Politecnico di Torino

Politecnico di Torino

Corso di laurea magistrale in Ingegneria Biomedica

A.a. 2025/2026

Sessione di laurea Marzo 2026

Sviluppo di un Sistema Multimodale di Supporto alle Decisioni per la Qualità dell'Imaging Prostatico

Un Approccio basato su Radiomica ed Ensemble Learning

Relatori:

Valentina Giannini

Gabriella Balestra

Samanta Rosati

Candidati:

Marco La Iacona

Sommario

La gestione clinica del carcinoma prostatico (PCa), si affida oggi in modo cruciale alla Risonanza Magnetica Multiparametrica (mpMRI). Tuttavia, l'affidabilità diagnostica della mpMRI è strettamente subordinata alla qualità tecnica delle immagini, la cui valutazione standardizzata tramite il sistema PI-QUAL v2 risulta spesso soggettiva e dispendiosa in termini di tempo. Questo studio esplora lo sviluppo di un sistema automatico di supporto alle decisioni capace di discriminare la qualità diagnostica integrando tecniche di Radiomica, Machine Learning e Deep Learning.

Utilizzando un dataset multicentrico di 802 pazienti provenienti dal progetto europeo ProCAncer-I, sono state analizzate le sequenze fondamentali T2w, ADC e DWI. La pipeline metodologica ha previsto l'estrazione di 1.056 feature radiomiche per modalità, adottando sia un approccio Whole-Volume per catturare artefatti globali, sia un'analisi ROI-based focalizzata sul tessuto prostatico. Per far fronte all'elevata dimensionalità e allo sbilanciamento dei dati, sono stati implementati algoritmi di selezione come LASSO e Mutual Information e tecniche di bilanciamento sintetico (SMOTE) mentre, per garantire la massima robustezza e prevenire l'overfitting, i modelli sono stati addestrati e ottimizzati attraverso una strategia di Stratified 5-Fold Cross-Validation.

I risultati evidenziano come il Machine Learning classico offra una notevole stabilità in scenari di risorse limitate. Il modello Multi-Layer Perceptron ha raggiunto una Balanced Accuracy (BACC) media di 0,604 sulla modalità ADC, mentre il Random Forest si è distinto sulle sequenze T2w con una BACC di 0,593. L'integrazione multimodale tramite strategie di Ensemble Learning ha mostrato la superiorità del Soft Voting rispetto allo Stacking gerarchico, ottenendo una BACC di 0,683 in validazione. Sul test set indipendente, l'Ensemble ha raggiunto una BACC di 0,641 e una Specificità di 0,533, dimostrandosi più efficace dei modelli unimodali nel filtrare le immagini non diagnostiche e riducendo il tasso di falsi positivi.

Sebbene le architetture di Deep Learning 3D abbiano mostrato potenzialità interessanti - specialmente con l'uso di oversampling tramite rumore gaussiano (BACC 0,617) - hanno risentito maggiormente della scarsità di volumi annotati rispetto all'approccio radiomico. In conclusione, lo studio suggerisce che un sistema ibrido basato su feature ingegnerizzate e fusione multimodale possa costituire uno strumento di supporto al controllo qualità della mpMRI, affiancando il radiologo nella riduzione della soggettività valutativa e nel miglioramento dell'efficienza dei flussi di lavoro clinici nella diagnosi del tumore alla prostata.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione e Contesto Clinico

1.1 Il Carcinoma Prostatico

Il carcinoma della prostata (PCa) costituisce una delle neoplasie più diffuse nella popolazione maschile con oltre 161.000 nuove diagnosi, solo negli Stati Uniti, nel 2017 [1]. Tuttavia, questi dati epidemiologici, per quanto rilevanti, descrivono solo parzialmente la complessità della patologia, la quale si distingue per una marcata eterogeneità biologica e prognostica che rende estremamente variabile il decorso clinico. Mentre in molti casi la malattia presenta un andamento indolente, senza impatto significativo sulla qualità di vita; in altri, invece, evolve rapidamente verso forme aggressive e potenzialmente letali, nonostante il trattamento [1]. Questa variabilità fenotipica riflette una profonda eterogeneità molecolare, guidata da differenti vie di segnalazione, tra cui il pathway del recettore degli androgeni (AR-signaling), che contribuiscono a rendere la prognosi del singolo paziente complessa e spesso imprevedibile [2].

Alla luce di tale complessità, la gestione clinica moderna si fonda sull'identificazione delle diverse tappe dell'evoluzione del tumore, che spaziano dal tumore localizzato ormono-sensibile fino al carcinoma metastatico resistente alla castrazione (mCRPC) [1]. La progressione verso la resistenza alla castrazione rappresenta lo stadio finale della patologia, associato a un netto peggioramento della prognosi. Ne deriva l'importanza cruciale di una diagnosi non solo accurata, ma capace di supportare una stratificazione del rischio precoce per guidare scelte terapeutiche personalizzate.

1.2 Limiti dello Screening Tradizionale

Sebbene l'introduzione del test del *Prostate-Specific Antigen* (PSA) abbia rivoluzionato la diagnosi precoce, il suo utilizzo massivo come strumento di screening è oggi oggetto di dibattito a causa di limitazioni intrinseche [3]. La criticità principale risiede nella scarsa specificità del PSA per il carcinoma clinicamente significativo (csPCa): i livelli sierici possono infatti aumentare anche in presenza di condizioni benigne come l'iperplasia prostatica (BPH) o infiammazioni [2].

Questa mancanza di specificità ha generato storicamente delle problematiche di gestione clinica. Da un lato, l'alto tasso di falsi positivi ha portato a un eccesso di biopsie, esponendo molti pazienti a procedure invasive spesso non necessarie. Dall'altro, si è assistito al fenomeno dell'*overtreatment*: la diagnosi di tumori a basso rischio (Gleason Score = 6), che non avrebbero impattato la sopravvivenza del paziente, ha spesso condotto a trattamenti radicali con conseguenti effetti collaterali evitabili [3]. È stimato, infatti, che oltre il 40% dei pazienti diagnosticati presenti una malattia di basso grado [2].

Di fronte a queste evidenze, l'approccio diagnostico si è evoluto verso l'integrazione della Risonanza Magnetica Multiparametrica (mpMRI). Studi recenti suggeriscono che l'uso della MRI come strumento di triage possa mitigare i problemi di sovradiagnosi, fungendo da filtro efficace per selezionare i pazienti che necessitano realmente di biopsia [3, 4, 5]. A differenza delle metodiche tradizionali, la mpMRI combina informazioni anatomiche e funzionali, offrendo una visione più dettagliata dell'aggressività biologica della lesione.

1.3 Risonanza Magnetica Multiparametrica

La Risonanza Magnetica Multiparametrica (mpMRI) rappresenta l'evoluzione funzionale dell'imaging morfologico convenzionale. Il termine "multiparametrico" definisce la capacità della metodica di campionare diverse proprietà biofisiche del tessuto prostatico all'interno della stessa sessione diagnostica, integrando l'alta risoluzione spaziale delle sequenze anatomiche (T2-weighted) con le informazioni funzionali sulla mobilità delle molecole d'acqua. Quest'ultima viene indagata tramite le sequenze pesate in diffusione (DWI) e quantificata oggettivamente mediante le mappe del Coefficiente di Diffusione Apparente (ADC), che forniscono un parametro numerico diretto della cellularità tissutale [6].

Questa sinergia tra anatomia e dati quantitativi ha permesso di superare la bassa specificità del PSA e i limiti di campionamento della biopsia, elevando la mpMRI a strumento di triage fondamentale per la rilevazione delle neoplasie clinicamente significative (Gleason Score ≥ 7) [7].

In accordo con le linee guida *PI-RADS v2.1* [6], l'esame multiparametrico si basa sull'integrazione di tre pesature fondamentali, ciascuna destinata a fornire informazioni specifiche sullo stadio della malattia.

Analisi Morfologica: T2-weighted Imaging (T2W)

Le sequenze pesate in T2 rappresentano il riferimento anatomico per lo studio della prostata (Figura 1.1). Il contrasto di queste immagini dipende dal contenuto d'acqua dei tessuti e dalla struttura ghiandolare. In condizioni fisiologiche, la Zona Periferica (PZ) appare marcatamente iperintensa grazie all'elevato contenuto di fluido nei dotti acinari [8]. La presenza di un adenocarcinoma altera questa architettura: la proliferazione cellulare sostituisce gli spazi ricchi di fluido con tessuto solido e stroma compatto, determinando una caduta del segnale che appare come un'area focale ipointensa. La sequenza T2W è la determinante principale per la valutazione delle lesioni nella Zona di Transizione (TZ), dove l'eterogeneità dovuta alla BPH rende l'analisi più complessa [8, 9].

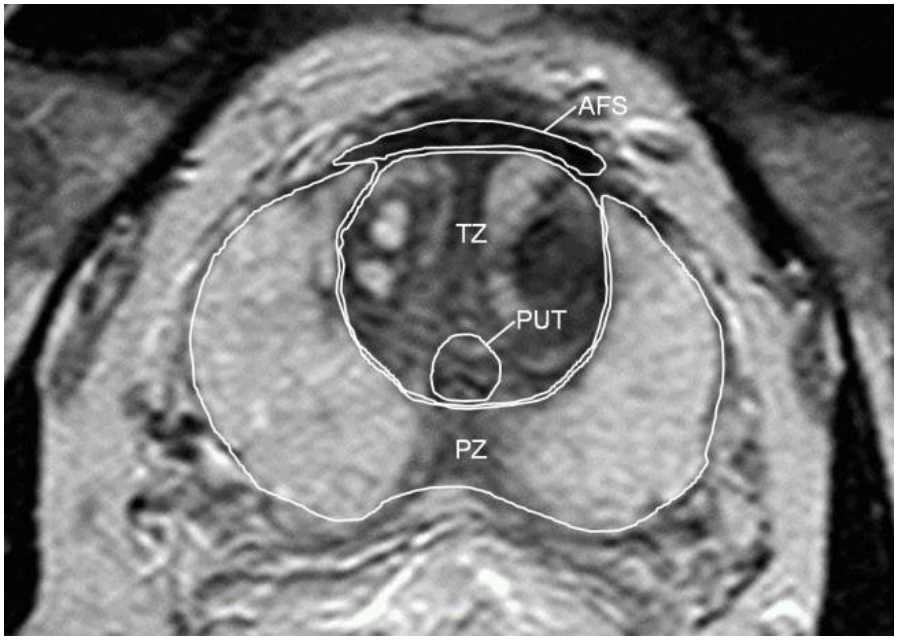


Figura 1.1: Rappresentazione schematica dell'anatomia zonale prostatica in T2-weighted. La Zona Periferica (PZ) iperintensa circonda la Zona di Transizione (TZ), sede tipica dell'iperplasia benigna. Immagine adattata da Busè et al. [8].

Analisi Funzionale: Diffusione (DWI) e Mappe ADC

La *Diffusion Weighted Imaging* (DWI) è la sequenza cardine per la rilevazione dei tumori nella zona periferica. Essa valuta il movimento microscopico dei moti Browniani delle molecole d'acqua: nei tessuti tumorali, l'alta densità cellulare riduce lo spazio extracellulare, limitando la libertà di movimento dell'acqua ("restrizione della diffusione"). Visivamente, questo fenomeno si traduce in un segnale elevato nelle immagini acquisite ad alti valori di b ($b \geq 1400\, s/mm^2$). Per oggettivare questa osservazione e rimuovere potenziali artefatti, si generano le mappe del Coefficiente di Diffusione Apparente (ADC). In queste mappe quantitative, il tumore appare come un'area scura (ipointensa). È ampiamente dimostrata una correlazione inversa tra il valore numerico dell'ADC e l'aggressività del tumore; infatti, tessuti più densi e aggressivi mostrano valori di ADC più bassi (Figura 1.2) [10].

Valutazione Vascolare: Dynamic Contrast Enhanced (DCE)

La sequenza DCE studia la neoangiogenesi tumorale attraverso l'iniezione di mezzo di contrasto come il Gadolinio. I vasi tumorali, essendo immaturi e permeabili, mostrano tipicamente una cinetica di impregnazione rapida e precoce seguita da un veloce lavaggio, a differenza del tessuto sano che si impregna più lentamente. Sebbene il ruolo della DCE sia stato ridimensionato nell'ultima versione del PI-RADS divenendo secondario rispetto alla DWI nella zona periferica, essa mantiene una funzione dirimente nei casi dubbi, aiutando a discriminare le lesioni equivoche e aumentando la sensibilità complessiva dell'esame [7]. Nello studio qui trattato, queste sequenze non vengono prese in considerazione per via della complessità aggiuntiva data dall'iniezione del mezzo di contrasto introdotta nell'esame.

1.4 Standardizzazione e Limiti: Il Sistema PI-RADS

Prima dell'introduzione di standard condivisi, l'interpretazione della RM prostatica soffriva di una marcata soggettività, con una scarsa riproducibilità tra diversi centri e radiologi. Per colmare questa lacuna, nel 2012 l'*European Society of Urogenital Radiology* (ESUR) ha sviluppato il Prostate Imaging – Reporting and Data System (PI-RADS v1) [11], attualmente giunto alla versione 2.1 (2019) [6].

Il PI-RADS assegna un punteggio su una scala Likert da 1 a 5 per identificare le lesioni clinicamente significative, definendo quali pazienti necessitino di una biopsia mirata e quali possano proseguire con il monitoraggio attivo. Il sistema, dunque, assegna un punteggio crescente in base alla probabilità che una determinata lesione presenti un tumore significativo.

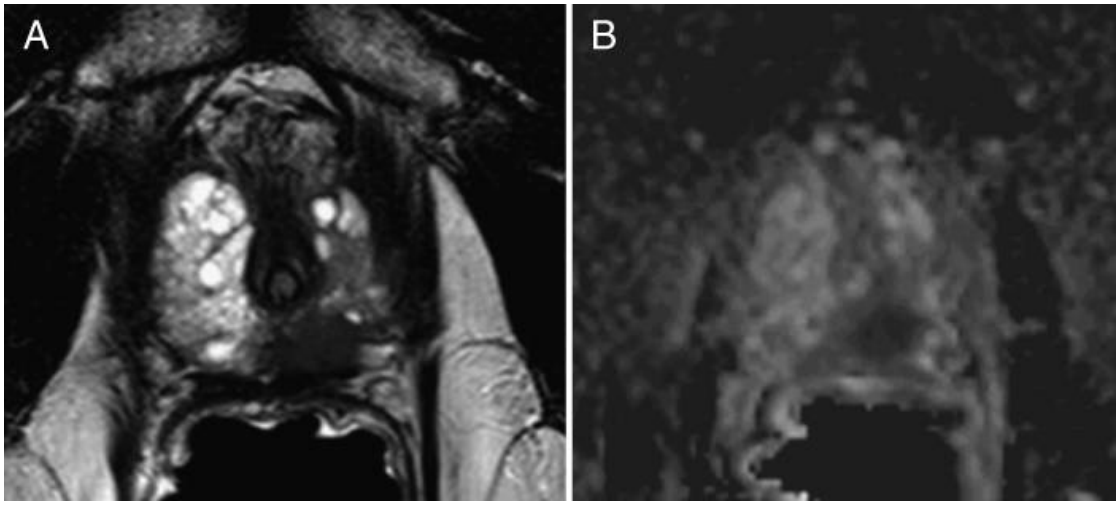


Figura 1.2: Manifestazione tipica di un adenocarcinoma prostatico in RM multiparametrica. A sinistra, la sequenza T2-weighted mostra una lesione ipointensa circoscritta nella zona periferica sinistra. A destra, la corrispondente mappa ADC evidenzia una marcata restrizione della diffusione (segnale scuro) in corrispondenza della lesione, confermando la natura ad alta cellularità del tessuto. Immagine adattata da Padhani et al. [10].

PI-RADS Versione 2.1 (2019)

L'aspetto più tecnico e rilevante del PI-RADS v2.1 è l'abbandono di una valutazione globale generica in favore di un approccio zona-specifico [6]. Il sistema definisce una sequenza dominante diversa a seconda della localizzazione anatomica della lesione, riconoscendo che le diverse sequenze hanno sensibilità diverse nelle varie zone della ghiandola.

Nella Zona Periferica la sequenza dominante è la DWI/ADC. Un'area focale di restrizione definisce il punteggio base mentre la sequenza DCE interviene solo come supporto per dirimere i casi dubbi con valore PI-RADS pari a 3, elevando il punteggio a 4 se positiva. Al contrario nella Zona di Transizione la sequenza dominante è la T2W. Qui, l'eterogeneità causata dall'iperplasia benigna rende la diffusione meno specifica. La DWI viene usata come supporto per analizzare noduli sospetti già individuati morfologicamente.

Sebbene il PI-RADS abbia standardizzato il linguaggio radiologico, la sua affidabilità crolla se le immagini di input non rispettano precisi standard tecnici. Un'immagine affetta da artefatti da movimento o distorsioni geometriche può simulare una lesione (Falso Positivo) o nasconderne una reale (Falso Negativo), rendendo inapplicabile l'algoritmo di scoring. Di conseguenza, una qualità diagnostica adeguata è un prerequisito importante per utilizzare il PI-RADS; è in questo

contesto che si inserisce la necessità di uno strumento dedicato alla valutazione tecnica: il PI-QUAL.

1.5 Il Problema della Qualità dell'Immagine: PI-QUAL

Come discusso nella sezione precedente, l'affidabilità del sistema PI-RADS è intrinsecamente dipendente dalla qualità tecnica delle immagini di input. Artefatti da movimento, distorsioni geometriche nella DWI o un basso rapporto segnale-rumore (SNR) possono compromettere l'interpretazione diagnostica, portando a falsi negativi o a biopsie inutili [12]. Per colmare il vuoto normativo lasciato dal PI-RADS che richiedeva qualità adeguata senza definirla metricamente, il *Precision Prostate Project* ha introdotto il *Prostate Imaging Quality (PI-QUAL) Score*, il primo sistema standardizzato per la valutazione della qualità diagnostica della mpMRI [13].

A differenza del PI-RADS, il PI-QUAL valuta l'idoneità tecnica dell'esame con l'obiettivo di fornire al radiologo uno strumento oggettivo per decidere se una scansione può essere refertata con confidenza o se deve essere ripetuta. In letteratura è stata dimostrata una forte correlazione tra il punteggio PI-QUAL (v1) e la performance diagnostica, dove la qualità dell'immagine ha una ricaduta diretta sulla rilevazione del tumore e sulla pianificazione della biopsia [12, 14]. Di seguito sono riportati i dettagli delle due versioni di PI-QUAL esistenti e le loro differenze.

PI-QUAL Versione 1 (2020)

La prima versione proponeva una scala a 5 punti basata sulla valutazione rigorosa di 34 criteri tecnici specifici per T2W, DWI e DCE [13]. Sebbene molto dettagliata, questa versione presentava un'eccessiva complessità che richiedeva troppo tempo per integrare il sistema nel flusso di lavoro clinico. Inoltre, il sistema penalizzava drasticamente gli esami in cui mancava una sequenza, assegnando automaticamente un punteggio basso anche se le sequenze T2W e DWI erano diagnostiche, caso tipico dei protocolli biparametrici. Per questi motivi il sistema è stato ripensato e aggiornato alla versione due.

PI-QUAL Versione 2 (2024)

La versione aggiornata, utilizzata come riferimento concettuale in questo studio, introduce una semplificazione radicale. Abbandonando la scala a 5 punti, il PI-QUAL v2 adotta un approccio funzionale a 3 categorie, focalizzandosi sulle sequenze dominanti T2W e DWI. La logica segue un modello a "semaforo" per guidare l'azione clinica (Figura 1.3):

- **PI-QUAL 1 (Non Diagnostico - Rosso):** La qualità è insufficiente. Una o più sequenze chiave sono gravemente compromesse da artefatti che impediscono l'analisi anatomica o funzionale. L'esame deve essere ripetuto.
- **PI-QUAL 2 (Diagnostico Sub-ottimale - Giallo):** Sono presenti artefatti, ma non tali da impedire l'interpretazione poiché le strutture chiave e le eventuali lesioni sono visibili, sebbene con un livello di confidenza ridotto. L'esame può essere refertato da un radiologo esperto.
- **PI-QUAL 3 (Ottimale - Verde):** Tutte le sequenze rispettano i criteri di eccellenza (FOV corretto, assenza di movimento, alta risoluzione) e dunque l'esame è refertabile con massima confidenza.

Questa nuova classificazione è particolarmente rilevante per lo sviluppo di sistemi di Intelligenza Artificiale, in quanto semplifica il problema da una regressione complessa a un task di classificazione ternaria o binaria, più robusto da modellare computazionalmente.

1.6 Stato dell'Arte e Motivazione dello Studio

L'introduzione del PI-QUAL ha fornito un linguaggio comune per la valutazione della qualità, ma la sua applicazione rimane manuale, soggettiva e dispendiosa in termini di tempo. Un primo tentativo di mitigare queste problematiche è stato proposto da Giganti et al. (2021), i quali hanno sviluppato un software di valutazione semi-automatizzato dedicato al PI-QUAL (Figura 1.4). Sebbene lo studio abbia dimostrato che un supporto guidato può migliorare l'accordo interlettore e velocizzare la compilazione della scheda di valutazione, l'approccio rimane intrinsecamente dipendente dall'input dell'operatore umano [15]. Di conseguenza, negli ultimi anni, la ricerca si è concentrata sullo sviluppo di sistemi di *Computer-Aided Quality Assessment* (CA-QA) completamente automatici, sfruttando due paradigmi principali: l'analisi delle texture Radiomica e il Deep Learning.

I primi tentativi di automatizzare il controllo qualità si sono basati sull'estrazione di metriche fisiche convenzionali (SNR, CNR). Tuttavia, come evidenziato nelle review sulla RM quantitativa prostatica, queste misure globali sono spesso insufficienti per garantire l'affidabilità diagnostica richiesta dai protocolli avanzati [16, 17]. Successivamente, l'attenzione si è spostata verso la Radiomica, con l'obiettivo di estrarre descrittori statistici avanzati capaci di quantificare l'eterogeneità e la disorganizzazione dell'immagine causata dagli artefatti. Studi recenti hanno proposto l'utilizzo di metriche *No-Reference*, come l'algoritmo BRISQUE (*Blind/Referenceless Image Spatial Quality Evaluator*), combinato con feature di Haralick. Questi approcci hanno mostrato risultati promettenti nella classificazione binaria (Accettabile e Non Accettabile), ma faticano a catturare la complessità

PI-QUAL v2 scoring sheet

MRI without intravenous contrast medium

T2-WI	DWI	PI-QUAL score	Remarks	General clinical implication
≤ 2	≤ 2	1	-	Inadequate scan: scan should be repeated
3 or 4	3 or 4	2	No: ≤ 2 /4 for T2-WI and DWI	Acceptable scan: consider repeat scan
4	4	3	Full scores for T2-WI and DWI	Optimal scan: scan of optimal diagnostic quality

Multiparametric MRI

T2-WI	DWI	DCE	PI-QUAL score	Remarks	General clinical implication
≤ 2	≤ 2	+	1	-	Inadequate scan: scan should be repeated
3 or 4	3 or 4		2	No: ≤ 2 /4 for T2-WI and DWI	Acceptable scan: consider repeat scan
4	4	-	3	Full scores for T2-WI and DWI	Optimal scan: scan of optimal diagnostic quality

'+' : **both** criteria for DCE are satisfied and at least one sequence (either T2-WI or DWI) must score **4/4**

'-' : either **only one** criterion or **no criteria** for DCE are satisfied

Please (✓) if present:

T2-WI

Essential requirement before proceeding *(equals 0/4 if not met):* **Slice thickness: 3 mm**	
Axial T2-WI: adequate signal-to-noise ratio (SNR) in all parts of the images	
Axial T2-WI: ability to clearly delineate relevant structures in the prostate	
Axial T2-WI: absence of significant artefacts in the prostatic region	
Sagittal OR coronal: adequate SNR and image resolution AND absence of significant artefacts	
Total score for T2-WI	**/ 4**

DWI

Essential requirement before proceeding *(equals 0/4 if not met):* **Slice thickness: ≤ 4 mm**, **High *b* value sequence (≥ 1,400 s/mm²)**, **ADC map using at least two *b* values up to 1,000 s/mm²**	
Adequate contrast and SNR on high *b* value images	
Adequate range of contrast to differentiate TZ/BPH from PZ on the ADC maps	
Absence of significant artefacts in the prostatic region	
Anatomical matching of the ADC map / high *b* value sequence to the axial T2-WI	
Total score for DWI	**/ 4**

DCE

Essential requirement before proceeding *(equals '–' if not met):* **Slice thickness: 3 mm**, **Temporal resolution: ≤ 15 seconds**, **Fat saturation** *(or include post-processing, e.g. subtraction / heat maps)*	
Absence of significant artefacts in the prostatic region and appropriate bolus enhancement	
Ability to identify anatomical structures (e.g. capsular vessels or pudendal artery)	
Total score for DCE *('+' only when **both** criteria are met)*	**+ / -**

PI-QUAL score ☐ **1** ☐ **2** ☐ **3**

Figura 1.3: Scoring Sheet del sistema PI-QUAL v2. La scala ridotta a 3 punti (1: Insufficiente, 2: Sufficiente, 3: Ottimale) facilita l'adozione clinica e pone l'enfasi sulla qualità delle sequenze dominanti T2W e DWI, rendendo il sistema applicabile anche ai protocolli biparametrici. Immagina adattata da De Rooij et al. [12].

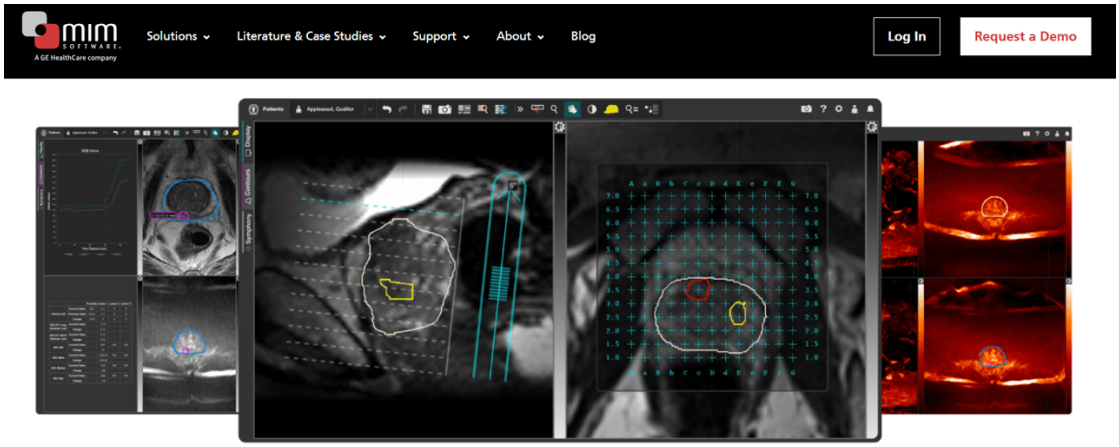


Figura 1.4: Interfaccia del software MIM® Symphony Dx v. 7.1.2. Il sistema di valutazione assistita proposto da Giganti et al. guida il radiologo attraverso i criteri del PI-QUAL, ma la valutazione di ogni singolo item richiede ancora l'input manuale dell'operatore, non risolvendo il problema della soggettività e del carico di lavoro. Immagine adattata da Giganti et al. [15].

semantica dei criteri anatomici del PI-QUAL v2 come la visibilità della capsula o dei fasci neuro-vascolari [18].

Con l'aumento della disponibilità computazionale, le Reti Neurali Convoluzionali (CNN) sono diventate il gold standard per la CA-QA, grazie alla loro capacità di apprendere feature gerarchiche direttamente dai dati grezzi. Uno dei primi lavori significativi è stato condotto da Esses et al. (2018), i quali hanno addestrato una CNN 2D per valutare la presenza di artefatti da movimento in risonanza epatica, dimostrando un'accuratezza paragonabile a quella di due radiologi esperti [17]. Sebbene non specifico per la prostata, questo studio ha validato il concetto che le CNN sono sensibili alla degradazione dei bordi e al *blurring* tipico del movimento. Più recentemente, la ricerca si è focalizzata specificamente sul dominio prostatico, adottando architetture 3D per sfruttare la natura volumetrica del dato MRI.

- **Belue et al. (2024):** Hanno proposto l'uso di una DenseNet-121 3D addestrata su un dataset multicentrico di oltre 1.000 pazienti. Il loro modello ha raggiunto un'accuratezza dell'87.9% nella classificazione binaria della qualità T2W, con un accordo inter-osservatore "sostanziale" ($\kappa = 0.70$) rispetto agli esperti umani [19].

- **Gloe et al. (2025):** Nel lavoro più recente e vasto ad oggi disponibile (1.412 esami), gli autori hanno sviluppato un sistema basato su DenseNet-169 per ridurre il tasso di richiamo (*recall*) dei pazienti. Il loro modello ha ottenuto

un'Area Sotto la Curva (AUC) di 0.88 nel distinguere scansioni diagnostiche da quelle non diagnostiche. Un risultato chiave di questo studio è la dimostrazione che l'IA tende a essere più "severa" degli umani, identificando artefatti sottili che potrebbero sfuggire a una prima revisione visiva [20].

Nonostante i promettenti risultati riportati in letteratura, lo stato dell'arte attuale rivela criticità strutturali che motivano la presente indagine. In primo luogo, i modelli di Deep Learning attualmente proposti tendono a by-passare la rigorosa griglia di valutazione del PI-QUAL. Le classificazioni prodotte in output, infatti, non riflettono fedelmente i singoli criteri dello score, ma si riducono a una stima della qualità globale dell'immagine. Questo approccio rende la valutazione automatica difficilmente standardizzabile e ne compromette la reale adattabilità clinica, poiché il radiologo riceve un responso slegato dalle linee guida internazionali di riferimento.

In secondo luogo, persiste il dilemma dell'interpretabilità. Mentre le CNN agiscono come "scatole nere" ad alta accuratezza ma scarsa trasparenza decisionale, l'approccio radiomico offre il vantaggio di estrarre *feature* fisicamente spiegabili, un requisito fondamentale per l'accettazione clinica di questi sistemi. La maggior parte delle architetture Deep Learning, come quella presentata da Belue et al. (2024), viene addestrata in modalità *End-to-End* utilizzando direttamente il punteggio di qualità globale come target. Nessuno di questi modelli verifica esplicitamente il rispetto dei singoli criteri definiti dalle linee guida, come la visibilità della capsula prostatica o dei fasci neuro-vascolari; di conseguenza, il Deep Learning fornisce un voto finale, ma fallisce nello spiegare quale specifico limite tecnico abbia compromesso l'esame.

Ad oggi, inoltre, mancano in letteratura studi che confrontino direttamente l'efficacia relativa di queste due metodologie — Machine Learning su feature radiomiche e Deep Learning *End-to-End* — applicate al medesimo dataset multicentrico per il task specifico del PI-QUAL v2. Negli studi citati risulta spesso carente una valutazione che vada oltre la semplice Accuratezza globale, trascurando l'analisi del bilanciamento tra Sensibilità e Specificità, che rappresenta invece un parametro cruciale per evitare falsi allarmi e scongiurare inutili prolungamenti dei tempi di scansione. L'obiettivo primario di questo lavoro è colmare tali lacune, valutando sperimentalmente quale strategia risulti più robusta, interpretabile e affidabile in un contesto clinico reale, caratterizzato da sfide operative quali il forte sbilanciamento delle classi e la presenza di variabilità inter-osservatore.

Capitolo 2

Pre-elaborazione dei Dati e Feature Extraction Radiomica

2.1 Dataset e Selezione della Coorte

I dati analizzati in questo studio costituiscono un sottoinsieme del database ProstateNet, una raccolta di immagini di Risonanza Magnetica multiparametrica acquisite nell'ambito del progetto europeo ProCAncer-I [21]. Provenendo da 12 diversi centri clinici europei, l'archivio garantisce una notevole eterogeneità nei protocolli e nella strumentazione.

Questa varietà rispecchia fedelmente la realtà clinica quotidiana, includendo esami condotti con scanner dei principali produttori mondiali (Siemens, Philips, GE) [21]. Sebbene tale diversità rappresenti una sfida per l'analisi automatica, essa è cruciale per sviluppare algoritmi che siano al contempo robusti e generalizzabili.

Per valutare la qualità e la possibile valutazione diagnostica degli esami mpMRI, è stato isolato un campione rappresentativo pari a circa il 6% dell'intero archivio, identificando una coorte iniziale di 1.050 esami. I criteri di inclusione scelti per il dataset richiedono la disponibilità della triade diagnostica composta da immagini pesate in T2, mappe ADC e immagini pesate in diffusione DWI.

Un ulteriore criterio di inclusione è stato il non utilizzo della bobina endorettale (ERC). Questa scelta mira a preservare l'omogeneità del dataset poichè l'ERC induce significative deformazioni morfologiche della prostata per compressione e altera le caratteristiche del segnale. Studi recenti su questo stesso database hanno evidenziato come le acquisizioni con ERC generino una distribuzione delle feature radiomiche drasticamente diversa rispetto alle bobine di superficie, causando un

domain shift che rischierebbe di compromettere le prestazioni dei modelli di IA [21]. Inoltre, diverse evidenze suggeriscono che, con le moderne macchine a 3 Tesla, l'uso dell'ERC non apporti sempre un beneficio diagnostico tale da giustificarne l'impiego [22, 23].

Grazie allo screening sui 1.050 pazienti iniziali, sono stati esclusi 27 esami. Nel dettaglio, 10 casi appartenevano a un campionamento casuale per uno studio preliminare; un paziente presentava un catetere in situ che oscurava la visione; 17 esami presentavano errori tecnici di conversione, nello specifico, 9 privi di sequenza DWI, 4 con DWI non convertibile e 3 con T2w non convertibile.

In conclusione, il dataset finale ammesso alla valutazione della qualità comprende 1.023 esami.

2.2 Metodologia di Valutazione della Qualità

La valutazione della diagnosticità è stata condotta seguendo i criteri PI-QUAL e le linee guida PI-RADS, ed è stata affidata ad un gruppo di 13 radiologi provenienti da 10 centri clinici differenti. Al fine di bilanciare le competenze, il gruppo è stato suddiviso in base all'esperienza: 6 radiologi esperti (*Reader 1*) e 7 radiologi non esperti (*Reader 2*).

Il protocollo operativo di valutazione della qualità prevede una prima fase di valutazione indipendente, seguita, in caso di convergenza, dalla conferma del voto. Le eventuali discordanze vengono invece risolte da un terzo lettore esperto (*Adjudicator*), chiamato a validare il giudizio di uno dei due lettori iniziali. Nonostante l'ampiezza della coorte iniziale di 1.023 casi, il ritiro di uno dei medici durante lo studio ha imposto una riduzione del campione; di conseguenza, il dataset finale comprende 853 esami MRI.

Analizzando la concordanza tra i radiologi, 523 esami sono stati giudicati unanimemente come "Diagnostici" e 46 come "Non Diagnostici". I restanti 284 casi, risultati discordanti, hanno richiesto l'intervento dirimente dell'*Adjudicator*.

A seguito del processo di revisione e aggiudicazione, la distribuzione qualitativa finale evidenzia una netta prevalenza di dati diagnostici. L'82% del dataset è stato classificato come diagnostico, mentre il restante 18% è risultato non diagnostico.

2.3 Strutturazione del Dataset per l'Analisi Computazionale

A valle della selezione clinica, la fase operativa si è concentrata sulla trasformazione dell'archivio grezzo in una struttura dati coerente, pronta per l'estrazione automatica delle feature. Poiché le immagini non richiedevano pre-processing iniziale

per non modificare la qualità dell'imamgine e influenzare la valutazione, è stata implementata una pipeline di data curation in Python per mappare univocamente i volumi NIfTI. Per ciascuno degli 853 esami validati, lo script ha associato la triade di sequenze multiparametriche alla corrispondente maschera di segmentazione della prostata (T2w Mask), generando un registro strutturato dei percorsi file.

Contestualmente, le valutazioni qualitative descritte nella Sez. 2.2, sono state convertite in etichette computazionali. I giudizi finali sono stati codificati in un vettore binario, dove la classe 1 identifica i volumi valutati come "Diagnostici" e la classe 0 quelli "Non Diagnostici". Prima dell'elaborazione finale, l'intero dataset è stato sottoposto a una verifica di integrità tecnica tramite ispezione visiva, garantendo che nessun artefatto tecnico o corruzione del file inficiasse il calcolo delle feature descritto nella sezione successiva.

