% dei 66 casi negativi.
  - *Test set*: 10% dei 220 casi positivi con maschera manuale più il 10% dei 66 casi negativi.
- "Configurazione Ottimale": set formati da casi positivi e negativi. I negativi sono circa pari al 20% dei positivi.
  - *Train set primo training*: 80% dei 205 casi positivi con solo maschera AI più circa il 80% dei 42 casi negativi più i casi del *Train set Human*.

  - *Validation set primo training*: 20% dei 205 casi positivi con solo maschera AI più il 20% degli 42 casi negativi più i casi del *Validation set Human*.
  - *Train set Human*: 70% dei 220 casi positivi con maschera manuale più il 70% dei 46 casi negativi.
  - *Validation set Human*: 20% dei 220 casi positivi con maschera manuale più il 20% dei 46 casi negativi.
  - *Test set*: 10% dei 220 casi positivi con maschera manuale più il 10% dei 46 casi negativi.

- "Esclusione casi dubbi": set formati con la stessa configurazione della prova "Configurazione Ottimale" in cui sono stati esclusi i casi positivi con maschera vuota.

Nella tabella 5.3 sono riportati i valori di Dice calcolati sullo stesso Test set interno.

Tabella 5.3: Tabella di valutazione dell'effetto della configurazione del dataset sulle prestazioni del modello. Per ciascuna prova è riportato il valore medio ± std del Dice calcolato sul Test set e sui relativi subset. In verde è evidenziata la prova ritenuta migliore.

Metrica : Dice	Numero casi	Solo casi positivi	Inclusione negativi	Configurazione ottimale	Esclusione casi dubbi
Isup 2	14	0.22 ± 0.21	0.26 ± 0.24	0.28 ± 0.26	0.29 ± 0.25
Isup 3	6	0.28 ± 0.22	0.31 ± 0.32	0.32 ± 0.35	0.34 ± 0.34
Isup 4	4	0.18 ± 0.26	0.25 ± 0.30	0.27 ± 0.32	0.28 ± 0.29
Isup 5	3	0.23 ± 0.25	0.26 ± 0.31	0.27 ± 0.35	0.28 ± 0.35
Piccola dim.	7	0.04 ± 0.15	0.06 ± 0.17	0.08 ± 0.19	0.08 ± 0.18
Grande dim.	20	0.32 ± 0.22	0.36 ± 0.24	0.35 ± 0.29	0.37 ± 0.28
Set completo	27	0.25 ± 0.21	0.27 ± 0.23	0.29 ± 0.28	0.30 ± 0.29

Le analisi effettuate hanno evidenziato come l'inclusione dei casi negativi (prova: *Inclusione negativi* tabella 5.3) consenta di migliorare le performance di predizione del modello riducendone la tendenza nell'identificare erroneamente lesioni tumorali nei pazienti sani.

I risultati dimostrano inoltre come il corretto bilanciamento delle classi nei set consenta di ottimizzare l'addestramento del modello ottenendo prestazioni migliori e più stabili (prova: *Configurazione ottimale* tabella 5.3).

Infine, dalle analisi e dai controlli effettuati sono emersi alcuni casi in cui il report diagnostico segnalava la presenza di lesioni tumorali che tuttavia non erano presenti nelle relative maschere AI e/o manuali che risultavano essere vuote. Sebbene pochi in numero, l'esclusione di questi casi (i cui ID sono riportati nella tabella 5.4), ha permesso di ottenere un leggero miglioramento delle prestazioni (prova: *Esclusione casi dubbi* tabella 5.3). Infatti, durante l'addestramento di un modello, è fondamentale eliminare tutti i casi "dubbi" e fornire degli input "corretti" ovvero, nel caso specifico, delle immagini dove il tessuto sano ed il tessuto tumorale siano classificati con un livello di precisione, sensibilità e accuratezza quanto più alto possibile.

Tabella 5.4: Elenco degli ID dei casi esclusi.

ID casi esclusi
11157 ; 10577 ; 10882 ; 10104 ; 11350 ; 11143 ; 10555 ; 10094
10811 ; 10397 ; 10658 ; 11243 ; 10508 ; 10660 ; 10636 ; 11296

Nella Figura 5.14, è presentata una rappresentazione grafica della divisione del Dataset. Ogni set è accompagnato da un diagramma a torta che ne descrive la composizione. Le percentuali riportate nella figura sono quelle della *configurazione ottimale*. Per agevolarne la consultazione, ciascuna coppia set-diagramma è stata incorniciata e numerata.

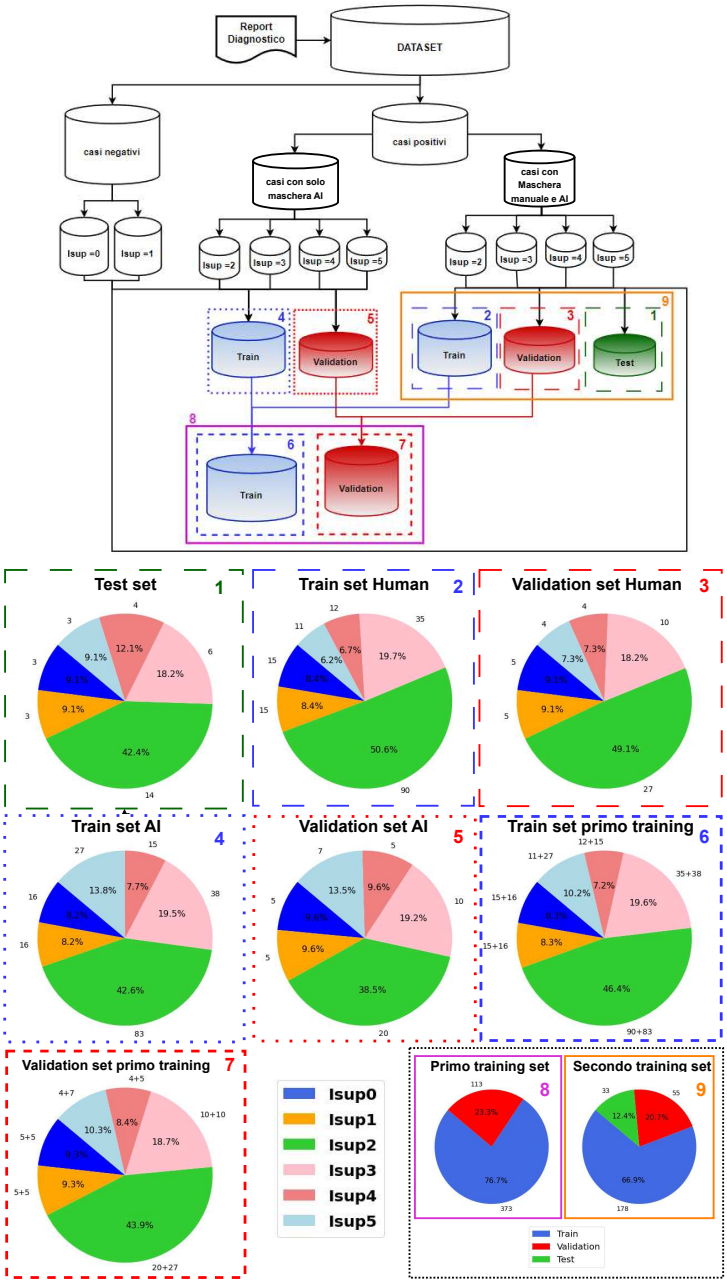


Figura 5.14: Visualizzazione grafica della divisione del Dataset. Ogni set è accompagnato da un diagramma a torta che ne descrive la composizione. Le percentuali riportate sono quelle relative alla *configurazione ottimale*, mentre i numeri neri indicano il numero esatto dei casi inseriti. Per agevolare la consultazione, ciascuna coppia set-diagramma è stata numerata e incorniciata.

Resample

Per la scelta del metodo di interpolazione si rimanda ai risultati del confronto tra l'interpolazione lineare e quella bicubica descritti nella sottosezione 4.4.1: Metodi di interpolazione. Per una migliore valutazione dell'impatto dello Spacing, adottato per effettuare il resample iniziale, si sono riportati nella figura 5.15 i boxplot dei valori di Dice e Distanza di Hausdorff (95° percentile) calcolati sui rispettivi Validation set.

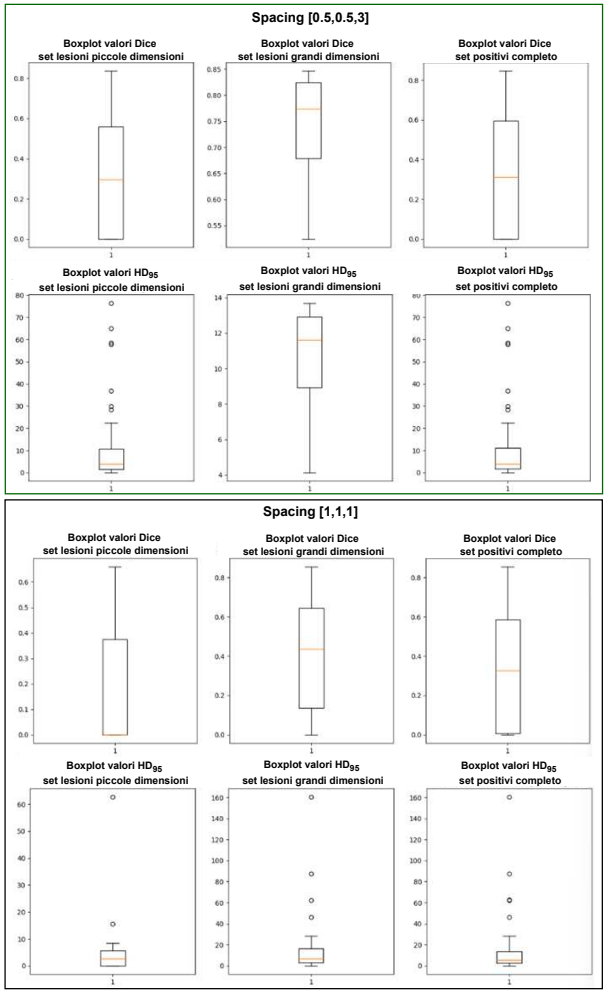


Figura 5.15: Box plot dei valori di Dice e distanza di Hausdorff (95° percentile) ottenuti sui rispettivi Validation set. Nel riquadro verde sono riportati i plot relativi all'analisi effettuata mediante spacing pari a (0.5,0.5,3). Nel riquadro nero sono riportati i plot relativi all'analisi effettuata mediante spacing pari a (1,1,1).

Dall'osservazione dei boxplot emerge che l'utilizzo di uno spacing di (0.5, 0.5, 3) consente di ottenere prestazioni più stabili. Pertanto, in linea con quanto riportato in letteratura per questa specifica task, è stata selezionata la configurazione pari a (0.5, 0.5, 3) per eseguire il resample delle immagini.

Omogeneizzazione spaziale

Di seguito sono presentati i risultati delle metriche utilizzate per valutare quantitativamente l'impatto della funzione di omogeneizzazione spaziale sulle prestazioni del modello. Le performance sono state valutate e confrontate utilizzando tutti i risultati e i relativi boxplot. Per agevolarne il confronto, nella tabella 5.6 sono riportate le metriche di Dice e Distanza di Hausdorff (95° percentile) calcolate sui relativi subset, mentre nella tabella 5.5 sono riportati i valori di tutte le metriche calcolate sui Validation set completi.

Tabella 5.5: Tabella di valutazione dell'impatto dell'operazione di omogeneizzazione spaziale delle immagini sulle prestazioni del modello. Le metriche sono state calcolate sui relativi Validation set. A *Riferimento* sono associati i valori ottenuti facendo uso della pipeline di *pre-processing 2 iniziale*. A *Prova* sono associati i valori ottenuti implementando nella pipeline la funzione di omogeneizzazione spaziale. In verde è evidenziata la prova considerata migliore. Per ciascun set è riportato, nella colonna *Num.*, il numero di casi contenuti.

Analisi omog. spaziale		**Riferimento**		**Prova omogeneizzazione spaziale**	
Classe	**Num.**	**Dice**	**HD_{95}**	**Dice**	**HD_{95}**
Isup 2	42	0.30 ± 0.28	9.20 ± 15.70	0.38 ± 0.27	7.57 ± 9.35
Isup 3	20	0.36 ± 0.28	7.45± 7.41	0.41 ± 0.29	8.48 ± 10.46
Isup 4	9	0.24 ± 0.31	14.98 ± 23.82	0.32 ± 0.30	15.70 ± 21.66
Isup 5	11	0.43 ± 0.27	7.08 ± 6.82	0.40 ± 0.30	7 ± 4.98
Piccola dim.	21	0.11 ± 0.19	5.39± 13.87	0.16 ± 0.22	4.88 ± 9.22
Grande dim	61	0.40 ± 0.28	10.41± 14.63	0.46 ± 0.26	9.89 ± 11.94
Set completo	82	0.33 ± 0.29	9.12±14.60	0.38 ± 0.28	8.61 ± 11.52

Tabella 5.6: Tabella di valutazione dell'impatto dell'operazione di omogeneizzazione spaziale delle immagini sulle prestazioni del modello. Le metriche sono state calcolate sui relativi Validation set completi. A *Riferimento* sono associati i valori ottenuti facendo uso della pipeline di *pre-processing 2 iniziale*. A *Prova* sono associati i valori ottenuti implementando nella pipeline la funzione di omogeneizzazione spaziale. In verde è evidenziata la prova considerata migliore.

Set completo	Dice	HD_{95}	RVD	Precisione	Sensibilità
Riferimento	0.33 ± 0.29	9.12 ± 14.60	1.51 ± 5.29	0.43 ± 0.35	0.43 ± 0.37
Prova	0.38 ± 0.28	8.61 ± 11.52	1.73 ± 5.53	0.45 ± 0.34	0.47 ± 0.34

In assenza di questa operazione in letteratura viene spesso eseguita l'operazione di crop centrale. Questa, oltre a ridurre le dimensioni delle ROI, mira a uniformare le dimensioni delle scansioni ADC e DWI rispetto alle T2W, in quanto le prime due presentano tipicamente una finestra di acquisizione più ampia. Tuttavia, come discusso nel dettaglio nella sezione 4, l'operazione di crop risulta essere poco robusta, può generare diverse criticità e non risolve il problema del disallineamento. I risultati ottenuti dimostrano come la funzione implementata consente di risolvere in modo efficiente ed efficace il problema del disallineamento delle scansioni e di conseguenza migliorare le prestazioni del modello *M2*.

Crop

Di seguito sono riportati i risultati delle analisi effettuate durante il processo di ottimizzazione dell'operazione di crop. In particolare si è deciso di riportare il confronto tra le seguenti prove:

- *Prova di Riferimento*. Modello addestrato sui set ottenuti implementando la seguente pipeline di pre-processing: Resample (0.5,0.5,3); Omogeneizzazione spaziale; Crop centrale [z,320,320] con z pari al numero slice del volume; Resize [16,256,256]; Normalizzazione; Combinazione nel volume 4D ("rgb").

- *Prova 1*: l'obiettivo è valutare l'impatto, sulle prestazioni del modello, dell'utilizzo di una finestra più piccola per eseguire il crop dei volumi. La finestra adoperata in questa prova ha le seguenti dimensioni: [z,150,150], dove z corrisponde al numero di slice del volume.

- *Prova 2*: l'obiettivo è valutare l'impatto, sulle prestazioni del modello, dell'operazione di *crop centrato sulla prostata*. I valori riportati sono stati ottenuti implementando nella pipeline la funzione di crop ottimizzata che esegue, se necessario, lo shift delle coordinate del centro. La finestra utilizzata ha dimensioni [z',150,150] con z' pari al numero slice contenenti la ghiandola.

- *Prova 3*: l'obiettivo è valutare l'impatto, sulle prestazioni del modello, dell'operazione di *crop centrato sulla ROI e di dimensioni proporzionali alla prostata*. La finestra utilizzata in questa prova ha dimensioni [z',d,d] con *z'* pari al numero slice contenenti la ghiandola e *d* proporzionale alle dimensioni della prostata.

Durante l'analisi sperimentale, le performance sono state valutate e confrontate utilizzando tutti i risultati e i relativi boxplot.

Per agevolarne il confronto, nelle tabelle 5.7 e 5.8 sono riportati i valori delle metriche di Dice e Distanza di Hausdorff (95° percentile) calcolati sui relativi subset, mentre nella tabella 5.9 sono riportati i valori di tutte le metriche calcolate sugli interi set di validazione.

Tabella 5.7: Tabella di valutazione stratificata, tramite Dice, delle prove effettuate durante il processo di ottimizzazione dell'operazione di crop. I valori riportati sono stati calcolati sul Validation set e sui relativi subset. Il *riferimento* è la prova in cui viene effettuato il crop centrale [z,320,320]. *Prova 1*: crop centrale [z,150,150]. *Prova 2*: crop ottimizzato [z',150,150] centrato sulla prostata. *Prova 3*: crop ottimizzato [z',d,d] centrato sulla prostata, con d proporzionale alle dimensioni della ghiandola. La prova migliore è evidenziata in verde. *Num.* indica il numero di casi. I valori in riportati in arancione si riferiscono al set di riferimento, mentre quelli in nero alle restanti prove. Le celle contenenti un unico valore fanno riferimento a tutte le prove effettuate.

Metrica: Dice	Num. casi	Riferimento:	Prova 1	Prova 2	Prova 3
Isup 2	42	0.38 ± 0.27	0.46±0.28	0.47 ± 0.30	0.49 ± 0.29
Isup 3	20 ; 19	0.41 ± 0.29	0.42±0.23	0.49 ± 0.25	0.52 ± 0.26
Isup 4	9	0.32 ± 0.30	0.28±0.28	0.30 ± 0.25	0.30 ± 0.31
Isup 5	11	0.40 ± 0.30	0.41±0.26	0.45 ±0.23	0.48 ± 0.25
Piccola dim.	21 ; 20	0.16 ± 0.22	0.23±0.26	0.26 ± 0.26	0.25 ± 0.23
Grande dim.	61	0.46 ± 0.26	0.49±0.25	0.52 ± 0.25	0.55 ± 0.27
Set completo	82 ; 81	0.38 ± 0.28	0.42±0.27	0.45 ± 28	0.47 ± 0.29

Tabella 5.8: Tabella di valutazione stratificata, tramite Distanza di Hausdorff (95° percentile), delle prove effettuate durante il processo di ottimizzazione dell'operazione di crop. I valori riportati sono stati calcolati sul Validation set e sui relativi subset. Il *riferimento* è la prova in cui viene effettuato il crop centrale [z,320,320]. *Prova 1*: crop centrale [z,150,150]. *Prova 2*: crop ottimizzato [z',150,150] centrato sulla prostata. *Prova 3*: crop ottimizzato [z',d,d] centrato sulla prostata, con d proporzionale alle dimensioni della ghiandola. La prova migliore è evidenziata in verde. *Num.* indica il numero di casi. I valori in arancione si riferiscono al set di riferimento, mentre quelli riportati in nero alle restanti prove. Le celle contenenti un unico valore fanno riferimento a tutte le prove effettuate.

Metrica: HD_{95}	Num. casi	Riferimento:	Prova 1	Prova 2	Prova 3
Isup 2	42	7.57 ± 9.35	19.03 ± 26.11	24.35 ± 27.17	20.69 ± 25.82
Isup 3	20 ; 19	8.48 ± 10.46	23.42 ± 17.61	22.07 ± 17.69	21.66 ± 19.56
Isup 4	9	15.70 ± 21.66	26.09 ± 23.60	25.79 ± 20.14	36.67 ± 31.65
Isup 5	11	7 ± 4.98	24.53 ± 20.19	28.22 ± 19.30	32.79 ± 32.38
Piccola dim.	21 ; 20	4.88 ± 9.22	13.69 ± 15.82	27.52 ± 23.59	22.84 ±21.37
Grande dim.	61	9.89 ± 11.94	24.18 ± 24.95	23.51 ± 23.47	24.83 ± 28.48
Set completo	82 ; 81	8.61 ± 11.52	21.59 ± 23.48	24.50 ± 23.56	24.34 ± 26.91

Tabella 5.9: Tabella di valutazione sui Validation set completi delle prove effettuate durante il processo di ottimizzazione dell'operazione di crop. A *Riferimento* sono associati i valori ottenuti applicando il crop centrale [z,320,320] con z pari al numero di slice del volume. A *Prova 1* sono associati i valori ottenuti applicando il crop centrale [z,150,150]. A *Prova 2* sono associati i valori ottenuti applicando l'operazione ottimizzata di crop [z',150,150] centrato sulla prostata, con z' pari al numero di slice contenenti la ghiandola. A *Prova 3* sono associati i valori ottenuti applicando l'operazione ottimizzata di crop [z',d,d] centrato sulla prostata, con d proporzionale alle dimensioni della ghiandola. In verde è evidenziata la prova considerata migliore.

Set completo	**Dice**	**HD_{95}**	**RVD**	**Precisione**	**Sensibilità**
Riferimento	0.38 ± 0.28	8.61 ± 11.52	1.73 ± 5.53	0.45 ± 0.34	0.47 ± 0.34
Prova 1	0.42 ± 0.27	21.59 ± 23.48	2.05 ± 4.58	0.46 ± 0.32	0.58 ± 0.35
Prova 2	0.45 ± 0.28	24.50 ± 23.56	3.33 ± 7.00	0.46 ± 0.32	0.64 ± 0.33
Prova 3	0.47 ± 0.29	24.34 ± 26.91	3.64 ± 7.74	0.48 ± 0.33	0.68 ± 0.31

Nella figura 5.16 è riportato un esempio visivo dell'effetto dell'operazione di crop implementata sui volumi. Per semplificarne la visualizzazione si è riportata una slice per ciascun volume. In particolare nella figura sono confrontate le scansioni senza l'applicazione del crop, con l'applicazione del crop di dimensioni [z',150,150] centrato sulla prostata mediante la funzione ottimizzata (*prova 2*) ed infine con l'applicazione del crop con una finestra di dimensioni [z',d,d] (*prova 3*).

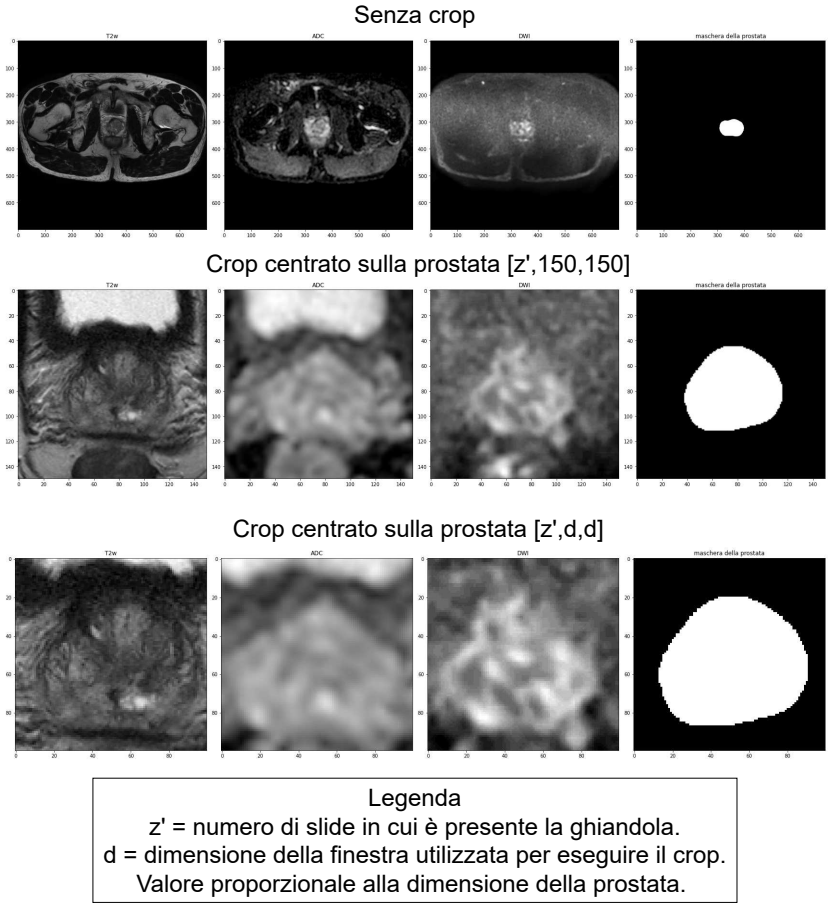


Figura 5.16: Esempio dell'effetto dell'operazione di crop implementata. Nella figura sono confrontate le scansioni senza l'applicazione del crop, con l'applicazione del crop di dimensioni [z',150,150] centrato sulla prostata mediante la funzione ottimizzata ed infine con l'utilizzo di una finestra di dimensioni [z',d,d].

I risultati ottenuti evidenziano come, oltre ai benefici legati alla robustezza del metodo, la funzione implementata ed ottimizzata consenta di migliorare le prestazioni di predizione del modello. In particolare, l'applicazione di una finestra di dimensioni personalizzate sul singolo caso si è

rilevata particolarmente vantaggiosa per i casi presentati l'intero addome (figura 5.16).

L'implementazione dell'operazione di crop centrato sulla prostata ha permesso inoltre l'identificazione e l'esclusione di alcuni casi in cui le lesioni identificate non si sovrappongono spazialmente con la regione dell'immagine occupata dalla prostata. Un esempio è riportato nella figura 5.17 in cui viene mostrata una slice della scansione T2w sovrapposta alla maschera AI della lesione.

Figura 5.17: Esempio di un caso escluso dal dataset in cui la lesione identificata non si sovrappone con la la regione dell'immagine occupata dalla prostata. L'immagine è stata ottenuta sovrapponendo la scansione T2w con la maschera AI della lesione rappresentata in verde. L'immagine riportata nell'esempio è quella del paziente con ID 11472.

Di seguito è riportata la tabella aggiornata contenente tutti gli ID dei casi esclusi dal dataset 5.10.

Resize

Di seguito sono presentati i risultati delle metriche utilizzate per valutare quantitativamente l'impatto, sulle prestazioni del modello, della dimensione z usata per il Resize. Durante l'analisi sperimentale, le performance sono state valutate e confrontate utilizzando tutti i risultati e i relativi boxplot. Per agevolarne la lettura, i valori delle metriche di Dice e Distanza di Hausdorff (95° percentile) calcolate su tutti i set sono riportate nella tabella 5.11,

Tabella 5.10: Elenco degli ID di tutti i casi esclusi dal dataset. Gli ID evidenziati in blu indicano i casi in cui la lesione identificata nella maschera non si sovrappone con la regione dell'immagine occupata dalla prostata. In nero sono riportati gli ID dei casi esclusi contenenti la maschera vuota.

ID casi esclusi
11157 ; 10577 ; 10882 ; 10104 ; 11350 ; 11143 ; 10555 ; 10094 ; 10811 ; 10397 ; 10658 ; 11243 ; 10508 ; 10660 ; 10636 ; 11296; 10775 ; 10717 ; 10811 ; 11472 ; 10140 ; 11169 ; 11050

mentre i valori di tutte le metriche calcolate sull'intero set di validazione sono riportati nella tabella 5.12.

Tabella 5.11: Tabella di valutazione dell'impatto, sulle prestazioni del modello, della dimensione z usata per il Resize. Le metriche sono state calcolate sui relativi Validation set. A *Riferimento* sono associati i valori delle metriche calcolate facendo uso del Resize pari a [16,256,256]. A *Prova 1* sono associati i valori delle metriche calcolate facendo uso del Resize pari a [32,256,256]. In verde è evidenziata la prova considerata migliore. Per ciascun set inoltre è riportato, nella colonna *Num.*, il numero di casi contenuti.

Analisi Resize		**Riferimento** Resize [16,256,256]		**Prova 1** Resize [32,256,256]	
Classe	**Num.**	**Dice**	**HD_{95}**	**Dice**	**HD_{95}**
Isup 2	42	0.49 ± 0.29	20.69 ± 25.82	0.53 ± 0.30	18.67 ± 25.41
Isup 3	19	0.52 ± 0.26	21.66 ± 19.56	0.53 ± 0.26	21.12 ± 19.74
Isup 4	9	0.30 ± 0.31	36.67 ± 31.65	0.41 ± 0.33	35.26 ± 36.17
Isup 5	11	0.48 ± 0.25	32.79 ± 32.38	0.51 ± 0.27	28.18 ± 23.98
Piccola dim.	20	0.25 ± 0.23	22.84 ±21.37	0.30 ± 0.27	20.91 ± 24.53
Grande dim	61	0.55 ± 0.27	24.83 ± 28.48	0.58 ± 0.27	22.86 ± 26.55
Set completo	81	0.47 ± 0.29	24.34 ± 26.91	0.51 ± 0.29	22.38 ± 26.08

Tabella 5.12: Tabella di valutazione dell'impatto, sulle prestazioni del modello, della dimensione z usata per il Resize. Le metriche sono state calcolate sui relativi Validation set completi. A *Riferimento* sono associati i valori delle metriche calcolate facendo uso del Resize pari a [16,256,256]. A *Prova 1* sono associati i valori delle metriche calcolate facendo uso del Resize pari a [32,256,256]. In verde è evidenziata la prova considerata migliore.