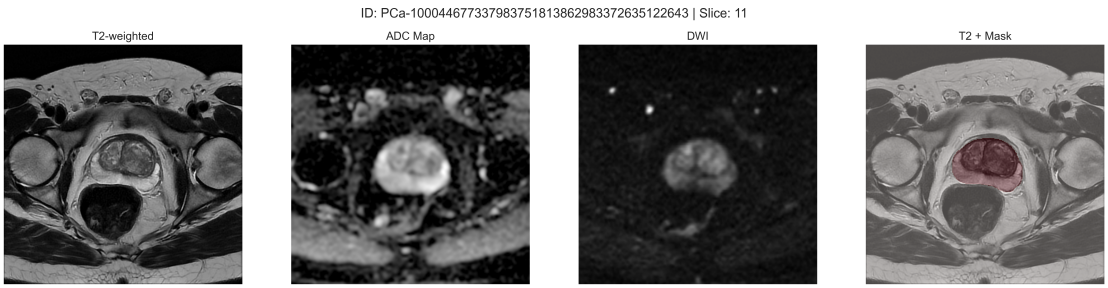


Figura 2.1: Rappresentazione di un caso tipo incluso nel dataset. Da sinistra a destra: sequenza anatomica T2-weighted, mappa ADC, immagini pesate in diffusione DWI e visualizzazione della maschera di segmentazione prostatica sovrapposta all'anatomia T2w. La coerenza spaziale tra le sequenze è stata verificata tramite ispezione visiva.

2.4 Estrazione delle Feature Radiomiche

Il processo di quantificazione delle immagini è stato eseguito utilizzando la libreria open-source PyRadiomics (v3.0.1) [24], operando in conformità con gli standard definiti dall'IBSI (*Image Biomarker Standardization Initiative*) [25]. Al fine di ottenere una caratterizzazione completa, che tenesse conto sia della qualità globale dell'acquisizione sia delle specificità testurali dell'organo bersaglio, è stata implementata una duplice strategia di estrazione, articolata in due pipeline distinte:

1. **Approccio Whole-Volume:** Mirato a valutare l'intero volume di acquisizione per rilevare rumore di fondo e artefatti.
2. **Approccio ROI-Based:** Focalizzato esclusivamente sulla regione anatomica della prostata e basato su maschera di segmentazione.

2.4.1 Definizione dei volumi di analisi: Approccio Whole-Volume e ROI-Based

Poiché l'algoritmo di estrazione impone l'uso di una maschera binaria per delimitare la regione di calcolo, è stata implementata una procedura automatica per generare una maschera sintetica integrale; per ogni volume NIfTI in ingresso, lo script di elaborazione ha generato una matrice di dimensioni identiche all'immagine originale contenente esclusivamente valori pari a 1. Questo approccio (*Whole-Volume Masking*) ha forzato l'algoritmo a calcolare le statistiche su tutti i voxel disponibili della matrice tridimensionale, permettendo una valutazione della qualità dell'immagine nella sua interezza.

Parallelamente l'estrazione è stata condotta utilizzando le maschere di segmentazione fornite nel dataset per caratterizzare la texture specifica del tessuto prostatico. In questa configurazione, il calcolo delle feature è stato ristretto esclusivamente ai voxel compresi nella ROI della prostata, per valutare, ad esempio, se le sequenze siano effettivamente in grado di mostrare i dettagli cruciali per la refertazione, come la distinzione tra zona periferica e zona di transizione [6].

2.4.2 Configurazione dei Parametri di Estrazione

La configurazione di PyRadiomics è stata definita per massimizzare la sensibilità alle variazioni di texture e intensità indicative della qualità dell'immagine. In accordo con il file di configurazione utilizzato (*.yaml*), sono stati applicati i seguenti parametri:

- **Spaziatura e Risoluzione:** È stata mantenuta la risoluzione spaziale nativa delle immagini, disattivando il ricampionamento (`resampledPixelSpacing: null` e `interpolator: sitkNearestNeighbor`). Questa scelta è fondamentale per evitare che l'interpolazione alteri artificialmente la grana del rumore o la nitidezza dei bordi originali.
- **Normalizzazione:** È stata applicata una normalizzazione Z-score con un fattore di scala pari a 100, per rendere confrontabili le intensità di segnale tra scanner di produttori differenti.
- **Discretizzazione:** Si è scelto di utilizzare, per discretizzare i dati, un numero fisso di bin pari a 64, in modo tale da garantire una quantizzazione omogenea dei livelli di grigio. Tale valore è stato selezionato a seguito di un'analisi preliminare sulla distribuzione delle intensità del dataset che presentava un range medio pari a circa 1850. Un numero di bin inferiore avrebbe comportato una perdita di risoluzione testurale, mentre una discretizzazione più fine avrebbe generato matrici eccessivamente sparse, aumentando la sensibilità al rumore.

- **Resegmentazione:** Al fine di mitigare l'impatto di artefatti radiologici e rumore impulsivo, è stata applicata una procedura di resegmentazione basata sui percentili utilizzando un intervallo relativo $[0.01, 0.90]$.

2.4.3 Filtri Applicati e Classi di Feature

Per approfondire l'indagine oltre ld immagini originali, la pipeline ha integrato l'applicazione di filtri spaziali e frequenziali mirati a isolare specifici pattern di tessitura. Nello specifico, è stata eseguita una decomposizione Wavelet sulle tre dimensioni spaziali (H=High-Pass, L=Low-Pass), generando le otto possibili combinazioni di frequenza (es. LLL, LLH, HHL) per separare il rumore ad alta frequenza dalle disomogeneità a bassa frequenza. Parallelamente, sono stati applicati il filtro Gradiente, per quantificare la magnitudo delle variazioni di intensità, e il filtro *Laplacian of Gaussian* (LoG) con sigma $\sigma = [2.0, 3.0]$ mm, funzionale alla rilevazione di dettagli fini e grossolani.

La Figura 2.2 illustra l'applicazione della decomposizione Wavelet su una caso del dataset, mentre la Figura 2.3 riporta il risultato dell'applicazione dei filtri Gradiente e LoG.

Sulle immagini originali e filtrate sono state estratte le feature statistiche del Primo Ordine, descrittive della distribuzione globale delle intensità dei voxel.

Congiuntamente, per quantificare l'eterogeneità spaziale del tessuto, sono state calcolate le feature testurali di Secondo Ordine derivate da quattro matrici standard, schematizzate in Figura 2.4: la *Gray Level Co-occurrence Matrix* (GLCM) [26], per la valutazione di contrasto e omogeneità locale; la *Gray Level Run Length Matrix* (GLRLM) [27], indicativa della granularità direzionale; la *Gray Level Size Zone Matrix* (GLSZM) [28], per il rilevamento di zone a intensità uniforme; e la *Gray Level Dependence Matrix* (GLDM) [29], relativa alla complessità delle dipendenze nel vicinato. Un riepilogo dettagliato delle classi di feature estratte e della loro numerosità è riportato in Tabella 2.1.

Considerando l'applicazione dei filtri Wavelet a 8 canali, Gradiente e LoG con due valori di σ su ciascuna delle 3 modalità, lo spazio delle feature per ogni paziente risulta ad alta dimensionalità. Nello specifico, per ogni singola modalità vengono calcolate 1.056 feature (88 feature base $\times$ 12 mappe filtrate), portando il vettore descrittivo finale del paziente a un totale di 3.168 variabili. Questa elevata complessità computazionale suggerisce l'utilizzo di tecniche di selezione delle feature e di riduzione della dimensionalità che saranno discusse nel capitolo 3.

2.4.4 Implementazione Computazionale e Parallelizzazione

Per poter processare l'intero volume tridimensionale di tutti i pazienti presenti nel dataset, per tre modalità diverse, il carico computazionale risulta elevato.

Tabella 2.1: Riepilogo delle feature radiomiche estratte per singola immagine. Il totale effettivo per paziente è dato dalla moltiplicazione di questo set base per il numero di immagini derivate (Originale, Wavelet, LoG, Gradiente) e per le modalità (T2w, ADC, DWI).

Classe di Feature	Descrizione	N. Features
First Order	Statistiche di distribuzione (istogramma)	18
GLCM	Gray Level Co-occurrence Matrix	24
GLRLM	Gray Level Run Length Matrix	16
GLSZM	Gray Level Size Zone Matrix	16
GLDM	Gray Level Dependence Matrix	14
Totale per Immagine		**88**

Per ottimizzare i tempi di esecuzione, lo script di estrazione è stato sviluppato in Python sfruttando un'architettura di calcolo parallelo basata sulla libreria `concurrent.futures`. Utilizzando un `ProcessPoolExecutor` configurato con 10 worker simultanei, il sistema ha elaborato i volumi in parallelo, serializzando progressivamente i risultati. L'output finale è stato consolidato in un dataset strutturato contenente le feature estratte per tutte le modalità, pronto per le successive analisi statistiche.

2.5 Analisi Statistica Esplorativa e Ottimizzazione del Dataset

2.5.1 Integrità dei Dati e Gestione dei Valori Mancanti

Preliminarmente alla fase di modellazione predittiva, la matrice delle feature estratte è stata sottoposta a una verifica di coerenza strutturale e numerica. Il primo step ha riguardato la verifica tecnica dell'estrazione. Nonostante la selezione clinica iniziale di 853 pazienti descritta nella Sez. 2.2, durante l'esecuzione della pipeline di Feature Extraction si sono registrate criticità legate alla gestione delle risorse hardware. Nello specifico, per 9 casi caratterizzati da volumi di acquisizione particolarmente estesi, l'algoritmo ha generato errori di allocazione della memoria (*Memory Allocation Error*), impedendo il completamento del calcolo delle feature testurali ad alta dimensionalità. Di conseguenza, il pool effettivo di pazienti

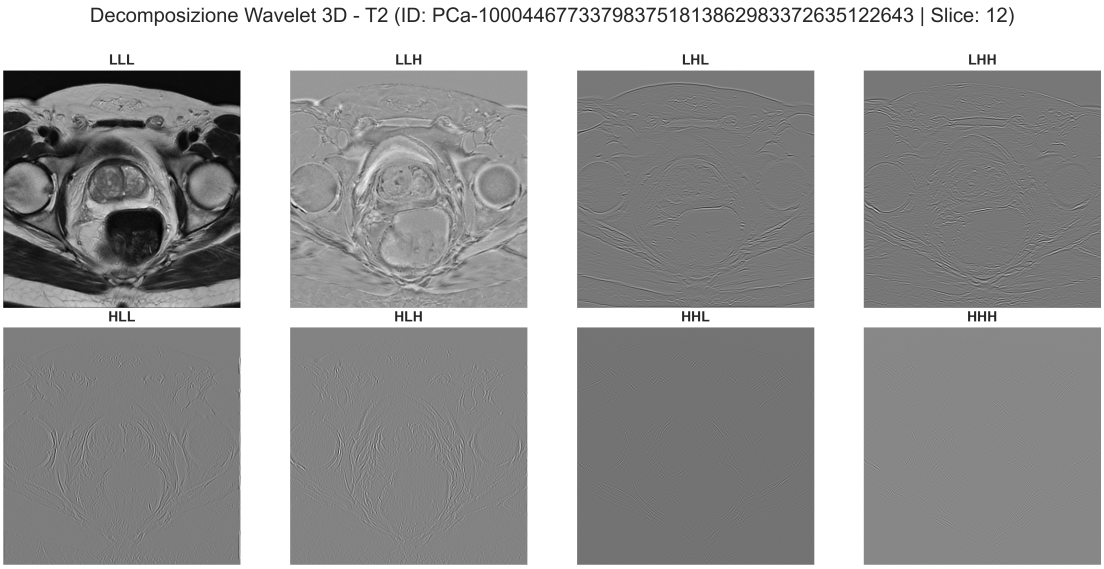


Figura 2.2: L'immagine originale viene decomposta attraverso filtri passa-basso (L) e passa-alto (H) lungo le tre direzioni spaziali (x, y, z), generando 8 sotto-bande. La combinazione LLL (Low-Low-Low) rappresenta l'approssimazione smussata dell'immagine originale, trattenendo le informazioni strutturali grossolane. Le altre combinazioni (es. HHL, HHH) catturano i dettagli ad alta frequenza, come il rumore, i bordi fini e le tessiture direzionali non visibili all'occhio umano.

correttamente processati si è assestato a 844 unità. Il dettaglio numerico dei passaggi di selezione e la composizione finale del dataset sono schematizzati in Tabella 2.2.

Per garantire l'omogeneità dell'analisi multiparametrica, si è reso necessario un allineamento delle coorti. Dei 844 casi iniziali, sono stati inclusi nello studio soltanto i pazienti per i quali fosse disponibile il set completo di sequenze e la maschera di segmentazione. A seguito di questa operazione di filtraggio si è ottenuto un dataset finale di 802 pazienti.

Successivamente, si è proceduto alla ricerca di valori mancanti o corrotti (*Missing Values*) all'interno dell'intera matrice dei dati. Su un totale di oltre 2,5 milioni di punti dati (802 pazienti $\times$ 3.168 feature), sono state rilevate solamente due istanze di valori *Not a Number* (NaN), imputabili a sporadiche instabilità numeriche nel calcolo delle feature trasformate Wavelet sull'immagine T2w (`wavelet-HHH-skewness`, `wavelet-LHL-energy`)

Data la natura puntuale e la marginalità statistica dell'evento, è stata applicata una strategia di imputazione conservativa (*Mean Imputation*): i valori mancanti sono stati sostituiti con la media aritmetica della rispettiva feature calcolata sull'intera popolazione. Tale approccio ha permesso di completare la matrice senza

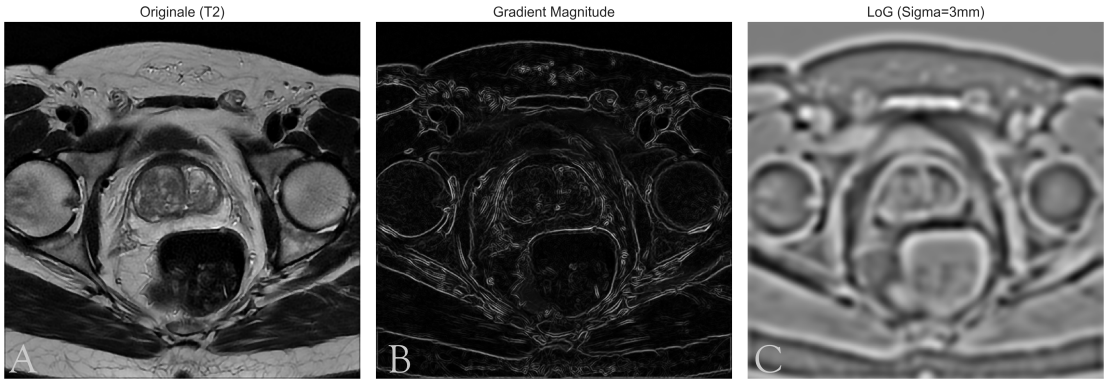


Figura 2.3: Visualizzazione dell'applicazione dei filtri spaziali sull'immagine T2-weighted. (A) Immagine originale assiale della prostata. (B) Magnitudine del Gradiente (*Gradient Magnitude*), che enfatizza i bordi e le variazioni improvvise di intensità, utile per visualizzare i bordi dei componenti. (C) *Laplacian of Gaussian* (LoG) con $\sigma = 3mm$. Questo filtro agisce come un rilevatore di "blob", evidenziando regioni di intensità omogenea a una specifica scala spaziale, permettendo di analizzare la tessitura a diverse granulosità (fine, media, grossolana).

introdurre bias significativi nella distribuzione dei dati.

Tabella 2.2: Dettaglio della composizione del dataset al termine della pipeline di estrazione. Il numero finale di pazienti inclusi nello studio (intersezione) è determinato dalla disponibilità simultanea di tutte le tre sequenze diagnostiche e delle relative maschere valide.

Metrica	**Conteggio**
Pazienti totali processati	853
Feature estratte (per modalità)	1.056
Disponibilità per sequenza	
Volumi ADC elaborati	842
Volumi DWI elaborati	842
Volumi T2w elaborati	804
Dataset Finale (Intersezione T2 ∩ ADC ∩ DWI)	**802**

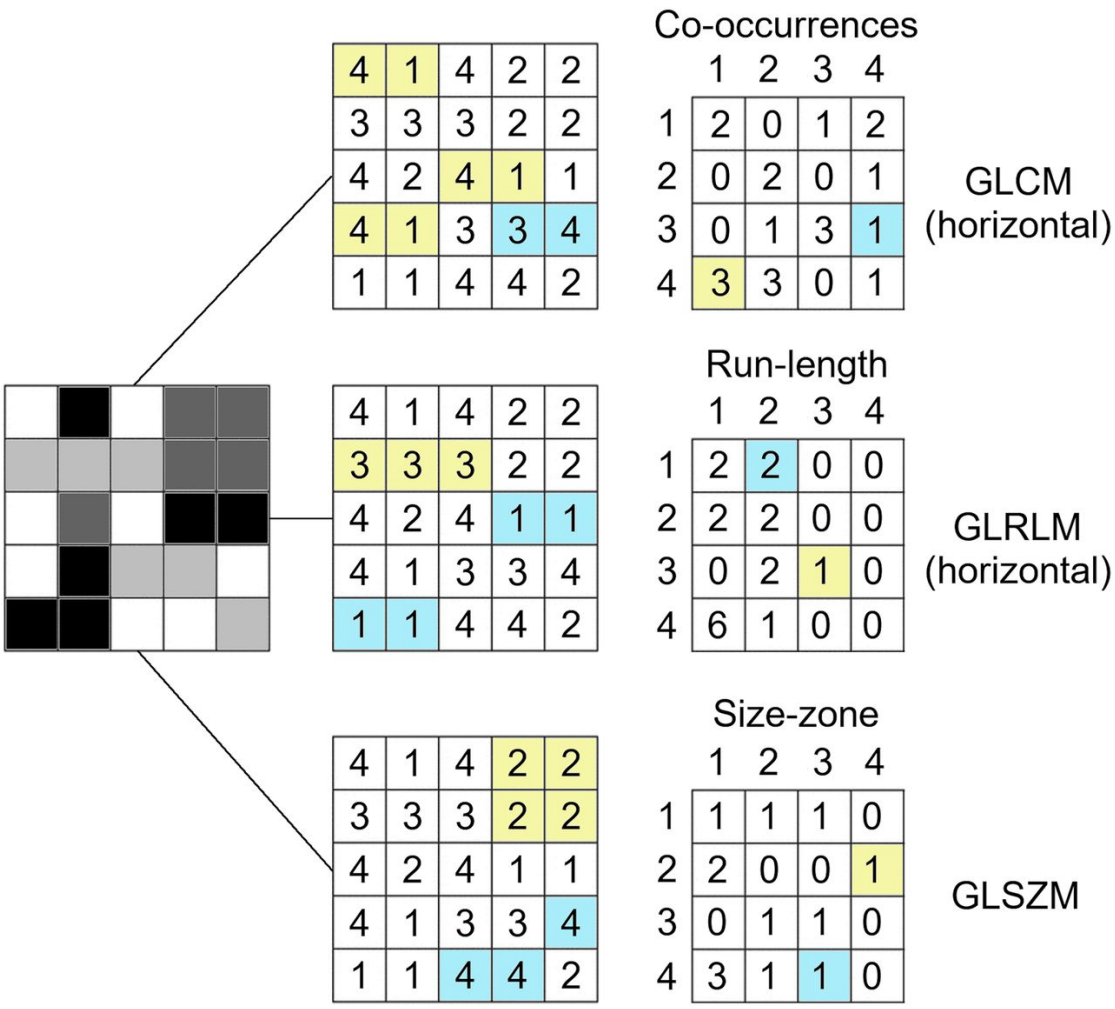


Figura 2.4: Esempio concettuale del calcolo delle matrici di texture di secondo ordine. A sinistra, una Regione di Interesse (ROI) discretizzata in 4 livelli di grigio. A destra, la trasformazione in matrici numeriche che quantificano le relazioni spaziali: (1) GLCM: conta le co-occorrenze di coppie di pixel adiacenti; (2) GLRLM: analizza la lunghezza delle sequenze consecutive di pixel identici; (3) GLSZM: mappa le zone connesse di pixel uniformi. Queste matrici costituiscono la base matematica per l'estrazione delle feature testurali. Immagine adattata da Mayerhoefer et al. [30].

2.5.2 Visualizzazione Globale: Overlay di Massa delle Densità

Per visualizzare l'intero spazio delle feature simultaneamente, viene generato un grafico di Overlay di Massa delle Densità (*Density Overlay*, Figura 2.5). In questa rappresentazione, ogni linea sottile corrisponde a una singola feature del dataset, e ciò permette di verificare la coerenza statistica del dataset e identificare eventuali anomalie sistematiche.

Al fine di rendere visivamente confrontabili metriche caratterizzate da unità di misura e range eterogenei, ogni feature è stata standardizzata e tracciata la sua funzione di densità di probabilità (PDF). Osservando il grafico derivante dalla sovrapposizione di tutte le curve, la concentrazione della "massa" delle feature attorno a una distribuzione normale standard con media 0 e deviazione standard 1 si evince che la maggior parte delle feature radiomiche ha un comportamento statistico coerente e ben condizionato. L'ampiezza del fascio di curve evidenzia la variabilità intrinseca del dataset ed eventuali linee che si discostano marcatamente dal pattern centrale segnalano gruppi di feature con comportamento anomalo, potenzialmente critiche per i classificatori.

Come osservabile in Figura 2.5, le modalità ADC e DWI mostrano una sovrapposizione molto compatta, indice di un'elevata stabilità del segnale estratto. Le sequenze T2, pur mantenendo una forma a campana, mostrano una dispersione leggermente maggiore, riflettendo la maggiore eterogeneità testurale di queste immagini anatomiche.

2.5.3 Associazione del Ground Truth e Finalizzazione del Dataset

Al termine della fase di *Data Cleaning*, si è proceduto all'integrazione delle matrici numeriche con le informazioni cliniche derivate dalla valutazione della qualità dei radiologhi (Sezione 2.2), al fine di associare a ciascun paziente la relativa classe target. L'unione tra il vettore delle feature e l'etichetta diagnostica è avvenuta mediante un'operazione di *left join* utilizzando l'identificativo univoco del paziente (`patient_id`) come chiave primaria. Successivamente è stata effettuata una verifica di completezza per identificare eventuali disallineamenti tra i dati di imaging e i dati clinici. L'analisi ha rilevato la presenza di un singolo paziente privo di etichetta diagnostica associata in tutte le modalità. In accordo con i criteri di inclusione, tale paziente è stato rimosso dal dataset per garantire l'integrità dell'addestramento supervisionato.

Il risultato finale di questa procedura è la generazione di quattro dataset definitivi, ciascuno composto da 801 pazienti e 1.056 feature radiomiche.

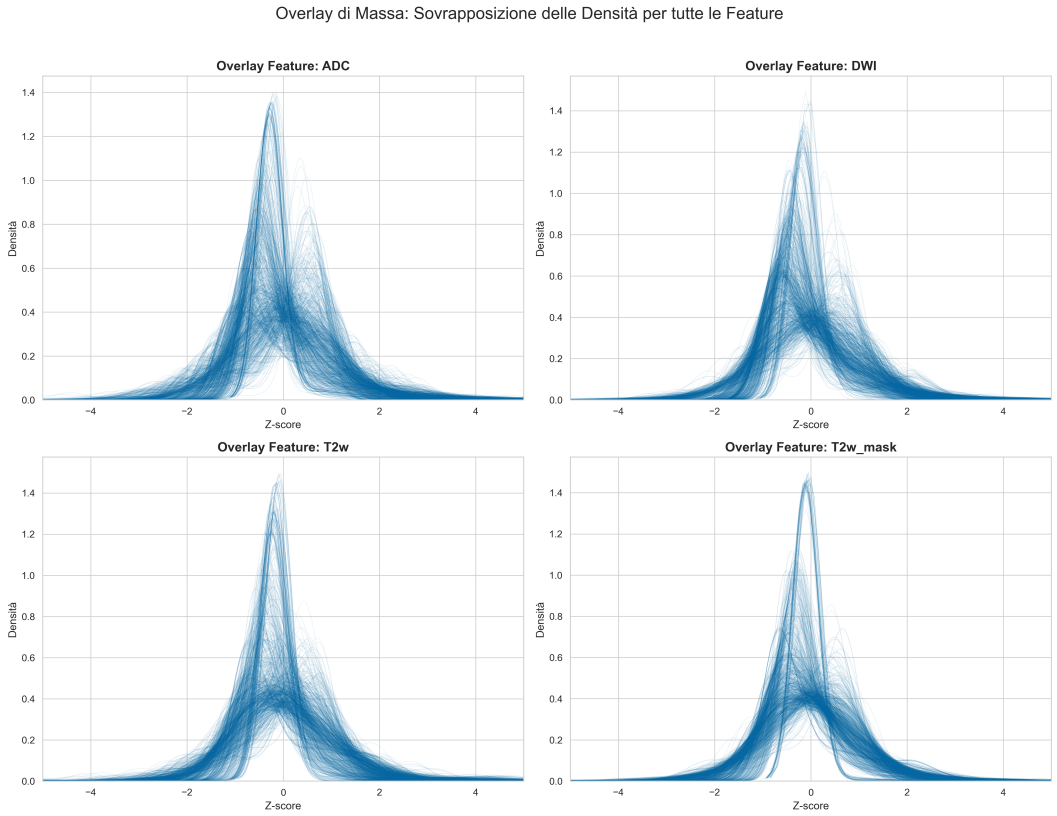


Figura 2.5: Visualizzazione globale delle feature radiomiche normalizzate. La sovrapposizione coerente delle curve suggerisce l'assenza di gravi anomalie di scala o artefatti di estrazione sistematici nel dataset.

2.6 Divisione in Construction Set e Test Set

Al fine di garantire una valutazione robusta e prevenire il fenomeno di *data leakage*, il dataset è stato ripartito seguendo una strategia di *Stratified Split* con proporzioni 80:20. L'80% dei dati (640 pazienti) è stato destinato alla fase di sviluppo e ottimizzazione dei modelli (Construction Set), mentre il restante 20% (161 pazienti) è stato segregato come Test Set esterno per la valutazione finale delle prestazioni.

La ripartizione dei dati è stata condotta a livello di singolo paziente per garantire la corrispondenza tra le sequenze multiparametriche dello stesso soggetto e rispettando la distribuzione originale delle classi di qualità. Il dettaglio quantitativo della suddivisione è riassunto nella Tabella 2.3.

Tabella 2.3: Ripartizione dei pazienti dopo lo split iniziale. Il Construction Set viene utilizzato per la Cross-Validation, mentre il Test Set rimane invisibile ai modelli fino alla fase finale.

Subset	**Numero Pazienti (N)**	**Percentuale (%)**
Construction Set (Train + Val)	640	79.9%
Test Set (Indipendente)	161	20.1%
Totale	**801**	**100.0%**

2.7 Strategia di Validazione: k-Fold Cross Validation

Per eliminare la dipendenza delle prestazioni del modello da una particolare suddivisione casuale, è stata implementata una strategia di *Stratified k-Fold Cross-Validation* [31] applicata al Construction Set con $k = 5$. Questo consente al modello di raggiungere un equilibrio ottimale tra capacità di apprendimento e generalizzazione, minimizzando il rischio che l'elevata dimensionalità delle feature radiomiche conduca a fenomeni di overfitting [32].

Il cuore di questa metodologia risiede in un processo ciclico nel quale il Construction Set viene ripartito in cinque sotto-gruppi (fold) di uguali dimensioni, preservando in ciascuno la distribuzione originale tra esami diagnostici e non diagnostici. Durante ciascuna delle cinque iterazioni, il modello viene addestrato su un set di addestramento composto da quattro fold (pari all'80% del totale), mentre il fold rimanente (20%) viene segregato e utilizzato esclusivamente come Validation Set per monitorare le prestazioni e guidare l'ottimizzazione degli iperparametri. Attraverso la rotazione sistematica del fold di validazione, ogni singolo campione del dataset viene impiegato esattamente una volta per la validazione, garantendo una copertura totale e imparziale del dataset.

Un prodotto fondamentale di questa architettura è la generazione delle previsioni *Out-of-Fold (OOF)*. Poiché ogni campione viene trattato come dato "invisibile" nel momento in cui appartiene al fold di validazione, le probabilità predette dal modello in quel frangente possono essere memorizzate in modo sistematico. Al termine del ciclo completo, si ottiene un set di previsioni riferito all'intero Construction Set, con la particolarità cruciale che ogni predizione è stata generata da un classificatore che non ha mai processato quel determinato paziente in fase di addestramento. Questa metodologia non è fine a se stessa, ma costituisce il presupposto tecnico per la costruzione dei modelli Ensemble tramite *Stacked Generalization* [33]. L'impiego delle previsioni OOF permette infatti di addestrare un meta-modello di secondo

livello basandosi sulle risposte fornite dai classificatori di base. Questo approccio evita fenomeni di *overfitting* o *data leakage*: il meta-modello impara a combinare i pesi delle diverse modalità MRI basandosi su stime di errore realistiche e non su dati già memorizzati. In questo studio, la solidità delle predizioni OOF ha rappresentato la base analitica per sviluppare e validare le strategie di Ensemble Learning discusse nel Capitolo 4.

2.8 Data Processing Pipeline

Le analisi statistiche che seguono, come le considerazioni sulla ridondanza delle feature e sulla distribuzione degli outlier, sono state condotte sull'intero Construction Set a scopo esplorativo. Questo approccio ha permesso di valutare le caratteristiche globali del dataset e di guidare le scelte progettuali della pipeline. Tuttavia, ogni operazione di trasformazione dei dati come la rimozione delle correlazioni, la normalizzazione e il bilanciamento è stata poi replicata in modo dinamico all'interno di ogni singolo fold di validazione sul Training Set, garantendo che il modello operasse sempre su dati nuovi e non contaminati.

2.8.1 Analisi della Ridondanza e Selezione delle Feature

La letteratura evidenzia come lo spazio delle feature radiomiche, in particolare quando espanso tramite l'applicazione di filtri Wavelet o LoG, sia intrinsecamente caratterizzato da un'elevata ridondanza. Molte variabili, infatti, tendono a catturare aspetti simili della tessitura, risultando in una forte collinearità tra i descrittori. Tale fenomeno, spesso riferito come "maledizione della dimensionalità" (*curse of dimensionality*), può destabilizzare i modelli di Machine Learning e incrementare drasticamente il rischio di overfitting, compromettendo la capacità di generalizzazione su dati esterni [34].

Per quantificare tale ridondanza, è stata calcolata la matrice di correlazione di Pearson (ρ) sul Construction Set di ciascuna modalità. L'analisi preliminare ha evidenziato l'esistenza di un sottoinsieme di variabili caratterizzato da una dipendenza lineare quasi perfetta. Come riportato in Tabella 2.4, sebbene la percentuale relativa di coppie fortemente correlate ($|\rho| > 0.99$) possa apparire contenuta, in termini assoluti essa si traduce in migliaia di variabili che trasportano la medesima informazione (fino a oltre 9.000 coppie nel caso dell'ADC). È doveroso precisare che tale grado di correlazione non è attribuibile esclusivamente a caratteristiche biologiche del tessuto, ma deriva in parte dalla definizione matematica stessa delle feature implementate in PyRadiomics. Come documentato nelle specifiche della libreria [35], diverse metriche presentano una dipendenza analitica diretta. Ad esempio, *Energy* e *Root Mean Squared* sono funzionalmente legate, così come *Compactness* e

Sphericity. Di conseguenza, molte variabili trasportano la medesima informazione fisica pur essendo formalmente distinte.

Al fine di ridurre questa sovrapposizione informativa preservando il massimo rigore metodologico, è stato implementato un filtro di rimozione non supervisionato operante dinamicamente all'interno della pipeline di *k-Fold Cross-Validation*. Durante ogni iterazione della procedura, la matrice di correlazione di Pearson viene ricalcolata esclusivamente sul Training Set del fold corrente. Esaminando a coppie le variabili estratte, per ogni coppia che presenta un coefficiente di correlazione assoluta superiore alla soglia predefinita $|\rho| > 0.99$, il sistema procede a mantenerne unicamente una, rimuovendo l'altra in quanto portatrice di un'informazione strettamente ridondante e linearmente dipendente. La maschera di selezione così generata viene successivamente applicata al corrispondente set di validazione e al Test Set indipendente, garantendo che nessuna informazione derivante dai dati di test possa influenzare la fase di riduzione dimensionale.

Tabella 2.4: Analisi quantitativa della collinearità pre-selezione condotta sul Construction Set globale a scopo esplorativo. Per ciascun sotto-dataset è riportato il numero di coppie di feature che presentano una correlazione di Pearson superiore alla soglia di 0.99 (altamente correlate). Il numero totale di coppie uniche analizzate è 557.040.

Dataset (Modalità)	**Coppie Correlate ($\|\rho\| > 0.99$)**	**Incidenza (%)**
ADC (Whole-Volume)	8.966	1.61%
DWI (Whole-Volume)	3.836	0.69%
T2w (Whole-Volume)	6.146	1.10%
T2w (Prostate Mask)	6.074	1.09%

2.8.2 Analisi della Distribuzione e Gestione degli Outlier

A valle della pulizia tecnica, è stata condotta un'analisi distributiva per identificare eventuali valori anomali (*outliers*) calcolando lo Z-score per ogni feature all'interno del Construction Set. Sono stati considerati outlier i valori che si discostavano dalla media per oltre 3 deviazioni standard ($|z| > 3$), come illustrato nei grafici di distribuzione riportati nelle Figure 2.6, 2.7, 2.8 e 2.9.

L'incidenza di tali outlier pari a circa il 3-4% della coorte risulta significativamente superiore a quanto atteso in una distribuzione normale standard, dove la probabilità teorica di valori oltre 3σ è inferiore allo 0.3% [36]. Questo eccesso di valori estremi indica una distribuzione dei dati a "code pesanti" (*heavy-tailed*),

confermando che la variabilità osservata non è puramente stocastica, ma riflette una complessità fenotipica reale che il modello deve essere in grado di gestire [37].

Alla luce di queste evidenze, in questo studio si è scelto metodologicamente di non rimuovere i pazienti sulla base di questo criterio statistico. Tale decisione è cruciale nel contesto della valutazione della qualità poiché i valori estremi delle feature radiomiche spesso non rappresentano errori di calcolo, bensì sono i descrittori quantitativi di artefatti d'immagine come rumore di fondo e ghosting da movimento o di variazioni nei protocolli di acquisizione tra i diversi centri. L'eliminazione di questi casi avrebbe comportato la perdita di esempi fondamentali per l'addestramento del modello, rischiando di rimuovere proprio le immagini non diagnostiche che l'algoritmo deve imparare a discriminare rispetto agli esami ottimali.

2.8.3 Preparazione per la Modellazione: Standardizzazione e Bilanciamento (SMOTE)

Normalizzazione delle Feature (Z-score Scaling)

E' stata applicata una procedura di standardizzazione (*Z-score Normalization*) per evitare che variabili con scale numeriche più ampie dominassero impropriamente il processo di apprendimento utilizzando la classe `StandardScaler` della libreria `scikit-learn`. Ogni feature x è stata traslata e scalata affinché presentasse media nulla ($\mu = 0$) e deviazione standard unitaria ($\sigma = 1$):

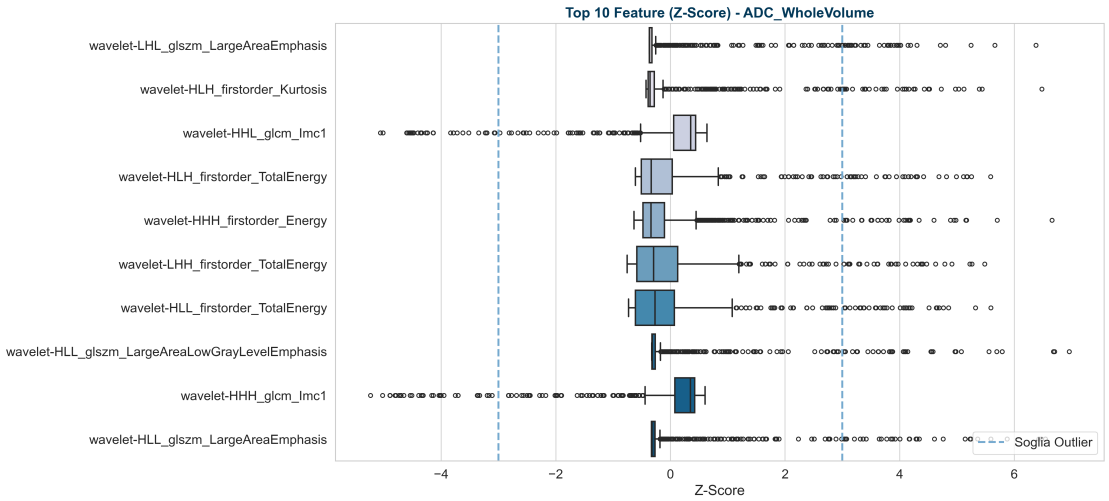
$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

Per garantire il rigore metodologico, i parametri di normalizzazione (μ e σ) sono stati calcolati sul Training Set di ogni fold e successivamente applicati per trasformare il Validation Set e il Test Set, simulando uno scenario reale in cui i dati futuri non sono noti a priori. Gli oggetti scaler addestrati dei casi migliori sono stati salvati in formato `.pkl` per garantire la riproducibilità della pipeline di inferenza su nuovi dati.