Set completo	**Dice**	**HD_{95}**	**RVD**	**Precisione**	**Sensibilità**
Riferimento	0.47 ± 0.29	24.34 ± 26.91	3.64 ± 7.74	0.48 ± 0.33	0.68 ± 0.31
Prova 1	0.51 ± 0.29	22.38 ± 26.08	2.51 ± 6.40	0.54 ± 0.33	0.69 ± 0.34

I risultati evidenziano che l'impiego della dimensione 32 per l'asse z consente di ottenere prestazioni superiori rilevandosi adeguata per migliorare la qualità della segmentazione delle lesioni.

Bias Field Correction

Nella tabella 5.13 sono riportati i valori delle metriche calcolate sui validation set completi utili per una valutazione quantitativa dell'impatto, sulle prestazioni del modello, dell'applicazione del filtro N4BiasFieldCorrection impiegato per effettuare la correzione della polarizzazione del campo magnetico.

Tabella 5.13: Tabella di valutazione dell'impatto dell'applicazione del filtro N4ITK per effettuare la bias field correction sulle prestazioni del modello. Le metriche sono state calcolate sui relativi Validation set completi. A *Riferimento* sono associati i valori delle metriche calcolate senza l'applicazione del filtro. In verde è evidenziata la prova considerata migliore.

Set completo	**Dice**	**HD_{95}**	**RVD**	**Precisione**	**Sensibilità**
Bias field correction	0.50 ± 0.29	34.47 ± 35.97	4.75 ± 9.52	0.46 ± 0.33	0.77 ± 0.28
Riferimento	0.51 ± 0.29	22.38 ± 26.08	2.51 ± 6.40	0.54 ± 0.33	0.69 ± 0.34

I risultati dell'analisi sperimentale indicano che l'applicazione della correzione del campo magnetico non è necessaria per il dataset in questione. Le linee guida di acquisizione infatti consigliano l'utilizzo di una "small shim box" [16] per garantire un campo magnetico omogeneo intorno alla prostata e ridurre al minimo ulteriori artefatti generati da presenza di aria o movimenti intestinali migliorando la qualità delle immagini. Di conseguenza, seguendo quanto fatto in altri studi, tale operazione è stata esclusa dalla pipeline di *pre-processing 2*. Questa decisione è principalmente motivata dai costi computazionali e dai tempi necessari per l'applicazione del filtro. Tale correzione infatti richiede un notevole dispendio di risorse CPU e prolunga i tempi di pre-processing di circa 15-20 minuti per ciascun caso. Inoltre, questi costi non sono giustificati dalle prestazioni, che risultano paragonabili e leggermente inferiori a quelle ottenute senza la correzione.

Pipeline di *pre processig 2* finale

Di seguito sono elencate le operazioni implementate nelle pipeline di *pre-processing 2* definite in seguito al processo di ottimizzazione.

Pipeline di *pre-processing 2* delle scansioni T2w ADC e DWI:

1. *Resample (0.5,0.5,3)* [7] con interpolazione lineare;
2. *Omogeneizzazione spaziale;*
3. *Crop* centrato sulla prostata [z',d,d] mediante funzione ottimizzata;
4. *Resize [32,256,256]* [8]
5. *Normalizzazione* delle scansioni : Zscore + min-max scaling;
6. *Formazione del volume 4D.*

Pipeline di *pre-processing 2* delle maschere delle lesioni utilizzate per l'addestramento e la valutazione delle performance del modello:

1. *Resample (0.5,0.5,3)* [9] con interpolazione NearestNeighbor;

[7] (x,y,z)
[8] (z,x,y)
[9] (x,y,z)

2. *Omogeneizzazione spaziale*;

3. *Crop* centrato sulla prostata [z',d,d] mediante funzione ottimizzata;

4. *Resize [32,256,256]*; [10]

5. *Thresholding con soglia 0,5*.

Nella fase di *inference* viene fatto uso delle maschere prostatiche generate dal modello *M1*.

5.4.2 Modello *M2*

In questa sottosezione sono riportati e commentati i risultati relativi alle analisi sperimentali effettuate nella fase di sviluppo e ottimizzazione del modello *M2*.

Approccio Patch

Durante il processo di ottimizzazione dell'algoritmo si sono esaminati gli effetti del *metodo Patch*. Inizialmente, questo approccio ha mostrato risultati migliori e più stabili rispetto a quelli ottenuti mediante metodo tradizionale che fa uso dell'intero volume. Tuttavia, una volta ottimizzata la pipeline di pre-processing, i risultati delle diverse prove effettuate hanno indicato che l'impiego delle patch non comporta un miglioramento delle performance di predizione del modello.

Architettura

Per quanto concerne l'ottimizzazione dell'architettura, tra le varie configurazioni testate si è riportato il confronto tra l'architettura del modello iniziale, utilizzato nella fase di ottimizzazione della pipeline di pre-processing, e l'architettura finale definita in seguito al processo di ottimizzazione. Nella tabella 5.14 sono riportati i parametri riassuntivi delle due architetture confrontate. Le due architetture si distinguono soltanto per i valori dello stride e hanno in comune la combinazione ottimale dei parametri di training.

[10] (z,x,y)

Tabella 5.14: Tabella riassuntiva dei parametri delle architetture dei modelli *M2* confrontati.

Architettura iniziale	model = UNet(spatial_dims=3, in_channels=3, out_channels=1, channels=(16,32,64,128), strides= ((2,2,2),(2,2,2),(2,2,2), num_res_units=2, norm=Norm.BATCH)
Architettura finale	model = UNet(spatial_dims=3, in_channels=3, out_channels=1, channels=(16,32,64,128), strides= ((2,2,2),(2,2,1),(2,2,2), num_res_units=2, norm=Norm.BATCH)
Parametri in comune	
Loss function	DiceFocalLoss (include_background=False, to_onehot_y=False, sigmoid=False, softmax=False, other_act=None, squared_pred=True, jaccard=False, reduction='mean', smooth_nr=1e-05, smooth_dr=1e05, batch=True, gamma=2, focal_weight=None, weight=None, lambda_dice=1, lambda_focal=1)
Ottimizzatore	Adam(lr=5e-4, weight_decay=5e-06)
Batch size	8
Data Augmentation	None
Patience	10

Per agevolare la valutazione quantitativa delle performance, nella tabella 5.15 sono riportati i valori di Dice e Distanza di Hausdorff (95° percentile) calcolate su tutti i subset, mentre nella tabella 5.16 sono riportati i valori di tutte le metriche calcolate sull'intero set di validazione.

Tabella 5.15: Tabella di valutazione stratificata dell'impatto dell'architettura sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni del modello con *architettura iniziale* con quelle del modello con *architettura finale*. I valori delle metriche sono stati calcolati sull'intero Validation set e sui relativi subset. In verde è evidenziata la prova considerata migliore.

Analisi Architettura		**Riferimento** Architettura iniziale		**Prova Architettura finale**	
Classe	**Num.**	**Dice**	**HD_{95}**	**Dice**	**HD_{95}**
Isup 2	42	0.53 ± 0.30	18.67 ± 25.41	0.58 ± 0.28	15.29 ± 22.37
Isup 3	19	0.53 ± 0.26	21.12 ± 19.74	0.56 ± 0.26	19.35 ± 19.45
Isup 4	9	0.41 ± 0.33	35.26 ± 36.17	0.39 ± 0.32	33.54± 33.85
Isup 5	11	0.51 ± 0.27	28.18 ± 23.98	0.53 ± 0.28	23.70 ± 21.12
Piccola dim.	20	0.30 ± 0.27	20.91 ± 24.53	0.37 ± 0.27	16.51 ± 20.06
Grande dim.	61	0.58 ± 0.27	22.86 ± 26.55	0.61 ± 0.26	20.37± 24.91
Set completo	81	0.51 ± 0.29	22.38 ± 26.08	0.55 ± 0.23	19.41 ± 23.87

Tabella 5.16: Tabella di valutazione dell'impatto dell'architettura sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni del modello con *architettura iniziale* con quelle del modello con *architettura finale*. I valori delle metriche sono stati calcolati sull'intero Validation set. In verde è evidenziata la prova considerata migliore.

Set completo	**Dice**	**HD_{95}**	**RVD**	**Precisione**	**Sensibilità**
Architettura iniziale	0.51 ± 0.29	22.38 ± 26.08	2.51 ± 6.40	0.54 ± 0.33	0.69 ± 0.34
Architettura finale	0.55 ± 0.29	19.41 ± 23.87	1.99 ± 6.30	0.57 ± 0.33	0.67 ± 0.30

Il confronto tra le due architetture, *iniziale* e *finale*, mostra che la modifica apportata all'architettura, riguardante la variazione dei valori di stride lungo l'asse z, porta a un miglioramento delle prestazioni del modello producendo segmentazioni di maggiore qualità.

Loss Funciton

Per lo sviluppo del modello è stata implementata la funzione di costo DiceFocalLoss dalla libreria MONAI. Questa è stata preferita rispetto ad altre opzioni disponibili, come ad esempio la BCELoss o la DicelLoss (descritte in 3.5), in quanto ha permesso di ottenere risultati migliori. Per un confronto numerico, è possibile consultare la tabella 5.17.

Tabella 5.17: Tabella di valutazione dell'impatto della Loss function sulle performance del modello. I valori delle metriche sono stati calcolati sull'intero Validation set. In verde è evidenziata la prova considerata migliore.

Set completo	Dice	HD_{95}	RVD	Precisione	Sensibilità
DiceLoss	0.50 ± 0.22	26.34 ± 22.18	3.21 ± 4.30	0.42 ± 0.23	0.65 ± 0.36
BCELoss	0.52 ± 0.13	28.16 ± 27.21	4.41 ± 2.22	0.43 ± 0.25	0.59 ± 0.15
DiceFocalLoss	0.55 ± 0.29	19.41 ± 23.87	1.99 ± 6.30	0.57 ± 0.33	0.67 ± 0.30

Parametri primo training

Le analisi condotte hanno evidenziato come l'utilizzo dell'ottimizzatore Adam fosse il più adatto per la task in questione. Questo infatti ha permesso di ottenere prestazioni migliori e più stabili rispetto a quelli ottenuti utilizzando altri ottimizzatori.

Nella tabella 5.18 sono riporti i valori medi di Dice calcolati sul Validation set mediante le principali configurazioni di learning rate (lr) e weight decay (wd) testate. I valori riportati in tabella sono stati ottenuti facendo uso dell'architettura finale.

Tabella 5.18: Tabella dei valori medi di Dice calcolati sul validation set, utilizzando diverse combinazioni di learning rate (lr) e weight decay (wd). Questi risultati sono stati ottenuti utilizzando il modello con *architettura finale*.

Prova	DICE	Configurazione
1	0.4619	Lr: 5e-03, wd: 1e-03
2	0.5041	Lr: 5e-03, wd: 5e-04
3	0.4748	Lr: 5e-03, wd: 1e-04
4	0.5137	Lr: 5e-03, wd: 5e-05
5	0.5036	Lr: 5e-03, wd: 1e-05
6	0.5264	Lr: 5e-03, wd: 5e-06
7	0.5160	Lr: 1e-03, wd: 1e-03
8	0.4967	Lr: 1e-03, wd: 5e-04
9	0.5432	Lr: 1e-03, wd: 1e-04
10	0.5621	Lr: 1e-03, wd: 5e-05
11	0.5492	Lr: 1e-03, wd: 1e-05
12	0.5304	Lr: 1e-03, wd: 5e-06
13	0.5119	Lr: 5e-04, wd: 1e-03
14	0.5082	Lr: 5e-04, wd: 5e-04
15	0.5295	Lr: 5e-04, wd: 1e-04
16	0.5174	Lr: 5e-04, wd: 5e-05
17	0.3917	Lr: 5e-04, wd: 1e-05
18	0.5876	Lr: 5e-04, wd: 5e-06

Tra i parametri di training si è testato anche l'impatto del *Batch size* sulle prestazioni del modello. Nella tabella 5.19 sono riporti i valori delle metriche calcolate sui Validation set completi utilizzando i valori di Batch che hanno permesso di ottenere i risultati più stabili. La combinazione di lr e wd identificata nello step precedente si è rilevata essere quella ottimale anche al variare dei valori di Batch.

Tabella 5.19: Tabella di valutazione dell'impatto del batch size sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni dei modelli addestrati utilizzando batch differenti. I valori delle metriche sono stati calcolati sull'intero Validation set. In verde è evidenziata la prova considerata migliore.

Set completo	Dice	HD_{95}	RVD	Precisione	Sensibilità
Batch size: 6	0.50 ± 0.29	34.37 ± 35.97	4.75 ± 9.52	0.46± 0.33	0.77 ± 0.28
Batch size: 8	0.55 ± 0.29	19.41 ± 23.87	1.99 ± 6.30	0.57 ± 0.33	0.67 ± 0.30
Batch size: 10	0.54 ± 0.27	21.72 ± 24.35	1.45± 4.09	0.59± 0.34	0.62 ± 0.27

I risultati sperimentali evidenziano come per il primo training l'utilizzo di un batch pari ad 8 consenta di ottenere prestazioni migliori.

Data augmentation primo training

Per la valutazione quantitativa dell'impatto delle operazioni di data augmentation sulle prestazioni del modello, nella tabella 5.20 sono riportati i valori delle metriche calcolate sul Validation set delle principali prove effettuate.

Tabella 5.20: Tabella di valutazione dell'impatto delle operazioni di data augmentation sulle prestazioni del modello. I valori delle metriche sono stati calcolati sull'intero Validation set. In verde è evidenziata la prova considerata migliore.

Set completo	Dice	HD_{95}	RVD	Precisione	Sensibilità
Riferimento	0.55 ± 0.29	19.41 ± 23.87	1.99 ± 6.30	0.57 ± 0.33	0.67 ± 0.30
Rotazione (±6°)	0.58 ± 0.28	17.18 ± 21.47	1.29 ± 3.98	0.61 ± 0.30	0.68 ± 0.30
Rotazione (±6°) & Flip verticale	0.52 ± 0.28	24.14 ± 27.47	3.05 ± 7.25	0.51 ± 0.32	0.73 ± 0.31
Rotazione (±6°) & Shift (2° e 3° canale)	0.37 ± 0.26	49.22 ± 40.79	6.59 ± 10.97	0.28 ± 0.24	0.73 ± 0.34
Rotazione (±6°) & Zoom in-out	0.53 ± 0.29	21.51 ± 24.89	2.94 ± 7.33	0.51 ± 0.31	0.76 ± 0.30

Come si evince dalla tabella, tra le operazioni implementate soltanto la rotazione random di (± 6°) con probabilità del 50% ha avuto effetti positivi sul training del modello. Per una migliore valutazione dell'impatto di questa

operazione, nella tabella 5.21 sono riportati i valori delle metriche di Dice e Distanza di Hausdorff (95° percentile) calcolate su tutti i subset.

Tabella 5.21: Tabella di valutazione stratificata dell'impatto dell'operazione di rotazione random sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni del modello con e senza l'applicazione dell'operazione di data augmentation. Le metriche sono state calcolate sul Validation set completo e sui relativi subset. In verde è evidenziata la prova considerata migliore.

Analisi Data augmentation		**Riferimento**		**Prova Rotazione random (±6°)**	
Classe	**Num.**	**Dice**	**HD_{95}**	**Dice**	**HD_{95}**
Isup 2	42	0.58 ± 0.28	15.29 ± 22.37	0. 60 ± 0.29	10.83 ± 12.31
Isup 3	19	0.56 ± 0.26	19.35 ± 19.45	0.59 ± 0.24	21.78 ± 21.22
Isup 4	9	0.39 ± 0.32	33.54± 33.85	0.43 ± 0.29	36.08 ± 37.46
Isup 5	11	0.53 ± 0.28	23.70 ± 21.12	0.60 ± 0.25	17.98 ± 20.14
Piccola dim.	20	0.37 ± 0.27	16.51 ± 20.06	0.41 ± 0.29	16.98 ± 17.78
Grande dim	61	0.61 ± 0.26	20.37± 24.91	0.63 ± 0.25	17.24 ± 22.54
Set completo	81	0.55 ± 0.23	19.41 ± 23.87	0.58 ± 0.28	17.18 ± 21.47

Parametri secondo training

Nella tabella 5.22 sono riportati i valori medi di Dice calcolati sul Validation set mediante le principali configurazioni di learning rate (lr) e weight decay (wd) testate durante la fase di ottimizzazione del fine tuning del modello. I valori riportati sono stati ottenuti facendo uso di un batch size pari a 9 e del modello con con architettura finale.

Tabella 5.22: Tabella dei valori medi di Dice calcolati sul Validation set, utilizzando diverse combinazioni di learning rate (lr) e weight decay (wd). I risultati sono stati ottenuti utilizzando il modello con *architettura finale*.

Prova	DICE	Configurazione
1	0.4142	Lr: 5e-03, wd: 1e-03
2	0.4771	Lr: 5e-03, wd: 5e-04
3	0.4951	Lr: 5e-03, wd: 1e-04
4	0.4900	Lr: 5e-03, wd: 5e-05
5	0.4960	Lr: 5e-03, wd: 1e-05
6	0.5161	Lr: 5e-03, wd: 5e-06
7	0.5116	Lr: 1e-03, wd: 1e-03
8	0.5263	Lr: 1e-03, wd: 5e-04
9	0.5155	Lr: 1e-03, wd: 1e-04
10	0.5084	Lr: 1e-03, wd: 5e-05
11	0.5042	Lr: 1e-03, wd: 1e-05
12	0.5047	Lr: 1e-03, wd: 5e-06
13	0.5206	Lr: 5e-04, wd: 1e-03
14	0.5223	Lr: 5e-04, wd: 5e-04
15	0.5171	Lr: 5e-04, wd: 1e-04
16	0.5187	Lr: 5e-04, wd: 5e-05
17	0.5176	Lr: 5e-04, wd: 1e-05
18	0.5294	Lr: 5e-04, wd: 5e-06

Nella tabella 5.23 sono riporti i valori delle metriche calcolate sul Test set completo utilizzando i valori di Batch che hanno permesso di ottenere i risultati più stabili. La combinazione di lr e wd identificata nello step precedente si è rilevata essere quella ottimale anche al variare dei valori di Batch provati.

Tabella 5.23: Tabella di valutazione dell'impatto del batch size sulle prestazioni del modello. Nella tabella vengono confrontate le prestazioni dei modelli addestrati utilizzando valori di batch differenti. I valori delle metriche sono stati calcolati sull'intero Test set. In verde è evidenziata la prova considerata migliore.

Set completo	Dice	HD_{95}	RVD	Precisione	Sensibilità
Batch size: 7	0.46 ± 0.26	30.43 ± 31.19	1.10 ± 1.71	0.41± 0.26	0.63 ± 0.34
Batch size: 8	0.48 ± 0.29	20.35 ± 25.63	0.26 ± 0.93	0.53 ± 0.30	0.55 ± 0.35
Batch size: 9	0.49 ± 0.26	22.66 ± 27.38	0.96 ± 2.49	0.50 ± 0.28	0.62 ± 0.34
Batch size: 10	0.47 ± 0.25	25.80 ± 30.77	0.82 ± 1.09	0.42 ± 0.25	0.65 ± 0.34

I risultati sperimentali evidenziano come per il secondo training l'utilizzo di un batch pari a 9 consenta di ottenere prestazioni migliori. Si specifica che i valori dei parametri sono stati settati facendo uso del Validation set e successivamente valutati sul Test set.

Effetto Data augmentation secondo training

I risultati dei numerosi test condotti indicano che l'implementazione delle operazioni di data augmentation, contrariamente alle attese, non contribuisce in modo significativo al miglioramento dell'efficacia del processo di fine-tuning del modello. Questo potrebbe essere attribuito alla struttura poco profonda del modello utilizzato, la quale potrebbe limitarne la capacità di apprendimento. Un modello con una struttura più profonda potrebbe essere in grado di catturare relazioni più complesse nei dati e trarre vantaggio dalle operazioni di data augmentation migliorando le prestazioni complessive di predizione. Questa combinazione tuttavia determina anche un aumento significativo dei costi e dei tempi computazionali dei training senza garantirne l'efficacia.

Impatto del *Fine Tuning* sulle prestazioni del modello

Di seguito sono riportate le metriche di valutazione, calcolate sul Test set interno, utili per il confronto delle seguenti prove:

- Prova 1: modello *M2* ottenuto mediante unico training effettuato utilizzando solo i set con con le annotazioni manuali (tabella 5.24).

- Prova 2: modello *M2* ottenuto mediante training effettuato utilizzando solo i set con le annotazioni AI (tabella 5.25).

- Prova 3: Performance del modello *M2* dopo il fine Tuning effettuato utilizzando le annotazioni manuali (tabella 5.26). Il modello è stato pre-addestrato utilizzando le maschere AI.

Tabella 5.24: Tabella di valutazione delle performance del modello *M2* ottenuto mediante unico training effettuato utilizzando solo i set con con le annotazioni manuali. Le metriche sono state calcolate sul Test set interno.

Prova 1	Num.	Dice	HD_{95}	RVD	Precisione	Sensibilità
Isup 2	14	0.54 ± 0.24	21.29 ± 30.02	-0.09 ± 0.45	0.61 ± 0.27	0.55 ± 0.29
Isup 3	6	0.50 ± 0.32	17.94 ± 12.55	0.20 ± 0.72	0.55 ± 0.24	0.61 ± 0.41
Isup 4	4	0.32 ± 0.32	7.83 ± 8.16	-0.39 ± 0.64	0.59 ± 0.01	0.36 ± 0.38
Isup 5	3	0.40 ± 0.27	47.81 ± 29.51	0.96 ± 0.42	0.33 ± 0.24	0.53 ± 0.27
Piccola dim.	7	0.29 ± 0.31	25.10 ± 42.88	-0.32 ± 0.85	0.56 ± 0.30	0.35 ± 0.41
Grande dim.	20	0.56 ± 0.24	19.98 ± 19.33	0.20 ± 0.49	0.56 ± 0.25	0.61 ± 0.27
Set completo	27	0.48 ± 0.29	21.50 ± 28.52	0.04 ± 0.66	0.56 ± 0.26	0.53 ± 0.34

Tabella 5.25: Tabella di valutazione delle performance del modello *M2* ottenuto mediante training effettuato utilizzando solo i set con le annotazioni AI. Le metriche sono state calcolate sul Test set interno.

Prova 2	Num.	Dice	HD_{95}	RVD	Precisione	sensibilità
Isup 2	14	0.51 ± 0.27	18.92 ± 27.69	0.43 ± 1.25	0.55 ± 0.27	0.57± 0.32
Isup 3	6	0.46 ± 0.29	12.30 ± 8.95	0.39 ± 0.84	0.44 ± 0.17	0.64 ± 0.42
Isup 4	4	0.32 ± 0.32	8.53 ± 9.02	-0.29 ± 0.73	0.54 ± 0.02	0.39 ± 0.41
Isup 5	3	0.43 ± 0.17	30.00 ± 17.25	1.00 ± 0.99	0.35 ± 0.12	0.66 ± 0.30
Piccola dim.	8	0.24 ± 0.32	16.31 ± 35.36	0.27 ± 0.29	0.44 ± 0.29	0.31 ± 0.41
Grande dim.	19	0.55 ± 0.20	17.49 ± 13.39	0.42 ± 0.22	0.51 ± 0.22	0.68 ± 0.28
Set completo	27	0.46 ± 0.28	17.14 ± 22.29	0.38 ± 1.12	0.50 ± 0.23	0.57 ± 0.37

Tabella 5.26: Tabella di valutazione delle performance del modello *M2* dopo il fine Tuning.Le metriche sono state calcolate sul Test set interno.

Prova 3	Num.	Dice	HD_{95}	RVD	Precisione	sensibilità
Isup 2	14	0.52 ± 0.23	21.10 ± 32.12	1.25 ± 3.31	0.52 ± 0.27	0.61 ± 0.31
Isup 3	6	0.59 ± 0.23	14.04 ± 10.685	0.60 ± 0.73	0.51 ± 0.23	0.75 ± 0.31
Isup 4	4	0.32 ± 0.33	30.26 ± 24.63	0.09 ± 0.65	0.53 ± 0.36	0.39 ± 0.40
Isup 5	3	0.43 ± 0.20	36.98 ± 21.32	1.53 ± 0.87	0.31 ± 0.17	0.70 ± 0.23
Piccola dim.	8	0.37 ± 0.30	34.25 ± 42.66	1.94 ± 4.23	0.47 ± 0.33	0.48 ± 0.41
Grande dim.	19	0.55 ± 0.22	17.78 ± 14.80	0.55 ± 0.83	0.51 ± 0.25	0.68 ± 0.28
Set completo	27	0.49 ± 0.26	22.66 ± 27.38	0.96 ± 2.49	0.50 ± 0.28	0.62 ± 0.34

Il confronto dei risultati ottenuti evidenzia i benefici della tecnica implementata in particolare per l'identificazione e la segmentazione delle lesioni di piccole dimensioni. Nella figura 5.18 sono riportati, per le tre prove effettuate, i boxplot dei valori di Dice calcolati sul subset delle lesioni di piccola dimensione del Test set.

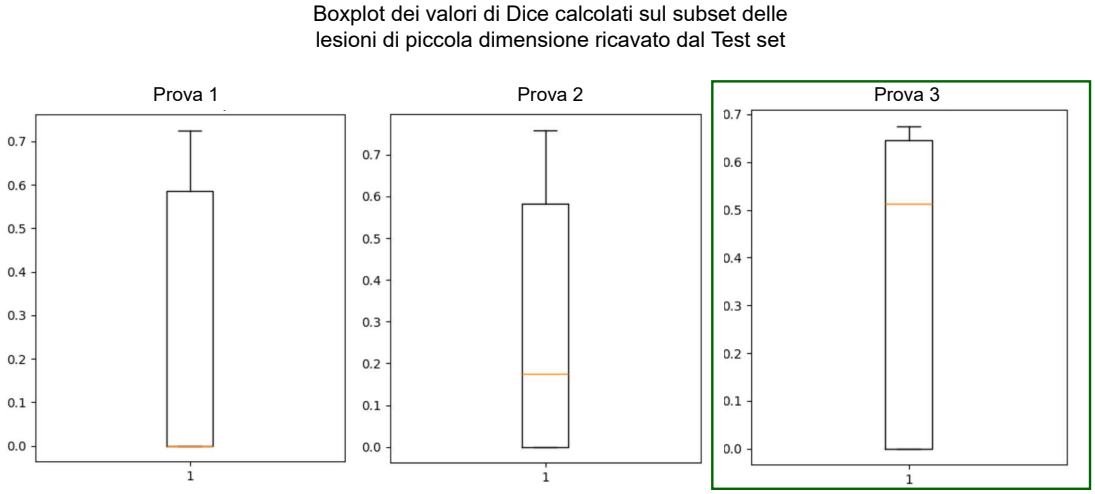


Figura 5.18: Boxplot dei valori di Dice calcolati, per le tre prove effettuate, sul subset delle lesioni di piccola dimensione del Test set. Prova 1: unico training effettuato utilizzando i set con le annotazioni manuali. Prova 2: training effettuato utilizzando solo i set con le annotazioni AI. Prova 3: Performance del modello in seguito al processo di fine Tuning. In verde è evidenziata la prova considerata migliore.

I boxplot dei valori di Dice consentono una visualizzazione più chiara dei

benefici che il metodo del fine tuning ha apportato sui casi aventi lesioni di piccole dimensioni. Si precisa che gli stessi benefici si sono riscontrati sui set, più numerosi, di Training e Validation.

Modello *M2* Finale

Per fornire una visione complessiva delle caratteristiche del modello UNet *M2* implementato, la cui struttura finale è descritta nella figura 5.19, si sono riportati i parametri principali nella tabella 5.27. Nella tabella 5.28 sono invece riportati i parametri settati per il primo ed il secondo training.