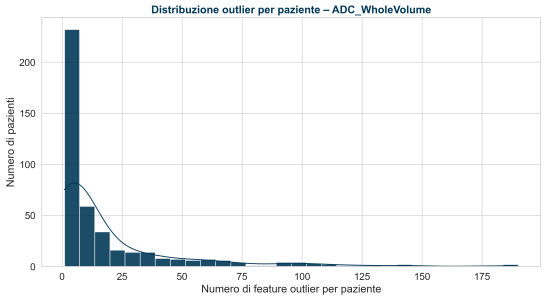
Ribilanciamento delle Classi (SMOTE)

L'analisi della variabile target nel Construction Set ha confermato un forte sbilanciamento delle classi, con una prevalenza di casi diagnostici pari a 81.7% del totale, rispetto ai casi non diagnostici pari a 18.3%. Nell'ambito della valutazione della qualità delle immagini e dell'allenamento dei diversi classificatori questo può portare a gravi problemi di overfitting e valori di accuratezza irrealisticamente elevati.

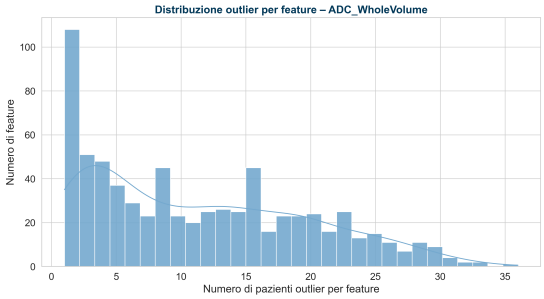
Per mitigare questo fenomeno, è stata adottata la tecnica SMOTE (*Synthetic Minority Over-sampling Technique*) [38], un algoritmo che genera nuovi dati sintetici



(a) Distribuzione Z-Score delle prime 10 feature.



(b) Outlier per Paziente

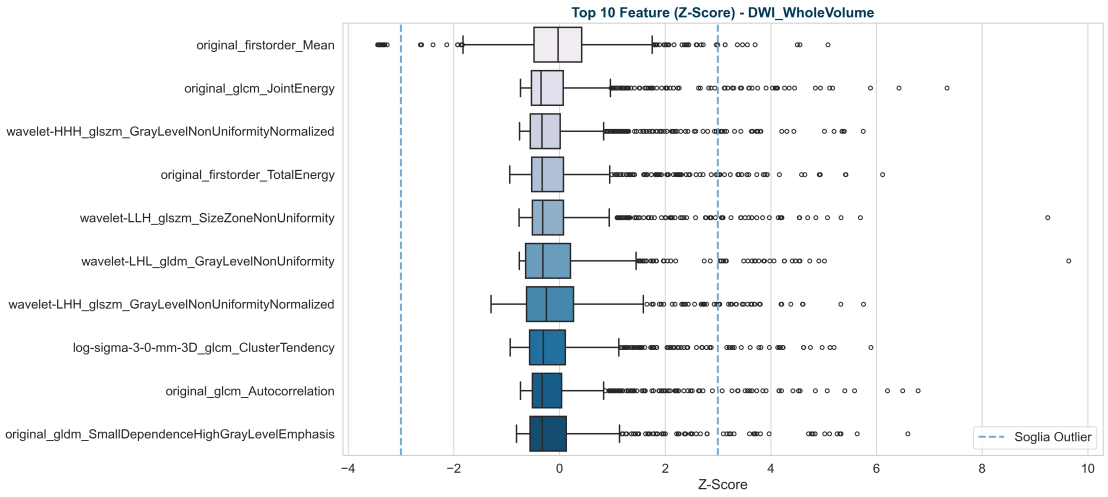


(c) Outlier per Feature

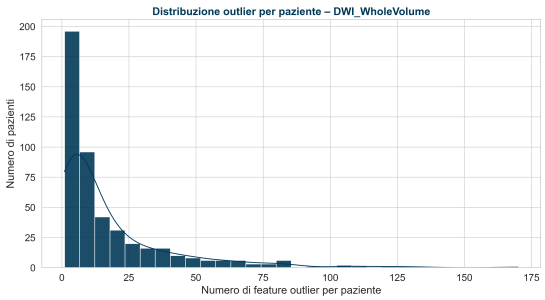
Figura 2.6: Analisi degli outlier per la modalità ADC. (a) I boxplot evidenziano valori estremi in feature legate all'energia e all'omogeneità spaziale, correlati a disomogeneità di campo o rumore diffuso. (b-c) La distribuzione a coda lunga conferma che una minoranza significativa di pazienti presenta marcate deviazioni rispetto alla popolazione media.

tramite interpolazione vettoriale tra i k vicini più prossimi (k-NN) della classe minoritaria nello spazio delle feature.

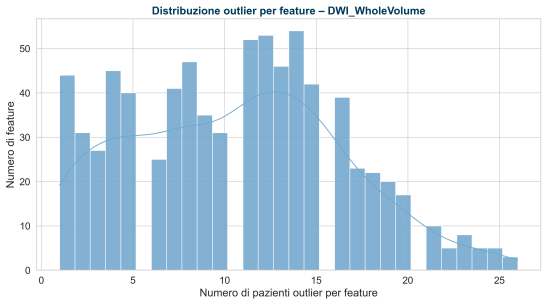
L'algoritmo è stato testato sul Construction Set per ogni modalità, configurando una strategia di campionamento pari a 0.6 (`sampling_strategy=0.6`). Questa configurazione ha portato la numerosità della classe minoritaria al 60% della classe maggioritaria, modificando la distribuzione finale in un rapporto di 37.5% per la Classe 0 e di 62.5% per la Classe 1. L'intervento ha comportato la generazione di circa 147 campioni sintetici per modalità portando il Training Set da 480 a 627 campioni, densificando la regione dello spazio delle feature associata alle immagini



(a) Distribuzione Z-Score delle prime 10 feature.



(b) Outlier per Paziente

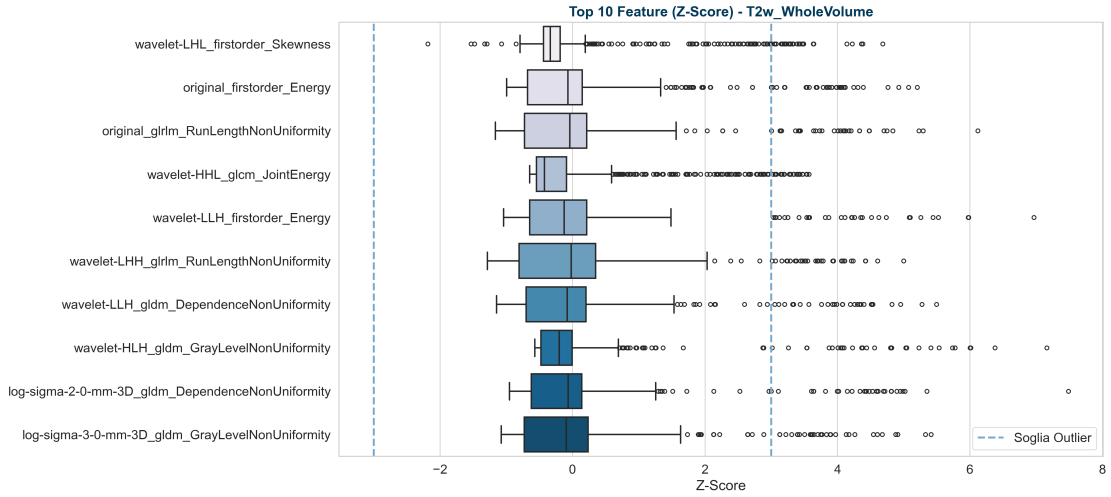


(c) Outlier per Feature

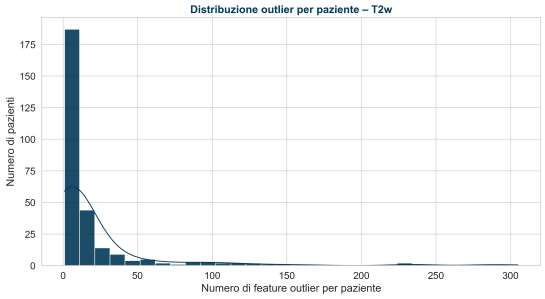
Figura 2.7: Analisi degli outlier per la modalità DWI. (a) Le feature testurali mostrano una dispersione significativa oltre la soglia di 3σ. (b-c) Gli istogrammi evidenziano come gli outlier non siano limitati a pochi casi isolati, ma rappresentino una caratteristica intrinseca della variabilità del segnale.

non diagnostiche senza introdurre un eccessivo rumore artificiale. Questo processo verrà poi replicato sui singoli Training Set di ogni fold dove i campioni sintetici sono confinati.

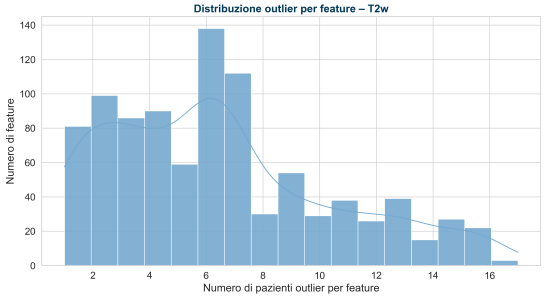
In conclusione, la pipeline di elaborazione descritta in questa sezione permette di trasformare il dataset clinico in una struttura dati robusta, pronta per la fase di modellazione predittiva.



(a) Distribuzione Z-Score delle prime 10 feature.

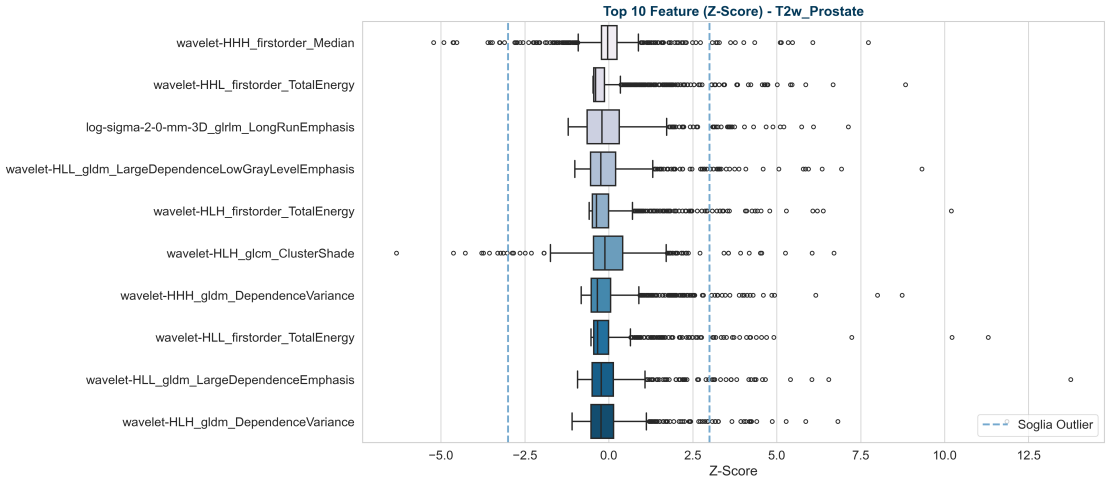


(b) Outlier per Paziente

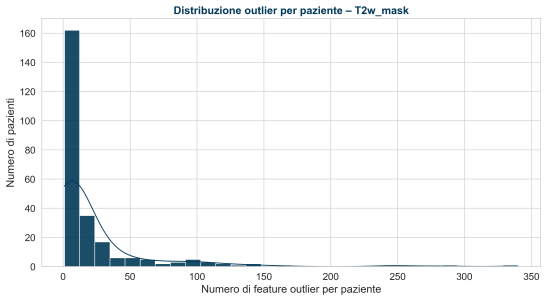


(c) Outlier per Feature

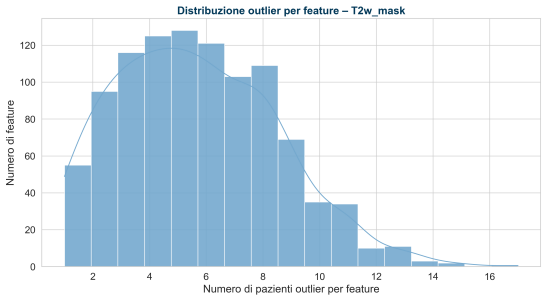
Figura 2.8: Analisi degli outlier per T2w (a) Si notano valori estremi nelle feature di asimmetria, probabilmente influenzati dal forte contrasto tra il tessuto anatomico e il fondo dell'immagine. (b-c) La frequenza degli outlier suggerisce la presenza di artefatti o variazioni di FOV tra i centri.



(a) Distribuzione Z-Score delle prime 10 feature.



(b) Outlier per Paziente



(c) Outlier per Feature

Figura 2.9: Analisi degli outlier per la maschera prostatica. (a) Focalizzando l'analisi sul solo organo, emergono feature legate alla dipendenza spaziale, riflettendo l'eterogeneità tissutale intrinseca del carcinoma prostatico. (b-c) A differenza dell'analisi sull'intero volume, qui la distribuzione degli outlier appare meno influenzata da artefatti di fondo e più correlata alla variabilità biologica inter-paziente. La presenza di code lunghe indica che specifici sottogruppi di pazienti manifestano pattern di tessitura unici, che è fondamentale preservare per il training del modello.

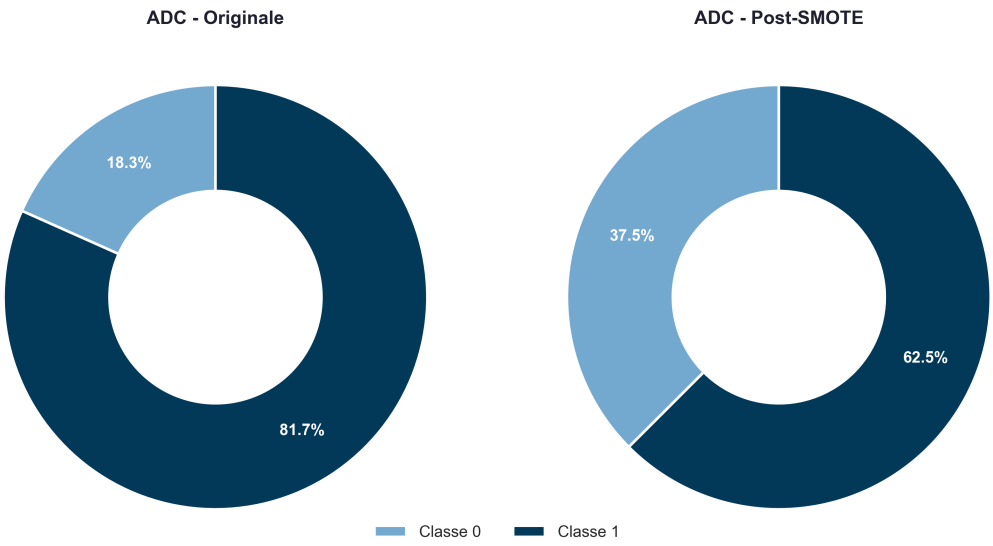


Figura 2.10: Effetto dell'applicazione dell'algoritmo SMOTE sulla distribuzione delle classi nel Construction Set per la modalità ADC. A sinistra, la distribuzione originale mostra lo sbilanciamento nativo. A destra, la distribuzione post-processing evidenzia l'arricchimento della classe minoritaria tramite generazione sintetica, raggiungendo un rapporto più bilanciato.

Capitolo 3

Strategie di Selezione delle Feature e Riduzione della Dimensionalità

Il dataset radiomico consolidato nelle fasi precedenti si presenta inizialmente con un'elevata dimensionalità, caratterizzato da un vettore descrittivo di 1056 feature per ciascuna modalità di imaging, e tra le 700 e 800 feature a seguito della rimozione delle feature altamente correlate. In letteratura è noto il rischio teorico legato alla "maledizione della dimensionalità" (*curse of dimensionality*), che suggerisce come un numero eccessivo di variabili rispetto ai campioni possa condurre a *overfitting* [34]. Tuttavia, ridurre drasticamente le feature a priori potrebbe comportare una perdita di informazione preziosa per la valutazione della qualità delle immagini.

Per tale ragione, l'approccio adottato si basa su un'indagine empirica condotta sul Construction Set globale. Questa analisi preliminare ha lo scopo di identificare quali strategie di selezione siano più idonee a estrarre i segnali rilevanti, definendo così i protocolli che sono stati poi integrati dinamicamente nella pipeline di validazione.

È opportuno sottolineare che i valori numerici e i sottoinsiemi di feature discussi in questa sezione rappresentano il risultato di un'analisi esplorativa condotta a scopo conoscitivo sulla coorte di sviluppo. Tuttavia, per garantire il massimo rigore metodologico, la selezione effettiva delle variabili è stata ricalcolata all'interno di ciascuno dei cinque fold della cross-validazione. In questo modo, ogni modello ha operato esclusivamente su un set di feature identificato sui propri dati di addestramento, assicurando che la capacità di generalizzazione venisse testata su dati di validazione e di test rigorosamente indipendenti.

Nelle sezioni successive vengono analizzate diverse strategie di selezione Filter, Embedded e Wrapper per valutare se una riduzione della dimensionalità possa

portare un effettivo beneficio in termini di separabilità delle classi e performance predittive.

3.1 Visualizzazione dello Spazio delle Feature

Per valutare la distribuzione intrinseca dei dati e la separabilità tra le classi, sono state impiegate due tecniche di riduzione della dimensionalità a scopo di visualizzazione:

- **Principal Component Analysis (PCA):** Tecnica lineare che proietta i dati in uno spazio ridotto preservando la massima varianza globale [39]. Questa tecnica è stata utilizzata per valutare la presenza di direzioni di massima variabilità intrinseca nello spazio radiomico e per identificare strutture globali non influenzate da fenomeni locali o non lineari.

- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** Tecnica non lineare orientata alla conservazione delle relazioni locali e alla scoperta di cluster ben definiti nello spazio dei dati [40]. Il t-SNE è stato utilizzato come complemento alla PCA perché è in grado di evidenziare pattern non lineari, sottogruppi e separazioni sottili tra classi che le proiezioni lineari non riescono a catturare. Questa tecnica permette di visualizzare eventuali gruppi naturali derivanti dalla combinazione di feature radiomiche anche quando lo spazio originale è altamente complesso e non lineare.

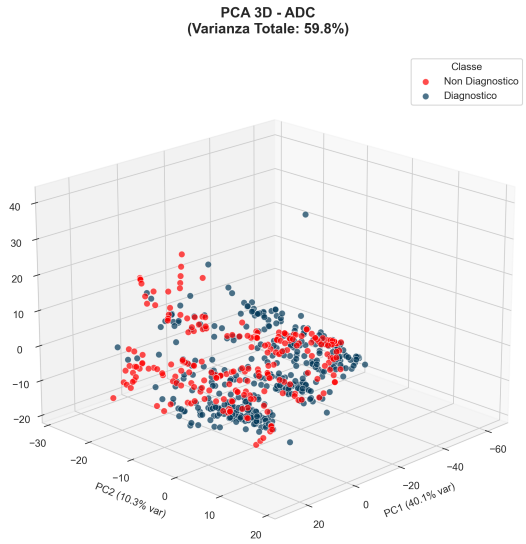
L'analisi visiva è stata condotta sia attraverso una valutazione bidimensionale comparativa (Figure 3.1 e 3.2) sia grazie a un'esplorazione tridimensionale specifica (Figura 3.3). In particolare, per le modalità ADC e DWI, i grafici t-SNE evidenziano la presenza di cluster locali, ma non mostrano un confine di separazione netto tra le due classi. Inoltre, la sovrapposizione osservata nella PCA suggerisce che la relazione tra le feature radiomiche e la qualità dell'immagine è di natura prevalentemente non lineare, suggerendo l'utilizzo di classificatori complessi nelle fasi successive.

3.2 Metodologie di Selezione delle Feature

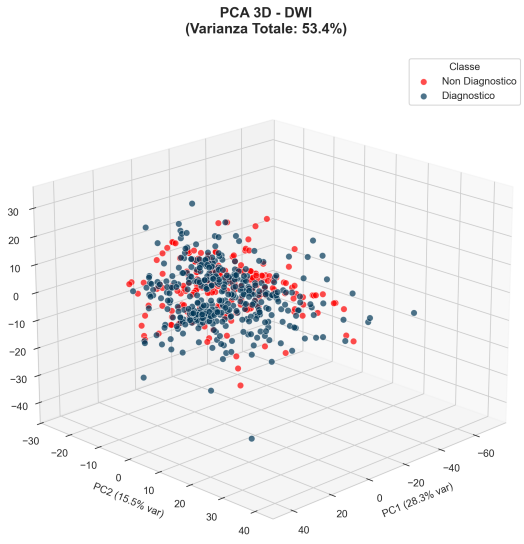
Per identificare le feature radiomiche più rilevanti, sono stati confrontati approcci *Filter* univariati e multivariati, *Embedded* e *Wrapper*. Questa diversificazione permette di compensare i bias intrinseci di ogni singolo algoritmo.

Figura 3.1: Analisi Globale dello spazio delle feature tramite PCA (Prime due componenti). Ogni punto rappresenta un paziente. La forte sovrapposizione tra i punti rossi (Non Diagnostico) e blu (Diagnostico) conferma che la varianza globale del dataset non è sufficiente, da sola, a separare linearmente le classi di qualità.

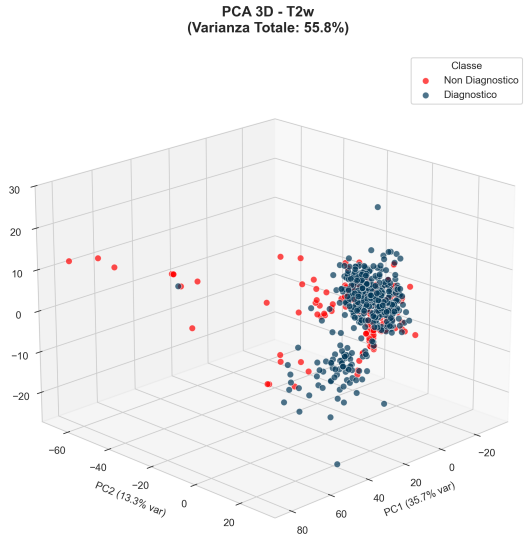
Figura 3.2: Mappa delle feature non lineare (t-SNE). A differenza della PCA, questa tecnica preserva la struttura locale dei dati. Si nota la formazione di sottogruppi e una distribuzione più complessa, indicando che l'informazione discriminante risiede nelle relazioni non lineari tra le variabili radiomiche.



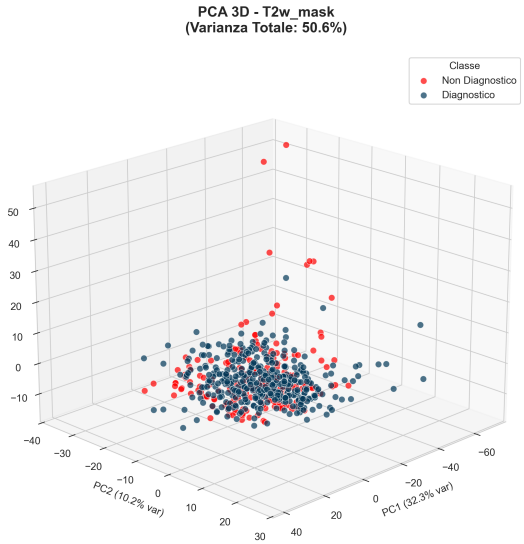
(a) ADC (Whole Volume)



(b) DWI (Whole Volume)



(c) T2-weighted (Whole Volume)



(d) T2-weighted (Prostate Mask)

Figura 3.3: Visualizzazione tridimensionale dello spazio delle feature (PC1, PC2, PC3). **(a-b)** Le sequenze di diffusione mostrano una nuvola di punti con forte sovrapposizione tra le classi (Diagnostico in azzurro, Non Diagnostico in rosso). **(c-d)** Confronto T2-weighted: l'analisi sull'intero volume (c) presenta una maggiore dispersione dovuta al background, mentre la restrizione alla maschera prostatica (d) condensa i dati in un cluster centrale più compatto, confermando il beneficio della segmentazione nel ridurre la varianza non informativa.

3.2.1 Approcci Filter

I metodi Filter valutano la rilevanza delle feature basandosi sulle loro proprietà statistiche intrinseche rispetto alla variabile target, indipendentemente dal modello di classificazione che verrà utilizzato successivamente.

Analisi Univariata: ANOVA e Mutual Information

In prima istanza, sono stati applicati due test statistici univariati per un ranking preliminare:

- **ANOVA F-test:** Valuta la varianza tra le medie delle classi [41]. Sono state selezionate le feature con un punteggio $F > 10$. In caso di scarsità di risultati, è stato applicato un criterio di *fallback* selezionando le prime 100 feature;
- **Mutual Information (MI):** Misura la dipendenza non lineare tra variabili basandosi sull'entropia dell'informazione [42]. Sono state trattenute le feature con un punteggio $MI > 0.05$.

Dal punto di vista quantitativo, il test ANOVA ha selezionato un set di feature robusto, oscillando tra le 117 variabili significative per DWI e le 168 per T2w, con una maggiore selettività nel caso della maschera prostatica con 75 feature. Parallelamente, l'analisi basata sulla MI ha prodotto sottoinsiemi di dimensioni comparabili con 152 feature per DWI e 170 per T2w. Tale evidenza suggerisce che il dataset sia ricco di relazioni informative catturabili sia linearmente con F-test che non linearmente tramite MI.

Tuttavia, sul piano qualitativo, emergono differenze più significative tra i due metodi, riassunte in Tabella 3.1. Analizzando la composizione per famiglia, si delineano delle preferenze da parte dei metodi. Mentre l'ANOVA tende a includere una quota rilevante di feature filtrate LoG fino al 26% in DWI e T2 Mask, la MI mostra una predilezione marcata per le trasformate *Wavelet* che raggiungono il 75% in ADC e T2 Mask. Nel caso del volume T2w, MI valorizza significativamente le feature di Gradiente con una percentuale del 15.3% contro il 3.0% dell'ANOVA. Ciò conferma che la MI è particolarmente sensibile a pattern complessi e variazioni di contorno, ovvero il filtro Gradiente, che sfuggono all'analisi della varianza classica. In sintesi, i risultati smentiscono l'ipotesi di una marcata differenza quantitativa tra i due metodi poichè ANOVA e MI convergono verso un numero comparabile di feature, differenziandosi piuttosto per la tipologia di pattern estratti prediligendo pattern lineari o non lineari.

Un dato trasversale che emerge con forza da entrambe le selezioni è la dominanza numerica della famiglia Wavelet, che costituisce frequentemente oltre il 60% del subset finale. Tale prevalenza è parzialmente influenzata da un bias di frequenza intrinseco alla pipeline di estrazione. La libreria PyRadiomics, infatti, genera

nativamente un numero di feature Wavelet otto volte superiore rispetto alle feature Originali a causa delle 8 decomposizioni spettrali per ogni feature base.

Di conseguenza, l'elevata percentuale di Wavelet selezionate potrebbe riflettere semplicemente la loro sovrabbondanza nel dataset di partenza, piuttosto che una reale preferenza algoritmica legata al contenuto informativo. Proprio per discernere tra la semplice abbondanza numerica e la reale importanza biologica, nella sezione successiva (Sezione 3.3.2) verrà introdotto il Fattore di Arricchimento (*Enrichment Factor*), una metrica normalizzata concepita per correggere questo bias di frequenza e quantificare l'effettiva selettività dei metodi verso specifiche classi di feature.

Analisi Multivariata: Random Forest Importance

Per catturare interazioni più complesse e non lineari, è stato utilizzato un classificatore Random Forest. Questo algoritmo calcola l'importanza di ciascuna feature (*Gini Importance*) misurando quanto ogni variabile contribuisca a ridurre l'impurità dei nodi durante la costruzione degli alberi decisionali [43].

Il modello, composto da 500 alberi decisionali, è stato addestrato sull'intero spazio delle feature per calcolarne il ranking di importanza. Per ottenere un sottoinsieme compatto ma rappresentativo, si è analizzato il decadimento del punteggio di Gini Importance. Come illustrato in Figura 3.4 per la modalità ADC, l'andamento della curva non mostra un gomito netto o plateau, bensì una decrescita graduale. Tale comportamento si è rivelato coerente e sovrapponibile per tutte le sequenze analizzate, indicando che l'informazione radiomica è distribuita su un ampio spettro di variabili piuttosto che concentrata in poche feature dominanti. Alla luce di ciò, si è scelto di applicare una soglia uniforme, selezionando le prime 100 feature per ciascuna modalità.

L'analisi delle feature con importanza più alta mostrata in Tabella 3.2 rivela un cambio di paradigma rispetto all'analisi univariata. Mentre le Wavelet rimangono dominanti nelle sequenze di diffusione, nelle sequenze morfologiche T2w il classificatore Random Forest attribuisce un peso determinante alle feature filtrate con LoG. In particolare, nel dataset T2 Prostate Mask, 4 delle prime 5 feature più importanti sono derivate da filtri LoG con $\sigma = 2.0$ e 3.0 mm. Questo suggerisce che, una volta isolato l'organo, la qualità diagnostica è fortemente correlata alla presenza di aree di intensità omogenea e alla loro uniformità locale, che il filtro LoG è specificamente progettato per esaltare. Al contrario, nell'ADC, il modello si affida maggiormente a statistiche del Primo Ordine come Media e valore massimo filtrate Wavelet per valutare l'intensità globale del segnale.

Tabella 3.1: Confronto dettagliato della selezione per famiglia. Oltre ai valori percentuali, viene riportata la classe di rappresentatività (Marginale, Moderata, Significativa, Dominante). Si noti come la MI valorizzi le feature di Gradiente nelle sequenze anatomiche (T2w) molto più dell'ANOVA, che le considera marginali.

Famiglia	**ANOVA (F-Test)**	**Mutual Information (MI)**
ADC Whole Volume		
Totale Selezionate	**152 feature**	**135 feature**
Original	Moderata (9.9%)	Moderata (6.7%)
Wavelet	Dominante (65.1%)	Dominante (74.8%)
LoG (σ)	Significativa (16.4%)	Moderata (7.4%)
Gradient	Moderata (8.6%)	Moderata (11.1%)
DWI Whole Volume		
Totale Selezionate	**117 feature**	**152 feature**
Original	Moderata (6.0%)	Moderata (9.2%)
Wavelet	Dominante (50.4%)	Dominante (65.8%)
LoG (σ)	Significativa (26.5%)	Moderata (12.5%)
Gradient	Significativa (17.1%)	Moderata (12.5%)
T2-weighted Whole Volume		
Totale Selezionate	**168 feature**	**170 feature**
Original	Moderata (7.1%)	Moderata (10.0%)
Wavelet	Dominante (66.7%)	Dominante (63.5%)
LoG (σ)	Significativa (23.2%)	Moderata (11.2%)
Gradient	Marginale (3.0%)	Significativa (15.3%)
T2-weighted Prostate Mask		
Totale Selezionate	**75 feature**	**72 feature**
Original	Moderata (6.7%)	Moderata (6.9%)
Wavelet	Dominante (46.7%)	Dominante (75.0%)
LoG (σ)	Significativa (26.7%)	Moderata (11.1%)
Gradient	Significativa (20.0%)	Moderata (6.9%)

3.2.2 Approccio Embedded: LASSO Regression

Come metodo Embedded, è stata utilizzata la tecnica LASSO (*Least Absolute Shrinkage and Selection Operator*). Si tratta di un modello di regressione logistica

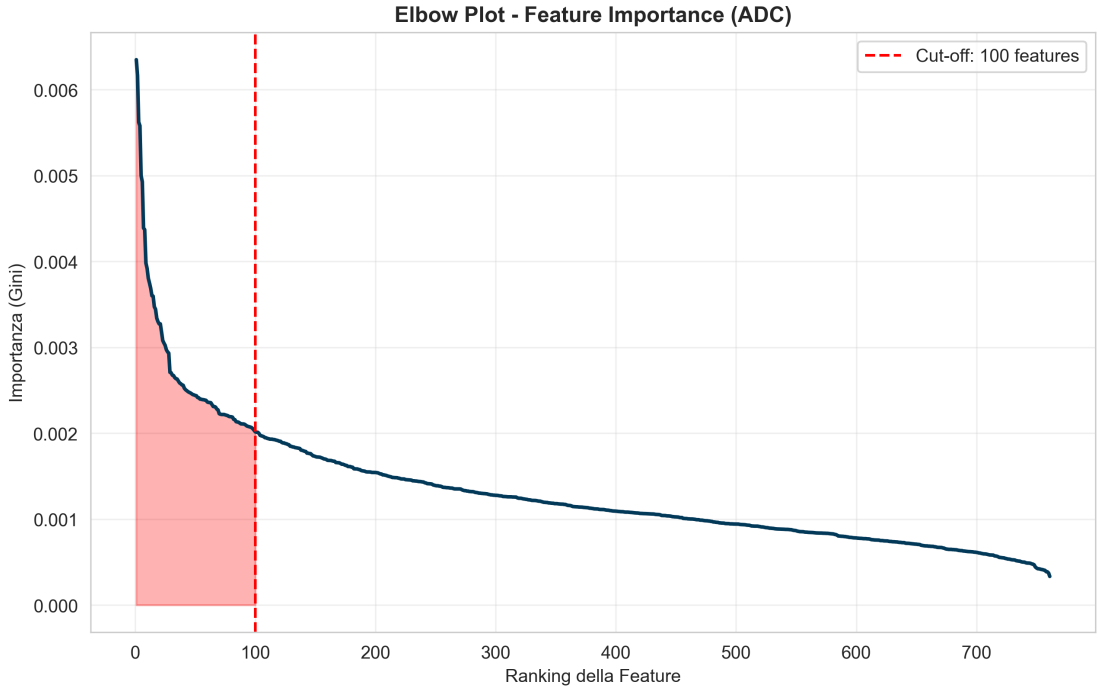


Figura 3.4: Ranking dell'importanza delle feature per la modalità ADC, riportata come esempio rappresentativo dell'andamento osservato in tutte le sequenze. Il grafico mostra un decadimento graduale senza un punto di flesso evidente. La linea tratteggiata rossa indica il cut-off applicato (100 Features), selezionato per isolare il nucleo di variabili a maggior impatto predittivo.

penalizzata che introduce un termine di regolarizzazione L_1 alla funzione di costo:

$$Loss = \sum (y - \hat{y})^2 + \lambda \sum |\beta_j| \tag{3.1}$$

La proprietà geometrica fondamentale della penalità L_1 è la capacità di forzare esattamente a zero i coefficienti β delle feature ridondanti o poco informative, operando contestualmente una selezione delle variabili e una regolarizzazione del modello per prevenire l'overfitting [44].

L'algoritmo è stato configurato con un parametro di regolarizzazione C pari a 0.1, ottenendo una riduzione della dimensionalità. Come riportato in Tabella 3.3, il modello ha selezionato un nucleo compatto di circa 40 feature per modalità su oltre 700 iniziali.

Nel contesto di un modello lineare penalizzato, il segno dei coefficienti assume un significato interpretabile, in quanto ogni feature contribuisce in modo additivo alla probabilità di appartenenza alla classe positiva:

Tabella 3.2: Le 5 feature con maggiore Gini Importance identificate dal Random Forest. L'ADC premia statistiche di intensità (First Order), mentre la T2 Mask è dominata da feature testurali complesse estratte tramite filtro LoG, indicativo di una sensibilità alla struttura granulare del tessuto.