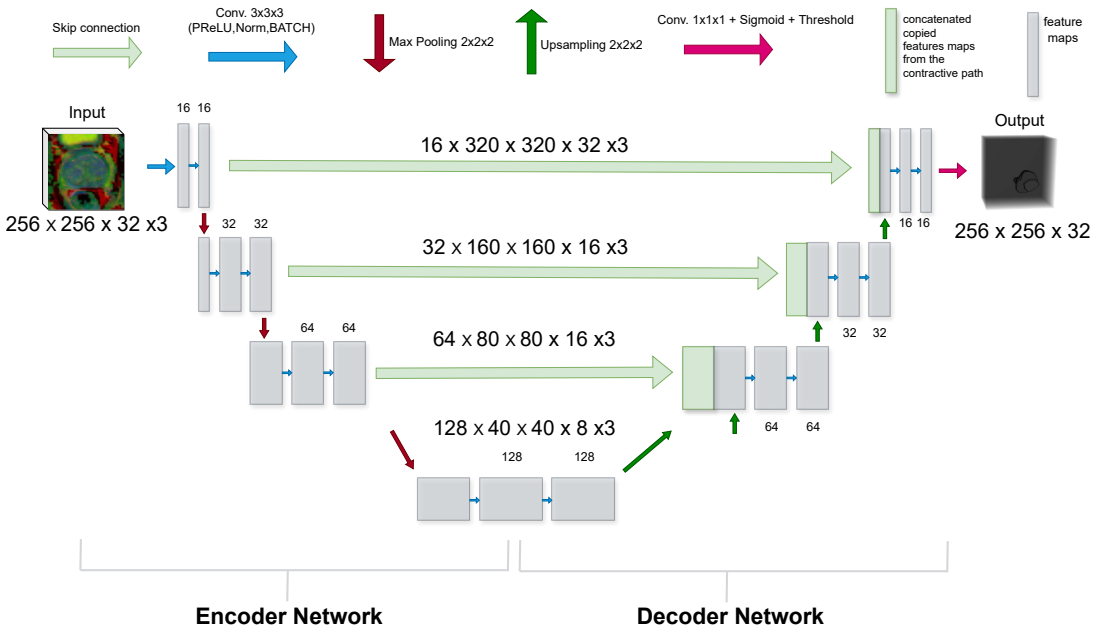


Figura 5.19: Rappresentazione grafica dell'architettura del modello UNet *M2*.

Tabella 5.27: Parametri principali dell'architettura UNet *M2*.

Parametri del Modello	
spatial_dims	3
in_channels	3
out_channels	1
channels	(16, 32, 64, 128)
strides	((2, 2, 2), (2, 2, 1), (2, 2, 2))
kernel_size	(3, 3, 3)
num_res_units	2
norm	Norm.BATCH
padding	(1, 1, 1)
activation function	PReLU
dim input volume	256x256x32x3
dim output volume	256x256x32

Tabella 5.28: Parametri utilizzati per il training del modello UNet *M2* finale. Dove non specificato i parametri sono uguali per entrambi i training.

Parametri di training	
BATCH_SIZE 1° training	8
BATCH_SIZE 2° training	9
num_iterations (max)	300 epoche
patience	20
Loss function	DiceFocalLoss (include_background=False, to_onehot_y=False, sigmoid=False, softmax=False, other_act=None, squared_pred=True, jaccard=False, reduction='mean', smooth_nr=1e-05, smooth_dr=1e05, batch=True, gamma=2, focal_weight=None, weight=None, lambda_dice=1, lambda_focal=1)
optimizer	Adam(lr=5e-4, weight_decay=5e-06)
data augmentation 1° training	rotazione random ±6°
data augmentation 2° training	none

5.4.3 *Post processing 2*

Nella tabella 5.29 sono riportati i risultati dell'analisi sperimentale condotta per l'ottimizzazione del valore di threshold.

Tabella 5.29: Tabella di valutazione dell'impatto, sulla qualità della segmentazione, del valore soglia utilizzato per effettuare la binarizzazione della maschera delle lesioni generate da *M2*. le metriche sono state calcolate sul Validation set. In verde è evidenziata la prova ritenuta migliore.

Prova	Dice	HD_{95}	RVD	Precisione	sensibilità
Soglia = 0.5	0.49± 0.30	30.66± 40.11	4.29 ± 8.41	0.48 ± 0.37	0.72 ± 0.31
Soglia = 0.85	0.49± 0.29	29.18± 39.26	2.14 ± 5.25	0.49 ± 0.37	0.71 ± 0.32
Soglia = 0.995	0.49± 0.26	24.38± 36.76	1.84 ± 4.48	0.56 ± 0.38	0.59 ± 0.31

Dall' analisi sperimentale è emerso che un valore soglia più alto rispetto a "0.5" consente di eliminare eventuali predizioni ambigue. In particolare, per questa specifica task e per questo modello, si è identificato mediante approccio iterativo un valore soglia ottimale pari a "0.995".

Come analizzato nella sezione 1.3, le lesioni tumorali possono presentare una vasta gamma di forme e dimensioni. Questa varietà di caratteristiche complica l'implementazione di operazioni di post-processing mirate a migliorare la qualità della segmentazione in modo automatico. Nello specifico i contorni irregolari complicano l'implementazione di operazioni di post processing standard come erosione o dilatazione. Queste inoltre potrebbero avere effetti differenti su lesioni di dimensioni diverse e, nello specifico, potrebbero risultare troppo invasive per quelle di piccola dimensione. Inoltre operazioni come la dilatazione potrebbe compromettere la qualità della segmentazione causando la fusione delle lesioni multi focali.

5.4.4 Algoritmo finale di segmentazione delle csPCa

In questa sottosezione sono analizzate le performance finali dell'algoritmo implementato per la segmentazione delle lesioni.

Di seguito sono riportati i valori delle metriche di valutazione ottenute su *Train set Human* (tabella 5.30) *Validation set Human* (tabella 5.31) e *Test set* interno (tabella 5.32) utili per una valutazione complessiva dell'intero l'algoritmo ottimizzato.

Tabella 5.30: Tabella di valutazione delle performance dell'intero algoritmo sul *Train set Human*.

Train set	Num.	Dice	HD_{95}	RVD	Precisione	sensibilità
Isup 2	89	0.46 ± 0.28	37.78 ± 47.02	1.91 ± 6.66	0.58 ± 0.40	0.48 ± 0.28
Isup 3	35	0.48 ± 0.27	33.19 ± 50.21	0.77 ± 3.54	0.71 ± 0.36	0.49 ± 0.31
Isup 4	12	0.53 ± 0.28	55.03 ± 12.15	-0.02 ± 1.04	0.67 ± 0.39	0.52 ± 0.31
Isup 5	11	0.65 ± 0.10	17.00 ± 21.56	-0.05 ± 0.66	0.79 ± 0.21	0.63 ± 0.20
Piccola dim.	37	0.42± 0.33	32.63 ± 44.31	5.12 ± 9.79	0.40 ± 0.36	0.63 ± 0.36
Grande dim.	110	0.51 ± 0.25	37.86 ± 61.94	0.06 ± 1.44	0.72 ± 0.36	0.45 ± 0.24
Set completo	147	0.48 ± 0.27	36.54 ± 58.05	1.34 ± 5.52	0.64 ± 0.38	0.50 ± 0.29

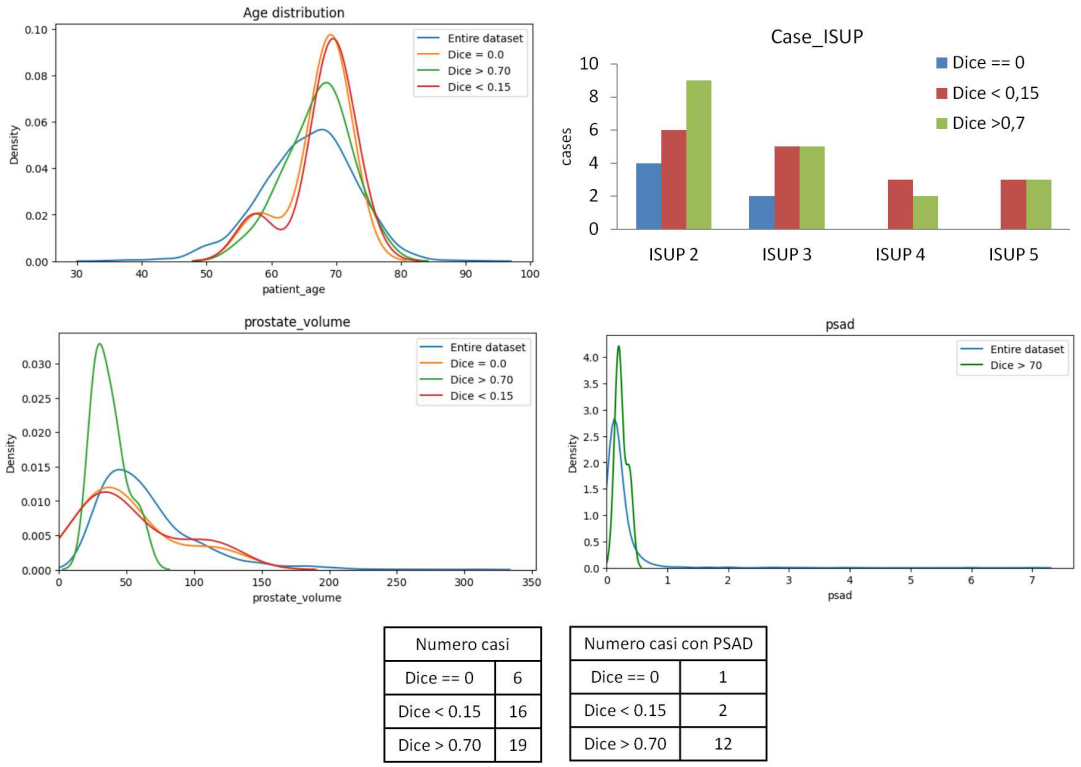
Tabella 5.31: Tabella di valutazione delle performance dell'intero algoritmo sul *Validation set Human*.

Validation set	Num.	Dice	HD_{95}	RVD	Precisione	sensibilità
Isup 2	27	0.47 ± 0.28	27.53 ± 41.97	2.22 ± 5.41	0.55 ± 0.39	0.56 ± 0.34
Isup 3	10	0.56 ± 0.25	13.72 ± 13.96	0.54 ± 1.67	0.70 ± 0.33	0.57 ± 0.29
Isup 4	4	0.44 ± 0.18	17.72 ± 15.74	2.61 ± 2.05	0.32 ± 0.16	0.84 ± 0.08
Isup 5	4	0.46 ± 0.21	36.49 ± 45.12	1.76 ± 3.38	0.59 ± 0.39	0.58 ± 0.18
Piccola dim.	11	0.34 ± 0.28	39.61 ± 56.49	5.24 ± 6.76	0.26 ± 0.24	0.65 ± 0.40
Grande dim.	34	0.53 ± 0.24	19.46 ± 25.64	0.74 ± 2.63	0.66 ± 0.36	0.57 ± 0.27
Set completo	45	0.49 ± 0.26	24.38 ± 36.76	1.84 ± 4.48	0.56 ± 0.38	0.59 ± 0.31

Tabella 5.32: Tabella di valutazione delle performance dell'intero algoritmo sul *Test set*.

Test set	Num.	Dice	HD_{95}	RVD	Precisione	sensibilità
Isup 2	14	0.54 ± 0.27	28.05 ± 30.02	1.53 ± 2.50	0.55 ± 0.33	0.64 ± 0.29
Isup 3	6	0.55 ± 0.22	31.09 ± 39.54	0.69 ± 1.41	0.67 ± 0.37	0.62 ± 0.26
Isup 4	4	0.34 ± 0.25	43.68 ± 32.14	15.84 ± 25.24	0.53 ± 0.47	0.52 ± 0.15
Isup 5	3	0.42 ± 0.31	31.72 ± 20.27	4.73 ± 7.26	0.56 ± 0.41	0.35 ± 0.22
Piccola dim.	7	0.28 ± 0.28	39.17 ± 27.03	13.04 ± 19.43	0.20 ± 0.23	0.63 ± 0.41
Grande dim.	20	0.58 ± 0.22	28.75 ± 33.50	0.59 ± 1.89	0.71 ± 0.33	0.57 ± 0.21
Set completo	27	0.50 ± 0.27	31.45 ± 32.27	3.82 ± 11.42	0.58 ± 0.38	0.59 ± 0.28

I risultati dimostrano come la divisione stratificata del dataset consenta di ottenere performance paragonabili indipendentemente dalla classe ISUP di appartenenza raggiungendo l'obiettivo per il quale questa strategia è stata implementata. Tuttavia, i valori presentano anche un'elevata deviazione standard. Dall'analisi qualitativa effettuata non si sono evidenziate differenze riscontrabili da un occhio inesperto. Facendo uso del report diagnostico si è cercato di identificarne i fattori alla base di questo comportamento confrontando i dati relativi ai casi con Dice inferiori a 0.15 con quelli dei casi con Dice superiore a 0,7. Per questa analisi si sono tenuti in considerazione tutti i casi positivi del *Validation set Human* e del *Test set* con l'obiettivo di effettuare una valutazione più accurata e robusta utilizzando un set più numeroso. Nella figura 5.20 sono riportati i grafici risultanti dal confronto effettuato.



Numero casi	
Dice == 0	6
Dice < 0.15	16
Dice > 0.70	19

Numero casi con PSAD	
Dice == 0	1
Dice < 0.15	2
Dice > 0.70	12

Figura 5.20: Analisi e confronto delle caratteristiche dei casi in cui si sono ottenute ottime e cattive performance (Dice==0, Dice>70 e Dice<0.15). I casi considerati appartengono al Validation e al Test set.

Dall'analisi di questi grafici emerge che la discrepanza nelle performance non può essere attribuibile a fattori come l'età o classe ISUP di appartenenza. Sebbene infatti i casi con valori di Dice pari a 0 appartengano principalmente alla classe 2 e 3, una percentuale dei casi con Dice inferiore a 0.15 appartiene anche alla classe 4 e 5. I casi in cui si sono ottenute performance migliori presentano tipicamente un volume più piccolo. Questo fattore tuttavia potrebbe essere attribuibile all'efficacia della pipeline di pre-processing piuttosto che alle caratteristiche della lesione o dell'immagine. Per quanto riguarda la distribuzione dei valori di *psad*, come atteso, i casi in cui si sono ottenute performance soddisfacenti (Dice > 0.70) presentano dei valori tipicamente superiori a quelli dell'intero dataset. Un dato interessante emerso da questa analisi è che soltanto per 2 dei 16 casi con dice inferiore a 0.15 viene riportato, nel report diagnostico, il valore di psad. Come trattato nella sezione 1.4 questo è un parametro tipicamente utilizzato durante la diagnosi del paziente ed è molto utile per avere una visione più chiara e completa

del caso clinico. Di conseguenza, questa mancanza potrebbe essere indice di una classificazione e di una maschera poco accurata, possibilità che, data la complessità della task e l'elevata variabilità inter/intra-operatore, non è possibile escludere. Infine, le cause del basso valore dei coefficienti di Dice potrebbero essere molteplici:

- La metrica stessa potrebbe essere intrinsecamente più sensibile nei casi in cui i veri positivi siano pochi, influenzando quindi la valutazione complessiva.