Modalità	Feature Name	Importance
ADC	original_firstorder_Maximum	0.0063
	wavelet-HHH_firstorder_Mean	0.0062
	wavelet-LLL_gldm_DependenceNonUniformityNormalized	0.0056
	wavelet-LLL_gldm_SmallDependenceLowGrayLevelEmphasis	0.0056
	original_firstorder_Mean	0.0050
DWI	wavelet-HHL_firstorder_Energy	0.0051
	log-sigma-3-0-mm-3D_firstorder_90Percentile	0.0045
	wavelet-LHL_glcm_MCC	0.0043
	wavelet-LLH_glcm_InverseVariance	0.0043
	wavelet-HLH_glcm_Imc2	0.0040
T2w (Whole)	log-sigma-3-0-mm-3D_firstorder_Variance	0.0055
	wavelet-LLH_glszm_SizeZoneNonUniformity	0.0055
	wavelet-LHL_firstorder_Mean	0.0054
	wavelet-LLH_gldm_DependenceEntropy	0.0053
	wavelet-HHL_glszm_SmallAreaHighGrayLevelEmphasis	0.0048
T2w (Mask)	log-sigma-3-0-mm-3D_glszm_GrayLevelNonUniformity	0.0062
	log-sigma-2-0-mm-3D_glszm_LowGrayLevelZoneEmphasis	0.0060
	wavelet-LHH_glcm_ClusterShade	0.0060
	log-sigma-2-0-mm-3D_glrlm_LowGrayLevelRunEmphasis	0.0050
	log-sigma-2-0-mm-3D_glszm_HighGrayLevelZoneEmphasis	0.0050

- **Segno Positivo ($\beta > 0$):** Feature associate alla classe "Diagnostico". Valori elevati di queste metriche come ad esempio la Varianza dei livelli di grigio in DWI, indicano una buona qualità dell'immagine.
- **Segno Negativo ($\beta < 0$):** Feature associate alla classe "Non Diagnostico". Valori alti suggeriscono la presenza di artefatti o rumore.

L'analisi dei coefficienti rivela dettagli clinici interessanti. Nelle sequenze T2w, la feature con il peso negativo più forte (-0.53) è l'Autocorrelazione, suggerendo che pattern ripetitivi artificiali tipici del *ringing* o del *ghosting* sono forti predittori di scarsa qualità. Al contrario, nelle mappe DWI, un'alta varianza dei livelli di grigio con un peso positivo di $+0.48$ è indicativa di un segnale diagnostico ricco di contrasto e informazione. Questo comportamento è schematizzato in Tabella 3.4.

Tabella 3.3: Riepilogo della selezione LASSO con $C = 0.1$. Il metodo ha ridotto lo spazio delle feature del 95%, trattenendo circa 40 variabili per modalità.

Modalità	**Feature Selezionate**	**Feature Scartate ($\beta = 0$)**
ADC	38	726
DWI	42	760
T2w (Whole)	43	673
T2w (Mask)	43	658

Tabella 3.4: Le feature più determinanti identificate dal LASSO. Il "Peso" (β) indica la forza e la direzione dell'associazione: pesi negativi penalizzano la qualità indicando artefatti, pesi positivi la premiano.

Modalità	**Nome Feature**	**Peso (β)**	**Significato**
ADC	wavelet-LLH_glcm_Imc2	+0.26	Pro-Diagnostico
	log-sigma-2.0...glszm_GrayLevelNonUniformity	-0.37	Pro-Non Diagnostico
DWI	wavelet-HLL_glszm_GrayLevelVariance	+0.48	Pro-Diagnostico
	wavelet-LHL_glcm_MCC	-0.34	Pro-Non Diagnostico
T2w (Whole)	wavelet-HHL_glrlm_RunEntropy	+0.34	Pro-Diagnostico
	wavelet-LHL_glcm_Autocorrelation	-0.53	Pro-Non Diagnostico
T2w (Mask)	wavelet-HLH_glszm_SizeZoneNonUniformity	+0.36	Pro-Diagnostico
	gradient_gldm_LargeDependenceHighGray...	-0.48	Pro-Non Diagnostico

3.2.3 Approccio Wrapper: RFECV

Infine, è stato adottato un metodo Wrapper basato sull'Eliminazione Ricorsiva delle Feature con Cross-Validazione (RFECV), derivato dall'approccio RFE di Guyon et al. [45]. In questa configurazione, un estimatore Random Forest viene addestrato sul set completo di feature e, a ogni iterazione, le variabili con minore importanza vengono rimosse. L'integrazione con una *Stratified 5-Fold Cross-Validation* permette di identificare automaticamente il numero ottimale di feature che massimizza la metrica target (Balanced Accuracy), garantendo che la selezione non dipenda da un singolo split casuale dei dati.

I risultati ottenuti presenti in Tabella 3.5 offrono uno spunto di riflessione cruciale sulla natura del dataset. Contrariamente all'approccio LASSO, che ha isolato un nucleo ristretto di variabili, l'RFECV ha ritenuto necessario mantenere quasi l'intera totalità delle feature disponibili per massimizzare la performance.

Nello specifico, l'algoritmo ha selezionato tra le 642 e le 720 feature per modalità, scartando meno del 10-15% delle variabili iniziali.

Questo comportamento suggerisce che, per un classificatore non lineare come il Random Forest, l'informazione utile alla diagnosi non è concentrata in poche feature dominanti ma è distribuita su tutto lo spazio delle features. Ogni feature sembra apportare un contributo incrementale alla capacità discriminante complessiva, portando l'algoritmo a non effettuare tagli drastici.

Tabella 3.5: Sintesi della selezione RFECV. L'algoritmo ha mantenuto la quasi totalità delle feature, indicando che la massima accuratezza si ottiene sfruttando l'interazione complessa tra centinaia di variabili piuttosto che selezionandone poche.

Modalità	**Feature Selezionate**	**Comportamento**
ADC	685	Conservativo (High Retention)
DWI	720	Conservativo (High Retention)
T2w (Whole)	642	Conservativo (High Retention)
T2w (Mask)	698	Conservativo (High Retention)

3.3 Analisi Comparativa della Selezione

L'applicazione dei cinque metodi di selezione (RF, LASSO, RFECV, ANOVA, MI) ha prodotto sottoinsiemi di feature marcatamente eterogenei sia in termini di numerosità che di composizione. Come riassunto in Tabella 3.6, si osserva una notevole variabilità nel numero di variabili trattenute, che spaziano dalle poche decine dei metodi Filter e Random Forest alle diverse centinaia del metodo Wrapper RFECV.

Questa eterogeneità di risultati motiva la scelta metodologica adottata nel capitolo successivo. Invece di selezionare a priori un unico set di feature, verranno addestrati modelli predittivi paralleli su ciascun sottoinsieme, lasciando che siano le metriche di validazione finali a decretare quale strategia di selezione offra il miglior bilanciamento tra complessità e accuratezza clinica.

Questa diversità è stata analizzata sotto tre profili complementari: la sovrapposizione del contenuto informativo tramite Jaccard Index, la tipologia delle feature con Enrichment Factor e la capacità di preservare la struttura topologica dei dati tramite ispezione visiva con t-SNE.

Tabella 3.6: Confronto del numero di feature selezionate da ciascun algoritmo per modalità. Si osserva l'estrema variabilità delle dimensioni dei subset dalla selezione contenuta della tecnica LASSO ($\sim$ 40 feature) alla ritenzione quasi totale dell'RFECV ($\sim$ 700 feature).

Modalità	Iniziali	LASSO	RF	ANOVA	MI	RFECV
ADC	764	38	100	152	135	685
DWI	802	42	100	117	152	720
T2w (Whole)	716	43	100	168	170	642
T2w (Mask)	701	43	100	75	72	698

3.3.1 Analisi della Sovrapposizione: Jaccard Index

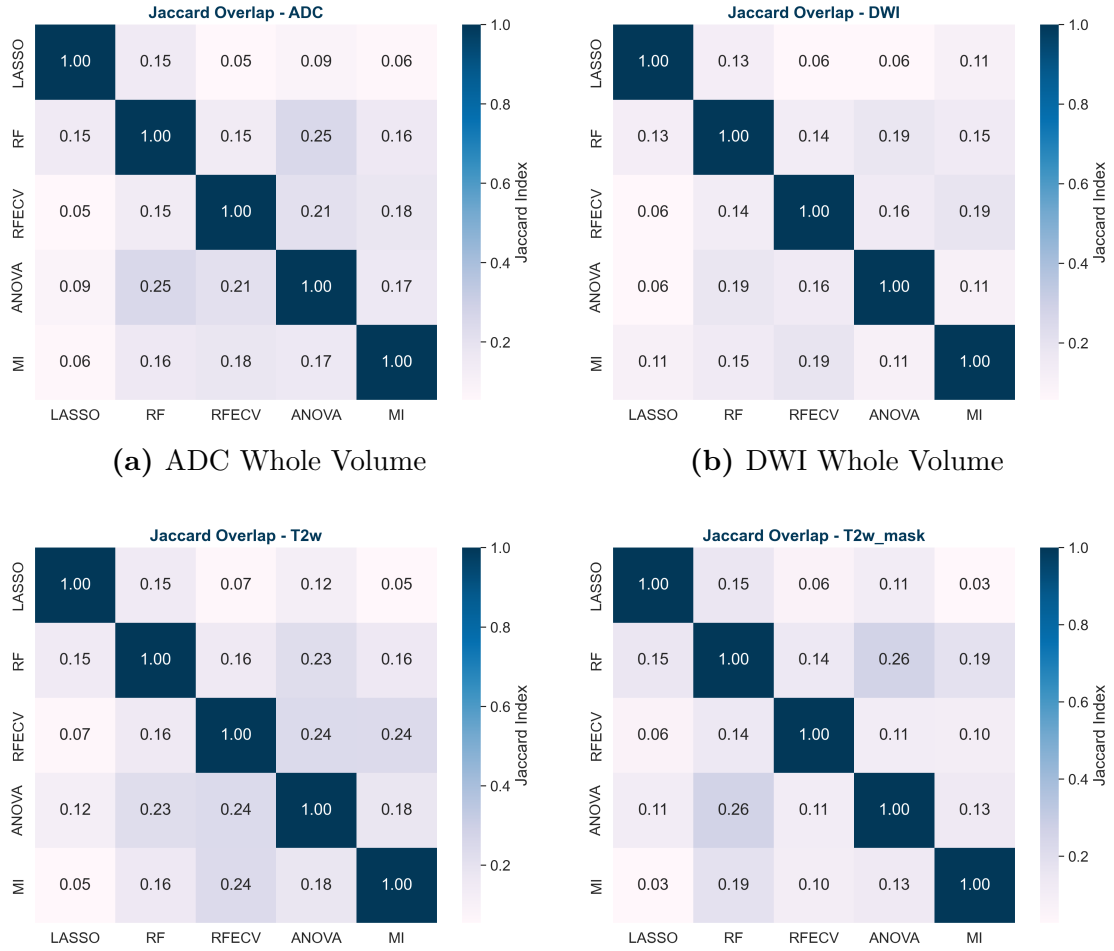
Per valutare il grado di accordo tra le diverse tecniche di selezione, è stato calcolato l'Indice di Jaccard (J). Tale metrica quantifica la similarità tra due insiemi di feature selezionati (A e B) come il rapporto tra la loro intersezione e la loro unione:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.2}$$

Questo indice varia da 0, nel caso in cui non si abbia nessuna feature in comune, a 1 in presenza di selezione identica. Le matrici di sovrapposizione risultanti mostrate in Figura 3.5 permettono di verificare se metodi diversi convergano verso le stesse variabili radiomiche o se, al contrario, estraggano informazioni complementari. Dall'analisi dei risultati emerge un valore medio di J generalmente basso, inferiore a 0.4, tra le diverse coppie di algoritmi. Questo risultato conferma la forte eterogeneità delle strategie di selezione adottate. Ogni metodo tende a identificare un sottoinsieme di feature distintivo, suggerendo che le diverse tecniche catturano aspetti complementari dell'informazione radiomica piuttosto che ridondanti.

3.3.2 Analisi della Composizione Tipologica: Fattore di Arricchimento

Oltre alla sovrapposizione delle singole feature, è stata analizzata la tipologia delle variabili selezionate. Poiché un semplice conteggio delle feature selezionate risulterebbe distorto a causa dello sbilanciamento intrinseco nella generazione delle feature è stato calcolato il Fattore di Arricchimento Per correggere questo bias di frequenza e identificare le famiglie realmente informative.



(a) ADC Whole Volume **(b)** DWI Whole Volume

(c) T2 Whole Volume **(d)** T2 Prostate Mask

Figura 3.5: Analisi della sovrapposizione tra i set di feature selezionati. I valori bassi (colore chiaro) indicano che metodi diversi selezionano variabili diverse, confermando la complementarità degli approcci.

Questa metrica statistica viene mutuata dalla bioinformatica per l'analisi della sovra-rappresentazione di categorie [46]. Per ogni famiglia f e metodo di selezione m, l'*Enrichment Factor* (EF) è definito come il rapporto tra la proporzione osservata nel sottoinsieme selezionato e la proporzione attesa nel dataset originale (frequenza di background):

$$EF_{f,m} = \frac{P_{selected}(f)}{P_{background}(f)} = \frac{\frac{N_{sel}(f)}{N_{sel}(tot)}}{\frac{N_{orig}(f)}{N_{orig}(tot)}} \tag{3.3}$$

Dove:

- $N_{sel}(f)$ è il numero di feature della famiglia f selezionate dal metodo;
- $N_{orig}(f)$ è il numero totale di feature della famiglia f disponibili nel dataset completo.

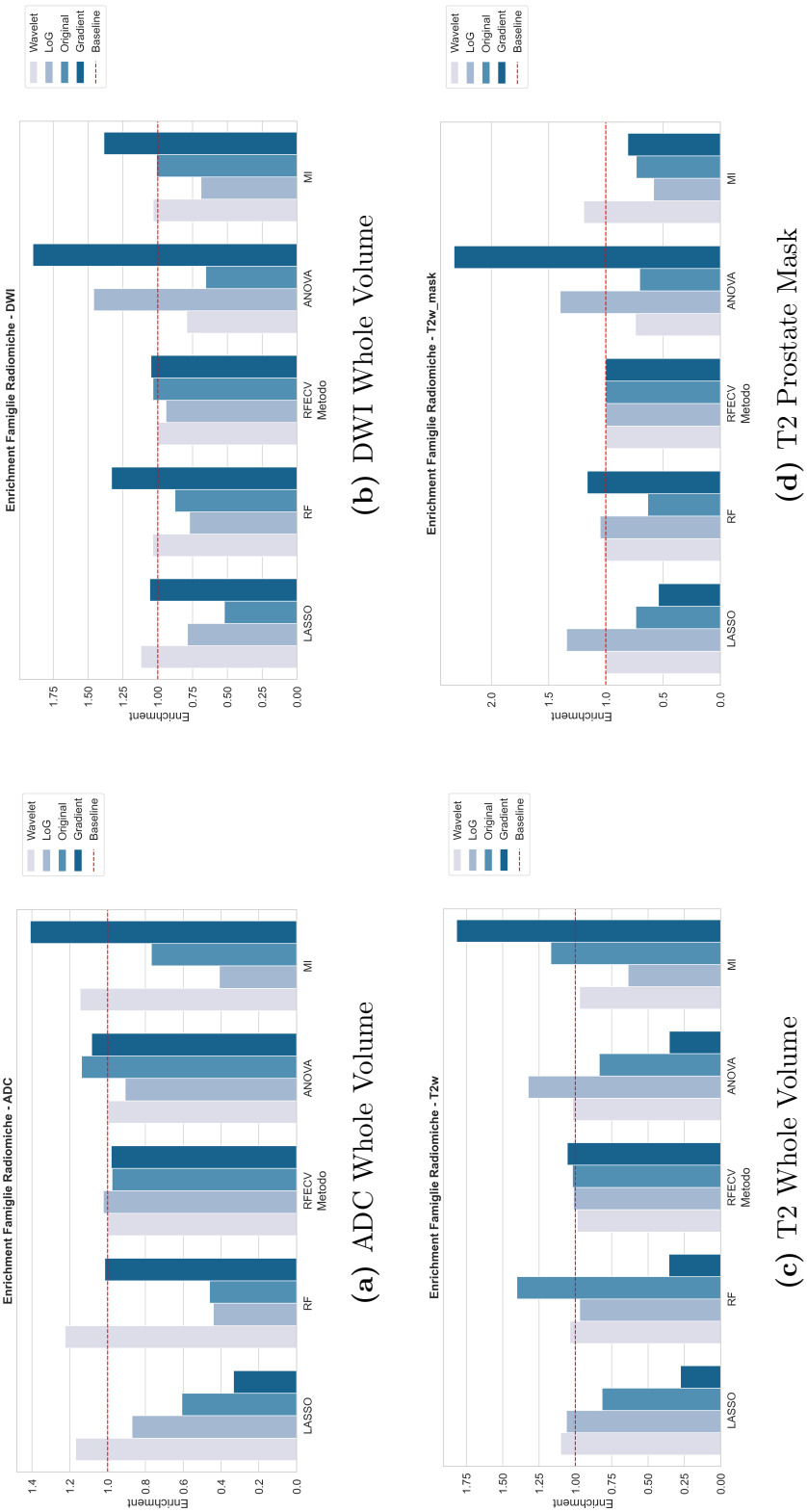
Valori prossimi all'unità ($EF \approx 1$) indicano una selezione neutra, che rispecchia la proporzione casuale di partenza. Al contrario, un $EF > 1$ denota una famiglia arricchita o sovra-rappresentata, in cui l'algoritmo seleziona le feature con una frequenza significativamente superiore al caso, rivelando una specifica preferenza per quel contenuto informativo. Viceversa, un $EF < 1$ caratterizza le famiglie impoverite o sotto-rappresentate, che vengono penalizzate o attivamente scartate dal metodo di selezione in quanto ritenute poco rilevanti.

I risultati di questa analisi per le quattro modalità sono riportati in Figura 3.6 dove emergono pattern distintivi. Le Trasformate Wavelet sebbene numericamente dominanti, presentano EF spesso vicino a 1 o leggermente superiore, indicando che la loro abbondanza è in parte dovuta al caso, ma in parte riflette un reale contenuto informativo nelle frequenze trasformate. Laplacian of Gaussian mostrano un forte arricchimento ($EF > 1.5$) specialmente nei metodi non lineari applicati alla maschera prostatica, confermando che la rilevazione di texture granulari è cruciale per la qualità dell'immagine in T2. Gradient, infine, risultano arricchite nella selezione tramite Mutual Information per le sequenze T2w, indicando che l'informazione sui bordi è altamente discriminante ma non linearmente correlata.

3.3.3 Impatto sulla Topologia dei Dati: t-SNE

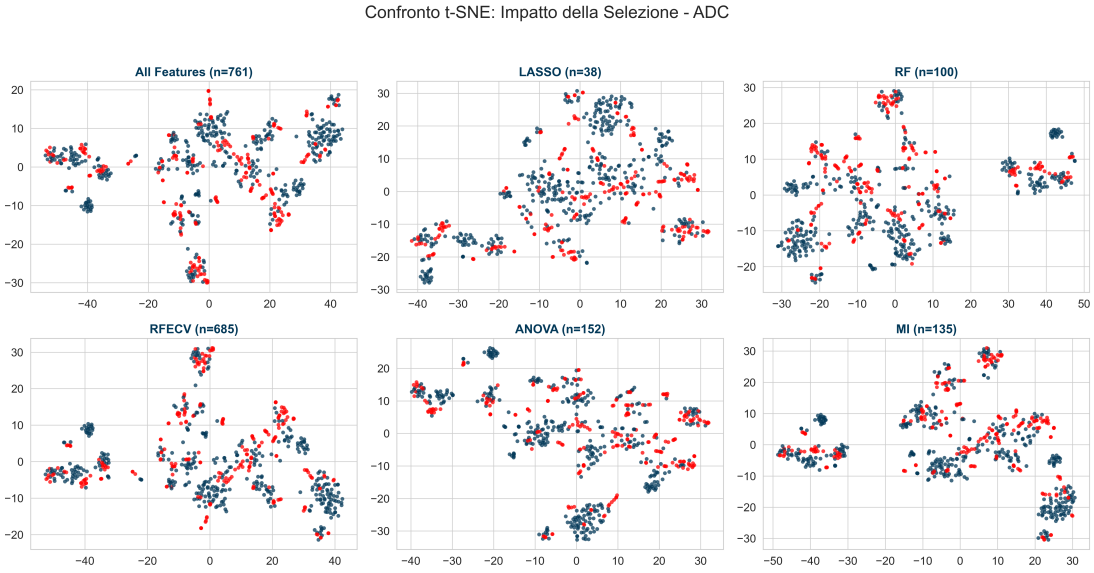
Infine, è stato verificato visivamente se la riduzione della dimensionalità preservasse la struttura dei dati. Le proiezioni t-SNE post-selezione sono state confrontate con quella originale come riportato nelle Figure 3.7 e 3.8. I metodi RFECV e ANOVA mantengono una topologia quasi identica all'originale, preservando la complessità globale. Al contrario, i metodi LASSO e RF Importance tendono a perdere parte della varianza globale, ma la struttura locale che separa le classi rimane visibile, suggerendo che le feature selezionate contengono effettivamente il nucleo dell'informazione discriminante.

Complessivamente, l'analisi delle diverse strategie di FS dimostra che non esiste un metodo di selezione perfetto a priori. Ogni algoritmo offre una prospettiva differente, enfatizzando aspetti complementari del segnale radiomico. Questa osservazione supporta l'approccio adottato nel capitolo successivo, in cui le diverse strategie di selezione vengono valutate direttamente nel contesto dei modelli di classificazione, consentendo una scelta guidata dalle prestazioni degli algoritmi.

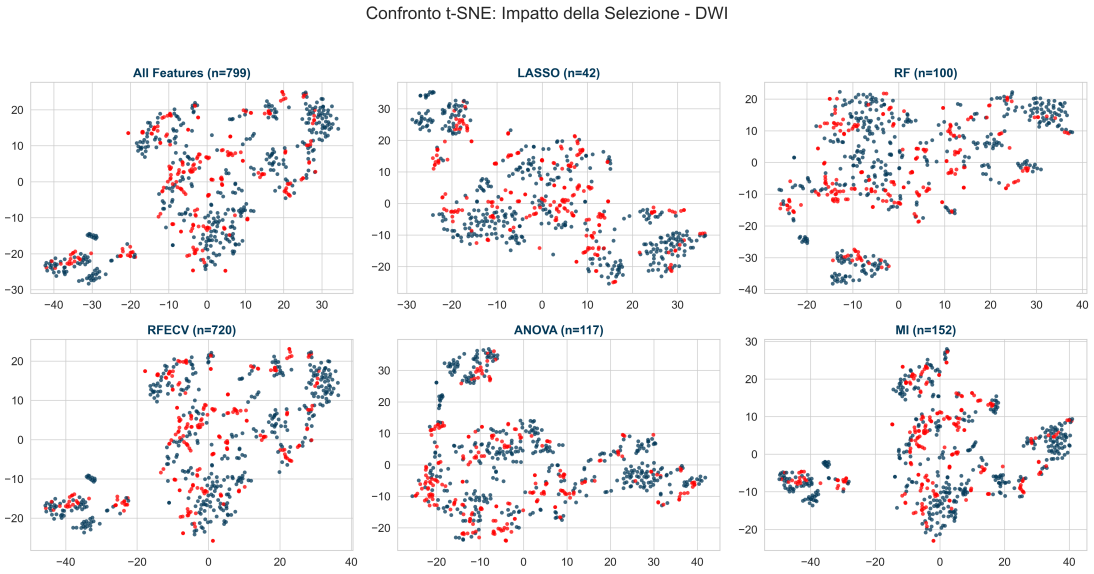


(a) ADC Whole Volume **(b)** DWI Whole Volume **(c)** T2 Whole Volume **(d)** T2 Prostate Mask

Figura 3.6: Analisi dell'Enrichment Factor. Le barre sopra la linea rossa ($EF = 1$) indicano famiglie di feature arrichhite dagli algoritmi. La famiglia Gradient presenta i valori di EF maggiori in diversi casi.

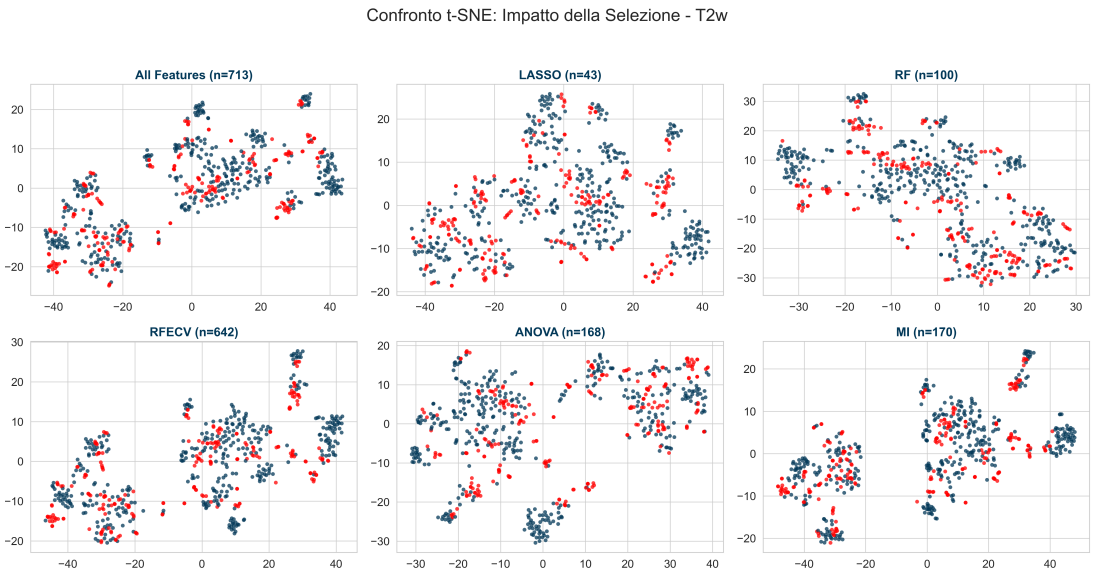


(a) ADC Whole Volume

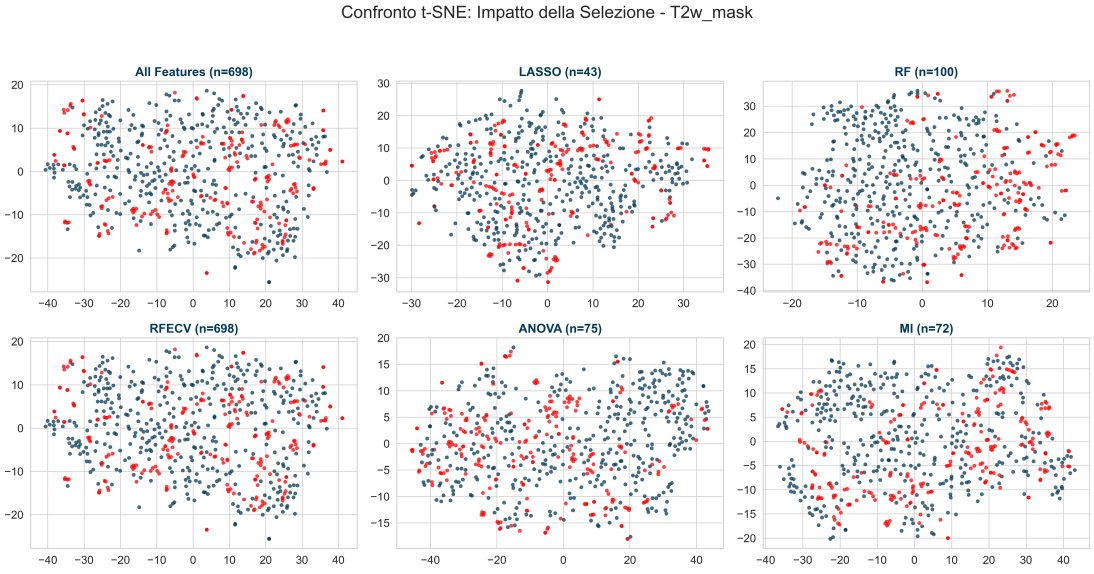


(b) DWI Whole Volume

Figura 3.7: Confronto della struttura dello spazio dei dati (t-SNE) prima e dopo la selezione delle feature per le sequenze di diffusione. I pannelli mostrano la proiezione originale in alto a sinistra in ogni griglia e quelle ottenute con i sottoinsiemi di feature selezionati dai vari metodi.



(a) T2 Whole Volume



(b) T2 Prostate Mask

Figura 3.8: Confronto della struttura dello spazio dei dati (t-SNE) per le sequenze morfologiche T2-weighted. Il confronto permette di valutare la conservazione della struttura globale dei dati attraverso le diverse strategie di selezione.

Capitolo 4

Classificazione e Valutazione dei Modelli

Dopo aver analizzato e valutato i sottoinsiemi di feature radiomiche più promettenti attraverso le tecniche di riduzione della dimensionalità discusse nel Capitolo 3, l'obiettivo di questa fase è tradurre tale contenuto informativo in predizioni diagnostiche accurate. In questo capitolo viene descritto lo sviluppo, l'addestramento e la validazione di modelli di Machine Learning (ML) per la classificazione automatica della qualità delle immagini RM prostatiche.

L'approccio sperimentale adottato compara le performance di diversi classificatori lineari e non lineari addestrati sui vari set di feature selezionate dai diversi metodi di FS. Questa strategia multiparametrica ha lo scopo di identificare la combinazione ottimale che massimizzi la capacità di distinguere tra scansioni diagnostiche e non diagnostiche, garantendo al contempo robustezza e generalizzabilità.

Ogni modello è stato addestrato e ottimizzato tramite 5-fold Cross-Validation all'interno del Construction Set, applicando dinamicamente in ogni fold le procedure di scalatura, oversampling (SMOTE) e selezione delle feature per prevenire ogni forma di *data leakage*. Il Test Set è stato mantenuto rigorosamente isolato, fungendo da banco di prova finale mai visto dai modelli durante lo sviluppo. Le metriche conclusive qui riportate, calcolate su questo set indipendente, forniscono pertanto una stima imparziale della reale capacità di generalizzazione del sistema su nuovi pazienti.

4.1 Metriche di Valutazione delle Performance

La valutazione dei modelli predittivi non si limita alla sola accuratezza globale, spesso fuorviante in presenza di dataset sbilanciati, ma ha esaminato un pannello di indicatori statistici derivati dalla Matrice di Confusione. Considerando la natura

binaria del problema, le classi sono state definite come segue: la Classe Positiva (1) rappresenta i casi "Diagnostici" mentre la Classe Negativa (0) quelli "Non Diagnostici". Di conseguenza, gli errori di classificazione assumono un peso clinico asimmetrico. La criticità maggiore è rappresentata dai Falsi Positivi, ovvero quelle immagini tecnicamente inadeguate che il sistema classifica erroneamente come "Diagnostiche", esponendo il radiologo al rischio di formulare diagnosi basate su dati corrotti o artefatti. Di contro, i Falsi Negativi portano allo scarto di un'immagine valida e incidono negativamente sull'efficienza del flusso di lavoro, portando a richieste di ri-acquisizioni non strettamente necessarie.

Sulla base di queste premesse, sono state adottate le seguenti metriche chiave:

1. **Balanced Accuracy (BA):** Metrica primaria utilizzata per l'ottimizzazione degli iperparametri e il ranking dei modelli. È definita come la media aritmetica tra Sensibilità e Specificità:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{4.1}$$

   A differenza dell'accuratezza tradizionale, la BA non è influenzata dalla frequenza delle classi, garantendo che il modello non predica sempre la classe maggioritaria.

2. **Specificità (True Negative Rate):** Misura la capacità del modello di identificare correttamente la Classe 0 ("Non Diagnostico"):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4.2}$$

   In questo contesto, la Specificità è un indicatore di sicurezza clinica. Massimizzare questo valore, infatti, significa minimizzare i Falsi Positivi, ovvero impedire che immagini di scarsa qualità superino il filtro di controllo.

3. **Sensibilità (Recall):** Misura la capacità di riconoscere correttamente la Classe 1 (Diagnostico):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.3}$$

   Un valore elevato indica che il sistema è efficiente nel preservare le immagini valide senza scartarle inutilmente.

4. **Area Under the Curve (ROC-AUC):** È una misura robusta della capacità di separazione globale del modello, indipendente dalla soglia di decisione scelta.

5. **Positive Predictive Value (PPV o Precision):** Rappresenta la probabilità che un'immagine classificata come "Diagnostica" dal sistema lo sia realmente.

$$PPV = \frac{TP}{TP + FP} \tag{4.4}$$

   Un alto PPV è fondamentale per il radiologo, in quanto garantisce che il flusso di lavoro automatizzato non sia inquinato da immagini di scarsa qualità erroneamente validate.

6. **Negative Predictive Value (NPV):** Rappresenta la probabilità che un'immagine scartata come "Non Diagnostica" sia effettivamente inutilizzabile.

$$NPV = \frac{TN}{TN + FN} \tag{4.5}$$

   Questo valore è critico per il tecnico di radiologia poiché un basso NPV implicherebbe troppi FN, costringendo a ripetizioni inutili dell'esame.

4.2 Modelli Supervisionati Utilizzati

Per garantire una valutazione esaustiva e indipendente dallo specifico bias induttivo di un singolo metodo, sono stati impiegati cinque algoritmi di apprendimento supervisionato. La selezione include tecniche di Ensemble *Bagging e Boosting*, metodi a Kernel e Reti Neurali, rappresentando lo stato dell'arte per la classificazione di dati tabulari complessi.

- **Random Forest (RF):** Tecnica di *Bagging* che costruisce parallelamente un elevato numero di alberi decisionali decorrelati. Ogni albero viene addestrato su un sottoinsieme casuale dei dati (*bootstrap*) e delle feature. La predizione finale avviene per votazione di maggioranza, riducendo drasticamente la varianza e il rischio di overfitting rispetto ai singoli alberi decisionali. È stato scelto per la sua robustezza al rumore e la capacità di gestire feature ridondanti [43].
- **Support Vector Machine (SVM):** Algoritmo che mappa i vettori di input in uno spazio dimensionale superiore tramite *Kernel Trick* per individuare l'iperpiano che massimizza il margine di separazione tra le classi. Le SVM sono particolarmente efficaci in scenari ad alta dimensionalità [47].
- **eXtreme Gradient Boosting (XGBoost):** Algoritmo di *Boosting* scalabile che costruisce gli alberi in modo sequenziale permettendo ad ogni nuovo modello di correggere gli errori residui commessi dai precedenti. XGBoost integra tecniche avanzate di regolarizzazione L_1 e L_2 per prevenire l'overfitting e gestisce nativamente i valori mancanti, risultando spesso il modello più performante su dati strutturati [48].

- **Light Gradient Boosting Machine (LightGBM):** Evoluzione del Gradient Boosting ottimizzata per l'efficienza computazionale su grandi dataset. Utilizza una strategia di crescita degli alberi *leaf-wise* invece che *level-wise* e algoritmi basati su istogrammi per discretizzare le feature continue. Questo approccio garantisce una velocità di addestramento superiore mantenendo un'elevata accuratezza predittiva [49].
- **Multilayer Perceptron (MLP):** Modello di rete neurale artificiale *feed-forward* composto da almeno tre strati di nodi: uno strato di input, uno o più strati nascosti (*hidden layers*) e uno strato di output. A differenza dei modelli lineari, l'MLP può distinguere dati non linearmente separabili grazie all'uso di funzioni di attivazione non lineari (es. ReLU) e all'algoritmo di retro-propagazione dell'errore (*Backpropagation*) per l'aggiornamento dei pesi [50].

4.3 Strategie di Ottimizzazione degli Iperparametri

Per massimizzare la capacità predittiva di ogni classificatore ed evitare una selezione arbitraria dei parametri, il processo di addestramento è stato integrato con tecniche di ottimizzazione sistematica. Invece di utilizzare configurazioni statiche, ogni fold della Cross-Validation ha incluso una fase di ricerca interna per identificare i parametri ottimali per quello specifico sottoinsieme di dati. L'obiettivo di questa procedura è garantire che ogni modello operi nella sua configurazione più performante prima di essere valutato sul fold di validazione esterno. Sono stati implementati e confrontati due approcci distinti:

4.3.1 Grid Search (Esplorazione Esaustiva)

L'approccio basato su Grid Search definisce una griglia predeterminata di valori per i parametri chiave di ogni modello. Il sistema valuta ogni possibile combinazione all'interno di questo spazio tramite una procedura di *Stratified Shuffle Split* interna al fold, garantendo l'identificazione del miglior set di parametri tra quelli proposti [51]. Nello specifico, la configurazione della griglia ha previsto:

- **Random Forest:** Ottimizzazione del numero di alberi ($n\_estimators \in [100, 200]$), della profondità massima ($max\_depth \in [None, 5, 10]$) e dei campioni minimi per foglia o split.
- **Support Vector Machine:** Ricerca del miglior compromesso di regolarizzazione attraverso il parametro $C \in [0.1, 1, 10]$ e diverse scale per il coefficiente del kernel γ.

- **XGBoost e LightGBM:** Ottimizzazione del *learning rate*, della profondità degli alberi e del *subsampling* per migliorare la generalizzazione.

- **Multilayer Perceptron:** Test di diverse architetture degli strati nascosti, tra cui configurazioni a singolo strato $(50,), (100,)$ e a doppio strato $(50, 50)$, unitamente al parametro di regolarizzazione α.

4.3.2 Optuna (Ottimizzazione Bayesiana)

In alternativa alla ricerca a griglia, è stata implementata l'ottimizzazione iperparametrica tramite Optuna, un framework basato su principi bayesiani [52]. A differenza della Grid Search, che procede per tentativi discreti, Optuna utilizza il campionamento *TPE (Tree-structured Parzen Estimator)* per apprendere dinamicamente dai risultati dei tentativi precedenti, restringendo progressivamente lo spazio di ricerca verso le regioni che massimizzano la Balanced Accuracy.

Nello specifico, per ogni modello è stato definito un set di iperparametri con i relativi intervalli di ricerca:

- **Random Forest:** per garantire la robustezza dell'ensemble, il numero di stimatori è stato campionato tra 100 e 1000 alberi, mentre la profondità massima è stata esplorata in un intervallo tra 5 e 50 livelli per prevenire l'overfitting. La qualità della separazione dei nodi è stata infine analizzata confrontando i criteri di impurità *'gini'* ed *'entropy'*.

- **Support Vector Machine:** l'ottimizzazione si è concentrata sul bilanciamento tra il margine di errore e la complessità, campionando il parametro di regolarizzazione C in scala logaritmica tra 0.1 e 100. Parallelamente, sono state valutate diverse proiezioni spaziali tramite kernel di tipo lineare, polinomiale, RBF e sigmoide, variando il coefficiente γ tra le opzioni *'scale'* e *'auto'*.+2

- **XGBoost e LightGBM:** l'attenzione è stata rivolta all'efficienza del boosting campionando il *learning rate* logaritmicamente tra 0.01 e 0.3. La complessità strutturale è stata regolata agendo sul numero di foglie e sulla profondità massima, mentre per incrementare la robustezza è stata esplorata la frazione di feature campionate (*colsample_bytree*) tra 0.5 e 1.0.

- **Multi-Layer Perceptron:** l'architettura della rete è stata definita testando configurazioni a uno o due strati nascosti con un numero di neuroni compreso tra 50 e 200. Per la stabilità dei pesi, il parametro di regolarizzazione $L2$ (α) è stato campionato tra 10^{-5} e 10^{-2}, mentre il tasso di apprendimento iniziale è stato esplorato nell'intervallo $[10^{-4}, 10^{-2}]$.

Ogni ottimizzazione è stata condotta per 100 iterazioni e applicando la tecnica di *pruning* per interrompere precocemente i trial non promettenti, ottimizzando così il tempo computazionale.

4.4 Valutazione dei Classificatori Ottimizzati

Nelle sezioni successive verranno analizzati in dettaglio i classificatori e i risultati migliori ottenuti per ogni combinazione Classificatore - Metodo di Feature Selection estratti al termine del ciclo di 5-fold Cross-Validation. Per completezza, i risultati completi di tutte le configurazioni esplorate sono riportati in Appendice A.

Prima di procedere all'analisi di dettaglio dei singoli algoritmi, è utile anticipare una tendenza generale emersa durante le sperimentazioni. Le performance di classificazione si sono assestate in un intervallo di Balanced Accuracy compreso prevalentemente tra il 55% e il 60%. In un dominio intrinsecamente complesso e rumoroso come la radiomica, dove il confine tra un'immagine diagnostica e una non diagnostica è spesso sfumato e soggetto a elevata variabilità inter-operatore, tali valori rappresentano un risultato predittivo atteso, seppur non definitivo. Nel complesso, i modelli dotati di maggiore flessibilità geometrica come le reti neurali hanno mostrato una lieve, ma costante, superiorità rispetto agli ensemble puramente *tree-based*, suggerendo che l'informazione sulla qualità dell'immagine risieda in relazioni multivariate complesse e non puramente gerarchiche.

4.4.1 Random Forest (RF)

La scelta di questo modello come punto di partenza è motivata dalla sua naturale robustezza nel gestire spazi ad alta dimensionalità e dalla capacità di tollerare la presenza di feature rumorose o ridondanti, caratteristiche tipiche dei dataset radiomici [43].

Per contrastare strutturalmente lo sbilanciamento delle classi, in tutte le configurazioni è stato mantenuto attivo il parametro di pesatura dinamica, che istruisce l'algoritmo a penalizzare maggiormente gli errori commessi sulla classe minoritaria ("Non Diagnostico").

Il confronto diretto tra i due framework di ottimizzazione riportato in Tabella 4.1 non rivela una superiorità assoluta dell'approccio bayesiano con Optuna rispetto alla ricerca esaustiva Grid Search. I due metodi si alternano infatti nell'individuazione dell'ottimo a seconda del dominio di imaging. Nelle sequenze ADC e T2w Mask, l'ottimizzazione dinamica di Optuna ha esplorato più efficacemente lo spazio degli iperparametri, ottenendo le performance migliori in sinergia con filtri di selezione strutturati RF-Importance e LASSO. Nelle sequenze DWI e T2w, al contrario, la griglia discreta della Grid Search ha intercettato configurazioni lievemente superiori,

portando il modello al suo picco prestazionale (0.593 ± 0.046 su T2w) utilizzando le feature selezionate tramite RF-Importance.

In conclusione, il Random Forest si configura come un classificatore stabile. Tuttavia, il lieve bias intrinseco tipico del bagging sembra limitare la capacità del modello di filtrare in modo aggressivo i falsi positivi, giustificando l'esplorazione successiva di metodi con paradigmi differenti.

Tabella 4.1: Performance medie del classificatore Random Forest. Si osserva un'alternanza di efficacia tra i due framework di ottimizzazione a seconda della modalità MRI analizzata.

Modalità	**Feature Selection**	**Ottimizzatore**	**Balanced Accuracy**
ADC	RF-Importance	Optuna	0.571 ± 0.042
DWI	ANOVA	Grid Search	0.577 ± 0.032
T2w (Whole)	RF-Importance	Grid Search	0.593 ± 0.046
T2w (Mask)	LASSO	Optuna	0.559 ± 0.046

4.4.2 Support Vector Machine (SVM)

Il secondo algoritmo analizzato è la Support Vector Machine, un modello che opera proiettando i dati in uno spazio multidimensionale per individuare l'iperpiano che massimizza il margine di separazione tra le classi (*Support Vectors*) [47]. L'utilizzo dell'SVM è particolarmente indicato in radiomica, data la sua naturale tolleranza a scenari in cui il numero di feature estratte è comparabile o superiore al numero di osservazioni cliniche disponibili.

A differenza dei modelli basati su alberi, l'SVM è estremamente sensibile alla scala e alla distribuzione geometrica dei dati. Pertanto, la fase di ottimizzazione iperparametrica si è concentrata sulla selezione dinamica della funzione Kernel testando il tipo Lineare, RBF, Polinomiale, Sigmoide. In concomitanza con la scelta del Kernel, l'ottimizzazione ha calibrato il parametro di regolarizzazione C in scala logaritmica, cercando il compromesso ideale tra la massimizzazione del margine e la tolleranza agli errori sul set di addestramento. Anche per questo classificatore è stato forzato il parametro il parametro di pesatura dinamica.

Il confronto tra le due strategie di tuning, presentato in Tabella 4.2, rivela che l'approccio esaustivo basato su Grid Search ha superato sistematicamente l'ottimizzazione bayesiana in tutte le modalità di imaging. Questo fenomeno suggerisce che, per spazi iperparametrici fortemente discreti e non convessi, una griglia fissa ben strutturata riesca a garantire maggiore stabilità rispetto alla ricerca

probabilistica TPE, la quale rischia di convergere prematuramente verso minimi locali.

L'analisi dei risultati aggregati mostra che l'SVM si attesta su una Balanced Accuracy media compresa tra il 55.1% e il 56.2%. In questa pipeline, l'SVM garantisce una classificazione geometricamente robusta ma con prestazioni globali lievemente inferiori rispetto al RF. Le configurazioni migliori confermano inoltre la forte necessità di applicare all'SVM dei filtri di feature selection parsimoniosi e basati su test statistici pregressi come MI e ANOVA.

Tabella 4.2: Performance medie del classificatore SVM. Per questo algoritmo, caratterizzato da uno spazio iperparametrico dominato dalla scelta discreta del Kernel, la Grid Search si è rivelata metodologicamente più solida dell'ottimizzazione bayesiana.

Modalità	**Feature Selection**	**Ottimizzatore**	**Balanced Accuracy**
ADC	Mutual Information	Grid Search	0.562 ± 0.048
DWI	ANOVA	Grid Search	0.559 ± 0.018
T2w (Whole)	Mutual Information	Grid Search	0.551 ± 0.047
T2w (Mask)	ANOVA	Grid Search	0.562 ± 0.039

4.4.3 eXtreme Gradient Boosting (XGBoost)

Tra i diversi metodi di *Boosting*, l'analisi è proseguita testando XGBoost. Questo algoritmo, a differenza degli approcci di *bagging*, costruisce il modello in modo sequenziale. Ogni nuovo albero decisionale viene addestrato con l'obiettivo di correggere gli errori residui commessi dai modelli precedenti, avvalendosi di meccanismi di regolarizzazione L_1 e L_2 per mitigare l'overfitting [48].

Per gestire il problema dello sbilanciamento delle classi, si è sfruttato il parametro interno *scale_pos_weight*, che modifica la *loss function* assegnando una penalità maggiore agli errori commessi sulla classe minoritaria. Questo peso è stato ricalcolato dinamicamente all'interno di ogni fold, utilizzando il rapporto esatto tra i campioni negativi e positivi.

Il confronto tra i framework di tuning evidenzia una chiara preferenza per l'ottimizzazione bayesiana. L'algoritmo Optuna si è rivelato infatti lo strumento ideale per navigare in modo efficiente lo spazio continuo e complesso dei parametri di boosting come il learning rate e la profondità degli alberi.

L'analisi delle performance riportata in Tabella 4.3 mostra come XGBoost si posizioni in una fascia prestazionale del tutto paragonabile ai modelli precedenti,

registrando valori medi di Balanced Accuracy tra il 55.5% e il 57%. Si evince inoltre una marcata sensibilità nella scelta del metodo di Feature Selection a seconda della modalità in esame. Sulle sequenze ADC, il modello raggiunge il suo massimo (0.570 ± 0.011) in combinazione con la selezione ricorsiva RFECV, mentre sulle sequenze anatomiche T2-weighted e sulla DWI, l'algoritmo premia l'analisi ANOVA.

Tabella 4.3: Performance medie del classificatore XGBoost. Optuna dimostra un'elevata efficacia nella calibrazione dei parametri continui tipici degli algoritmi di Gradient Boosting.

Modalità	**Feature Selection**	**Ottimizzatore**	**Balanced Accuracy**
ADC	RFECV	Optuna	0.570 ± 0.011
DWI	ANOVA	Optuna	0.555 ± 0.046
T2w (Whole)	ANOVA	Optuna	0.566 ± 0.024
T2w (Mask)	RF-Importance	Optuna	0.556 ± 0.058

4.4.4 Light Gradient Boosting Machine (LightGBM)

L'ultimo esponente della famiglia degli alberi decisionali testato è LightGBM, un algoritmo progettato per massimizzare l'efficienza computazionale su dataset complessi [49]. La peculiarità di LightGBM risiede nella sua strategia di crescita degli alberi *leaf-wise*, che seleziona ed espande il nodo con la massima riduzione della loss, a differenza della più simmetrica crescita *level-wise* tipica del Random Forest. Questo approccio tende a convergere più rapidamente, ma richiede una calibrazione rigorosa per prevenire l'overfitting.

L'analisi comparativa tra i framework di ottimizzazione in Tabella 4.4 smentisce la presunta superiorità assoluta dell'approccio bayesiano sui modelli di boosting, rivelando uno scenario ibrido. Nelle sequenze ADC e sulla maschera prostatica, Optuna ha intercettato le combinazioni migliori bilanciando dinamicamente il learning rate e la complessità dei nodi foglia. Al contrario, per le mappe di diffusione DWI e sull'intera sequenza T2w, Grid Search ha consentito al classificatore di registrare i suoi picchi prestazionali, suggerendo che in certi contesti radiomici uno spazio di esplorazione rigidamente delimitato offra maggiori garanzie.

Complessivamente, le prestazioni globali di LightGBM si attestano in una fascia intermedia con una Balanced Accuracy compresa tra il 55.4% e il 57.6%. Come ipotizzato per gli altri ensemble basati su alberi decisionali, la rigida logica gerarchica mostra limiti strutturali nel mappare la complessità delle texture non diagnostiche in questo specifico dominio.

Tabella 4.4: Performance medie del classificatore LightGBM. L'analisi mostra un sostanziale pareggio tra le strategie di tuning, ribadendo l'utilità metodologica di affiancare l'esplorazione esaustiva a quella probabilistica.

Modalità	Feature Selection	Ottimizzatore	Balanced Accuracy
ADC	Mutual Information	Optuna	0.565 ± 0.038
DWI	ANOVA	Grid Search	0.558 ± 0.043
T2w (Whole)	Mutual Information	Grid Search	0.576 ± 0.028
T2w (Mask)	RF-Importance	Optuna	0.556 ± 0.067

4.4.5 Multi-Layer Perceptron (MLP)

A completamento dell'indagine sperimentale, è stato valutato un modello basato su Reti Neurali Artificiali: il Multi-Layer Perceptron. L'utilizzo di reti *feed-forward* completamente connesse dota il sistema della flessibilità necessaria per astrarre rappresentazioni gerarchiche altamente non lineari dai dati radiomici, svincolandosi concettualmente dai limiti imposti dai partizionamenti lineari o ad albero [50].

La topologia della rete è stata definita dinamicamente all'interno della Cross-Validation. Attraverso le routine di Grid Search e Optuna, sono state esplorate architetture variabili da singoli strati a 50 o 100 neuroni, a configurazioni più profonde a doppio dense layer. Contestualmente, per stabilizzare l'ottimizzatore Adam ed evitare la rapida memorizzazione del rumore sui piccoli campioni diagnostici, sono stati calibrati il learning rate iniziale e il coefficiente di regolarizzazione L_2 (α).

L'analisi dei risultati in Tabella 4.5 conferma la validità di questo approccio geometricamente flessibile. L'MLP si è rivelato il modello mediamente più performante dello studio, registrando una Balanced Accuracy media di 0.604 ± 0.025 sulla modalità ADC.

Sulle sequenze T2-weighted, l'esplorazione bayesiana di Optuna ha individuato le configurazioni più solide, stabilizzandosi su una media del 58.3% sulla maschera. Nelle mappe di diffusione, invece, l'esplorazione esaustiva della Grid Search ha garantito una stabilità previsionale leggermente superiore.

Un dato di notevole rilevanza algoritmica emerge infine dall'accoppiamento con le tecniche di Feature Selection. In tre modalità su quattro, la rete neurale ha espresso il suo potenziale massimo elaborando i dati filtrati tramite Mutual Information. Tale ricorrenza ratifica l'esistenza di una forte sinergia tra la capacità di mappatura non lineare propria dell'MLP e le metriche di selezione basate sulla Teoria dell'Informazione.

Tabella 4.5: Performance medie del classificatore MLP. La flessibilità non lineare della rete neurale ha permesso di ottenere le capacità separative più elevate dello studio, dimostrando una notevole affinità con i filtri basati su MI.

Modalità	Feature Selection	Ottimizzatore	Balanced Accuracy
ADC	Mutual Information	Grid Search	0.604 ± 0.025
DWI	LASSO	Grid Search	0.590 ± 0.030
T2w (Whole)	Mutual Information	Optuna	0.574 ± 0.061
T2w (Mask)	Mutual Information	Optuna	0.583 ± 0.071

4.5 Sintesi Comparativa per Modalità

Al termine della fase di sperimentazione e ottimizzazione iperparametrica, è possibile tracciare un bilancio definitivo per identificare le migliori configurazioni per ciascuna delle quattro modalità di imaging analizzate. La selezione dei migliori classificatori è stata guidata dal valore medio di Balanced Accuracy ottenuto in 5-fold Cross-Validation, scegliendo tra i 5 fold, l'iterazione con valore di Balanced Acuuracy più elevato.

Per offrire una visione analitica dettagliata, i risultati sono stati suddivisi in base alla strategia di ottimizzazione adottata. In Tabella 4.6 vengono riassunti i vincitori identificati tramite Grid Search. Si osserva un netto predominio tecnico del MLP, che domina in tre modalità su quattro. La rete neurale raggiunge i picchi assoluti sulle mappe di diffusione, toccando il 64.1% su ADC con selezione Mutual Information e il 62.5% su DWI con selezione LASSO. Sull'intera sequenza anatomica T2w, il Random Forest mantiene il primato con un picco di 64.7%, confermando che per dati volumetrici strutturati la logica di bagging rimane estremamente competitiva.

In Tabella 4.7, invece, vengono mostrati i risultati ottenuti tramite ottimizzazione bayesiana dinamica dove si ottiene il risultato migliore per la maschera prostatica. Grazie all'esplorazione fine di Optuna, l'MLP combinato con la MI ha raggiunto il punteggio più alto dello studio sperimentale con una Balanced Accuracy del 67.5% nel miglior fold. Nelle modalità anatomiche DWI e T2w, Optuna ha favorito il Random Forest, portandolo a picchi rispettivamente del 61.8% e 64.3%.

In conclusione, la sintesi dei picchi prestazionali conferma che non esiste un algoritmo universale che vada bene per tutte le modalità, ma MLP e Random Forest, ottimizzati con i diversi framework, forniscono le prestazioni migliori sul dataset dello studio in oggetto. D'altra parte non emerge un chiaro pattern per quanto riguarda la migliore strategia di Feature Selection, in quanto ogni classificatore

Tabella 4.6: Sintesi delle migliori configurazioni identificate tramite Grid Search. I valori riportati indicano la Balanced Accuracy massima ottenuta nel miglior fold di validazione.

Modalità	Modello	Feature Selection	Balanced Accuracy (Best Fold)
ADC	MLP	Mutual Information	0.641
DWI	MLP	LASSO	0.625
T2w (Whole)	Random Forest	RF-Importance	0.647
T2w (Mask)	MLP	RFECV	0.649

Tabella 4.7: Sintesi delle migliori configurazioni identificate tramite Optuna. Spicca il risultato dell'MLP sulla maschera prostatica, con un valore di Balanced Accuracy pari a 67.5%.

Modalità	Modello	Feature Selection	Balanced Accuracy (Best Fold)
ADC	MLP	ANOVA	0.602
DWI	Random Forest	ANOVA	0.618
T2w (Whole)	Random Forest	RFECV	0.643
T2w (Mask)	MLP	Mutual Information	0.675

predilige una tecnica diversa.

4.6 Strategie di Ensemble Learning

Al termine della valutazione dei singoli classificatori, la ricerca si è orientata verso tecniche di *Ensemble Learning* con l'obiettivo di combinare le predizioni delle diverse modalità MRI. Sono state implementate e confrontate due architetture di fusione distinte: lo Stacking Generalization e il Soft Voting.

4.6.1 Stacking Generalization

La prima strategia testata è stata lo Stacking, strutturato su due livelli funzionali:

- **Level-0 (Base Learners):** Per ciascuna delle quattro modalità, è stato selezionato il classificatore vincitore identificato nelle sezioni precedenti riassunti in Tabella 4.6 per la pipeline Grid Search e in Tabella 4.7 per Optuna. Questi modelli agiscono come estrattori di meta-feature, trasformando i dati radiomici grezzi in probabilità di qualità ($P_{ADC}, P_{DWI}, P_{T2w}, P_{Mask}$). È fondamentale

sottolineare che queste probabilità provengono dalle predizioni OOF, garantendo che il meta-modello non veda mai dati su cui i base learner sono stati addestrati.

- **Level-1 (Meta-Learner):** Il classificatore di secondo livello è stato implementato utilizzando una Regressione Logistica configurata con pesatura delle classi per gestire la natura sbilanciata del dataset [53]. Per garantire una valutazione imparziale e prevenire il fenomeno del *meta-overfitting*, l'addestramento è avvenuto adottando una strategia di (*Internal Cross-Validation*) utilizzando la funzione `cross_val_predict` con schema *Stratified K-Fold* ($K = 5$). Questa procedura ha permesso di generare un vettore di probabilità di stacking (P_{stack}) per l'intero dataset di sviluppo, dove ogni singola predizione è stata ottenuta da un modello addestrato esclusivamente sui $K - 1$ fold restanti. In tal modo, le probabilità risultanti sono *unbiased*, rappresentando la reale capacità del meta-modello di generalizzare su dati non visti. Su tali probabilità è stata successivamente eseguita una procedura di *Threshold Tuning*. L'algoritmo ha iterato attraverso un range di possibili soglie decisionali $\tau \in [0.20, 0.80]$ con step di 0.01, selezionando il valore τ^* che massimizzasse la Balanced Accuracy. Questo approccio ha portato all'identificazione delle soglie operative di 0.37 per la pipeline Grid Search e 0.49 per Optuna.

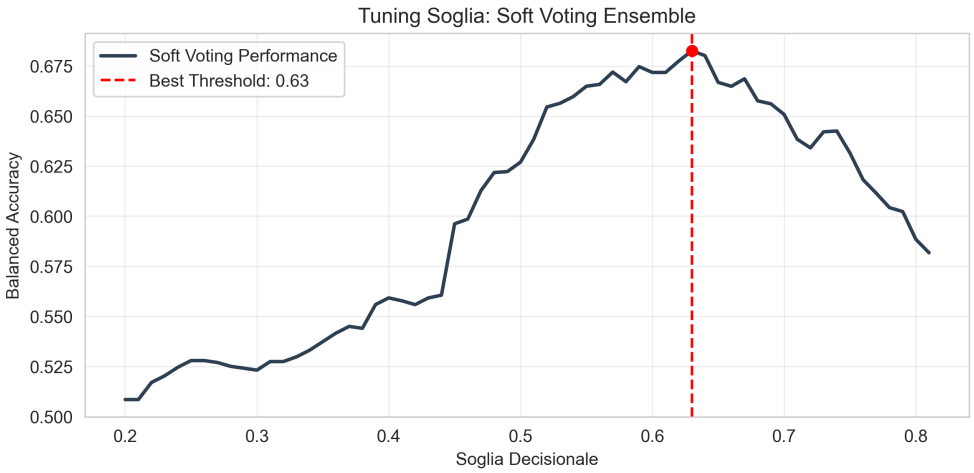
4.6.2 Soft Voting Ensemble

In alternativa all'approccio gerarchico, è stata implementata una strategia di Soft Voting. In questo schema, la probabilità finale di diagnosticità per il paziente i è calcolata come la media aritmetica delle confidenze predette dai quattro migliori classificatori base (P_m):

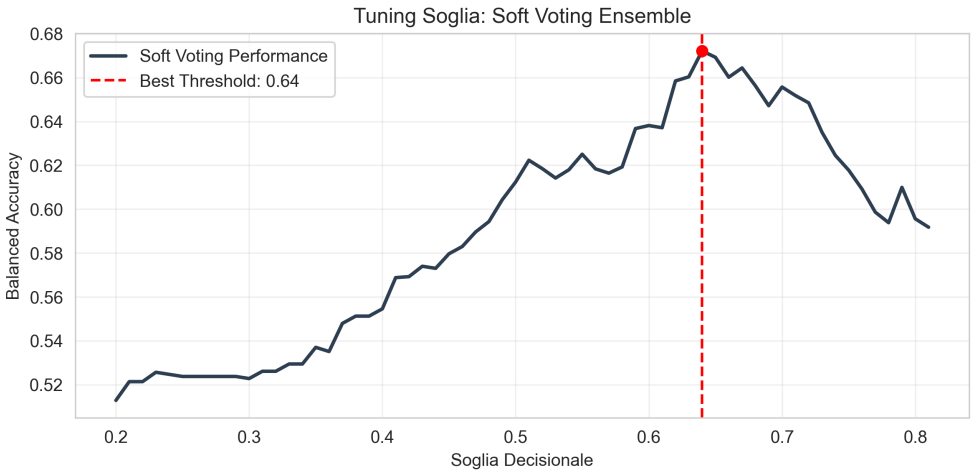
$$P_{ensemble}(x_i) = \frac{1}{M} \sum_{m=1}^{M} P_m(x_i) \tag{4.6}$$

A questo punto, per ottenere la classificazione sul paziente, è stata applicata una procedura di *Threshold Tuning* sul Validation Set, per poter selezionare la soglia migliore. Poiché la media delle probabilità tende a concentrarsi attorno a valori centrali, la soglia standard di 0.5 risulta spesso sub-ottimale in contesti sbilanciati. L'algoritmo di ottimizzazione ha analizzato l'intervallo di confidenza $[0.10, 0.90]$, selezionando una soglia operativa ottimale $\tau^* = 0.63$ e 0.64 rispettivamente per la pipeline Grid Search e Optuna, come mostrato nel confronto in Figura 4.1.

Il fatto che le due pipeline di ottimizzazione indipendenti convergano verso la stessa soglia decisionale conferma che la necessità di una soglia conservativa è una proprietà intrinseca del dataset e non un artefatto statistico.



(a) Grid Search Pipeline ($\tau^* = 0.63$)



(b) Optuna Pipeline ($\tau^* = 0.64$)

Figura 4.1: Analisi comparativa del Threshold Tuning per le due pipeline di ottimizzazione. In entrambi i casi, spostare la soglia decisionale verso valori più alti è fondamentale per massimizzare la Balanced Accuracy dell'Ensemble.

4.6.3 Confronto delle Strategie nella Fase di Validazione

Al termine della procedura di Cross-Validation, è stato effettuato un confronto diretto tra le prestazioni delle due strategie di ensemble applicate alle due pipeline di ottimizzazione. L'obiettivo di questa analisi comparativa è selezionare l'architettura

finale più robusta da applicare al Test Set, basandosi esclusivamente sulle metriche ottenute in validazione per evitare bias di selezione.

I risultati, riassunti in Tabella 4.8, evidenziano una gerarchia di performance nella quale l'approccio di Soft Voting supera lo Stacking gerarchico in entrambe le pipeline. Nello specifico, utilizzando la Grid Search, il Soft Voting ha ottenuto una Balanced Accuracy di 0.6826 contro lo 0.6768 dello Stacking, divario che risulta ancora più marcato nella pipeline Optuna dove si ha un valore di 0.6722 contro 0.6565. Questa dinamica suggerisce che l'introduzione di un *meta-learner*, come la Regressione Logistica, ha aggiunto complessità architetturale senza apportare un reale guadagno informativo, rischiando al contrario di indurre un lieve *overfitting* sui dati di validazione.

Oltre alla superiorità del metodo di aggregazione, emerge una maggiore stabilità dei modelli base ottimizzati tramite Grid Search, i quali hanno dimostrato una sinergia superiore in fase di assemblaggio rispetto a quelli derivati da Optuna. La configurazione migliore in assoluto risulta infatti essere il Soft Voting ottimizzato con Grid Search, che stacca di circa un punto percentuale la rispettiva controparte. Infine, un ulteriore elemento di rilievo riguarda la dinamica delle soglie decisionali e la calibrazione delle probabilità. Mentre il Soft Voting richiede una soglia di taglio più elevata per filtrare efficacemente i falsi positivi, lo Stacking tende a restituire probabilità compresse verso il basso, costringendo ad abbassare il limite di accettazione per poter recuperare un livello adeguato di sensibilità.

Tabella 4.8: Confronto delle performance dei metodi di Ensemble su predizioni OOF. Il Soft Voting basato su Grid Search emerge come la configurazione più performante e stabile.

Ottimizzazione	**Strategia Ensemble**	**Soglia (τ^*)**	**Balanced Accuracy**
Grid Search	Soft Voting	0.63	0.683
	Stacking (LogReg)	0.37	0.677
Optuna	Soft Voting	0.64	0.672
	Stacking (LogReg)	0.49	0.656

Alla luce di questi risultati, l'architettura basata su Grid Search con Soft Voting è stata selezionata come modello definitivo. Questa configurazione offre il miglior compromesso tra capacità discriminante e semplicità strutturale, minimizzando il rischio di instabilità su nuovi dati.

4.7 Valutazione Finale sul Test Set

L'architettura selezionata nella sezione precedente è stata infine valutata sul Test Set indipendente. Si ribadisce che tale campione, pari al 20% della coorte complessiva, è stato mantenuto completamente separato dalle fasi di addestramento, validazione incrociata e ottimizzazione delle soglie decisionali. Le metriche riportate in questa sezione costituiscono pertanto una stima non distorta della capacità di generalizzazione del modello in uno scenario clinico realistico.

La Tabella 4.9 riporta le performance ottenute dall'Ensemble sul Test Set, confrontandole con quelle dei modelli addestrati sulle singole modalità MRI. Inoltre, in Figura 4.2 vengono riportante le matrici di confusione dei vari classificatori e dei due metodi di Ensemble learning

Tabella 4.9: Confronto tra le performance dell'Ensemble e i singoli modelli sul Test Set. L'Ensemble mostra la Specificità più elevata di 0.533, migliorando la capacità di identificare immagini non diagnostiche pur mantenendo un'elevata accuratezza complessiva.

Modello	**B. Acc**	**Sens**	**Spec**	**PPV**	**NPV**	**AUC**
ADC (MLP)	0.580	0.893	0.267	0.842	0.364	0.571
DWI (MLP)	0.587	0.740	0.433	0.851	0.271	0.608
T2w (RF)	0.651	0.802	0.500	0.875	0.366	0.730
T2w Mask (MLP)	0.520	0.840	0.200	0.821	0.222	0.563
Soft Voting	0.641	0.748	0.533	0.875	0.320	0.710

Dall'analisi comparativa emerge un quadro coerente che permette di interpretare in modo integrato le performance dei diversi modelli. L'Ensemble raggiunge la Specificità più elevata pari a 0.533, superando le modalità di diffusione ADC con 0.267 e DWI con 0.433 e risultando leggermente superiore anche alla sequenza T2w con Specificità pari a 0.500. Poiché l'obiettivo prioritario di questa fase era ridurre il tasso di falsi positivi evitando che immagini tecnicamente non idonee venissero classificate come diagnostiche tale risultato assume una rilevanza operativa significativa. L'integrazione multiparametrica consente quindi un miglior bilanciamento tra sensibilità e capacità di filtro rispetto ai singoli classificatori, configurando l'Ensemble come soluzione più conservativa nel contesto del controllo qualità automatizzato.

Parallelamente, il modello Random Forest sulla sequenza T2w conferma una notevole robustezza individuale, con una Balanced Accuracy pari a 0.651 e un'AUC

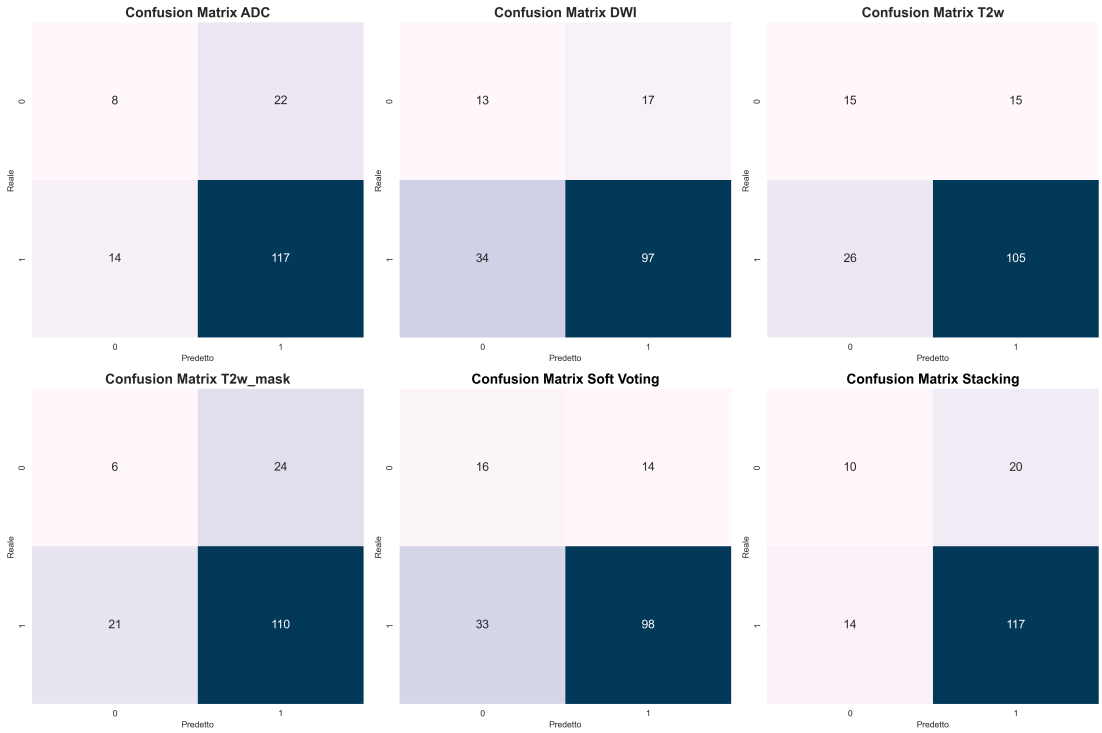


Figura 4.2: Matrici di Confusioni della pipeline di ottimizzazione Grid Search sul Test Set. Il Soft Voting mostra una capacità di identificazione dei casi Non Diagnostici superiore ai classificatori unimodali, raggiungendo una Specificità superiore al 53% in un contesto di classi sbilanciate.

di 0.730, valori superiori rispetto alle altre modalità considerate. Questo risultato suggerisce che l'informazione morfologica strutturale costituisca la componente più discriminativa per la valutazione della qualità dell'immagine. L'Ensemble, pur mostrando una Balanced Accuracy leggermente inferiore di 0.641, presenta un profilo operativo più prudente, privilegiando un incremento della Specificità a fronte di una moderata riduzione della Sensibilità, scelta coerente con l'obiettivo applicativo dello studio.

Infine, confrontando la Balanced Accuracy ottenuta in fase di validazione interna pari a 0.683 con quella osservata sul Test Set pari a 0.641, si rileva una riduzione di circa il 4%. Tale scostamento appare contenuto e compatibile con la variabilità attesa in contesti radiomici ad alta dimensionalità, suggerendo che la procedura di ottimizzazione non abbia introdotto fenomeni di overfitting rilevanti.

Il sistema mantiene una buona capacità di identificazione della classe diagnostica con una Sensibilità di 0.748 e migliora la rilevazione dei casi non diagnostici rispetto ai modelli unimodali. Nonostante permanga un margine di miglioramento nella

riduzione dei falsi positivi residui, i risultati supportano l'ipotesi che l'integrazione tra informazione morfologica e diffusiva, mediata da un Soft Voting calibrato, possa costituire un valido strumento di supporto al controllo qualità della mpMRI, affiancando il radiologo nella riduzione della soggettività valutativa e nel miglioramento dell'efficienza dei flussi di lavoro clinici nella diagnosi del tumore alla prostata.

Capitolo 5

Esplorazione di Approcci di Deep Learning

5.1 Introduzione e Stato dell'Arte

Parallelamente allo sviluppo della pipeline radiomica descritta nei capitoli precedenti, è stata condotta un'indagine sperimentale basata sul Deep Learning (DL). Negli ultimi decenni, l'avvento delle Reti Neurali Convoluzionali (CNN) ha ridefinito gli standard dell'imaging biomedico, sostituendo l'estrazione manuale delle feature con l'apprendimento automatico di rappresentazioni gerarchiche direttamente dai dati grezzi [54].