- Il disallineamento delle scansioni potrebbe essere un fattore significativo, specialmente per lesioni di piccole dimensioni, poiché può drasticamente ridurre il valore delle metriche calcolate.

- Alcune lesioni presentano una dimensione lungo l'asse Z molto piccola, rendendole più difficili da rilevare e segmentare accuratamente.

- Non è da escludere l'influenza di caratteristiche intrinseche delle immagini e delle lesioni difficilmente riscontrabili da un occhio inesperto.

Pertanto, per una valutazione accurata delle prestazioni dell'algoritmo, sarebbe opportuno che le lesioni identificate come false positive fossero supervisionate da un radiologo esperto al fine di escludere la possibilità di errori presenti nelle maschere manuali fornite. Infine non si sono rilevate differenze significative attribuibili alla data di acquisizione o Gleason Score.

Per una valutazione qualitativa delle performance, nella figura 5.21 è proposto un esempio in cui viene confrontata la maschera manuale fornita con quella ottenuta mediante l'algoritmo implementato. Il caso riportato come esempio ha ID 11465_1001489.

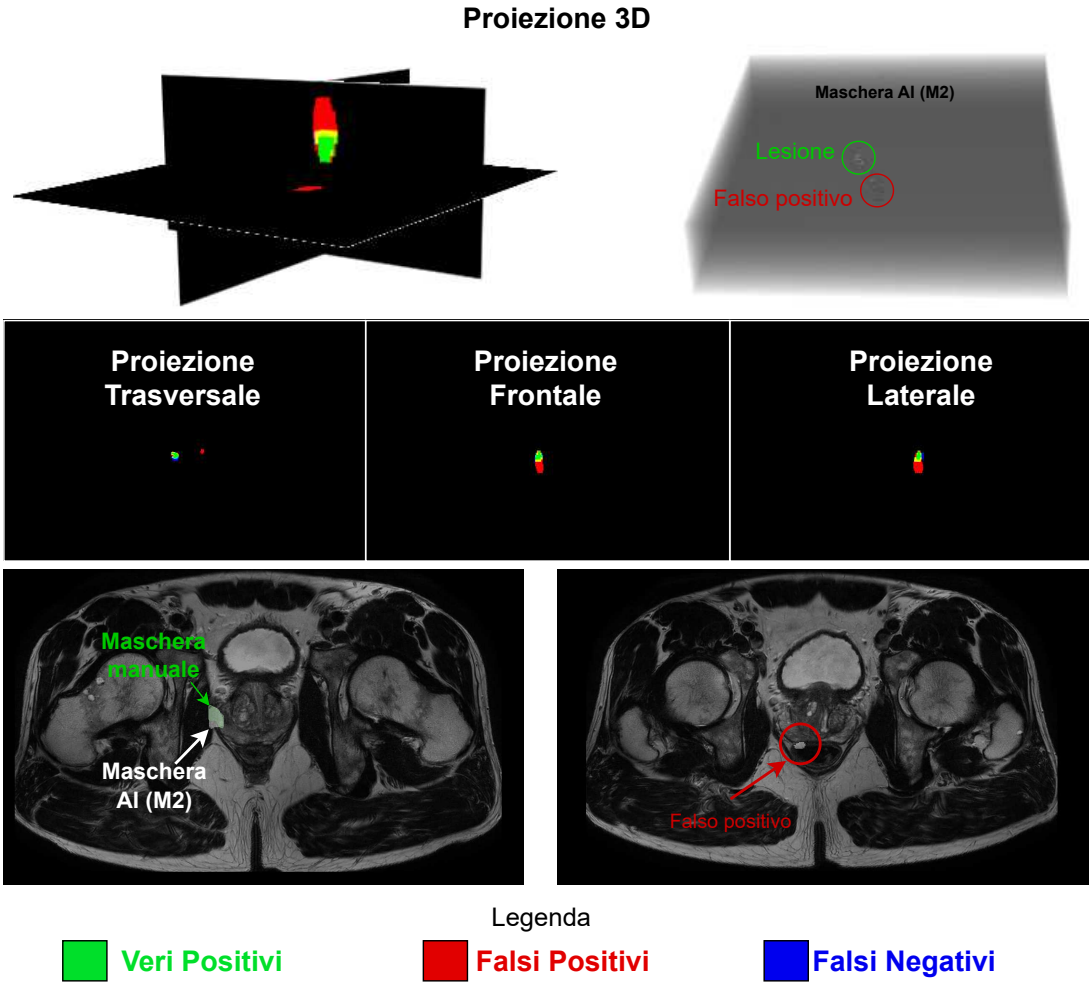


Figura 5.21: Confronto della maschera manuale e della maschera AI generata dall'algoritmo implementato. Il paziente preso in esame ha il seguente ID 11465_1001489. I pixel rappresentati in verde sono i *Veri Positivi*, quelli in rosso i *Falsi Positivi* e in blu i *Falsi Negativi*.

Valutazione dell'impatto del modello M1 sulle performance dell'algoritmo.

Per una valutazione quantitativa dell'impatto del modello *M1* sulle performance dell'algoritmo, nella tabella 5.33 sono confrontati i valori delle metriche di valutazione ottenute utilizzando, nella pipeline di *pre-processing 2*, le maschere prostatiche generate dal modello *M1* con i valori ottenuti utilizzando le maschere di "Guerbet23".

Tabella 5.33: Tabella di valutazione dell'impatto del modello *M1* (prova *Maschere prostatiche M1*) sulle performance dell'intero algoritmo. Nella prova *Maschere prostatiche fornite* sono riportate le metriche delle maschere delle lesioni ottenute facendo uso delle maschere di "Guerbet23" nella pipeline di *pre-processing 2*. Le metriche sono state calcolate sul Test set interno.

Prova	Dice	HD_{95}	RVD	Precisione	sensibilità
Maschere prostatiche M1	0.50 ± 0.27	31.45 ± 32.27	3.82 ± 11.42	0.58± 0.38	0.59 ± 0.28
Maschere prostatiche fornite	0.46 ± 0.23	24.14 ± 29.98	0.73 ± 3.26	0.68 ± 0.38	0.45 ± 0.25

Come atteso, le performance del modello *M1* influenzano le prestazioni dell'algoritmo implementato riducendo la precisione delle segmentazioni. I risultati tuttavia sono paragonabili a quelli ottenuti mediante l'utilizzo delle maschere di "Guerbet23" e dimostrano la validità del metodo implementato.

5.5 Confronto con la letteratura

Il confronto con la letteratura ha evidenziato una varietà di approcci e modelli utilizzati per la segmentazione delle lesioni. A differenza di quanto effettuato in diversi studi presenti in letteratura:

- Per lo sviluppo dell'algoritmo si è fatto uso solo del dataset della PICAI challenge. Altri gruppi di ricerca invece hanno fatto uso anche di set esterni (per es. [29] [30]).

- Si è utilizzata la *validazione hold-out* anziché il metodo della *cross-validazione*. La validazione hold-out presenta il vantaggio di essere semplice e veloce da implementare, ma ha lo svantaggio di una varianza elevata nelle prestazioni del modello a causa della dipendenza dalla specifica divisione dei dati. Per garantire una corretta rappresentazione di ciascuna classe nei set generati è stata eseguita una divisione stratificata del dataset. Inoltre, sono stati esclusi i casi dubbi, come i casi positivi con maschera vuota e quelli in cui la maschera della lesione non si sovrappone alla regione dell'immagine occupata dalla prostata.

- La pipeline di pre-processing è stata sottoposta ad un attento processo di analisi e ottimizzazione il quale è stato descritto nella sottosezione 5.1 e valutato nella 5.4.1.
- Per effettuare la segmentazione delle lesioni si è fatto uso solo del modello UNet *M2*. In letteratura sono state implementate strutture più complesse che prevedono l'uso di più modelli in cascata i cui output sono combinati tra loro per ottenere un unica maschera più precisa delle eventuali lesioni presenti nella prostata (sezione 2.2).
- I modelli utilizzati in letteratura (sezione 2.2) presentano architetture più profonde e strategie di addestramento molto più onerose a livello computazionale rispetto a quelle implementate.
- Contrariamente a quanto effettuato dai gruppi che hanno partecipato alla challenge, si sono riportati i valori di tutte le metriche utilizzate. Indipendentemente infatti dalla robustezza della metrica è importante poter avere dei valori di riferimento utili per una valutazione ed un confronto oggettivo del metodo proposto.

Capitolo 6
Sviluppo dell'algoritmo di classificazione dei pazienti

In questo capitolo, viene presentata la funzione sviluppata per fornire una diagnosi del paziente in termini probabilistici e per determinarne la classe di appartenenza, assegnando il valore 1 ai casi positivi e il valore 0 ai casi negativi.

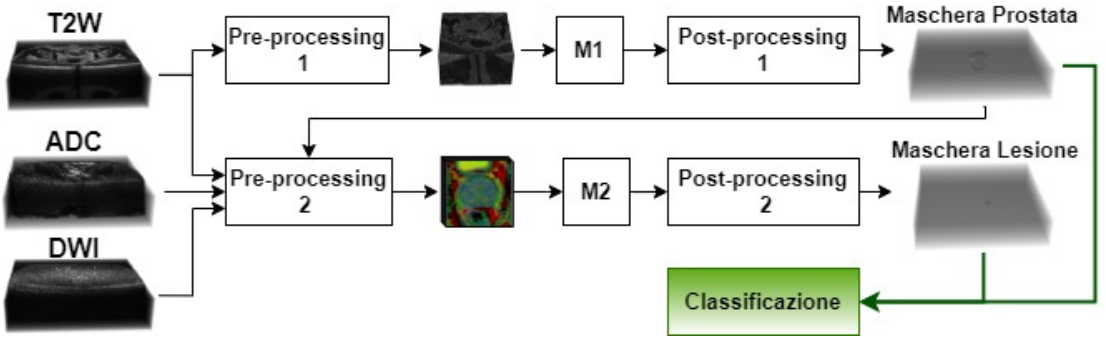


Figura 6.1: Flow chart del processo. I processi descritti nel capitolo seguente sono evidenziati in verde.

6.1 Sviluppo dell'algoritmo

L'algoritmo implementato per eseguire la classificazione del paziente si basa sull'utilizzo di quattro parametri:

- Dim_l: dimensione della lesione (in pixel) calcolata facendo uso della maschera della lesione generata da M2;
- Dim_p: dimensione della prostata (in pixel) calcolata facendo uso della maschera prostatica generata da M1;
- primo valore soglia "x": percentuale della dimensione della prostata utilizzata per calcolare la probabilità della positività del paziente;

- secondo valore soglia "x2": utilizzato per la diagnosi finale del paziente (classificazione binaria in positivo-negativo).

Nello specifico, si è definita una funzione che confrontata la dimensione della lesione identificata con la dimensione della prostata e, sulla base di questo confronto, viene generato un valore che indica la probabilità ("perc_classe") della positività del paziente. Secondo questo criterio pertanto più grande è la lesione identificata, maggiore è la probabilità che il paziente sia positivo. Se la dimensione della lesione è maggiore dell' x% del volume della prostata allora il paziente viene considerato positivo con probabilità unitaria. A sua volta la probabilità calcolata viene impiegata per determinare la classe di appartenenza del paziente per la diagnosi finale: 1 se positivo e 0 se negativo. Per eseguire questa classificazione si è determinato un secondo valore soglia "x2".

Nella figura 6.2 è riportato il flow chart descrittivo del processo di diagnosi del paziente.

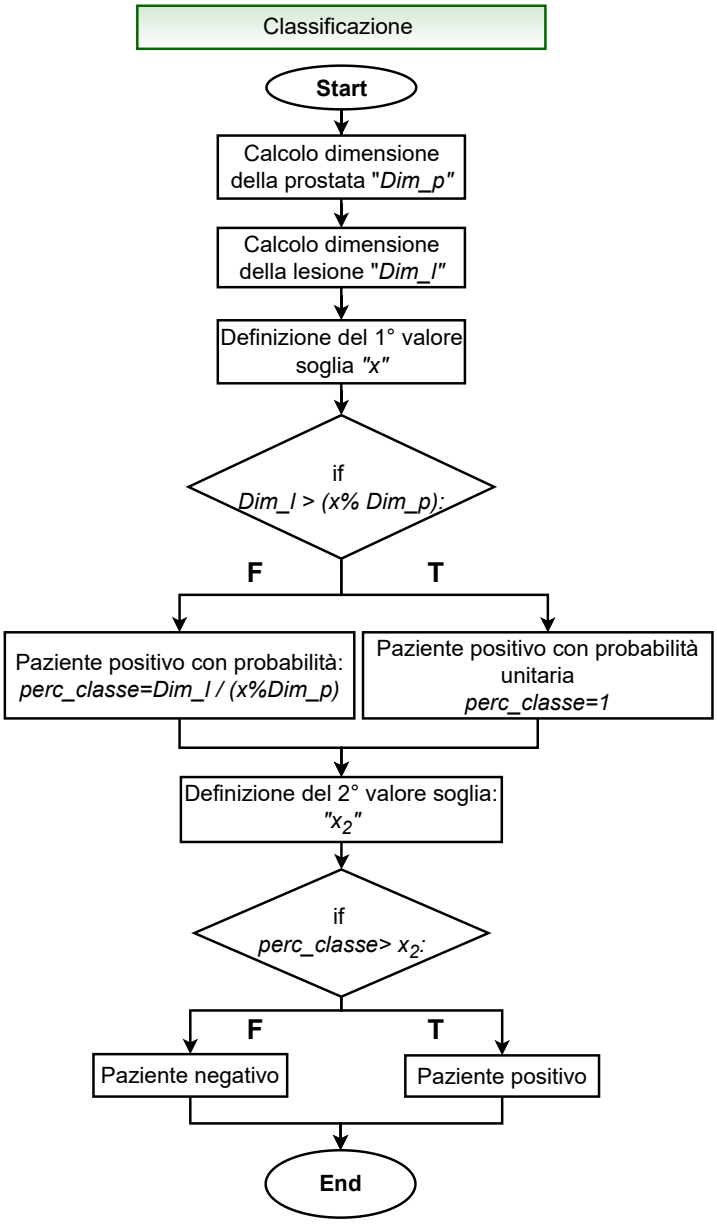


Figura 6.2: Flowchart del processo di diagnosi del paziente.