In tale contesto, la presente indagine si propone di valutare se questo paradigma *data-driven* possa effettivamente intercettare sfumature e artefatti qualitativi complessi che risultano invisibili agli algoritmi della radiomica tradizionale. Contestualmente, si intendeva confrontare le prestazioni delle CNN con i modelli di Machine Learning sviluppati in precedenza, per determinare quale paradigma risultasse più robusto rispetto ai dati disponibili.

La letteratura recente ha registrato un incremento significativo nell'applicazione di architetture 3D per l'automazione del controllo qualità in risonanza magnetica prostatica. Come riassunto nella Tabella 5.1, tre studi fondamentali hanno tracciato la direzione per l'uso di CNN volumetriche nella predizione degli score di qualità.

Il primo passo significativo è stato compiuto da Alis et al. (2023), i quali hanno condotto uno studio di fattibilità su un dataset pubblico (PI-CAI) di 700 scansioni, implementando un'architettura Inception-ResNet-V2 3D. Il loro lavoro ha evidenziato una netta dicotomia nelle prestazioni tra le diverse sequenze. Per le mappe ADC il modello ha raggiunto un accordo "buono" con i radiologi con $\kappa = 0.61$, per le immagini T2W l'accordo si è fermato a un livello "moderato" con

$\kappa = 0.42$, suggerendo che la complessità anatomica della T2w richieda capacità rappresentative superiori [55].

Superando i limiti dimensionali dei lavori precedenti, Belue et al. (2024) hanno successivamente addestrato una DenseNet-121 3D su una coorte multicentrica di 1.046 pazienti. Utilizzando un approccio di classificazione multiclasse, gli autori hanno riportato un'accuratezza globale del 73.9% e un accordo inter-osservatore "sostanziale" con $\kappa = 0.70$, confermando che l'uso di dense connections favorisce il riutilizzo delle feature e migliora la convergenza su volumi 3D complessi [19].

Infine, nel lavoro più recente disponibile in letteratura, Gloe et al. (2025) hanno focalizzato l'attenzione sulla riduzione delle ri-acquisizioni non necessarie. Impiegando una DenseNet-169 3D su un ampio dataset di 1.412 scansioni T2w assiali, hanno ottenuto una AUC di 0.88 e un'accuratezza del 78.3%. Il modello ha dimostrato un accordo "buono" con $\kappa = 0.66$ con i revisori esperti, validando definitivamente l'efficacia delle reti profonde quando supportate da una mole di dati massiva [20].

Sebbene questi lavori dimostrino che le architetture profonde possono raggiungere prestazioni di livello clinico, è doveroso sottolineare una criticità metodologica nel confronto diretto tra i risultati ottenuti in questo studio e quelli riportati in letteratura. La maggior parte degli studi citati in Tabella 5.1 utilizza l'Accuratezza Globale come metrica primaria. Tuttavia, in contesti di forte sbilanciamento delle classi, l'Accuratezza può risultare fuorviante. Nel presente lavoro, si è scelto di adottare la Balanced Accuracy ovvero la media aritmetica tra Sensibilità e Specificità come metrica di riferimento. Questa scelta offre una stima più onesta e clinicamente rilevante della capacità del modello di intercettare le immagini non diagnostiche.

5.2 Configurazione Sperimentale

Lo sviluppo e l'addestramento dei modelli di Deep Learning sono stati realizzati utilizzando il framework *PyTorch*. Il costo computazionale, derivante dall'elaborazione di volumi tridimensionali, è stato gestito accelerando i calcoli su una GPU NVIDIA GeForce GTX 1050 Ti. L'utilizzo di un hardware con vincoli di memoria (VRAM) specifici ha guidato parzialmente le scelte architetturali descritte in seguito, favorendo l'implementazione di modelli efficienti ("Light") rispetto a reti eccessivamente profonde.

Il dataset iniziale è stato ripartito destinando il 70% dei pazienti al Training Set, mentre il restante 30% è stato equamente diviso tra Validation Set e Test Set per la valutazione finale, ciascuno con una quota del 15%.

Tabella 5.1: Confronto tra i principali studi recenti basati su Deep Learning per la valutazione della qualità.

Studio	**Dataset**	**Architettura**	**Performance Chiave**
Alis et al. [55]	700 scansioni (Dataset pubblico PI-CAI)	Inception-ResNet-V2(3D)	**Kappa T2W:** 0.42 (Moderato) **Kappa ADC:** 0.61 (Buono)
Belue et al. [19]	1.046 pazienti (Multi-centrico)	DenseNet-121(3D)	**Accuratezza:** 73.9% **Kappa:** 0.70 (Sostanziale)
Gloe et al. [20]	1.412 scansioni (Solo T2w assiali)	DenseNet-169(3D)	**AUC:** 0.88 **Accuratezza:** 78.3% **Kappa:** 0.66 (Buono)

5.2.1 Preprocessing e Gestione Volumetrica

A differenza dell'approccio radiomico, che opera su feature estratte da regioni di interesse pre-segmentate, l'input per le reti neurali è costituito dai volumi immagine grezzi. Data la risoluzione eterogenea lungo l'asse Z, tipica di una studio multicentrico, è stata implementata una strategia di normalizzazione spaziale:

- **Selezione delle Slice:** Per ogni volume, sono state estratte le $N = 16$ slice centrali, scartando i pazienti che presentavano slice centrale priva dell'organo prostatico identificando i valori della maschera di segmentazione pari a zero nella slice centrale.

- **Input Multimodale (Early Fusion):** Le tre sequenze sono state co-registrate e impilate lungo la dimensione dei canali, generando un tensore di input 4D di dimensione $[3 \times 16 \times 256 \times 256]$.

- **Normalizzazione d'Intensità:** Ogni volume è stato normalizzato singolarmente tramite Z-score normalization per uniformare la dinamica dei segnali provenienti da scanner diversi.

5.2.2 Architetture Testate

Basandosi sugli esempi presenti in letteratura, sono state progettate e testate sei configurazioni architetturali, spaziando da semplici reti sequenziali a varianti customizzate delle famiglie ResNet e DenseNet, modulando il numero di parametri e la profondità dei blocchi convoluzionali. Tali configurazioni sono riassunte in Tabella 5.2.

Tabella 5.2: Dettaglio delle architetture 3D implementate. I modelli sono ordinati per complessità crescente. Le varianti "Light" e "Custom" sono state progettate specificamente per adattarsi alla ridotta dimensione del dataset.

Famiglia	**Variante**	**Dettagli Configurativi**
Baseline	Simple3D-CNN	Struttura con 3 Blocchi Convoluzionali Sequenziali. Filtri: [16, 32, 64]. Nessuna connessione residua.
ResNet (3D)	ResNet-Light	**Config:** `base_planes=32`, `layers=[1, 1, 1, 1]`. Versione miniaturizzata per massimizzare la regolarizzazione.
	ResNet-18	**Config:** `base_planes=64`, `layers=[2, 2, 2, 2]`. Architettura standard.
	ResNet-34	**Config:** `base_planes=64`, `layers=[3, 4, 6, 3]`. Versione profonda per catturare feature ad alta astrazione.
DenseNet (3D)	DenseNet-Light	**Config:** `growth=16`, `block=(4, 6, 8, 4)`. `init_feat=32`. Basso numero di feature map per ridurre i parametri.
	DenseNet-Medium	**Config:** `growth=24`, `block=(6, 8, 12, 6)`. `init_feat=48`. Aumento della capacità tramite growth rate maggiore.

Le varianti ResNet-Light e DenseNet-Light sono state introdotte riducendo il numero di filtri iniziali da 64 a 32 e la profondità dei blocchi. L'obiettivo era forzare la rete a imparare rappresentazioni compatte, riducendo i gradi di libertà per contrastare la memorizzazione del training set. Le varianti ResNet-34 e DenseNet-Medium rappresentano invece il tentativo di emulare le configurazioni dello stato dell'arte come quella di Belue et al. [19]), ipotizzando che la regolarizzazione fosse sufficiente a gestire la maggiore capacità del modello.

5.2.3 Protocollo di Training

L'addestramento è stato condotto in modalità supervisionata utilizzando le etichette di qualità PI-QUAL binarizzate come esposto in Sezione 2.2. Per mitigare il rischio di overfitting dovuto alla limitata dimensione del dataset, sono state adottate delle strategie di Data Augmentation 3D applicando trasformazioni geometriche di flip orizzontali, verticali e rotazioni coerenti su tutti i canali del volume. La funzione di costo adottata è stata la *Weighted Cross-Entropy Loss*, ponderata inversamente alla frequenza delle classi nel training set. Il protocollo di training ha incluso lutilizzo dell'algoritmo Adam, con un tasso di apprendimento iniziale fissato a 10^{-4}, e un meccanismo di *Early Stopping* che permette di monitorare la validation loss e interrompere l'addestramento dopo 5 epoche di mancato miglioramento.

5.3 Risultati Sperimentali: Fase Esplorativa

Come primo approccio, sono state testate diverse combinazioni di architetture e strategie di ottimizzazione per identificare una configurazione base stabile. I risultati ottenuti sul Test Set indipendente sono riassunti nella Tabella 5.3.

Tabella 5.3: Sintesi delle performance sul Test Set per le configurazioni testate. L'aumento della complessità del modello da ResNet-18 a DenseNet e l'uso di Data Augmentation portano ad un crollo della Specificità.

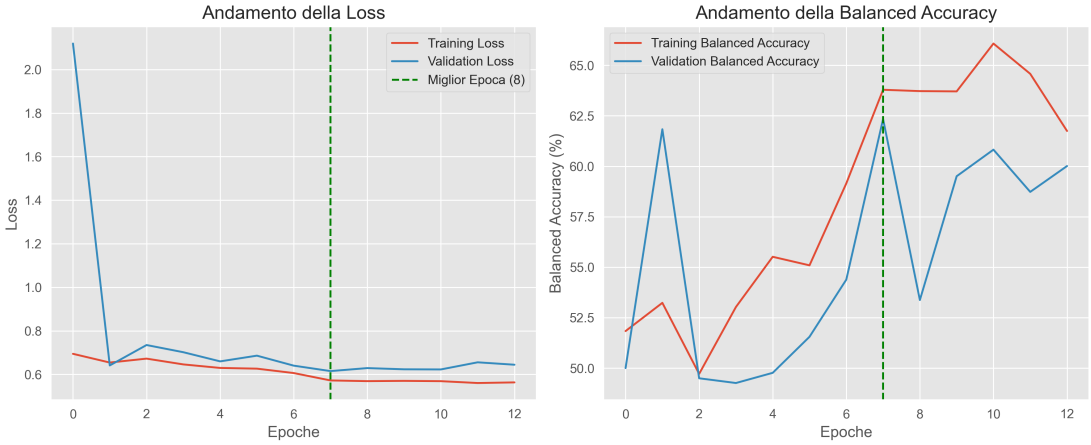
Architettura	Configurazione	Loss	BACC	Sens.	Spec.
ResNet-18	No Augmentation	CrossEntropy	56.2%	71.4%	40.9%
ResNet-34	+ Augmentation	CrossEntropy	52.6%	96.2%	9.1%
DenseNet-121	+ Augmentation	CrossEntropy	54.0%	94.3%	13.6%
DenseNet-121	+ Augmentation + AdamW	Focal Loss	52.4%	76.2%	31.8%

Vengono analizzati due casi esemplari per descrivere il comportamento generale dei modelli. Per motivi di sintesi non vengono riportati i grafici di addestramento di tutte le configurazioni testate, in quanto le dinamiche osservate sono risultate estremamente simili e affette dai medesimi problemi strutturali.

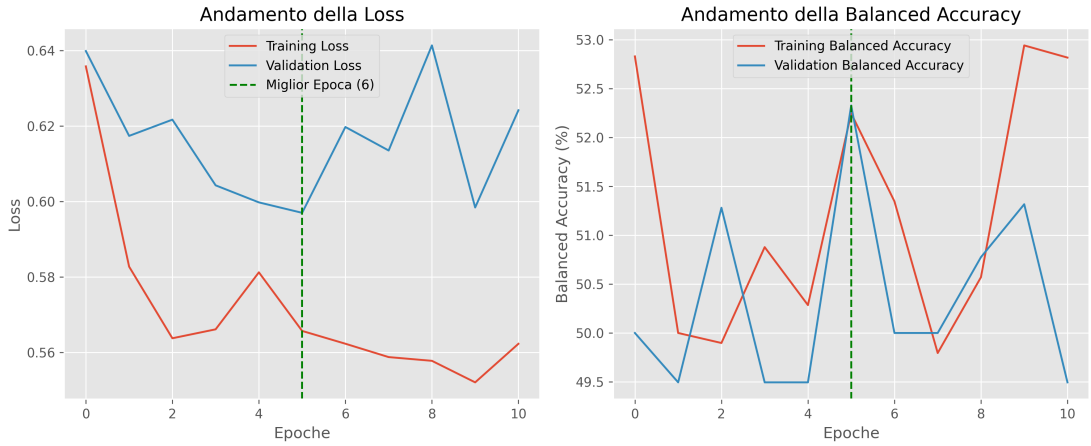
La Figura 5.1a mostra l'andamento della ResNet-18 senza Data Augmentation. Si osserva una rapida discesa della Training Loss che diverge precocemente dalla Validation Loss, sintomo di un overfitting immediato. Nonostante ciò, questa configurazione è stata l'unica a mantenere una minima capacità discriminante sulla classe negativa con una Specificità pari a circa il 41%.

Al contrario, aumentando la complessità della rete e introducendo la Data Augmentation, in Figura 5.1b è possibile notare un fenomeno di collasso modale.

Sebbene le curve sembrino più stabili, il modello ha imparato a minimizzare l'errore semplicemente predicendo sistematicamente la classe maggioritaria. Ciò spiega l'altissima Sensibilità con valori superiori al 94% contrapposta a una Specificità molto bassa, rendendo il modello clinicamente inutile come filtro di qualità.



(a) ResNet-18 senza Data Augmentation



(b) DenseNet-121 con Data Augmentation

Figura 5.1: (a) La ResNet-18 mostra un overfitting con divergenza delle loss. (b) La DenseNet, nonostante l'augmentation, non riesce a generalizzare, stallando su prestazioni casuali.

L'introduzione della Focal Loss ha parzialmente mitigato questo sbilanciamento, recuperando parte della Specificità, passando dal 13% al 32%, ma senza riuscire a portare la Balanced Accuracy sopra la soglia della casualità.

5.4 Strategie Avanzate di Bilanciamento

Alla luce della tendenza dei modelli a convergere sulla classe maggioritaria osservato nella prima fase, l'attenzione sperimentale si è spostata sul problema dello sbilanciamento delle classi. L'ipotesi investigativa era che la scarsa prevalenza di esempi non diagnostici impedisse alla rete di apprendere le caratteristiche distintive del degrado qualitativo. Sono state pertanto implementate e confrontate tre diverse strategie di ricampionamento per riequilibrare il training set, mantenendo fissa l'architettura base della ResNet-18.

I risultati complessivi sono riassunti nella Tabella 5.4.

Tabella 5.4: Confronto delle prestazioni sul Test Set per le diverse strategie di ricampionamento. L'Oversampling tramite iniezione di rumore gaussiano ha prodotto il miglior risultato assoluto, superando sia l'Undersampling che la generazione sintetica complessa.

Strategia	**Dettagli Tecnici**	**Train BACC**	**Test BACC**
Baseline	Nessun bilanciamento (Solo T2W)	70.8%	53.1%
Undersampling	Riduzione casuale della classe maggioritaria (Diagnostica) fino al rapporto 60/40.	63.8%	52.5%
Oversampling (Noise)	Duplicazione della classe minoritaria con aggiunta di *Gaussian Noise* ($\sigma \in [0.05, 0.2]$).	67.9%	61.7%
Oversampling (Synth)	Generazione di dati sintetici tramite simulazione fisica (*Motion Blur* + Rumore).	91.2%	50.0%

5.4.1 Undersampling

La prima strategia ha previsto la riduzione della classe maggioritaria per bilanciare il rapporto tra le classi. Tuttavia, i risultati sono rimasti pressoché identici alla

baseline. Eliminare campioni dalla classe maggioritaria ha ridotto la varietà di esempi visti dalla rete, senza però fornire informazioni aggiuntive su cosa costituisca un'immagine "Non Diagnostica".

5.4.2 Oversampling con Gaussian Noise

La seconda strategia ha previsto la duplicazione dei campioni minoritari con l'aggiunta di un leggero rumore gaussiano. I parametri utilizzati sono stati un range di deviazione standard $\sigma \in [0.05, 0.2]$ con media nulla, applicato indipendentemente su ogni canale. Un esempio di immagine è riportato in Figura 5.2a).

L'analisi delle curve di apprendimento in Figura 5.2b conferma l'efficacia di questa scelta. Rispetto ai tentativi precedenti, si osserva una convergenza più stabile e una divergenza tra le metriche ridotta tra Training e Validation, indice di una migliore generalizzazione.

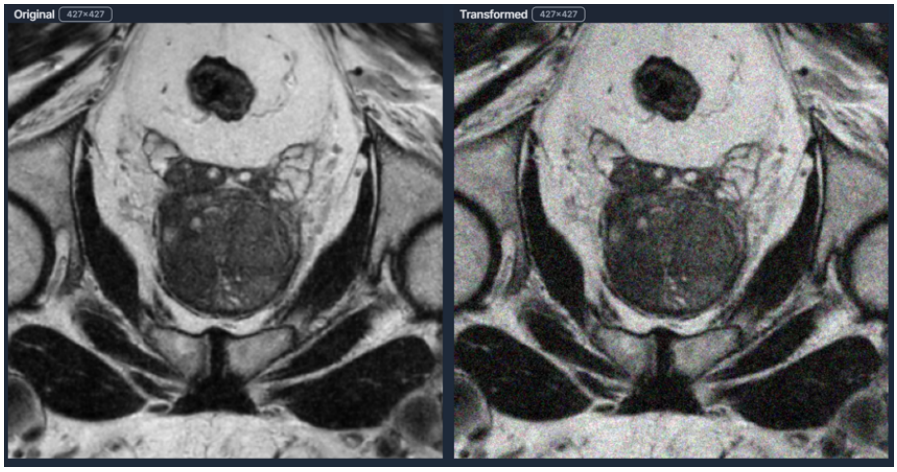
Questa tecnica si è rivelata la più efficace in assoluto, portando la Balanced Accuracy sul Test Set al 61.71%. A differenza della duplicazione semplice, l'iniezione di rumore ha agito come regolarizzatore, modificando leggermente le immagini duplicate e costringendo la rete a imparare feature più robuste. Sebbene ancora inferiore alla Radiomica, questo risultato dimostra che il Deep Learning può apprendere qualcosa se aiutato nella gestione dello sbilanciamento.

5.4.3 Approccio Sintetico

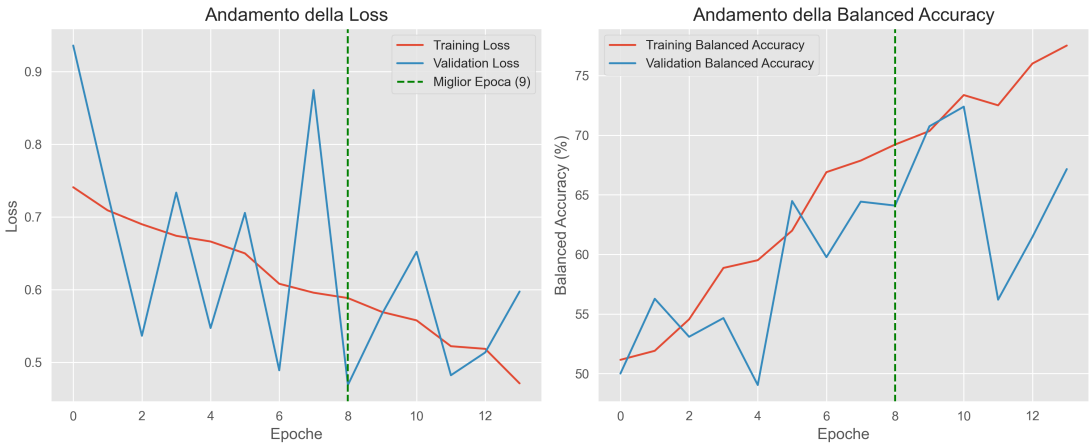
Per superare i limiti della duplicazione, è stato sviluppato un approccio basato sulla simulazione fisica degli artefatti. L'assunto di base è che la scarsa qualità nelle RM prostatiche sia dovuta principalmente a due fattori: movimento del paziente e basso rapporto segnale-rumore. Invece di duplicare le immagini esistenti, sono stati generati campioni "Non Diagnostici Sintetici" partendo da immagini di buona qualità e degradandole artificialmente tramite la libreria *Albumentations*. Il protocollo di degradazione visualizzato in Figura 5.3a prevede:

- **Motion Blur:** Simulazione del movimento lungo un asse casuale con kernel di dimensione variabile (`blur_limit` $= (3, 7)$).
- **Gaussian Noise:** Iniezione di rumore gaussiano con intensità inferiore rispetto all'esperimento precedente con $\sigma \in [0.05, 0.1]$, per simulare la granulosità tipica delle acquisizioni a basso SNR.

Le dinamiche di apprendimento mostrate in Figura 5.3b svelano il fallimento di questa strategia. Si nota che, mentre la Training Accuracy raggiunge rapidamente il 91%, la Validation Accuracy rimane bloccata al 50%. Questo conferma l'esistenza di un ampio *Domain Gap* poiché la rete ha imparato perfettamente a riconoscere



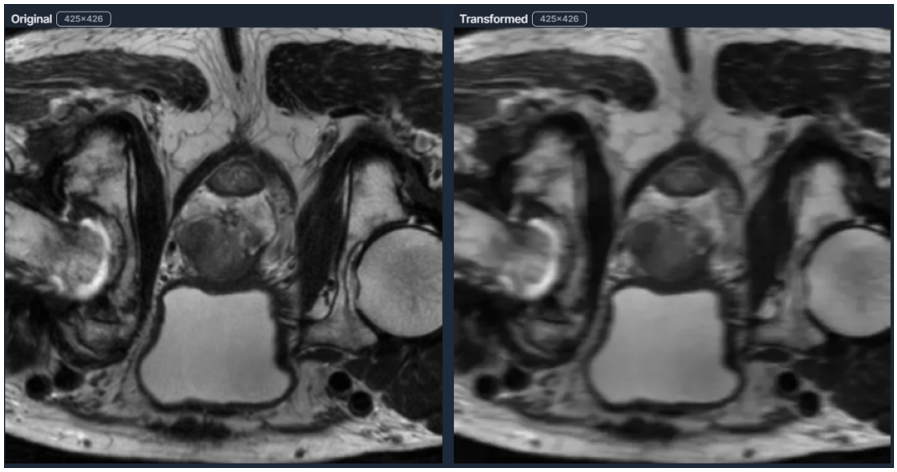
(a) Esempio di Augmentation



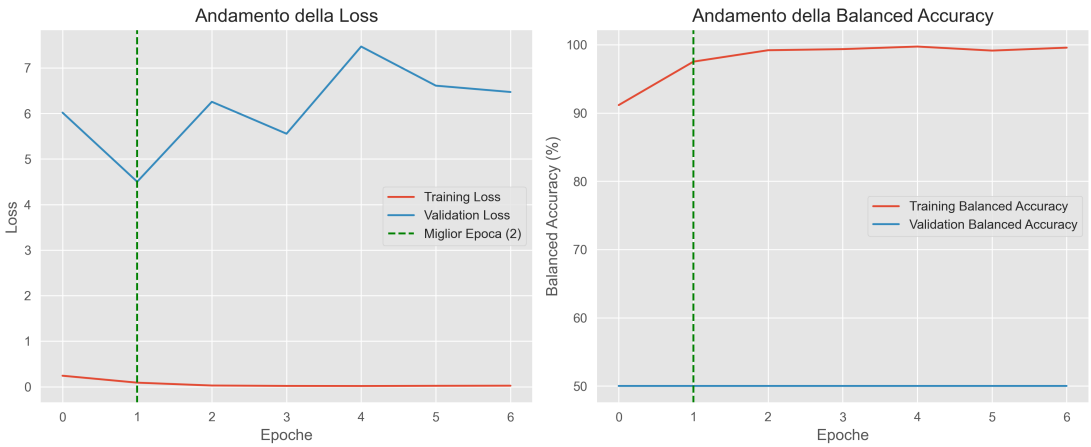
(b) Curve di Training

Figura 5.2: (a) Esempio di data augmentation mediante aggiunta di Rumore Gaussiano ($\sigma = 0.2$). (b) Le curve di addestramento mostrano una stabilità maggiore rispetto al baseline, con una Validation Accuracy che segue il trend del Training.

gli artefatti artificiali creati, ma non è riuscita a trasferire questa conoscenza sugli artefatti reali, rendendo l'augmentation sintetica inefficace.



(a) Generazione Sintetica



(b) Overfitting su Dati Sintetici

Figura 5.3: (a) A destra la versione degradata artificialmente. (b) Le curve mostrano un overfitting massivo: il modello apprende i dati sintetici (BACC Train $> 90\%$) ma fallisce completamente sui dati reali (BACC Val 50%).

5.5 Approccio Multi-Modale: Architettura Multi-Stream

Dopo aver analizzato i limiti delle singole sequenze, un nuovo tentativo sperimentale ha mirato a replicare il processo decisionale del radiologo, integrando le informazioni provenienti da tutte le sequenze disponibili sia morfologiche che funzionali. È stata pertanto sviluppata un'architettura Multi-Stream CNN basata su una strategia di

Late Fusion.

A differenza dell'approccio *Early Fusion* dove i canali vengono impilati all'input, in questa configurazione ogni modalità viene processata da un encoder dedicato. Come illustrato nello schema in Figura 5.4, il modello è composto da tre rami paralleli per classificare singolarmente l'immagine T2w complessiva, l'immagine DWI e la regione dell'immagine T2w dove è presente la prostata.

Ciascun encoder estrae un vettore di feature latenti di dimensione $[Batch, 512]$. Questi vettori vengono successivamente concatenati in un unico tensore denso di $512 \times 3 = 1536$ features e passati a un classificatore finale MLP composto da strati lineari con attivazione ReLU e Dropout ($p = 0.5$) per la predizione finale.

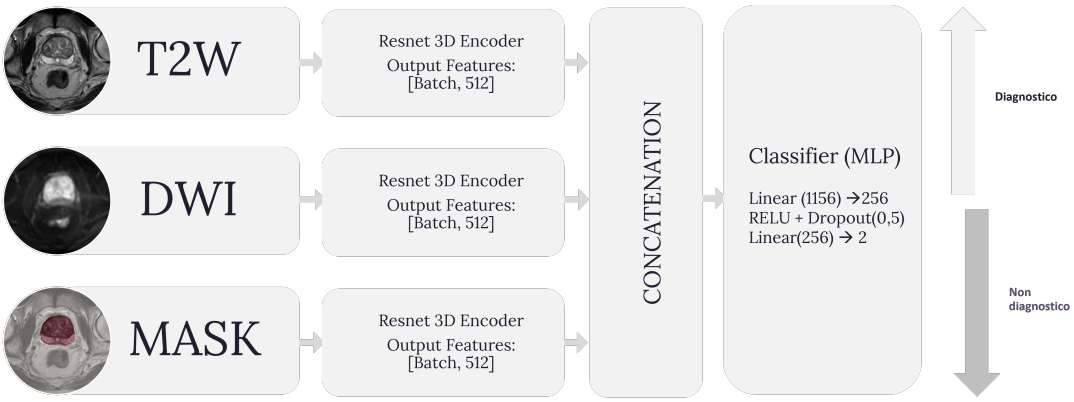


Figura 5.4: Schema dell'architettura Multi-Stream implementata. Le tre modalità vengono processate da encoder indipendenti ResNet 3D. Le feature estratte vengono fuse per concatenazione prima del classificatore finale, permettendo alla rete di apprendere rappresentazioni specifiche per ogni modalità.

Nonostante la maggiore complessità teorica e la ricchezza informativa, i risultati sperimentali presentati in Tabella 5.5 sanciscono il fallimento di questo approccio; l'aumento esponenziale del numero di parametri ha esacerbato il fenomeno della "Maledizione della Dimensionalità".

Come evidenziato dalla matrice di confusione, il modello a tre flussi ha mostrato un collasso totale sulla classe minoritaria, ottenendo una Balanced Accuracy del 41.35%, paradossalmente inferiore a un classificatore casuale.

Questo risultato conferma che, in assenza di un dataset di migliaia di volumi, l'aggiunta di modalità eterogenee non introduce informazione utile, ma solo rumore strutturato che impedisce l'ottimizzazione della Loss, rendendo preferibili modelli più semplici e focalizzati sulla sola sequenza morfologica dominante.

Tabella 5.5: Performance dell'approccio Multi-Stream sul Test Set. L'utilizzo congiunto di tutte le sequenze ha portato a un drastico calo delle prestazioni rispetto alla singola T2W. La colonna Confusion Matrix (CM) evidenzia come il modello Dual-Stream collassi sulla classe maggioritaria (TN=0), mentre il Triple-Stream introduca errori gravi (FN=18) senza guadagnare capacità discriminante sulla classe minoritaria.

Configurazione	**Test BACC**	**CM (TP/TN/FP/FN)**
Dual-Stream (T2W + DWI)	50.0%	104 / 0 / 19 / 0
Triple-Stream (T2W + DWI + Mask)	41.3%	86 / 0 / 19 / 18

5.6 Considerazioni Conclusive sul Deep Learning

In sintesi, l'indagine condotta con le reti neurali profonde ha dimostrato che, sebbene le CNN siano teoricamente superiori nell'estrazione di feature, esse richiedono una dataset bilanciato di maggiori dimensioni non disponibile in questo studio. La Radiomica, basandosi su feature ingegnerizzate a priori, si è rivelata più efficiente nel catturare i pattern di qualità su un dataset di dimensioni ridotte, superando le performance del Deep Learning. Questo risultato valida la scelta iniziale di puntare su un approccio ibrido guidato dalla conoscenza radiologica piuttosto che su un approccio puramente data-driven.

Capitolo 6

Conclusioni e Prospettive Future

In questo capitolo conclusivo vengono riassunti i risultati chiave dello studio e analizzate criticamente le limitazioni emerse con particolare attenzione all'analisi degli errori e all'interpretabilità. Vengono inoltre tracciate le direzioni per gli sviluppi futuri.

6.1 Sintesi dei Risultati Chiave

A seguito della raccolta dei risultati effettuata nei capitoli precedenti, vengono sintetizzati i dati dell'approccio basato sui classificatori radiomici e sul modello di Deep Larning. Per garantire una valutazione rigorosa, le migliori configurazioni derivanti da entrambi i paradigmi — rispettivamente il Soft Voting Ensemble radiomico e l'architettura convoluzionale tridimensionale ResNet18 con oversampling — sono state testate sul medesimo Test Set indipendente, mantenuto isolato durante l'intera fase di addestramento e validazione k-fold. I risultati delle prestazioni sono riportati nella Tabella 6.1.

Tabella 6.1: Confronto delle performance sul Test Set indipendente. Sebbene il Deep Learning mostri un'Accuratezza globale superiore, la Radiomica garantisce una Balanced Accuracy migliore e, soprattutto, una Specificità più alta.

Approccio	**Accuracy**	**B. Acc**	**Sens**	**Spec**	**NPV**	**PPV**
Radiomica	0.708	0.641	0.748	0.533	0.327	0.875
Deep Learning	0.783	0.609	0.886	0.333	0.400	0.853

Dall'analisi congiunta delle metriche considerate, il modello di Deep Learning potrebbe apparire complessivamente superiore, in virtù di un valore di Accuratezza globale pari al 78.3%, rispetto al 70.8% ottenuto dall'approccio radiomico. Tuttavia, in presenza di un marcato sbilanciamento tra le classi, tale metrica risulta scarsamente informativa. L'elevata accuratezza della rete neurale è infatti principalmente trainata da una Sensibilità molto alta, che riflette una tendenza sistematica a classificare la maggior parte dei campioni come appartenenti alla classe maggioritaria. Questo comportamento consente di ridurre l'errore complessivo, ma compromette l'obiettivo clinico principale del sistema, ovvero l'identificazione affidabile delle immagini non diagnostiche.

In questo contesto, la Specificità emerge come l'indicatore più rappresentativo della capacità del modello di valutare la qualità dell'immagine. Sotto questo profilo, l'approccio radiomico mostra una netta superiorità raggiungendo un valore di 53.3% rispetto al 33.3% del Deep Learning. L'estrazione a priori di feature geometriche e di texture, unite a una soglia decisionale calibrata all'interno dell'Ensemble, si è dimostrata più efficace nel distinguere i campioni compromessi da rumore o artefatti rispetto all'estrazione automatica delle rappresentazioni condotta dalla rete convoluzionale.

Queste differenze si riflettono anche nella Balanced Accuracy, che risulta complessivamente più elevata per la radiomica pari a 0.641 rispetto a 0.609 nel modello Deep. In scenari caratterizzati da una disponibilità limitata di dati annotati, modelli classici basati su feature ingegnerizzate e guidati dal dominio radiologico offrono una maggiore stabilità e capacità discriminativa rispetto alle architetture di Deep Learning. Queste ultime, al contrario, richiedono volumi di dati significativamente più ampi per poter ottimizzare un numero elevato di parametri senza incorrere in fenomeni di overfitting o in una degenerazione verso la classe maggioritaria.

Nel complesso, considerando il contesto clinico e i vincoli numerici affrontati in questo lavoro, l'Ensemble radiomico si configura come la soluzione più adeguata per l'implementazione di un sistema di supporto al controllo qualità automatico.

6.1.1 Implicazioni Cliniche e Interpretazione delle Metriche

Nel valutare l'applicabilità clinica del modello proposto, risulta inoltre fondamentale oltre l'analisi della Balanced Accuracy esposta, considerare il bilanciamento tra PPV e NPV.

Il principale punto di forza dell'Ensemble risiede nell'elevato PPV, pari all'87.5%. Quando il sistema classifica un volume come "Diagnostico", vi è un'alta probabilità che esso sia effettivamente tale permettendo di automatizzare l'accettazione di una quota significativa degli esami e ridurre il carico di lavoro del radiologo.

Al contrario, il basso NPV, associato a una Specificità del 53.3%, descrive il comportamento di un sistema deliberatamente prudente. Il modello tende a segnalare

come potenzialmente "Non Diagnostici" anche alcuni esami che risultano in realtà accettabili, generando un numero non trascurabile di falsi negativi. Questo atteggiamento conservativo è tuttavia preferibile in un contesto di supporto decisionale, poiché il costo operativo di una verifica aggiuntiva da parte dell'operatore umano è ampiamente inferiore al rischio clinico associato all'accettazione di immagini compromesse.

Alla luce di queste considerazioni, non è possibile implementare questo sistema come un filtro completamente autonomo ma, piuttosto, come uno strumento di supporto alla decisione. In questo modo si possono valutare i casi più affidabili e indirizzare quelli incerti verso una valutazione specialistica, lasciando il controllo finale al radiologo.

6.2 Spiegabilità e Analisi degli Errori

Al fine di comprendere la logica decisionale del modello e valutare le reali cause dei fallimenti predittivi, la valutazione quantitativa è stata integrata da un'indagine qualitativa basata su tecniche di eXplainable AI (XAI) e sull'analisi clinica dei casi misclassificati.

6.2.1 Interpretabilità Del Modello Radiomico

L'architettura radiomica è stata indagata utilizzando due approcci complementari: la Permutation Feature Importance [43, 56] per la spiegabilità globale e l'algoritmo LIME (Local Interpretable Model-agnostic Explanations) [57] per la spiegabilità locale su singoli pazienti.

La tecnica di Permutation Feature Importance valuta il decremento prestazionale del modello quando i valori di una specifica feature vengono rimescolati casualmente sul Test Set, eliminando la correlazione con la variabile target. Una caduta significativa delle performance definisce l'importanza di quella feature, specialmente nella sua capacità di filtrare il rumore [56]. Analizzando le caratteristiche più rilevanti per i diversi classificatori di base, emerge il dominio dei filtri Wavelet. Nello specifico, le feature che guidano maggiormente la classificazione sono Wavelet - GLDM - High Gray Level Dependecy, Wavelet - GLSZM - Zone Percentage, Mediana e Varianza applicate sempre sulle immagini Wavelet, rispettivamente per ADC, DWI, T2w e maschera prostatica.

La ricorrenza delle trasformazioni Wavelet suggerisce che gli artefatti responsabili del giudizio "Non Diagnostico" risiedono in specifiche bande di frequenza spaziale che solo la scomposizione multi-risoluzione Wavelet è in grado di isolare efficacemente dal rumore di fondo.