6.2 Risultati

Nella tabella 6.1 sono riportati i risultati dell'analisi sperimentale condotta per l'ottimizzazione del primo valore soglia.

Tabella 6.1: Tabella di valutazione dell'impatto del primo valore soglia sulle performance di classificazione dell'algoritmo applicato sul Validation set.

Percentuale corretti classificati Validation set (2° valore soglia = 50%)		
1° valore soglia	4%	5%
Positivi	66.66%	48.88%
Negativi	30%	30%
Set completo	60%	0.45%

L'analisi sperimentale ha evidenziato un valore ottimale pari al 4% del volume della prostata. Questo risulta essere un valore ragionevole considerate le dimensioni tipiche delle lesioni tumorali e la tendenza del modello M1 a sovra-segmentare la prostata.

Nella tabella 6.2 sono riportati i risultati dell'analisi sperimentale condotta per l'ottimizzazione del secondo valore soglia utilizzato per la diagnosi finale.

Tabella 6.2: Tabella di valutazione dell'impatto del secondo valore soglia sulle performance di classificazione dell'algoritmo applicato sul Validation set.

Percentuale corretti classificati Validation set (1° valore soglia = 4% volume prostata)		
2° valore soglia	50%	47%
Positivi	64.44%	73.33%
Negativi	30%	30%
Set completo	58%	65.45%

Per quanto concerne il secondo valore soglia si è scelto di classificare come positivi tutti i pazienti in cui si è stimata una probabilità di positività > del 47%. I risultati ottenuti evidenziano come l'utilizzo di un valore più basso rispetto al 50% consente di diagnosticare con maggiore precisione un numero maggiore di pazienti positivi.

Nella tabella 6.3 sono riportatati i valori delle performance di predizione dell'algoritmo implementato applicato sul Test set.

Tabella 6.3: Tabella di valutazione delle performance di classificazione dell'algoritmo applicato sul Test set.

Percentuale corretti classificati Test set	
Positivi	74%
Negativi	33,33%
Set completo	66,66%

Il criterio adottato accetta il compromesso di avere un numero maggiore di falsi positivi per poi effettuare successivi accertamenti, piuttosto che avere delle diagnosi negative errate con conseguenze più gravi per il paziente. Questo approccio mira a garantire un equilibrio tra la sensibilità nella rilevazione delle condizioni positive e la riduzione del rischio di diagnosi negative sbagliate.

Nell'analisi dei risultati ottenuti va sottolineato che, a differenza di quanto fatto da altri gruppi di ricerca [29] [33], non si è fatto uso del PSAD per affinare la diagnosi e la classificazione del paziente, poiché questo non è sempre disponibile nei dati clinici forniti. Inoltre, sebbene l'uso di parametri esterni possa essere utile per migliorare la precisione della classificazione, questi richiedono ulteriori esami (come le analisi del sangue) determinando un allungamento dei tempi diagnostici. L'obiettivo infatti è quello di implementare un algoritmo in grado di effettuare una diagnosi in tempi rapidi facendo uso esclusivamente delle scansioni bpMRI.

Contrariamente all'approccio proposto dagli organizzatori della challenge, la classificazione non viene effettuata facendo uso esclusivamente del massimo valore dei pixel presenti nella mappa di rilevamento delle lesioni generata. Tale metodo infatti risulterebbe molto sensibile alla presenza di falsi positivi e alle performance di predizione del modello M2.

Inoltre, è importante considerare che alcuni casi ISUP 1, classificati come negativi nel report diagnostico, possono presentare lesioni tumorali di basso grado con minima aggressività. Pertanto, la classificazione da parte del modello di questi casi come positivi non è considerabile del tutto errata.

Capitolo 7
Conclusioni

In questa sezione sono esaminati i limiti, l'efficacia, la robustezza del metodo implementato e le possibili direzioni future per il lavoro.

Una delle limitazioni evidenziate durante l'analisi sperimentale riguarda le scansioni che includono l'intera area addominale. In particolare si è osservato che queste, rispetto ai casi in cui è stata utilizzata una finestra di acquisizione più piccola, sono significativamente meno numerose. Le scansioni che includono l'intera area addominale, tipicamente con dimensioni post-resampling di 700x700 pixel, rappresentano una sfida per il modello M1 nella corretta identificazione e segmentazione della ghiandola prostatica. Sebbene sia possibile risolvere questa problematica mediante il crop dell'immagine, come implementato durante la fase di inference per questo specifico dataset, tale approccio non risulta essere sempre affidabile nella pratica clinica. Una possibile soluzione potrebbe essere l'inclusione nel dataset di un numero maggiore di acquisizioni che includano l'intero addome, al fine di migliorare la capacità del modello nel gestire questa tipologia di input. Nonostante le limitazioni legate alla qualità delle maschere utilizzate per il training del modello e all'impiego ridotto di risorse computazionali, l'algoritmo di segmentazione della prostata ha dimostrato di essere idoneo per il raggiungimento degli obiettivi prefissati.

Per quanto riguarda l'algoritmo di segmentazione delle lesioni, le analisi sperimentali effettuate ed il confronto con la letteratura evidenziano come il Dataset di cui si è fatto uso, in termini di numerosità, risulta essere appena sufficiente. Le lesioni tumorali infatti, come evidenziato anche nel capitolo 3, presentano caratteristiche eterogenee e i casi in esame possono essere molto differenti tra loro. L'uso di dataset più numerosi, come fatto da diversi gruppi di ricerca integrando set esterni, consentirebbe di allenare i modelli su set più eterogenei migliorandone le capacità di generalizzazione e discriminazione.

Tra le criticità emerge inoltre l'assenza di un metodo di registrazione delle scansioni. Essendo infatti un dataset reale, e di conseguenza molto eterogeneo, risulta essere complesso realizzare un algoritmo robusto che consenta

di eseguire questa operazione in modo del tutto automatico. Questo determina un disallineamento più o meno evidente delle scansioni, il quale, come riportato anche in letteratura [47], influisce negativamente sulle prestazioni del modello e sul valore delle metriche.

Nonostante il dataset fornito dagli organizzatori della challenge sia stato attentamente curato, si è reso necessario escludere alcuni casi "ambigui" relativi a pazienti positivi per il report diagnostico, ma con maschera della lesione nulla o con lesioni spazialmente non sovrapposte con la ROI. Al fine di escludere la possibilità di ulteriori errori nel dataset e di ottenere una valutazione più accurata delle prestazioni dell'algoritmo, sarebbe opportuna la supervisione di un radiologo esperto per la valutazione dei casi in cui l'output generato differisce dal ground truth.

Diversi studi presenti in letteratura suggeriscono che l'utilizzo di reti neurali più complesse e l'impiego di più modelli in cascata consentono di ottenere risultati migliori rispetto all'uso di un singolo modello UNet. Di conseguenza, si ipotizza che le soluzioni proposte, integrate in architetture più complesse, possano offrire prestazioni superiori rispetto a quelle attualmente ottenute. Tuttavia, le risorse computazionali disponibili per lo sviluppo del progetto non erano sufficienti per l'implementazione di reti di maggiore complessità.

Da studi clinici è emerso che il tumore prostatico presenta una maggiore propensione a svilupparsi nella zona PZ piuttosto che nella TZ [1]. Basandosi su tali evidenze, un potenziale miglioramento delle prestazioni del modello impiegato per l'identificazione e la segmentazione delle lesioni tumorali potrebbe derivare dall'implementazione di una "mappa di probabilità". Questa mappa, generata per ciascun caso, indicherebbe al modello le aree con una maggiore probabilità di ospitare una lesione. Tuttavia, questa strategia, proposta anche in letteratura [61], è attualmente limitata dall'utilizzo di dataset di dimensioni ridotte, il che potrebbe influenzare la validità del criterio definito, introducendo un potenziale bias nelle prestazioni dell'algoritmo. Inoltre, vi è da considerare anche i costi computazionali associati: l'idea di utilizzare la mappa generata per effettuare modifiche sulle scansioni stesse o come un'acquisizione aggiuntiva nel volume 4D (dato in input al modello M2 per la segmentazione delle lesioni), comporterebbe un aumento dei requisiti computazionali. Pertanto, oltre a valutare l'efficacia e la robustezza del metodo proposto, è essenziale condurre un processo di ottimizzazione per garantire che l'algoritmo risulti efficiente e non eccessivamente oneroso in termini di tempo e risorse.

Per quanto concerne la classificazione, il nuovo criterio introdotto non fa uso di parametri esterni, quali per esempio il PSAd, consentendo la diagnosi del paziente facendo uso esclusivamente delle scansioni bmMRI. La precisione della classificazione è strettamente correlata alla qualità delle maschere generate e migliorando le performance dei modelli è possibile migliorare la qualità della diagnosi. Il criterio introdotto tuttavia presenta un limite legato alla capacità nel diagnosticare la positività dei pazienti aventi lesioni di dimensioni molto piccole.

I risultati finali ottenuti sul test set interno (Dice: 0.50 ± 0.27; Distanza di Hausdorff (95°percentile): 31.45 pixel $\pm$ 32.27; RVD: 3.82 % $\pm$ 11.42; Precisione: 0.58 ± 0.38; Sensibilità: 0.59 ± 0.28; Percentuale corretti classificati positivi: 74%; Percentuale corretti classificati negativi: 33%) dimostrano la validità del metodo implementato rispetto alla soluzione base proposta dagli organizzatori della PI-CAI challenge (Dice: 0.30 ± 0.39; Distanza di Hausdorff (95°percentile): 2.74 pixel $\pm$ 4.24; RVD: -0.61 % $\pm$ 0.51; Precisione: 0.83 ± 0.10; Sensibilità: 0.30 ± 0.40; Percentuale corretti classificati positivi: 40%; Percentuale corretti classificati negativi: 83%). Tuttavia, è importante sottolineare che questo confronto presenta alcune limitazioni. In particolare, non è possibile escludere la possibilità che i casi inseriti nel test set siano stati utilizzati dagli organizzatori della challenge per addestrare e convalidare il modello. Dai risultati emerge un netto miglioramento del Dice medio e della sensibilità di rilevazione. Tuttavia, il modello implementato mostra una diminuzione della precisione e una distanza di Hausdorff più elevata. Questo potrebbe essere dovuto a diversi fattori quali la presenza di eventuali errori nel Ground Truth e/o la semplicità del modello e della strategia di addestramento adottata.

In conclusione: le prove sperimentali effettuate hanno dimostrato che l'introduzione delle nuove strategie e operazioni hanno permesso di migliorare la robustezza dell'algoritmo e la qualità delle segmentazioni raggiungendo gli obiettivi prefissati della tesi. In particolare, le principali innovazioni apportate riguardano:

- la formazione e la cura dei set di addestramento;
- l'ottimizzazione delle pipeline di pre-processing rese più efficaci e robuste;
- l'ottimizzazione dei processi di training dei modelli implementati;

- nuovo criterio di classificazione del paziente più robusto alla presenza di falsi positivi.

Sebbene l'algoritmo implementato risulti essere robusto e ben funzionante in contesti reali, è importante riconoscere come vi siano ancora ampi spazi di miglioramento e potenziali sviluppi futuri. Infatti le prestazioni ottenute sono fortemente influenzate dall'utilizzo di modelli UNet molto semplici e da architetture poco profonde.

L'elevata variabilità nelle performance indica la necessità di ulteriori analisi e miglioramenti per garantire una segmentazione delle lesioni di maggiore qualità e, di conseguenza, una diagnosi più affidabile. Questi sono essenziali per rendere l'algoritmo un'opzione affidabile e realmente utilizzabile nella pratica clinica. I risultati ottenuti aprono nuove prospettive nel campo della segmentazione delle lesioni prostatiche, contribuendo a colmare alcune lacune presenti nella letteratura esistente.

Bibliografia

[1] Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent Automatic Segmentation Algorithms of MRI Prostate Regions: A Review," *IEEE Access*, vol. 9, pp. 97878-97905, 2021. doi: 10.1109/ACCESS.2021.3090825.

[2] Istituto Italiano Edizioni Atlas. *Atlante del Corpo umano*.Immagine estratta dalla pagina 45. Disponibile online: file:///D:/DOWNLOAD/corpo_umano%20(1).pdf.

[3] "Prostata." Humanitas, *Humanitas*. Disponibile su: https://www.humanitas.it/enciclopedia/anatomia/apparato-riproduttivo/apparato-riproduttivo-maschile/prostata/#:~:text=La%20forma%20e%20le%20dimensioni,poi%20svilupparsi%20durante%20la%20pubert%C3%A0. Accesso Maggio 2024.

[4] "Dimensioni della Prostata." Sessa Francesco, *Dr. Francesco Sessa*. Disponibile su: https://sessafrancesco.it/dimensioni-prostata/. Accesso Maggio 2024.

[5] Sharma, M., Gupta, S., Dhole, B., & Kumar, A. (2017). The Prostate Gland. In *Textbook of Male Genitourethral Disorders* (pp. 17-35). doi: 10.1007/978-981-10-3695-8_2

[6] Eklund, Martin, et al. "Biopsia mirata o standard con RM nello screening del cancro alla prostata." *New England Journal of Medicine* 385.10 (2021): 908-920. DOI: https://doi.org/10.1056/NEJMoa2100852

[7] de Rooij, Maarten, et al. "Can Biparametric Magnetic Resonance Imaging of the Prostate Meet the Criteria for Prostate Cancer Screening Protocols?" *European Urology* (2020). DOI: https://doi.org/10.1016/j.eururo.2020.04.062

[8] Turkbey, Baris, et al. "Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2." *European Urology* 76.3 (2019): 340-351. DOI: https://doi.org/10.1016/j.eururo.2019.02.033

[9] Studio Urologico Gallo, *Prostata*, https://www.studiourologicogallo.it/prostata#:~:text=I%20dotti%20delle%20ghiandole%20prostatiche,eiaculatori%20fino%20base%20della%20vescica

[10] Urologo Genova. (2018, 25 Giugno). *Cosa indica il PI-RADS in un referto di risonanza magnetica multiparametrica prostatica?*. https://www.urologo-genova.it/articoli/pirads-pi-rads-risonanza-prostata-tumore-multiparametrica.htm Consultato a Maggio 2024.