A livello del singolo paziente, l'algoritmo LIME ha permesso di visualizzare il contributo puntuale di ogni feature nel spingere la predizione verso la classe

Diagnostica o Non Diagnostica. La Figura 6.1 mostra un esempio applicato alla modalità DWI dove si osserva come valori specifici di varianza e kurtosis nel dominio Wavelet agiscano come forze contrapposte che determinano la probabilità finale di classificazione. In una prospettiva di integrazione clinica, questo algoritmo è inteso come modulo esplicativo da integrare nella pipeline di valutazione per individuare i criteri radiomici specifici che hanno determinato tale valutazione caso per caso.

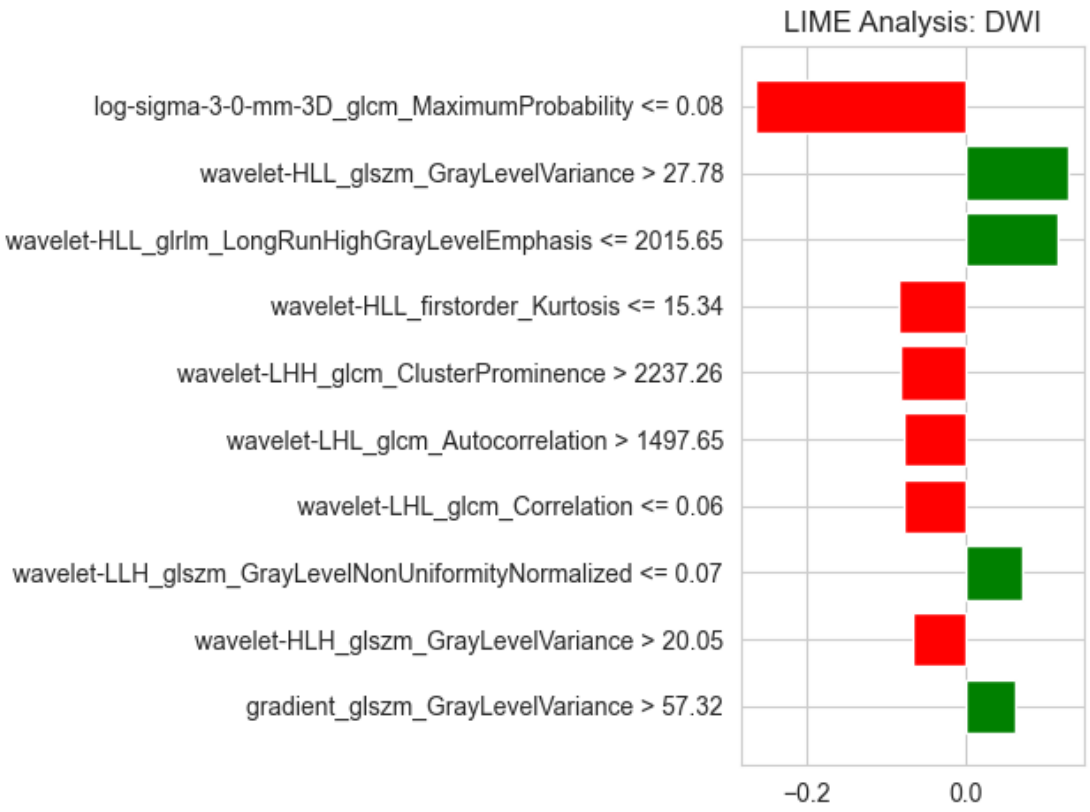


Figura 6.1: Esempio di spiegabilità locale tramite LIME per la modalità DWI del paziente PCa-1837604008857012103469204459438899697733. Il grafico illustra come la feature Maximum Probability valutata con filtro LoG abbia inciso profondamente sulla classificazione finale. In particolare, questa feature, influisce sulla classificazione dell'immagine come "Non Diagnostica" (Colore Rosso nel grafico), quando questa in realtà è "Diagnostica" (Colore Verde).

6.2.2 Indagine Statistica sulle Predizioni del Modello Radiomico

Al fine di investigare la relazione tra la qualità dell'immagine, valutata soggettivamente dai reader secondo i criteri PIQUAL, e le prestazioni del modello di classificazione, è stata condotta un'analisi di correlazione tra i punteggi assegnati alle singole domande e il comportamento predittivo dell'algoritmo. L'analisi è stata articolata valutando l'impatto della qualità sull'accuratezza complessiva del modello sull'intero Test Set e il bias predittivo nei soli casi misclassificati.

Nel primo scenario è stata valutata la correlazione tra ciascun item qualitativo e la variabile binaria Correct pari a 1 nel caso di predizione corretta, 0 altrimenti. Per fare ciò viene valutato il coefficiente di correlazione di Pearson approssimato a ϕ nel contesto binario come metrica principale di interpretazione, dove valori positivi indicano che una qualità elevata favorisce una classificazione corretta, mentre valori negativi segnalano una correlazione tra bassi punteggi qualitativi e l'aumento degli errori.

Parallelamente, per i punteggi totali delle sequenze DWI e T2 che vengono calcolati come somma aritmetica delle risposte binari, la relazione con l'accuratezza è stata stimata tramite la correlazione Point-Biserial (r_{pb}), specifica per misurare la dipendenza tra una variabile quantitativa e una binaria. Queste variabili discrete permettono infatti di valutare la qualità globale delle sequenze.

Il secondo livello di analisi ha ristretto il campo di osservazione ai soli pazienti misclassificati, indagando se la qualità dell'immagine influenzasse la direzione della decisione errata del modello. In questo caso, un valore positivo del coefficiente ϕ suggerisce che una qualità alta spinge il modello verso la predizione della classe positiva, mentre un valore negativo indica una tendenza sistematica verso la classe negativa.

L'analisi di correlazione, riportata in Tabella 6.2, è stata replicata separatamente per ciascuno dei tre radiologi valutatori, con l'obiettivo di isolare eventuali bias soggettivi e confermare la trasversalità delle osservazioni. L'indagine condotta sulle valutazioni del primo osservatore ha evidenziato una gerarchia nell'impatto dei difetti qualitativi, rivelando che la sequenza DWI rappresenta il principale determinante della performance del modello. Nello specifico, il criterio relativo all'assenza di artefatti significativi in DWI ha registrato la correlazione più alta con $\phi = 0.282$, indicando che la presenza di distorsioni in questa modalità aumenta la probabilità di errore dell'algoritmo. Anche il contrasto sulle immagini ad alto b-value ha mostrato un impatto rilevante ($\phi = 0.174$), suggerendo una difficoltà del modello nell'estrarre feature affidabili da immagini funzionali rumorose. Al contrario, per quanto riguarda la morfologia T2-weighted, un coefficiente di correlazione quasi nullo, prova che il classificatore performa costantemente indipendentemente dalla qualità percepita di tale sequenza.

Tali osservazioni puntuali trovano conferma nell'analisi sui punteggi totali aggregati riportati in Tabella 6.3. Mentre lo score complessivo della qualità DWI per il Reader 1 risulta significativamente correlato all'accuratezza ($r_{pb} = 0.224$, P-value $= 0.004$), lo score totale della T2w non raggiunge la significatività statistica ($r_{pb} = 0.093$, P-value $= 0.24$). Limitando l'indagine ai soli errori commessi dal modello, in Tabella 6.4 emerge inoltre un bias negativo debole ma presente per la DWI, con un coefficiente $\phi = -0.121$ che suggerisce una lieve tendenza sistematica a predire erroneamente la classe opposta in presenza di artefatti.

La validità di questi risultati è rafforzata dalla coerenza riscontrata nelle valutazioni degli altri due osservatori. Anche per il Reader 2, la qualità globale della sequenza DWI si conferma come l'unico predittore significativo dell'accuratezza del modello ($r_{pb} = 0.242$, P-value $= 0.002$), con un impatto addirittura superiore rispetto al primo valutatore, mentre la qualità T2w rimane non determinante con un P-value > 0.05. Analogamente, il Reader 3 ribadisce la criticità della componente funzionale, con una correlazione tra score DWI totale e correttezza della predizione che si attesta a $r_{pb} = 0.237$ e P-value $= 0.07$. In sintesi, l'analisi incrociata dimostra che il calo di performance del modello radiomico è strutturalmente legato alla qualità della sequenza DWI, la quale costituisce il principale collo di bottiglia per l'affidabilità del sistema, mentre l'architettura manifesta una notevole resilienza alle variazioni qualitative della sequenza morfologica T2w.

6.2.3 Confronto dei Risultati tra Approccio Radiomico e di Deep Learning

Valutando le predizioni sul Test Set, il Soft Voting Ensemble radiomico ha generato 47 errori totali su 161 casi, con 14 Falsi Positivi e 33 Falsi Negativi. È interessante confrontare questo comportamento con il modello di Deep Learning, che ha registrato 35 errori complessivi, con 20 Falsi Positivi e 15 Falsi Negativi. Analizzando l'intersezione tra le predizioni errate dei due paradigmi, si osserva che i modelli condividono 11 errori comuni. Per quantificare statisticamente questo grado di sovrapposizione, sono stati calcolati il coefficiente Kappa di Cohen (κ) e il coefficiente di correlazione Phi (ϕ) tra i risultati della classificazione dei due modelli. I risultati ottenuti, rispettivamente pari a -0.021 e -0.023, evidenziano un accordo pressoché nullo o di *Poor Agreement* secondo la scala di Landis e Koch [58], tra le due architetture. Questo dato suggerisce che Radiomica e Deep Learning non sono ridondanti, ma osservano il dato medico da prospettive diverse, fallendo su tipologie di pazienti distinte. Tale indipendenza degli errori è visualizzata nel diagramma di Venn in Figura 6.2, mentre l'analisi qualitativa degli 11 casi comuni verrà approfondita nella sezione 6.2.4.

Per verificare se la discrepanza nelle performance fosse statisticamente significativa, è stato applicato il Test di McNemar [59] sulle predizioni accoppiate. Il

Tabella 6.2: Correlazione Phi tra i singoli criteri PI-QUAL e l'accuratezza del modello sull'intero Test Set. Valori positivi elevati indicano che una migliore qualità dell'immagine è associata a una maggiore probabilità di successo del modello.

Criterio PI-QUAL	**Reader 1 (ϕ)**	**Reader 2 (ϕ)**	**Reader 3 (ϕ)**
Sequenza DWI / ADC			
Assenza di artefatti significativi	0.282	0.270	0.261
Contrasto e SNR (High b-value)	0.174	0.160	0.073
Range di contrasto (TZ vs PZ su ADC)	0.114	0.049	0.079
Matching anatomico (ADC vs T2w)	0.030	0.161	0.192
Sequenza T2w			
Delineazione strutture anatomiche	0.005	0.111	0.054
Assenza di artefatti significativi	0.064	0.043	0.135
SNR adeguato	0.102	0.123	0.204

Tabella 6.3: Sintesi dell'analisi di correlazione Point-Biserial (r_{pb}) tra i punteggi totali di qualità assegnati dai tre radiologi e l'accuratezza predittiva del modello. I valori in grassetto indicano una correlazione statisticamente significativa ($P < 0.05$). Per due valutatori su tre, la qualità della sequenza DWI raggiunge la significatività statistica.

Valutatore	**Sequenza DWI**		**Sequenza T2w**	
	r_{pb}	**P-Value**	r_{pb}	**P-Value**
Reader 1	**0.224**	0.004	0.093	0.243
Reader 2	**0.242**	0.002	0.128	0.105
Reader 3	0.237	0.078	0.192	0.156

Tabella 6.4: Correlazione Phi tra i criteri qualitativi e la direzione della predizione nei soli casi misclassificati. La prevalenza di valori negativi indica un bias sistematico poiché al diminuire della qualità, il modello tende erroneamente a predire la classe Diagnostica.

Criterio PI-QUAL	**Reader 1 (ϕ)**	**Reader 2 (ϕ)**	**Reader 3 (ϕ)**
Sequenza DWI / ADC			
Assenza di artefatti significativi	-0.121	-0.348	-0.177
Contrasto e SNR (High b-value)	-0.172	-0.192	-0.417
Range di contrasto (TZ vs PZ su ADC)	-0.390	-0.365	-0.331
Matching anatomico (ADC vs T2w)	-0.274	-0.250	-0.320
Sequenza T2w			
Delineazione strutture anatomiche	-0.020	-0.448	-0.265
Assenza di artefatti significativi	-0.211	-0.135	-0.222
SNR adeguato	-0.045	-0.250	+0.276

test ha restituito un P-Value di 0.155, valore ampiamente superiore alla soglia di significatività $\alpha = 0.05$. Questo risultato certifica che, nonostante le differenze nelle metriche di specificità, non esiste una differenza strutturale statisticamente significativa nella capacità di generalizzazione globale tra i due paradigmi.

6.2.4 Analisi dei Casi Critici

Per comprendere a fondo i limiti intrinseci dell'estrazione di feature quantitative, l'indagine qualitativa si è concentrata sull'intersezione degli errori tra l'approccio radiomico e la rete neurale convoluzionale. L'isolamento di questo sottoinsieme ha permesso di identificare 11 pazienti critici in cui entrambi i paradigmi predittivi hanno fallito. La visualizzazione della slice centtrale dei volumi T2w, DWI e ADC di questi casi è riportata in Figura 6.4.

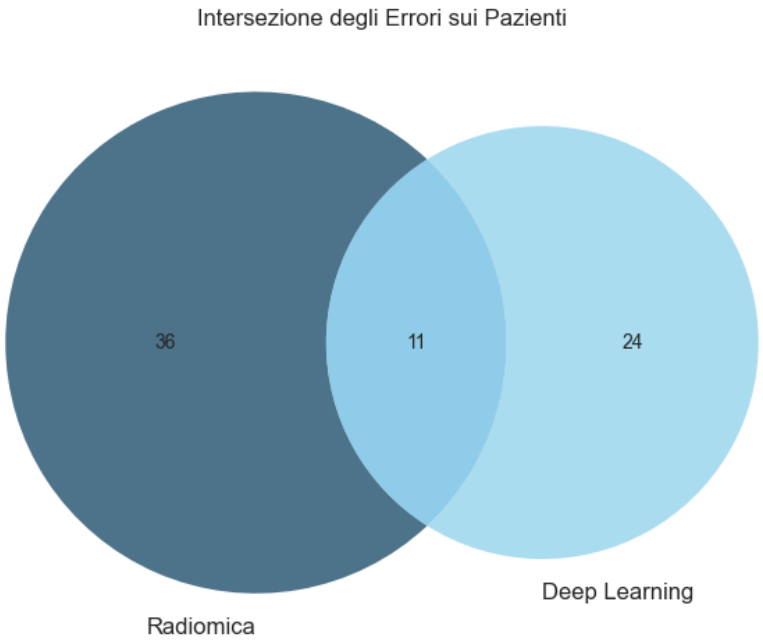


Figura 6.2: Analisi dell'intersezione degli errori tramite Diagramma di Venn. Su un totale di 161 pazienti, i due modelli condividono 11 errori comuni.

Il dato più rilevante emerso dall'analisi di questi 11 volumi condivisi risiede nell'asimmetria della direzione dell'errore. Si registrano infatti 9 Falsi Positivi e 2 Falsi Negativi. Questa tendenza condivisa a sovrastimare la qualità dell'immagine evidenzia come i due modelli vengano ingannati dalla medesima tipologia di artefatti. Ciò suggerisce che il fallimento non sia imputabile al limite di una specifica architettura, bensì alla natura intrinsecamente ambigua e ingannevole del dato radiologico stesso.

Per verificare l'ipotesi dell'ambiguità intrinseca, è stata indagata la storia diagnostica di questi pazienti incrociandola con i log di valutazione dei radiologi. È emerso che 9 degli 11 casi condivisi presentano una discordanza clinica iniziale, definita come la necessità di intervento del Reader 3 per dirimere un conflitto di giudizio tra i primi due valutatori. L'elevata incidenza di tale fenomeno ha suggerito di estendere questa verifica all'intero dataset degli errori commessi dai modelli, pari a 47 e 35 pazienti misclassificati rispettivamente per Radiomica e Deep Learning. I risultati confermano il trend poiché 28 errori su 47 pari al 59,6% e 18 su 35 pari al 51,4% appartengono alla categoria dei casi discordanti. Coerentemente con le premesse sul *Label Noise*, le immagini che traggono in inganno gli algoritmi sono, nella maggioranza dei casi, le stesse su cui l'occhio esperto dei medici fatica a trovare un consenso unanime sulla soglia di diagnosticità.

A riprova quantitativa di questa dinamica, sono state ricalcolate le metriche prestazionali di entrambi i modelli su un sottoinsieme del Test Set, dal quale sono stati esclusi tutti i pazienti rientranti nella categoria dei casi discordanti. L'obiettivo di questa valutazione è misurare le prestazioni degli algoritmi su una *Ground Truth* priva del rumore generato dalla variabilità inter-osservatore. I risultati di questa simulazione sono riportati nella Tabella 6.5.

Tabella 6.5: Metriche prestazionali ricalcolate escludendo i casi discordanti. I valori di *Balanced Accuracy* e di Specificità subiscono un notevole incremento per il modello Radiomico. Al contrario, il Deep Learning mantiene un bias strutturale verso la classe diagnostica.

Approccio	**Accuracy**	**B. Acc**	**Sens**	**Spec**	**PPV**	**NPV**
Radiomica	0.819	0.800	0.823	0.778	0.975	0.292
Deep Learning	0.838	0.609	0.885	0.333	0.934	0.214

L'esclusione delle ambiguità soggettive pari a 56 casi discordanti su 161 pazienti del Test Set ha un impatto rivelatore sulle performance. Il modello Radiomico fa registrare un balzo della Balanced Accuracy all'80.0% rispetto al 64.1% registrato sull'intero Test Set. Questo miglioramento è quasi interamente guidato dall'aumento della Specificità, che passa dal 53.3% al 77.8%. Ciò certifica che, quando la valutazione umana è unanime, le feature radiomiche riescono a intercettare i difetti tecnici con estrema efficacia.

Il modello di Deep Learning, pur mostrando un'Accuratezza globale apparentemente superiore pari a 83.8%, mantiene inalterata la Balanced Accuracy, dimostrando, diversamente dall'approccio radiomico un limite strutturale e un bias predittivo marcato.

L'analisi della distribuzione dei punteggi PI-QUALpresentata nella heatmap in Figura 6.3 offre un'ulteriore chiave di lettura su questi pazienti. Vengono messi a confronto i punteggi totali aggregati per le sequenze DWI e T2w, stratificando tali score per tipologia di errore e per i tre Reader. Da questa rappresentazione, non emerge alcuna separazione netta tra la distribuzione dei Falsi Positivi e dei Falsi Negativi. I due cluster di errore risultano infatti mescolati, distribuendosi lungo un range di punteggi prevalentemente borderline per entrambe le sequenze.

In conclusione, lo studio di questi casi critici certifica che il tetto prestazionale degli algoritmi attuali è fisiologicamente limitato dalla incertezza della valutazione umana. Qualsiasi futuro miglioramento dell'accuratezza non potrà prescindere da una ridefinizione più oggettiva e quantitativa della *Ground Truth* stessa, riducendo la variabilità inter-osservatore che attualmente si traduce in rumore ineliminabile durante la fase di addestramento dei modelli.

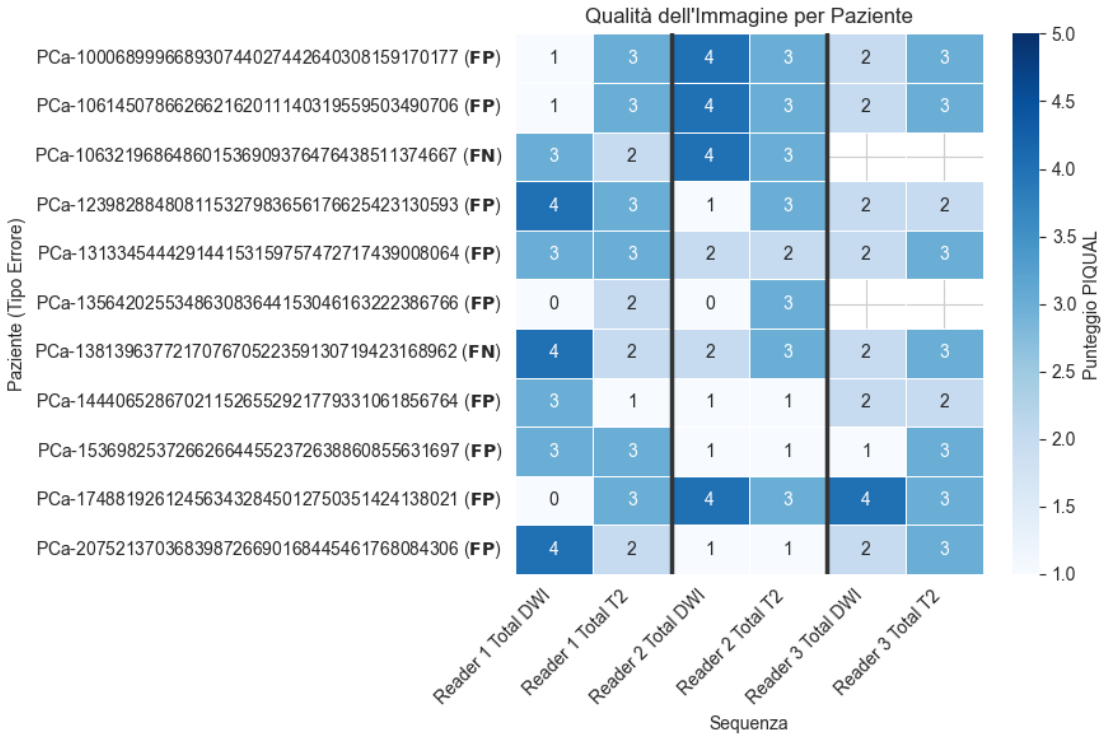


Figura 6.3: Analisi topologica quantitativa dei punteggi PI-QUAL totali per le sequenze DWI e T2w assegnati dai tre Reader agli 11 casi critici misclassificati. La rappresentazione evidenzia l'assenza di cluster definiti per tipologia di errore e la marcata variabilità inter-osservatore sugli stessi pazienti, confermando l'ipotesi di *Label Noise* intrinseco.

6.2.5 Analisi della Distribuzione delle Probabilità

A completamento dell'indagine, è stata condotta un'analisi quantitativa sulla confidenza predittiva del modello radiomico, studiando la distribuzione delle probabilità in uscita dai singoli classificatori di base e dal Soft Voting Ensemble finale. L'obiettivo è di valutare la calibrazione del modello sull'intero Test Set e identificare quali specifiche sequenze portino l'Ensemble in errore nei casi misclassificati.

La Figura 6.5 mostra l'analisi Waterfall disaggregata per le quattro modalità principali e per il Soft Voting finale, includendo tutti i pazienti misclassificati del Test Set. I pazienti sono ordinati, dal caso con la probabilità più alta a quello con la probabilità più bassa. Dall'analisi del Waterfall Plot emerge che le predizioni basate su DWI e sulla aschera prostatica influenzano in modo preponderante la predizione finale dell'Ensemble. Questi due classificatori tendono a generare valori di probabilità molto elevati ed errati, con probabilità prossime a 0.0 o 1.0.

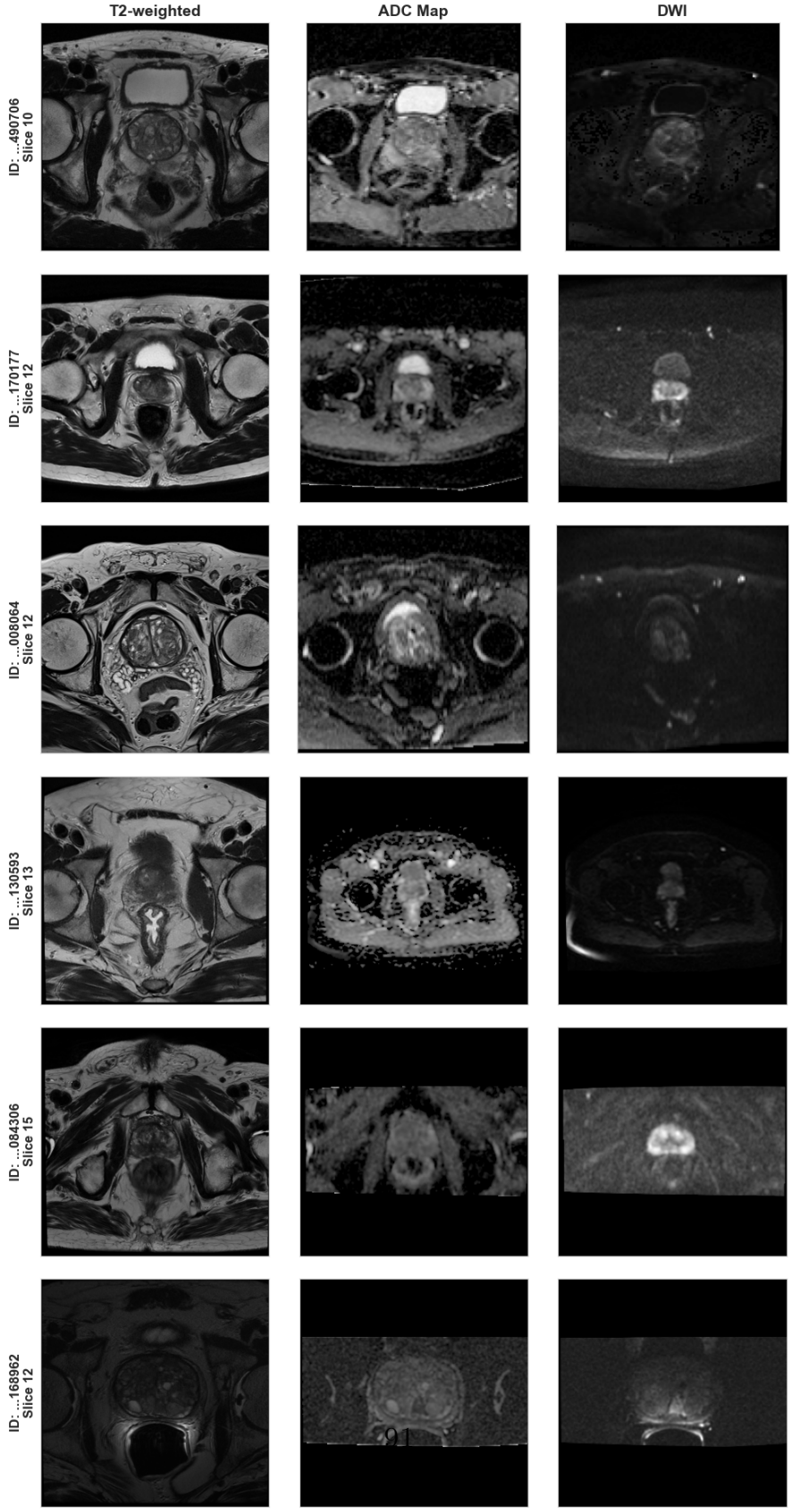


Figura 6.4: Rappresentazione visiva multi-parametrica (T2w, DWI, ADC) dei casi misclassificati da entrambi i modelli (Parte 1 di 2: primi 6 casi).

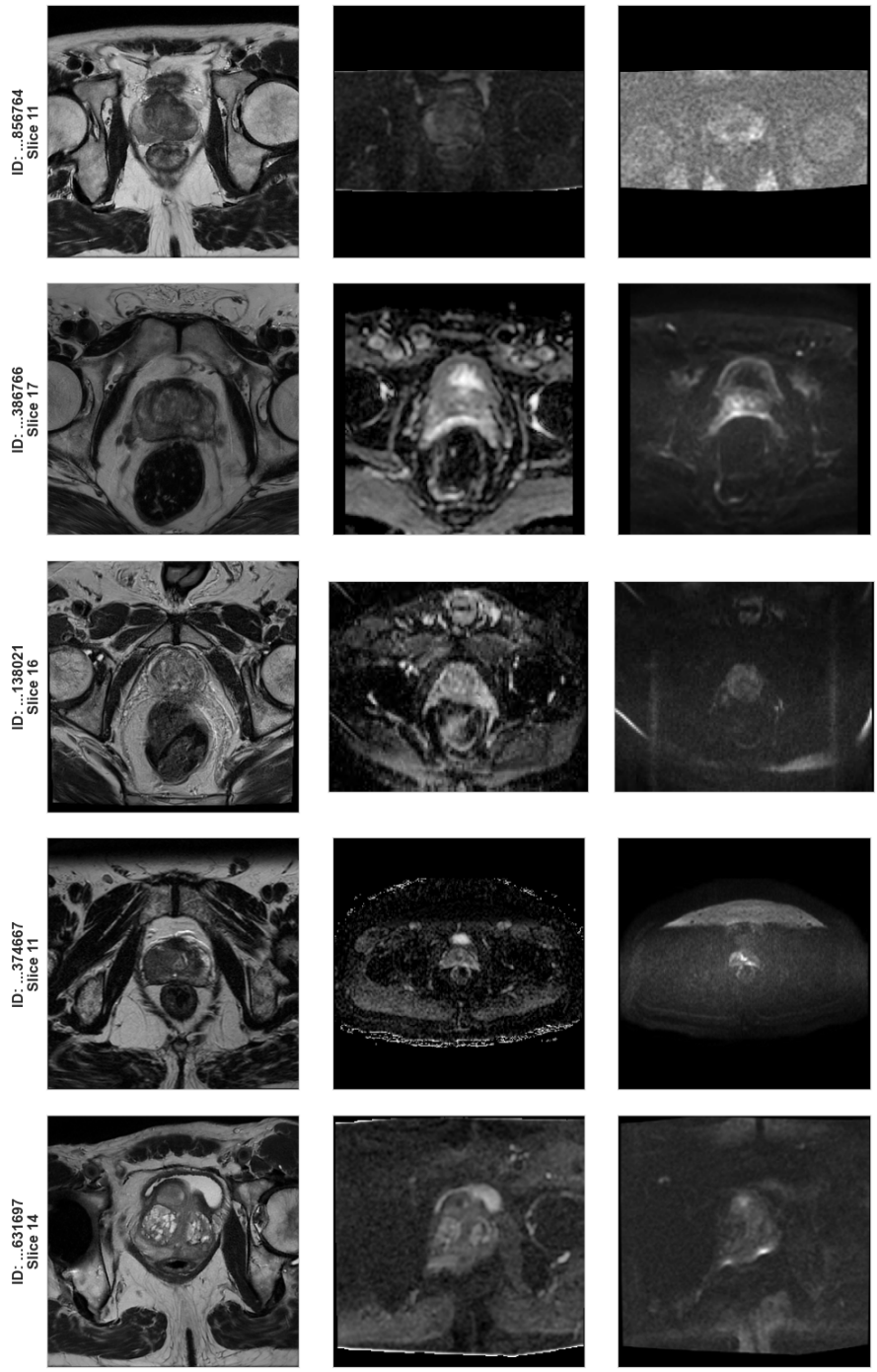


Figura 6.4: Campionario Visivo dei Casi Critici Condivisi (Parte 2 di 2). Rappresentazione visiva dei restanti 5 casi misclassificati.

Al contrario, le predizioni di ADC e T2w, seppur a loro volta inesatte nei casi misclassificati, presentano probabilità inferiori e più vicine alla soglia decisionale; traducendosi in classificazioni incerte e borderline, queste influenzano in misura minore il Soft Voting finale rispetto alle altre sequenze. In particolare, si osserva come l'errore del Soft Voting sia spesso guidato dall'estrema confidenza di una singola sequenza, che vince il confronto numerico con le altre modalità anche quando queste ultime presentano una tendenza corretta o incerta. Questa indagine permette di affermare che l'inclusione delle feature estratte dalla sola maschera di segmentazione, anziché apportare informazione aggiuntiva, sembra agire come un elemento di disturbo nei casi limite sbilanciando l'Ensemble verso l'*overconfidence* e la sovrastima della qualità. Considerando la fondamentale rilevanza clinica delle immagini DWI nella diagnosi prostatica, si è deciso di tentare una strategia di mitigazione di questo fenomeno conducendo un test di ottimizzazione, dove è stata oscurata la valutazione sulla maschera di segmentazione. I risultati di questa configurazione ridotta, confrontati con l'Ensemble originale a quattro vie, sono riportati nella Tabella 6.6.

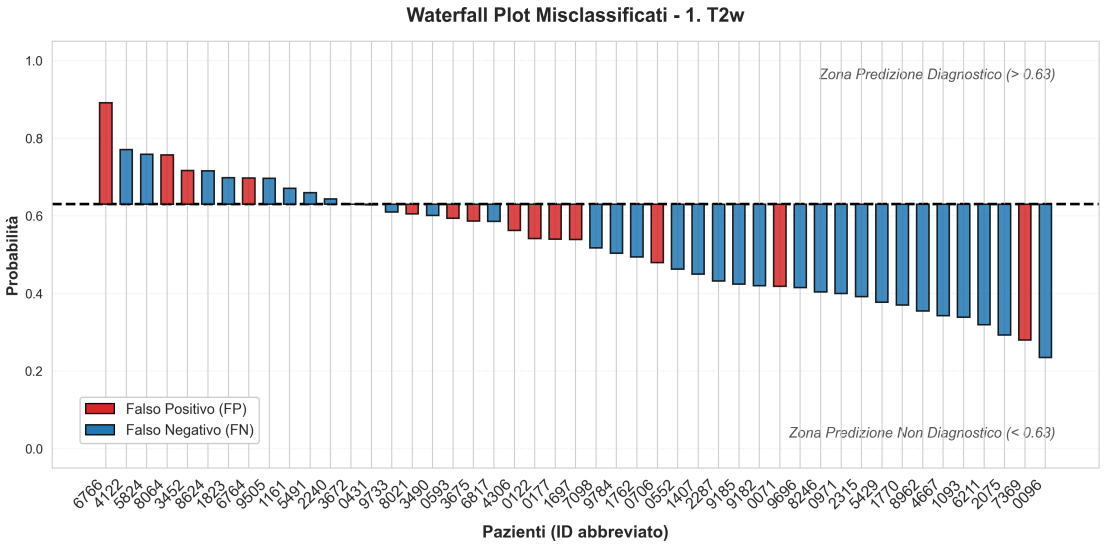
Tabella 6.6: Confronto prestazionale sul Test Set prima e dopo la rimozione del modello basato sulla Maschera T2w. L'eliminazione di questo elemento porta a un miglioramento della Specificità e quindi della Balanced Accuracy.

Soft Voting	**Accuracy**	**B. Acc**	**Sens**	**Spec**	**PPV**	**NPV**
Senza Maschera T2w	0.708	0.705	0.710	0.700	0.912	0.356
Completo	0.708	0.641	0.748	0.533	0.875	0.327

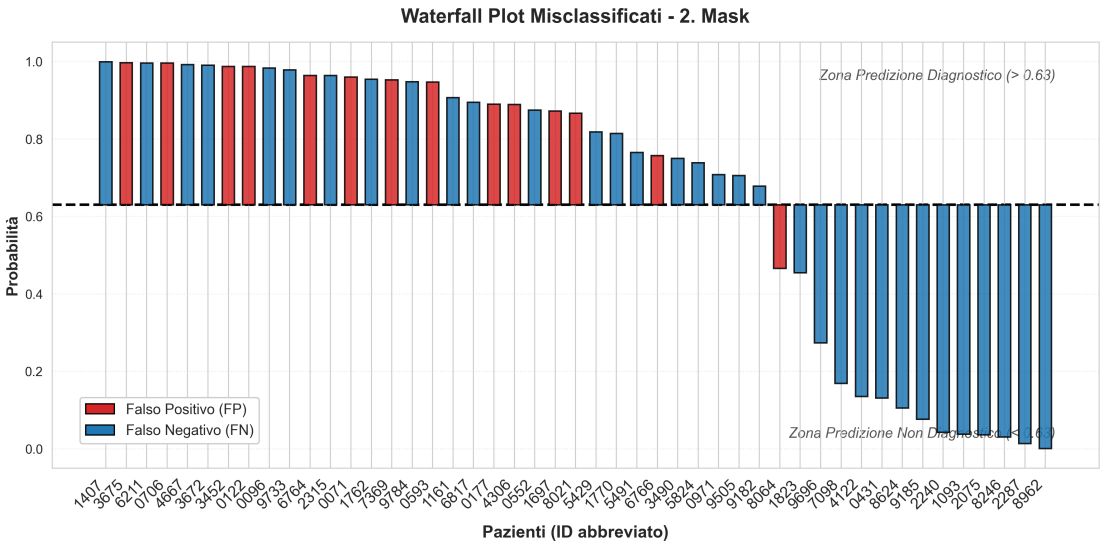
I dati confermano l'ipotesi formulata dall'analisi delle probabilità. Senza l'interferenza della maschera morfologica, l'Ensemble a tre vie registra un salto prestazionale, portando la Balanced Accuracy dal 64.1% al 70.5%. Il dato più rilevante è il crollo dei Falsi Positivi, che si traduce in un aumento della Specificità, la quale passa dal 53.3% al 70.0%. Tali risultati forniscono una linea guida per i futuri sviluppi della pipeline, indicando l'Ensemble a tre modalità come un'architettura più robusta e conservativa.