[11] Umberto Santaniello, *Patologie Organo per Organo: Prostata*, https://www.umbertosantaniello.it/index.php/il-dottore/patologie-organo-per-organo/prostata,

[12] Czarniecki M, Bickle I, Weerakkody Y, et al. Prostate Imaging-Reporting and Data System (PI-RADS). Reference article, Radiopaedia.org. rID: 27968. Article created: 1 Mar 2014, Marcin Czarniecki. Last revised: 13 Jun 2023, Ian Bickle. https://radiopaedia.org/articles/27968

[13] Urologo Genova. (2018). *Gleason score e tumore alla prostata: punteggio e prognosi*. https://www.urologo-genova.it/articoli/gleason-score-tumore-prostata-punteggio-prognosi.htm

[14] Urologo Genova. (2018). *Il grado di tumore alla prostata: il Grade Group (Grading ISUP) e il Gleason score*. https://www.urologo-genova.it/articoli/181204/grado-tumore-prostata-grade-group-grading-isup-gleason.htm

[15] Abdelrazek, A., Mahmoud, A. M., Joshi, V. B., Habeeb, M., Ahmed, M. E., Ghoniem, K., Delgado, A., Khater, N., Kwon, E., & Kendi, A. T. (2022). Recenti progressi nella diagnostica del cancro alla prostata (PCa). *Uro*, *2*(2), 109-121. https://doi.org/10.3390/uro2020014

[16] Engels, Rianne R.M., et al. "Multiparametric Magnetic Resonance Imaging for the Detection of Clinically Significant Prostate Cancer: What Urologists Need to Know. Part 1: Acquisition." *European Urology* 77.4 (2020): 457-468. DOI: 10.1016/j.eururo.2019.09.021

[17] Israël, Bas, et al. "Multiparametric Magnetic Resonance Imaging for the Detection of Clinically Significant Prostate Cancer: What Urologists Need to Know. Part 2: Interpretation." *European Urology* 77.4 (2020): 469-480. DOI: https://doi.org/10.1016/j.eururo.2019.10.024

[18] Collins, G. N., Lee, R. J., McKelvie, G. B., Rogers, A. C., & Hehir, M. (1993). Relationship between prostate specific antigen, prostate volume and age in the benign prostate. *British Journal of Urology*, *71*(4), 445-450. doi: 10.1111/j.1464-410x.1993.tb15990.x. PMID: 7684650.

[19] Lei He, Zhigang Peng, Bryan Everding, Xun Wang, Chia Y. Han, Kenneth L. Weiss, William G. Wee, *A comparative study of deformable contour methods on medical image segmentation*, *Image and Vision Computing*, **26**(2), 141-163, 2008, ISSN: 0262-8856, Keywords: Medical image segmentation, Deformable contour method, Snake, Level set, Comparative study, DOI: https://doi.org/10.1016/j.imavis.2007.07.010, URL: https://www.sciencedirect.com/science/article/pii/S0262885607001230,

[20] S. Menet, P. Saint-Marc, G. Medioni, *Active contour models: overview, implementation and applications*, *1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings*, Los Angeles, CA, USA, 1990, pp. 194-199, DOI: https://doi.org/10.1109/ICSMC.1990.142091, Keywords: Active contours, Tracking, Spline, Image segmentation, Intelligent robots, Deformable models, Convergence, Data mining, Image edge detection, Contracts.

[21] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203-211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[22] KnowledgeHut. (2023). *Introduction to k-Fold Cross-Validation in Machine Learning.* Recuperado de https://www.knowledgehut.com/blog/data-science/introduction-to-k-fold-cross-validation-in-machine-learning

[23] Turkbey, B., Shah, V. P., Pang, Y., et al. "Is apparent diffusion coefficient associated with clinical risk scores for prostate cancers that are visible on 3-T MR images?" *Radiology*, **258**, 488-495 (2011).

[24] Boesen, L., Chabanova, E., Logager, V., Balslev, I., & Thomsen, H. S. "Apparent diffusion coefficient ratio correlates significantly with prostate cancer Gleason score at final pathology." *Journal of Magnetic Resonance Imaging*, **42**(2), 446-453 (2015).

[25] Hambrock, T., Hoeks, C., Hulsbergen-van de Kaa, C., et al. "Prospective assessment of prostate cancer aggressiveness using 3-T diffusion-weighted magnetic resonance imaging-guided biopsies versus a systematic 10-core transrectal ultrasound prostate biopsy cohort." *European Urology*, **61**(1), 177-184 (2012).

[26] Giganti, F., Kirkham, A., Kasivisvanathan, V., et al. (2021). Understanding PI-QUAL for prostate MRI quality: a practical primer for radiologists. *Insights into Imaging*, *12*, 59. doi: 10.1186/s13244-021-00996-6.

[27] Zhang, H., & Qie, Y. (2023). Applying Deep Learning to Medical Imaging: A Review. *Applied Sciences*, *13*(18), 10521. doi: 10.3390/app131810521. Recuperato da https://www.mdpi.com/2076-3417/13/18/10521

[28] Anindo Saha, J. J. Twilt, Joeran S. Bosma, Bram van Ginneken, Doruk Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, Henkjan Huisman, *The PI-CAI Challenge: Public Training and Development Dataset (v2.0)*, Data set, Zenodo, version = v2.0, year = 2022, doi = 10.5281/zenodo.6624726, url = https://doi.org/10.5281/zenodo.6624726

[29] Debs, Noëlie, et al. "Deep learning for detection and diagnosis of prostate cancer from bpMRI and PSA: Guerbet's contribution to the PI-CAI 2022 Grand Challenge." Guerbet Research, Villepinte, France. *Email:* noelie.debs@guerbet.com, alexandre.routier@guerbet.com

[30] Yuan, Yuan, et al. "Prostate Imaging: Cancer AI (PI-CAI) Challenge 2022 Z-SSMNet: A Zonal-aware Self-Supervised Mesh Network for Prostate Cancer Detection and Diagnosis in bpMRI." *School of Computer Science, University of Sydney, Sydney, Australia; College of Science & Engineering, James Cook University, Cairns, Australia; Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China; Department of Urology, Nepean Hospital, Kingswood, Australia*

[31] Kan, Hongyu, et al. "Implementation Method of the PI-CAI Challenge (Swangeese Team) Technical Report." *Department of Computer Science and Technology, University of Science and Technology of China, Anhui, Hefei.* Emails: honeyk@mail.ustc.edu.cn, ql1an9@mail.ustc.edu.cn, shijun18@mail.ustc.edu.cn, han@ustc.edu.cn

[32] Li, Xinran, et al. "The Prostate Imaging: Cancer AI (PI-CAI) 2022 Grand Challenge (PIMed Team)." *Department of Radiology, Stanford University, Stanford CA 94305, USA; Department of Urology, Stanford University, Stanford CA 94305, USA; Institute for Computational and Mathematical Engineering, Stanford CA 94305, USA.* Emails: svesal@stanford.edu, mirabela.rusu@stanford.edu

[33] Karagöz, Ahmet, et al. "Prostate Lesion Estimation using Prostate Masks from Biparametric MRI." *Computer Engineering, Istanbul Technical University, Istanbul, Turkey; Acibadem Mehmet Ali Aydinlar University School of Medicine, Istanbul, Turkey; Department of Software Engineering and Applied Sciences, Bahcesehir University, Istanbul, Turkey; Faculty of Medicine, Sivas Cumhuriyet University, Sivas, Turkey*

[34] Barrett, T., Turkbey, B., & Choyke, P. L. "PI-RADS version 2: what you need to know." *Clinical Radiology*, **70**(11), 1165-1176 (2015). DOI: https://doi.org/10.1016/j.crad.2015.06.093

[35] Chenevert, T. L., Malyarenko, D. I., Newitt, D., Li, X., Jayatilake, M., Tudorica, A., Fedorov, A., Kikinis, R., Liu, T. T., Muzi, M., Oborski, M. J., Laymon, C. M., Li, X., Thomas, Y., Jayashree, K. C., Mountz, J. M., Kinahan, P. E., Rubin, D. L., Fennessy, F., Huang, W., Hylton, N., & Ross, B. D. "Errors in Quantitative Image Analysis due to Platform-Dependent Image Scaling." *Translational Oncology*, **7**(1), 65-71 (2014). DOI: https://doi.org/10.1593/tlo.13811

[36] Venderink, W., van Luijtelaar, A., van der Leest, M., Barentsz, J. O., Jenniskens, S. F. M., Sedelaar, M. J. P., Hulsbergen-van de Kaa, C., Overduin, C. G., & Fütterer, J. J. "Multiparametric magnetic resonance imaging and follow-up to avoid prostate biopsy in 4259 men." *BJU International*, **124**(5), 775-784 (2019). DOI: https://doi.org/10.1111/bju.14853

[37] Srivastava, Sudhir, et al. "Cancer overdiagnosis: a biological challenge and clinical dilemma." *Nature Reviews Cancer* 19.6 (2019): 349-358. DOI: https://doi.org/10.1038/s41568-019-0142-8

[38] Joeran S. Bosma, Anindo Saha, Matin Hosseinzadeh, Ivan Slootweg, Maarten de Rooij, Henkjan Huisman, *Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning–based Prostate Cancer Detection Using Biparametric MRI*, *Radiology: Artificial Intelligence*, volume = 5, number = 5, ISSN = 2638-6100, publisher = Radiological Society of North America (RSNA), year = 2023, month = September, url = http://dx.doi.org/10.1148/ryai.230031, DOI = 10.1148/ryai.230031

[39] Pulp Learning. "Train, Validation e Test: Cosa sono e Come si Usano nel Machine Learning." https://pulplearning.altervista.org/train-validation-test-cosa-sono-e-come-si-usano-nel-machine-learning/.

[40] Levity AI. *Difference Between Machine Learning and Deep Learning.* https://levity.ai/blog/difference-machine-learning-deep-learning.

[41] Prof. Gianluca Amato, *Dipartimento di Economia, Università "G. d'Annunzio" di Chieti-Pescara*, Ultimo aggiornamento: 22 mar 2022.

[42] Data Mining di Reti Neurali Artificiali, *https://ita.animalia-life.club/data-mining-di-reti-neurali-artificiali.*

[43] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[44] Sanjar, Karshiev, et al. *Improved U-Net: Fully Convolutional Network Model for Skin-Lesion Segmentation. Applied Sciences*, vol. 10, no. 10, 2020, pp. 3658. DOI: 10.3390/app10103658 https://www.mdpi.com/2076-3417/10/10/3658

[45] Anindo Saha, Jasper J. Twilt, Joeran S. Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, Henkjan Huisman, *Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol)*, 2022, doi:10.5281/zenodo.6667655

[46] Armato SG 3rd, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, Redmond G, Giger ML, Cha K, Mamonov A, Kalpathy-Cramer J, Farahani K. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging (Bellingham). 2018 Oct;5(4):044501. doi: 10.1117/1.JMI.5.4.044501. Epub 2018 Nov 10. PMID: 30840739; PMCID: PMC6228312.

[47] A. Saha, M. Hosseinzadeh, and H. Huisman, "End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction," *Medical Image Analysis*, vol. 73, p. 102155, 2021. DOI: 10.1016/j.media.2021.102155.

[48] Damiani, Ernesto. "L'intelligenza artificiale può compiere un salto di qualità grazie alle neuroscienze". Agenda Digitale, 11 Novembre 2021,

[49] S. R., and S. Patilkulkarni, "Visual speech recognition for small scale dataset using VGG16 convolution neural network," *Multimedia Tools and Applications*, vol. 80, 2021. DOI: 10.1007/s11042-021-11119-0.

[50] D. Palumbo, B. Yee, P. O'Dea, S. Leedy, S. Viswanath, and A. Madabhushi, "Interplay between bias field correction, intensity standardization, and noise filtering for T2-weighted MRI," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 5080–5083.

[51] Oracle, *What is Docker?*, https://www.oracle.com/it/cloud/cloud-native/container-registry/what-is-docker/.

[52] MONAI, *Medical Open Network for AI (MONAI)*, https://monai.io/.

[53] Haider, M. A., van der Kwast, T. H., Tanguay, J., Evans, A. J., Hashmi, A. T., Lockwood, G., & Trachtenberg, J. (2007). Combined T2-weighted and diffusion-weighted MRI for localization of prostate cancer. *AJR. American Journal of Roentgenology*, *189*(2), 323-328. doi: 10.2214/AJR.07.2211. PMID: 17646457.

[54] IBM, *Convolutional Neural Networks*, https://www.ibm.com/topics/convolutional-neural-networks.

[55] Barak Or. (2023). *On Common Split for Training, Validation, and Test Sets in Machine Learning*. Pubblicato in *Verso l'IA*, 18 aprile 2023. https://pub.towardsai.net/breaking-the-mold-challenging-the-common-splitfor-training.-validation-and-test-sets-in-machine-271fd405493d.

[56] Alain Horé and Djemel Ziou, *Image Quality Metrics: PSNR vs. SSIM*, In *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366-2369. https://doi.org/10.1109/ICPR.2010.579

[57] F. Milletari, N. Navab and S. -A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 2016, pp. 565-571, doi: 10.1109/3DV.2016.79.

[58] Jianping Li, Zhiming Cui, Shuai Wang, Jie Wei, Jun Feng, Shu Liao, & Dinggang Shen. (2021). Morphology-Guided Prostate MRI Segmentation with Multi-slice Association. In *Lecture Notes in Computer Science (LNIP, volume 12966)*. Springer.

[59] Ahmet Karagoz, Deniz Alis, Mustafa Ege Seker, Gokberk Zeybel, Mert Yergin, Ilkay Oksuz, & Ercan Karaarslan. (2023). Anatomically guided self-adapting deep neural network for clinically significant prostate cancer detection on bi-parametric MRI: a multi-center study. *Insights into Imaging*, **14**(1), 110. ISSN: 1869-4101. DOI: https://doi.org/10.1186/s13244-023-01439-0.

[60] Ocal, H., Barisci, N. (2022). Prostate Segmentation via Dynamic Fusion Model. *Arabian Journal for Science and Engineering*, **47**(8), 10211-10224. ISSN: 2191-4281. DOI: https://doi.org/10.1007/s13369-021-06502-w.

[61] Hosseinzadeh, M., Brand, P., & Huisman, H. (2019). Effect of adding probabilistic zonal prior in deep learning-based prostate cancer detection. In *International Conference on Medical Imaging with Deep Learning (MIDL)–Extended Abstract Track* (pp. 1026-1034). London, United Kingdom.

[62] Barrett, T., Turkbey, B., Choyke, P. L. (2015). PI-RADS version 2: what you need to know. *Clinical Radiology*, *70*(11), 1165-1176.

doi: 10.1016/j.crad.2015.06.093. Epub 2015 Jul 29. PMID: 26231470; PMCID: PMC6369533.