6.3 Prospettive Future

Un naturale sviluppo del presente lavoro consiste nell'implementazione del modello in un contesto clinico reale come strumento di supporto alle decisioni cliniche o in un'applicazione integrata direttamente nel flusso di lavoro PACS/RIS dei reparti

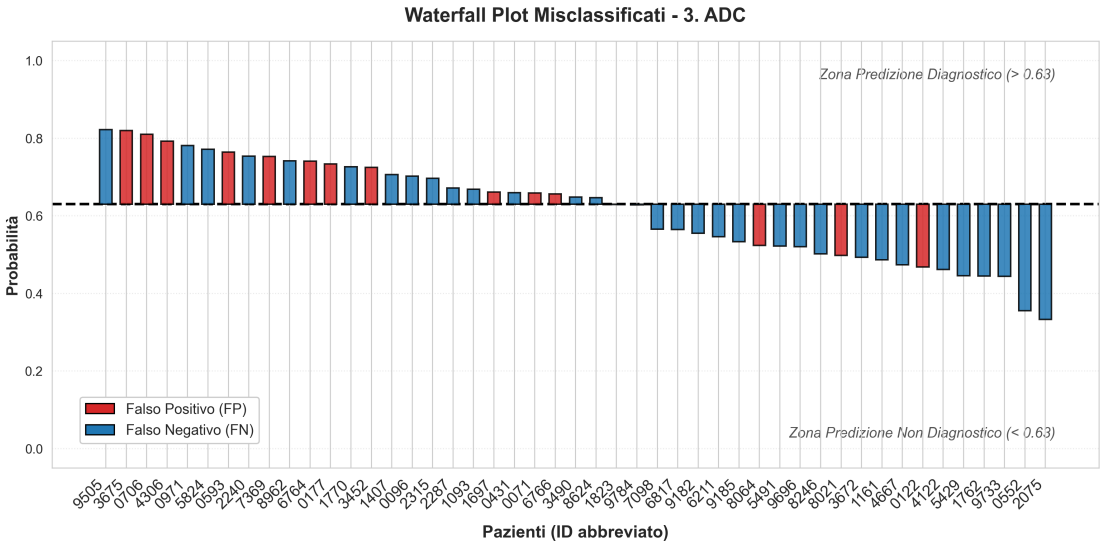


(a) T2w

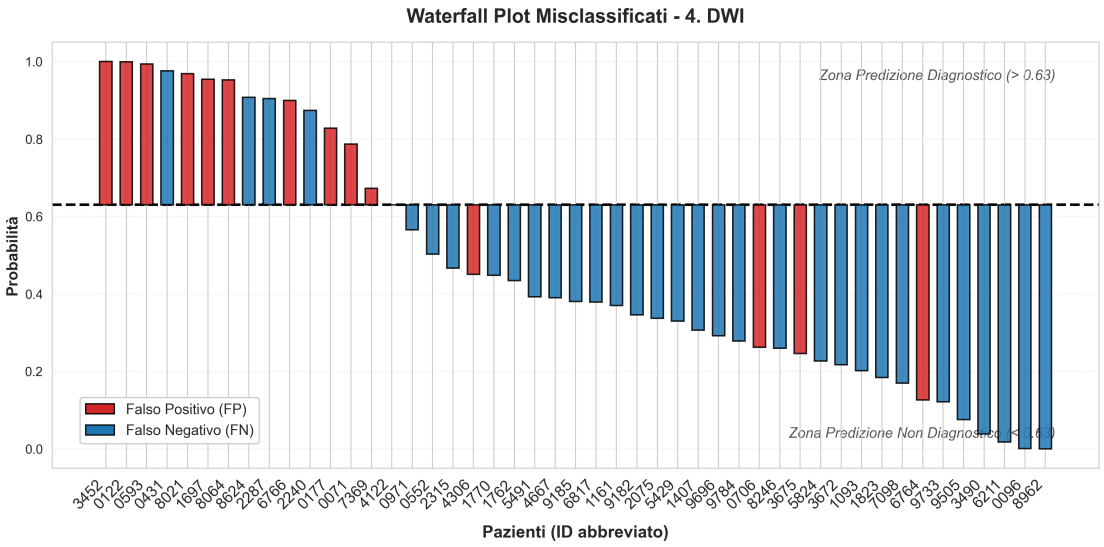


(b) Maschera T2w

Figura 6.5: Waterfall Plot dei Casi Misclassificati sul Test Set (Parte 1 di 3).

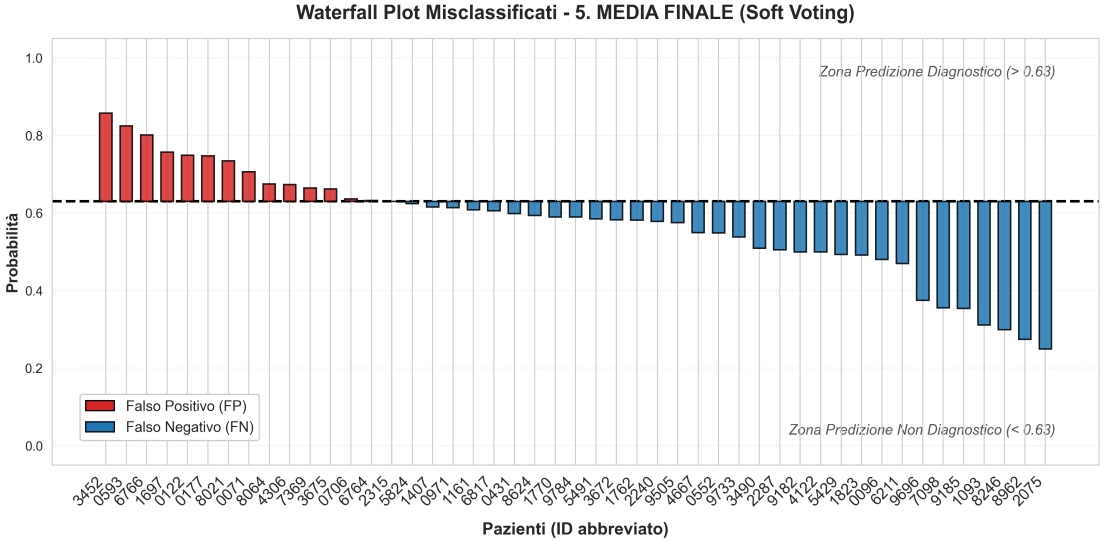


(c) ADC



(d) DWI

Figura 6.5: Waterfall Plot dei Casi Misclassificati sul Test Set (Parte 2 di 3).



(e) Media Finale (Soft Voting)

Figura 6.5: Waterfall Plot dei Casi Misclassificati sul Test Set (Parte 3 di 3). Analisi disaggregata delle probabilità predittive. Ogni grafico mostra l'andamento individuale per sequenza, con i pazienti ordinati in modo decrescente. Le barre rosse indicano i Falsi Positivi, quelle blu i Falsi Negativi.

di radiologia. In particolare, il principale collo di bottiglia della pipeline è rappresentato dal tempo richiesto per l'estrazione delle feature radiomiche, che incide significativamente sui tempi complessivi di elaborazione. I classificatori, una volta addestrati, presentano invece tempi di inferenza trascurabili. In prospettiva, sarà pertanto necessario ottimizzare il processo di estrazione delle feature, eventualmente riducendo il numero di variabili utilizzate o selezionando un sottoinsieme minimale ma altamente informativo, al fine di garantire un'applicazione compatibile con i tempi del workflow clinico.

Inoltre, l'interfaccia clinica non dovrà limitarsi a fornire un output binario, ma dovrà esporre chiaramente il livello di confidenza probabilistica, confermando la natura di supporto e non di sostituto del sistema rispetto al giudizio medico finale. La necessità di un output probabilistico si ricollega direttamente ai limiti intrinseci riscontrati in questo studio dove le criticità prestazionali non risiedono nelle architetture algoritmiche, bensì nella natura soggettiva della *Ground Truth* utilizzata per l'addestramento. La scala PI-QUAL, pur essendo il *gold standard* clinico attuale, soffre infatti di un'ineliminabile variabilità inter-osservatore.

Una prospettiva di ricerca dirompente consisterebbe nel ribaltare il paradigma, utilizzando la radiomica non solo per predire il punteggio PI-QUAL, ma per sostituirlo con un indice di qualità puramente quantitativo. Le feature matematiche,

essendo deterministiche e oggettive, potrebbero essere utilizzate per formulare nuovi parametri standardizzati in grado di misurare, ad esempio, il rapporto segnale-rumore, il contrasto tissutale e l'entità degli artefatti in modo continuo e riproducibile.

Per raggiungere questo obiettivo, sono richiesti dataset più ampi, distribuzioni di classe bilanciate e procedure di consenso tra radiologi al fine di garantire etichette di addestramento solide ed esenti da ambiguità.

6.4 Conclusione Generale

Il presente lavoro di tesi ha esplorato la possibilità di automatizzare il controllo qualità delle risonanze magnetiche prostatiche mediante approcci basati su Radiomica e Deep Learning. I risultati ottenuti mostrano che l'approccio radiomico multi-modale rappresenta, tra quelli testati, la soluzione più stabile e interpretabile. Tuttavia, le performance complessive evidenziano come il problema della classificazione della qualità diagnostica sia intrinsecamente complesso e fortemente influenzato dalla natura soggettiva delle etichette di riferimento.

L'analisi degli errori ha suggerito che una parte rilevante delle difficoltà del modello possa essere attribuita alla variabilità inter-osservatore e alla presenza di casi ambigui, più che a limiti strutturali delle architetture utilizzate.

In definitiva, il presente studio contribuisce a chiarire le criticità metodologiche e i vincoli strutturali che devono essere affrontati per rendere realmente affidabile un sistema automatico di valutazione della qualità. Lo sviluppo e il perfezionamento di questo framework non rispondono unicamente a un'esigenza di ricerca ingegneristica, ma a una necessità del Sistema Sanitario. La standardizzazione della qualità dell'imaging radiologico possiede il potenziale di abbattere i costi legati alle re-acquisizioni degli esami e di ottimizzare il carico di lavoro dei reparti. Delegando all'algoritmo la validazione puramente tecnica dell'immagine, l'intelligenza artificiale si propone di tutelare il tempo e l'attenzione del clinico, garantendogli un input visivo sempre ottimale su cui fondare diagnosi più rapide, precise e, in ultima analisi, salvavita.

Appendice A

Tabelle Complete dei Risultati Sperimentali

In questa appendice sono riportati i risultati dettagliati di tutte le configurazioni sperimentali discusse nel Capitolo 4. I risultati sono suddivisi per modalità di imaging (ADC, DWI, T2w Whole, T2w Mask) e per strategia di ottimizzazione: Grid Search e Optuna.

Le tabelle riportano la Balanced Accuracy media e la deviazione standard calcolate tramite 5-fold Cross-Validation. I modelli sono ordinati per performance decrescente.

A.1 Risultati Modalità ADC

ADC - Ottimizzazione: Grid Search

Tabella A.1: ADC - Grid Search (BACC Media ± Std)

Classificatore	Feature Selection	BACC
MLP	mutual_info	0.604 ± 0.025
MLP	rfecv	0.585 ± 0.030
MLP	rf	0.564 ± 0.042
SVM	mutual_info	0.562 ± 0.048
RandomForest	mutual_info	0.561 ± 0.047
MLP	lasso	0.559 ± 0.028
RandomForest	anova	0.558 ± 0.021
RandomForest	lasso	0.558 ± 0.048
XGBoost	lasso	0.557 ± 0.025

RandomForest	rf	0.557 ± 0.058
LightGBM	rfecv	0.555 ± 0.026
LightGBM	lasso	0.553 ± 0.024
RandomForest	rfecv	0.551 ± 0.050
LightGBM	anova	0.548 ± 0.018
LightGBM	rf	0.546 ± 0.029
XGBoost	rf	0.543 ± 0.034
LightGBM	mutual_info	0.543 ± 0.021
XGBoost	rfecv	0.542 ± 0.052
SVM	lasso	0.542 ± 0.050
MLP	anova	0.541 ± 0.032
XGBoost	mutual_info	0.538 ± 0.056
XGBoost	anova	0.537 ± 0.012
SVM	anova	0.536 ± 0.057
SVM	rf	0.522 ± 0.043
SVM	rfecv	0.513 ± 0.033

ADC - Ottimizzazione: Optuna

Tabella A.2: ADC - Optuna (BACC Media ± Std)

Classificatore	**Feature Selection**	**BACC**
MLP	rfecv	0.621 ± 0.041
MLP	mutual_info	0.612 ± 0.035
SVM	mutual_info	0.598 ± 0.048
MLP	lasso	0.584 ± 0.039
RandomForest	mutual_info	0.575 ± 0.042
MLP	anova	0.574 ± 0.029
MLP	rf	0.570 ± 0.051
RandomForest	rfecv	0.565 ± 0.044
SVM	lasso	0.562 ± 0.055
LightGBM	mutual_info	0.560 ± 0.028
XGBoost	rfecv	0.558 ± 0.040
RandomForest	lasso	0.555 ± 0.046
LightGBM	lasso	0.554 ± 0.030
XGBoost	lasso	0.552 ± 0.022
SVM	rfecv	0.548 ± 0.030
LightGBM	rf	0.545 ± 0.030
XGBoost	rf	0.544 ± 0.038

SVM	anova	0.542 ± 0.050
RandomForest	anova	0.540 ± 0.025
LightGBM	anova	0.538 ± 0.020
SVM	rf	0.535 ± 0.041
XGBoost	anova	0.532 ± 0.020
RandomForest	rf	0.530 ± 0.052
LightGBM	rfecv	0.568 ± 0.030
XGBoost	mutual_info	0.572 ± 0.030

A.2 Risultati Modalità DWI

DWI - Ottimizzazione: Grid Search

Tabella A.3: DWI - Grid Search (BACC Media ± Std)

Classificatore	**Feature Selection**	**BACC**
MLP	lasso	0.590 ± 0.030
RandomForest	anova	0.577 ± 0.032
RandomForest	rf	0.568 ± 0.033
MLP	anova	0.567 ± 0.062
SVM	anova	0.559 ± 0.018
LightGBM	anova	0.558 ± 0.043
RandomForest	mutual_info	0.555 ± 0.049
SVM	lasso	0.555 ± 0.058
XGBoost	anova	0.551 ± 0.044
XGBoost	mutual_info	0.542 ± 0.027
LightGBM	mutual_info	0.540 ± 0.024
LightGBM	rfecv	0.536 ± 0.043
RandomForest	rfecv	0.535 ± 0.079
MLP	rfecv	0.534 ± 0.031
RandomForest	lasso	0.534 ± 0.040
MLP	rf	0.530 ± 0.065
LightGBM	lasso	0.527 ± 0.012
SVM	rf	0.526 ± 0.036
XGBoost	lasso	0.525 ± 0.019
XGBoost	rfecv	0.524 ± 0.013
LightGBM	rf	0.518 ± 0.031
SVM	mutual_info	0.517 ± 0.010
SVM	rfecv	0.512 ± 0.036

MLP	mutual_info	0.507 ± 0.028
XGBoost	rf	0.504 ± 0.032

DWI - Ottimizzazione: Optuna

Tabella A.4: DWI - Optuna (BACC Media ± Std)

Classificatore	**Feature Selection**	**BACC**
MLP	lasso	0.612 ± 0.033
RandomForest	anova	0.589 ± 0.035
MLP	anova	0.585 ± 0.051
RandomForest	rf	0.582 ± 0.040
SVM	anova	0.574 ± 0.028
LightGBM	anova	0.570 ± 0.045
RandomForest	mutual_info	0.568 ± 0.052
SVM	lasso	0.565 ± 0.055
XGBoost	anova	0.560 ± 0.041
MLP	rfecv	0.558 ± 0.042
XGBoost	mutual_info	0.555 ± 0.030
LightGBM	mutual_info	0.552 ± 0.028
RandomForest	rfecv	0.548 ± 0.071
LightGBM	rfecv	0.545 ± 0.046
MLP	rf	0.542 ± 0.061
RandomForest	lasso	0.540 ± 0.045
SVM	rf	0.538 ± 0.040
XGBoost	lasso	0.535 ± 0.022
LightGBM	lasso	0.534 ± 0.018
XGBoost	rfecv	0.532 ± 0.018
SVM	mutual_info	0.528 ± 0.015
LightGBM	rf	0.525 ± 0.035
SVM	rfecv	0.522 ± 0.040
MLP	mutual_info	0.518 ± 0.032
XGBoost	rf	0.515 ± 0.035

A.3 Risultati Modalità T2w (Whole)

T2w - Ottimizzazione: Grid Search

Tabella A.5: T2w - Grid Search (BACC Media ± Std)

Classificatore	Feature Selection	BACC
RandomForest	rf	0.593 ± 0.046
RandomForest	rfecv	0.585 ± 0.042
LightGBM	mutual_info	0.576 ± 0.028
RandomForest	anova	0.571 ± 0.054
MLP	mutual_info	0.570 ± 0.052
MLP	rf	0.568 ± 0.039
XGBoost	rf	0.561 ± 0.019
RandomForest	lasso	0.560 ± 0.044
RandomForest	mutual_info	0.560 ± 0.050
XGBoost	mutual_info	0.557 ± 0.011
XGBoost	anova	0.557 ± 0.030
MLP	lasso	0.554 ± 0.067
SVM	mutual_info	0.551 ± 0.047
LightGBM	anova	0.550 ± 0.045
SVM	rf	0.544 ± 0.035
XGBoost	lasso	0.540 ± 0.024
LightGBM	rf	0.537 ± 0.013
MLP	anova	0.536 ± 0.042
MLP	rfecv	0.533 ± 0.062
XGBoost	rfecv	0.532 ± 0.037
LightGBM	lasso	0.532 ± 0.027
LightGBM	rfecv	0.527 ± 0.028
SVM	anova	0.512 ± 0.039
SVM	lasso	0.506 ± 0.028
SVM	rfecv	0.505 ± 0.020

T2w - Ottimizzazione: Optuna

Tabella A.6: T2w - Optuna (BACC Media ± Std)

Classificatore	Feature Selection	BACC
RandomForest	rf	0.605 ± 0.045
RandomForest	rfecv	0.598 ± 0.043
LightGBM	mutual_info	0.590 ± 0.031
MLP	mutual_info	0.588 ± 0.048
MLP	rf	0.585 ± 0.041

RandomForest	anova	0.582 ± 0.052
RandomForest	lasso	0.575 ± 0.045
XGBoost	rf	0.572 ± 0.022
RandomForest	mutual_info	0.570 ± 0.051
XGBoost	mutual_info	0.568 ± 0.015
MLP	lasso	0.565 ± 0.063
XGBoost	anova	0.565 ± 0.035
SVM	mutual_info	0.562 ± 0.046
LightGBM	anova	0.560 ± 0.043
SVM	rf	0.555 ± 0.038
XGBoost	lasso	0.552 ± 0.026
LightGBM	rf	0.548 ± 0.018
MLP	anova	0.545 ± 0.045
MLP	rfecv	0.542 ± 0.059
XGBoost	rfecv	0.540 ± 0.035
LightGBM	lasso	0.538 ± 0.025
LightGBM	rfecv	0.535 ± 0.032
SVM	anova	0.525 ± 0.041
SVM	lasso	0.518 ± 0.030
SVM	rfecv	0.512 ± 0.022

A.4 Risultati Modalità T2w (Mask)

T2w Mask - Ottimizzazione: Grid Search

Tabella A.7: T2w Mask - Grid Search (BACC Media ± Std)

Classificatore	**Feature Selection**	**BACC**
MLP	rfecv	0.580 ± 0.040
MLP	rf	0.567 ± 0.061
MLP	lasso	0.566 ± 0.058
SVM	anova	0.562 ± 0.039
MLP	anova	0.558 ± 0.057
LightGBM	rf	0.554 ± 0.050
SVM	mutual_info	0.552 ± 0.055
RandomForest	rf	0.552 ± 0.057
XGBoost	lasso	0.549 ± 0.047
RandomForest	lasso	0.544 ± 0.045
RandomForest	mutual_info	0.539 ± 0.056

SVM	rf	0.537 ± 0.058
XGBoost	rfecv	0.536 ± 0.046
RandomForest	anova	0.535 ± 0.032
SVM	lasso	0.535 ± 0.031
XGBoost	anova	0.534 ± 0.050
RandomForest	rfecv	0.533 ± 0.024
MLP	mutual_info	0.526 ± 0.076
LightGBM	anova	0.524 ± 0.038
LightGBM	mutual_info	0.523 ± 0.010
XGBoost	rf	0.522 ± 0.046
SVM	rfecv	0.518 ± 0.027
LightGBM	lasso	0.516 ± 0.027
LightGBM	rfecv	0.513 ± 0.030
XGBoost	mutual_info	0.497 ± 0.030

T2w Mask - Ottimizzazione: Optuna

Tabella A.8: T2w Mask - Optuna (BACC Media ± Std)

Classificatore	**Feature Selection**	**BACC**
MLP	rfecv	0.675 ± 0.051
MLP	rf	0.612 ± 0.055
MLP	lasso	0.598 ± 0.062
SVM	anova	0.575 ± 0.043
MLP	anova	0.572 ± 0.058
LightGBM	rf	0.565 ± 0.048
SVM	mutual_info	0.562 ± 0.051
RandomForest	rf	0.560 ± 0.055
XGBoost	lasso	0.558 ± 0.042
RandomForest	lasso	0.555 ± 0.048
SVM	rf	0.550 ± 0.052
RandomForest	mutual_info	0.548 ± 0.053
XGBoost	rfecv	0.545 ± 0.041
SVM	lasso	0.542 ± 0.035
RandomForest	anova	0.540 ± 0.035
XGBoost	anova	0.538 ± 0.045
RandomForest	rfecv	0.535 ± 0.028
MLP	mutual_info	0.532 ± 0.071
LightGBM	anova	0.530 ± 0.035

LightGBM	mutual_info	0.528 ± 0.015
XGBoost	rf	0.525 ± 0.042
SVM	rfecv	0.522 ± 0.030
LightGBM	lasso	0.520 ± 0.025
LightGBM	rfecv	0.518 ± 0.033
XGBoost	mutual_info	0.505 ± 0.033

Bibliografia

[1] Min Yuen Teo, Dana E. Rathkopf e Philip Kantoff. «Treatment of Advanced Prostate Cancer». In: *Annual Review of Medicine* 70.1 (2019), pp. 479–499. DOI: 10.1146/annurev-med-051517-011947 (cit. a p. 2).

[2] Sobia Wasim, Sang-Yoon Lee e Jaehong Kim. «Complexities of Prostate Cancer». In: *International Journal of Molecular Sciences* 23.22 (2022). ISSN: 1422-0067. DOI: 10.3390/ijms232214257. URL: https://www.mdpi.com/1422-0067/23/22/14257 (cit. alle pp. 2, 3).

[3] Christoph Würnschimmel, Thenappan Chandrasekar, Luisa Hahn, Tarik Esen, Shahrokh F Shariat e Derya Tilki. «MRI as a screening tool for prostate cancer: current evidence and future challenges». In: *World Journal of Urology* 41.4 (2023), pp. 921–928. DOI: 10.1007/s00345-022-03947-y (cit. a p. 3).

[4] E. T. d. Correia, A. Baydoun, Q. Li et al. «Emerging and anticipated innovations in prostate cancer MRI and their impact on patient care». In: *Abdominal Radiology* 49 (2024), pp. 3696–3710. DOI: 10.1007/s00261-024-04423-4. URL: https://doi.org/10.1007/s00261-024-04423-4 (cit. a p. 3).

[5] V. F. Muglia, L. Laschena, M. Pecoraro et al. «Imaging assessment of prostate cancer recurrence: advances in detection of local and systemic relapse». In: *Abdominal Radiology* 50 (2025), pp. 807–826. DOI: 10.1007/s00261-024-04412-7. URL: https://doi.org/10.1007/s00261-024-04412-7 (cit. a p. 3).

[6] Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke et al. «Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2». In: *European urology* 76.3 (2019), pp. 340–351 (cit. alle pp. 3–6, 15).

[7] Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager e Jurggen J Fütterer. «ESUR prostate MR guidelines 2012». In: *European radiology* 22 (2012), pp. 746–757 (cit. alle pp. 3, 5).

[8] Gabriele Busè et al. «Performance of Diffusion Kurtosis Imaging For Characterization of Prostate Lesions using 1.7T Magnetic Resonance Imaging Scanner». In: *EuroMediterranean Biomedical Journal* (nov. 2020). DOI: 10.3269/1970-5492.2020.15.46 (cit. a p. 4).

[9] John E McNeal. «The zonal anatomy of the prostate». In: *The Prostate* 2.1 (1981), pp. 35–49 (cit. a p. 4).

[10] A. R. Padhani et al. «Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations». In: *Neoplasia* 11.2 (2009), pp. 102–125. DOI: 10.1593/neo.81328 (cit. alle pp. 5, 6).

[11] Rajan T. Gupta, Kaivan A. Mehta, Baris Turkbey e Sadhna Verma. «PI-RADS: Past, present, and future». In: *Journal of Magnetic Resonance Imaging* 52.1 (2020), pp. 33–53. DOI: 10.1002/jmri.26896 (cit. a p. 5).

[12] Maarten de Rooij et al. «PI-QUAL version 2: an update of a standardised scoring system for the assessment of image quality of prostate MRI». In: *European Radiology* (2024). DOI: 10.1007/s00330-024-10642-w (cit. alle pp. 7, 9).

[13] M. de Rooij e J. O. Barentsz. «PI-QUAL v.1: the first step towards good-quality prostate MRI». In: *European Radiology* 32.2 (2022), pp. 876–878. DOI: 10.1007/s00330-021-08399-3 (cit. a p. 7).

[14] A. Woernle, C. Englman, L. Dickinson et al. «Picture perfect: the status of image quality in prostate MRI». In: *Journal of Magnetic Resonance Imaging* (2023). DOI: 10.1002/jmri.29025. URL: https://doi.org/10.1002/jmri.29025 (cit. a p. 7).

[15] Francesco Giganti, Sydney Lindner, Jonathan W. Piper, Veeru Kasivisvanathan, Mark Emberton, Caroline M. Moore e Clare Allen. «Multiparametric prostate MRI quality assessment using a semi-automated PI-QUAL software program». In: *European Radiology Experimental* 5.1 (2021), p. 48. DOI: 10.1186/s41747-021-00245-x (cit. alle pp. 8, 10).

[16] Nicola Schieda, Christopher S. Lim, Fatemeh Zabihollahy, Jorge Abreu-Gomez, Satheesh Krishna, Sungmin Woo, Gerd Melkus, Eran Ukwatta e Baris Turkbey. «Quantitative Prostate MRI». In: *Journal of Magnetic Resonance Imaging* 52 (2020), pp. 1207–1229. DOI: 10.1002/jmri.27191 (cit. a p. 8).

[17] Steven J. Esses, Xiaoguang Lu, Tiejun Zhao, Krishna Shanbhogue, Bari Dane, Mary Bruno e Hersh Chandarana. «Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture». In: *Magnetic Resonance Imaging* 47 (2018), pp. 723–728. DOI: 10.1002/jmri.25779 (cit. alle pp. 8, 10).

[18] S. Cengiz et al. «Radiomics based automated quality assessment for T2W prostate MR images». In: *Biomedical Signal Processing and Control* (2023). (Ref. inferred from content) (cit. a p. 10).

[19] Mason J Belue, Yan Mee Law, Jamie Marko, Evrim Turkbey, Ashkan Malayeri et al. «Deep Learning-Based Interpretable AI for Prostate T2W MRI Quality Evaluation». In: *Academic Radiology* (2024). DOI: 10.1016/j.acra.2023.12.025 (cit. alle pp. 10, 69–71).

[20] Jacob N Gloe, Eric A Borisch, Adam T Froemming, Akira Kawashima, Jordan D LeGout et al. «Deep learning for quality assessment of axial T2-weighted prostate MRI: a tool to reduce unnecessary rescanning». In: *European Radiology Experimental* 9.44 (2025). DOI: 10.1186/s41747-025-00584-z (cit. alle pp. 11, 69, 70).

[21] José Guilherme de Almeida et al. «Impact of Scanner Manufacturer, Endorectal Coil Use, and Clinical Variables on Deep Learning–assisted Prostate Cancer Classification Using Multiparametric MRI». In: *Radiology: Artificial Intelligence* 7.3 (2025). PMID: 39841063, e230555. DOI: 10.1148/ryai.230555. eprint: https://doi.org/10.1148/ryai.230555. URL: https://doi.org/10.1148/ryai.230555 (cit. alle pp. 12, 13).

[22] J. Gawlitza et al. «Impact of the use of an endorectal coil for 3 T prostate MRI on image quality and cancer detection rate». In: *Scientific Reports* 7.1 (2017), p. 40640. DOI: 10.1038/srep40640 (cit. a p. 13).

[23] Zarine K. Shah et al. «Performance Comparison of 1.5-T Endorectal Coil MRI with 3.0-T Nonendorectal Coil MRI in Patients with Prostate Cancer». In: *Academic Radiology* 22.4 (2015), pp. 467–474. DOI: 10.1016/j.acra.2014.12.006 (cit. a p. 13).

[24] Joost J.M. van Griethuysen et al. «Computational Radiomics System to Decode the Radiographic Phenotype». In: *Cancer Research* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339 (cit. a p. 14).

[25] Alex Zwanenburg, Martin Vallières, MA Abdalah, HJWL Aerts, V Andrearczyk, A Apte et al. «The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping». In: *Radiology* 295.2 (2020), pp. 328–338. DOI: 10.1148/radiol.2020191145 (cit. a p. 14).

[26] Robert M Haralick, Karthikeyan Shanmugam e Its' Hak Dinstein. «Textural features for image classification». In: *IEEE Transactions on Systems, Man, and Cybernetics* 3.6 (1973), pp. 610–621 (cit. a p. 16).

[27] Mary M Galloway. «Texture analysis using gray level run lengths». In: *Computer graphics and image processing* 4.2 (1975), pp. 172–179 (cit. a p. 16).

[28] Guillaume Thibault, Bernard Fertil, C Navarro, S Pereira, P Cau, N Levy, J Sequeira e JL Mari. «Texture indexes and gray level size zone matrix: application to cell nuclei classification». In: *Pattern Recognition and Information Processing (PRIP)* (2009), pp. 140–145 (cit. a p. 16).

[29] Changming Sun e William G Wee. «Neighboring gray level dependence matrix for texture classification». In: *Computer Vision, Graphics, and Image Processing* 23.3 (1983), pp. 341–352 (cit. a p. 16).

[30] Marius E. Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs e Gary Cook. «Introduction to Radiomics». In: *Journal of Nuclear Medicine* 61.4 (2020), pp. 488–495. ISSN: 0161-5505. DOI: 10.2967/jnumed.118.222893. eprint: https://jnm.snmjournals.org/content/61/4/488.full.pdf. URL: https://jnm.snmjournals.org/content/61/4/488 (cit. a p. 20).

[31] Mervyn Stone. «Cross-validatory choice and assessment of statistical predictions». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 111–133 (cit. a p. 23).

[32] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer Science & Business Media, 2009 (cit. a p. 23).

[33] David H Wolpert. «Stacked generalization». In: *Neural networks* 5.2 (1992), pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1 (cit. a p. 23).

[34] Visar Berisha, Christa Krantsevich, P. Robert Hahn, Sidney Hahn, Gautam Dasarathy, Pavan Turaga e Julie Liss. «Digital medicine and the curse of dimensionality». In: *NPJ Digital Medicine* 4.1 (ott. 2021), p. 153. DOI: 10.1038/s41746-021-00521-5 (cit. alle pp. 24, 32).

[35] J. J. M. van Griethuysen et al. *PyRadiomics Documentation: Feature Definitions*. https://pyradiomics.readthedocs.io/en/latest/features.html. Versione 3.0.1. 2020 (cit. a p. 24).

[36] George Casella e Roger L. Berger. *Statistical Inference*. 2nd. Duxbury Press, 2002 (cit. a p. 25).

[37] Nassim Nicholas Taleb. «The main consequences and how they link to the book». In: *Statistical Consequences of Fat Tails*. Consequence 1 and Consequence 5. Brooklyn, NY: STEM Academic Press, 2020. Cap. 3, pp. 34–38 (cit. a p. 26).

[38] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall e W. Philip Kegelmeyer. «SMOTE: synthetic minority over-sampling technique». In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357 (cit. a p. 26).

[39] Ian T. Jolliffe. *Principal Component Analysis*. 2ª ed. Springer, 2002 (cit. a p. 33).

[40] Laurens van der Maaten e Geoffrey Hinton. «Visualizing Data using t-SNE». In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605 (cit. a p. 33).

[41] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver e Boyd, 1925 (cit. a p. 37).

[42] Claude E. Shannon. «A Mathematical Theory of Communication». In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. a p. 37).

[43] Leo Breiman. «Random Forests». In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324 (cit. alle pp. 38, 52, 55, 82).

[44] Robert Tibshirani. «Regression Shrinkage and Selection via the Lasso». In: *Journal of the Royal Statistical Society: Series B* 58.1 (1996), pp. 267–288 (cit. a p. 40).

[45] Isabelle Guyon, Jason Weston, Stephen Barnhill e Vladimir Vapnik. «Gene selection for cancer classification using support vector machines». In: *Machine Learning* 46.1 (2002), pp. 389–422 (cit. a p. 42).

[46] Da Wei Huang, Brad T Sherman e Richard A Lempicki. «Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists». In: *Nucleic acids research* 37.1 (2009), pp. 1–13 (cit. a p. 45).

[47] Corinna Cortes e Vladimir Vapnik. «Support-vector networks». In: *Machine Learning* 20.3 (1995), pp. 273–297 (cit. alle pp. 52, 56).

[48] Tianqi Chen e Carlos Guestrin. «XGBoost: A scalable tree boosting system». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794 (cit. alle pp. 52, 57).

[49] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye e Tie-Yan Liu. «LightGBM: A highly efficient gradient boosting decision tree». In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 3146–3154 (cit. alle pp. 53, 58).

[50] David E Rumelhart, Geoffrey E Hinton e Ronald J Williams. «Learning representations by back-propagating errors». In: *Nature* 323.6088 (1986), pp. 533–536 (cit. alle pp. 53, 59).

[51] F. Pedregosa et al. «Scikit-learn: Machine Learning in Python». In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. a p. 53).

[52] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta e Masanori Koyama. «Optuna: A Next-generation Hyperparameter Optimization Framework». In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019, pp. 2623–2631 (cit. a p. 54).

[53] David R Cox. «The regression analysis of binary sequences». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232 (cit. a p. 62).

[54] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken e Clara I. Sánchez. «A survey on deep learning in medical image analysis». In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135 (cit. a p. 68).

[55] Deniz Alis, Mustafa Said Kartal, Mustafa Ege Seker, Batuhan Guroz et al. «Deep learning for assessing image quality in bi-parametric prostate MRI: A feasibility study». In: *European Journal of Radiology* 165 (2023), p. 110924. DOI: 10.1016/j.ejrad.2023.110924 (cit. alle pp. 69, 70).

[56] Aaron Fisher, Cynthia Rudin e Francesca Dominici. «All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously». In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1–81 (cit. a p. 82).

[57] Marco Tulio Ribeiro, Sameer Singh e Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: ACM, 2016, pp. 1135–1144 (cit. a p. 82).

[58] J. Richard Landis e Gary G. Koch. «The Measurement of Observer Agreement for Categorical Data». In: *Biometrics* 33.1 (1977), pp. 159–174 (cit. a p. 85).

[59] Quinn McNemar. «Note on the sampling error of the difference between correlated proportions or percentages». In: *Psychometrika* 12.2 (1947), pp. 153–157 (cit. a p. 85).